Information Fusion 131 (2026) 104189

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

Large multimodal models for low-resource languages: A survey

Marian Lupaşcu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, Radu Tudor Ionescu *

Department of Computer Science, University of Bucharest, Romania

ARTICLE INFO

Keywords:
Large multimodal models
Multimodal language models
Low-resource languages

ABSTRACT

In this survey, we systematically analyze techniques used to adapt large multimodal models (LMMs) for low-resource (LR) languages, examining approaches ranging from visual enhancement and data creation to cross-modal transfer and fusion strategies. Through a comprehensive analysis of 117 studies across 96 LR languages, we identify key patterns in how researchers tackle the challenges of limited data and computational resources. We categorize works into resource-oriented and method-oriented contributions, further dividing contributions into relevant sub-categories. We compare method-oriented contributions in terms of performance and efficiency, discussing benefits and limitations of representative studies. We find that visual information often serves as a crucial bridge for improving model performance in LR settings, though significant challenges remain in areas such as hallucination mitigation and computational efficiency. In summary, we provide researchers with a clear understanding of current approaches and remaining challenges in making LMMs more accessible to speakers of LR (understudied) languages. We complement our survey with an open-source repository available at: https://github.com/marianlupascu/LMM4LRL-Survey.

1. Introduction

Recent advancements in large multimodal models (LMMs) showcased remarkable capabilities in processing and understanding diverse types of data, including text, images, audio and video. Models like GPT-4V, KOSMOS-1 [1] and PaLM-E [2] achieved impressive performance levels across various multimodal tasks through their ability to simultaneously process and reason about multiple modalities. However, these developments have primarily focused on high-resource languages, particularly English, leaving a significant gap in supporting the world's many low-resource languages. The distinction between high-resource (HR) and low-resource (LR) languages is primarily determined by the availability of digital resources and training data. High-resource languages, such as English, Mandarin, and Spanish, benefit from extensive digital corpora, parallel texts, and annotated datasets. In contrast, low-resource or understudied languages, which constitute the majority of the world's languages, lack sufficient digital resources, standardized datasets, and computational tools. This disparity is particularly pronounced in multimodal contexts, where the scarcity of paired data across modalities (e.g. image-text pairs, audio-text alignments) poses additional challenges.

A recent analysis [3] identified 27% of languages as "Invisible Giants", i.e. demographically robust yet digitally absent, highlighting that resource scarcity is institutionally constructed rather than inherent. This distinction has practical implications, e.g. LMM development that treats data scarcity merely as technical risks can perpetuate the structural inequalities it ostensibly addresses. We therefore situate our analysis within the UNESCO International Decade of Indigenous Languages (2022–2032) and the CARE principles for Indigenous data governance [4], which emphasize community authority over linguistic data. Indeed, the very terminology "low-resource" has been critiqued as colonial and Eurocentric, obscuring the political decisions that produced linguistic marginalization [5].

The motivation for developing multimodal capabilities for LR languages is compelling. First, multimodal processing better reflects how humans naturally communicate and understand information through multiple sensory channels. Second, visual and audio cues can provide crucial contextual information that helps to overcome the limitations of scarce textual data. Third, many LR languages are primarily spoken rather than written, making multimodal approaches particularly relevant for their digital preservation and processing. However, developing multimodal systems for LR languages faces several significant challenges, including: (1) the scarcity of high-quality multimodal datasets in these languages, (2) the lack of standardized evaluation benchmarks, (3) the computational cost of training large-scale models with limited resources, and (4) the complexity of handling different writing systems, dialects, and cultural contexts. Moreover, the problem of catastrophic forgetting [6] when adapting pre-trained models to new languages and the challenge of maintaining performance across different modalities pose significant technical hurdles.

* Corresponding author.
E-mail address: radu.ionescu@fmi.unibuc.ro (R.T. Ionescu).

https://doi.org/10.1016/j.inffus.2026.104189
Received 4 August 2025; Received in revised form 24 December 2025; Accepted 26 January 2026
Available online 27 January 2026

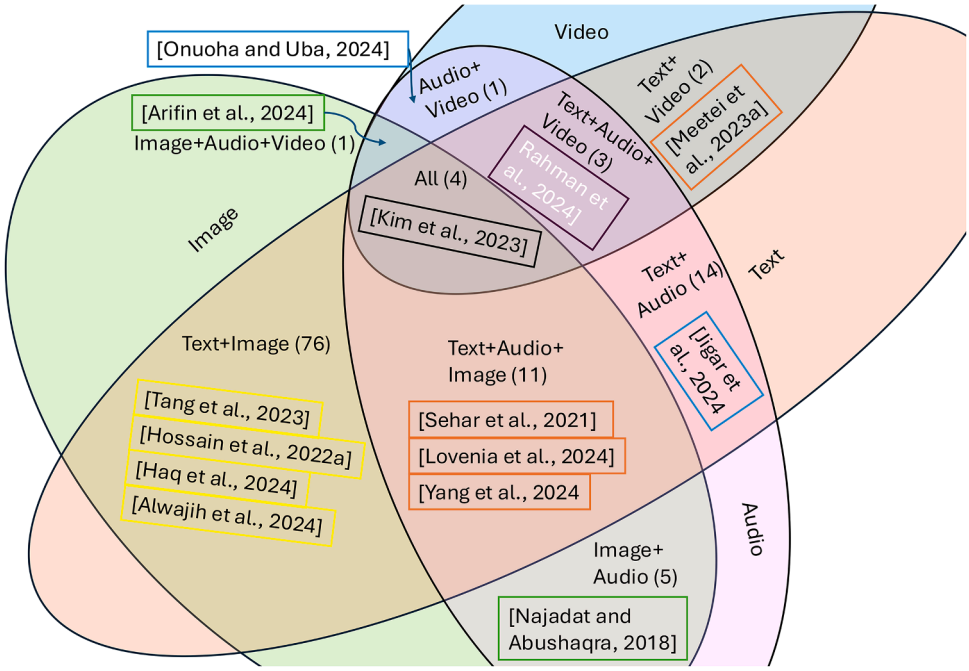

Fig. 1. A Venn diagram with the distribution of papers across different modality combinations used by LMMs for low-resource languages. Text + image is the dominant modality pair, while more complex video-inclusive combinations are less common. A selection of representative papers is included for each modality combination.

Literature selection process. We survey research articles from 2018 to 2025 that specifically study LMMs for LR languages. We begin our analysis with 2018 because one of the first large language models (LLMs), BERT [7], was introduced that year, marking a significant turning point in the development of modern language modeling techniques. We focus on works that go beyond simple cross-lingual transfer or translation, examining techniques that leverage multiple modalities to improve model performance.

We queried a broad set of digital libraries to ensure representative coverage: ACM Digital Library, IEEE Xplore, ACL Anthology, arXiv, SpringerLink, ScienceDirect, and Google Scholar. During this search, we specifically targeted venues known for frequent LR or multimodal contributions (e.g. ACL, EMNLP, NAACL, COLING, LREC, EACL, WMT, CVPR/ICCV/ECCV workshops, and INTERSPEECH) to ensure that relevant conference and workshop publications are captured. For each of these digital libraries, we formulated several keyword combinations capturing (i) multimodality, (ii) low-resource aspects, and (iii) language or task types. To further improve coverage, we performed a backward and forward search to identify additional relevant work from the reference lists of included papers, and we used citation links to identify more recent follow-up studies.

Finally, we merged all retrieved records and applied a manual two-stage screening process. We began by reviewing titles and abstracts to remove clearly irrelevant work (e.g. single modality studies or studies exclusively targeting high-resource languages such as English, Mandarin, or Spanish). We then examined the main contributions of the remaining papers to assess whether they were a suitable fit for our survey. Ultimately, a study was included in our survey if it matched all of the following criteria: (a) is multimodal (at least two input modalities), (b) focuses on low-resource languages (at least one of the targeted languages was an underrepresented language), and (c) proposes, adapts or evaluates a multimodal model.

Research focus distribution across LR languages. Our survey reveals several interesting patterns in how researchers approached multimodal learning for LR languages. As shown in Fig. 1, text-image combinations dominate the research landscape, appearing in 76 papers (65% of surveyed works), while more complex combinations incorporating audio and video remain less explored. In addition, the distribution of research focus across languages is notably uneven, as illustrated in Fig. 2, with Hindi (31 papers), Arabic (23) and Bengali (21 papers) receiving significant focus, whereas 42 other languages are each represented by a single study. On the one hand, this striking disparity highlights the need for a broader coverage of understudied languages in multimodal research. On the other hand, it also warrants critical examination of the factors influencing the distribution of research studies across languages.

We identify six interacting factors that explain this disparity (see Table 1). First, **institutional research capacity** plays a dominant role: Rungta et al. [8] demonstrated that NLP publications are heavily concentrated in North America, Western Europe, and China, with minimal representation from Africa and South America. Languages spoken in regions with established NLP research communities (e.g. Hindi in India, Arabic in the Middle East) benefit from existing infrastructure, funding, and researcher networks. Second, **speaker population** shows surprising variability: while one might expect larger speaker populations to attract more research, this correlation is weak. For instance, Swahili has approximately 200 million speakers, yet remains underrepresented compared with Malayalam (38 million speakers, 19 papers). Third, **digital resource availability** creates a self-reinforcing cycle: languages with existing datasets attract more research, which produces more datasets [9]. Ranathunga et al. [10] showed that even within the same resource class [9], languages from higher-GDP regions receive disproportionately more research attention. Fourth, **script and typological proximity** to high-resource languages facilitates transfer learning research, e.g. Hindi benefits from shared Devanagari script resources, while languages with unique scripts (e.g. Ge'ez for Amharic) face additional barriers. Fifth, **geographic location** determines access to NLP venues and research networks: languages spoken in regions hosting major conferences (Asia, Middle East, Europe) receive more attention than those in Sub-Saharan Africa or Oceania. Sixth, **geopolitical interest** drives strategic investment: Arabic NLP surged post-9/11, Ukrainian gained attention after 2022, and US-China AI competition benefits research on Mandarin Chinese, but not on minority languages within China.

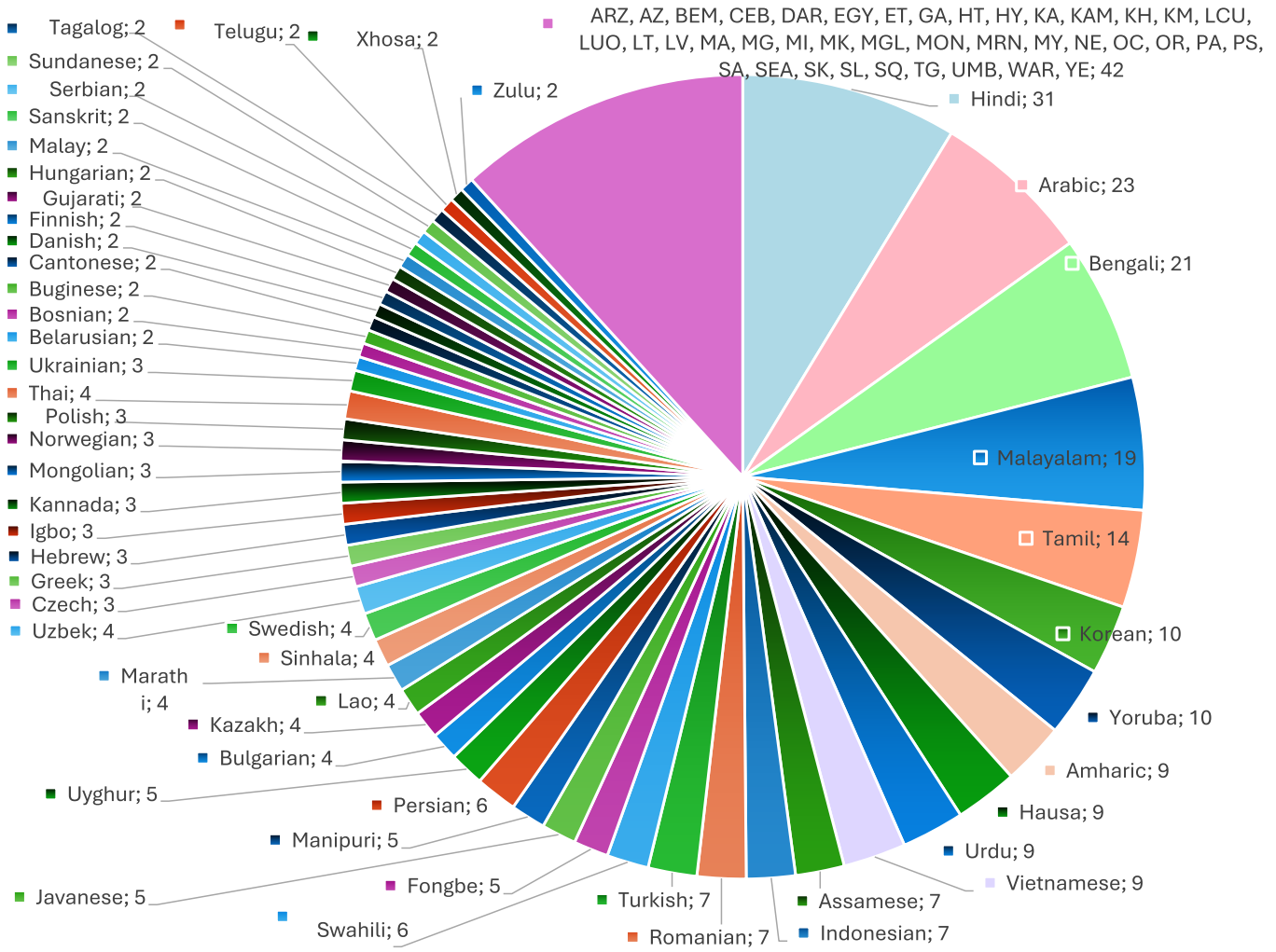

Fig. 2. Distribution of papers across 96 low-resource languages, representing 117 papers. Hindi leads with 31 studies, followed by Arabic (23), Bengali (21), Malayalam (19), Tamil (14), Korean and Yoruba (with 10 papers each). The remaining languages have less than 10 papers each. Languages with only one paper (42 languages) are listed using ISO 639-1 codes. The data highlights the disparity in research focus among LR languages, with a few languages receiving more focus, while many others remain understudied in the context of multimodal learning. Some papers simultaneously address multiple languages, contributing to the individual language counts. HR languages such as English, Chinese, Mandarin and Spanish are excluded from this chart.

Table 1
Factors explaining research disparity across low-resource languages in multimodal NLP. We categorize languages from our survey by paper count and analyze contributing factors.

Factor	**High Coverage** (10+ papers)	**Medium Coverage** (2–9 papers)	**Low Coverage** (1 paper)
Institutional capacity	Strong local NLP communities (India, Middle East)	Emerging research groups	Minimal local infrastructure
Speaker population	Variable (38M–600M)	Variable (2M–200M)	Often <1M, but not always
Existing resources	Multiple benchmarks, pre-trained models	Some datasets available	Little to no digital presence
Script accessibility	Shared with HR languages or well-supported	Moderate tool support	Often unique / unsupported scripts
Geographic location	Regions with NLP venues (Asia, Middle East)	Mixed	Often Sub-Saharan Africa, Oceania
Geopolitical interest	Strategic priority (Arabic post-9/11, Hindi for US-India relations, Korean for East Asia security)	Emerging strategic relevance (Ukrainian post-2022, Turkish for NATO relations)	No perceived strategic value; excluded from defense / intelligence funding
Example languages	Hindi, Arabic, Bengali, Malayalam	Swahili, Romanian, Turkish, Yoruba	Luo, Xhosa, Occitan, Maori

These factors have critical implications for researchers working on truly underrepresented languages. The 42 languages with single studies in our survey face a "cold start" problem: without existing benchmarks, baselines, or community momentum, new contributions are harder to contextualize and evaluate [11]. We observe that 88.4% of the world's languages (Class 0 defined by Joshi et al. [9]) have no representation in standard NLP resources whatsoever [9]. For researchers targeting these languages, we recommend: (1) prioritizing dataset creation with community involvement over model development, (2) leveraging typologically similar languages for transfer rather than defaulting to English, and (3) publishing in venues with explicit low-resource tracks (e.g. AfricaNLP, AmericasNLP, etc.) to build critical mass within language-specific research communities.

Research investment in specific languages is strongly influenced by geopolitical events and national security priorities. The clearest documented case is Arabic NLP, which experienced a dramatic surge in funding following September 11, 2001. Darwish et al. [12] documented the fact that "Arabic NLP gained increasing importance in the Western world especially after September 11. The USA funded large projects for companies and research centers to develop NLP tools for Arabic and its dialects". This investment wave (2001–2010) produced fundamental resources for machine translation, speech recognition, named entity recognition, and information extraction, that continue to underpin Arabic multimodal research today.

A similar pattern is emerging for Ukrainian. Prior to February 2022, Ukrainian was a moderately-resourced Slavic language, receiving limited attention in NLP research. The Russian invasion triggered rapid mobilization: the CLARIN Knowledge Centre for Ukrainian NLP (UkrNLP-Corpora) was established in 2023 [13], the Ukrainian Natural Language Processing Workshop (UNLP) expanded to four editions by 2025, and researchers developed numerous datasets for disinformation detection, sentiment analysis, and propaganda identification on Ukrainian social media [14]. This research is explicitly framed around information warfare: detecting "manipulative narratives" and "rhetorical manipulation techniques used to influence Ukrainian Telegram users" [15]. The geopolitical urgency has attracted Western funding and research attention that Ukrainian might not otherwise have received.

The US-China technology competition further illustrates how strategic rivalries shape NLP investment trajectories. China invested $125 billion in AI in 2025, representing 38% of global AI investment, with NLP receiving 11% of this allocation [16]. Chinese companies, including Baidu, Alibaba, and Tencent, are developing competitive large language models (Qwen, Yi, DeepSeek) trained on massive Chinese corpora, partly driven by the US export controls on advanced semiconductors [17]. This competition benefits Mandarin Chinese language resources, but does not extend to minority languages within China (Tibetan, Uyghur, Mongolian), which remain severely underrepresented despite large speaker populations. This indicates that geopolitical attention flows to languages of strategic interest to major powers, not necessarily to the most linguistically marginalized communities.

The patterns identified above reveal a troubling dynamic for truly underrepresented languages: research investment follows geopolitical salience rather than linguistic need. Languages become "high-resource" when powerful states perceive strategic value in processing them for intelligence gathering, countering disinformation, or economic competition. The 42 languages in our survey with only a single study lack this geopolitical visibility. For researchers working on such languages, this suggests that framing research around emerging strategic concerns (e.g. climate migration, regional stability, pandemic communication) may attract funding that purely linguistic motivations cannot.

Relation to other surveys and academic contributions. Some recent surveys have explored various aspects of multimodal language models. Zhao et al. [21] provided a comprehensive overview of LMM architectures, training strategies, and applications, while Wang et al. [24] focused on pre-training techniques and model architectures. Additional surveys have examined related areas, including LLMs [9,18–20,23,25,27,28,30], but they did not specifically address the unique challenges and solutions for LR languages in multimodal contexts. Alam et al. [29] explored LLMs for low-resource languages in multilingual, multimodal and dialectal settings, but they focused primarily on the capabilities of LLMs rather than presenting a comprehensive survey of techniques. To the best of our knowledge, our survey is the first to focus on multimodal learning for understudied languages.

As shown in Table 2, our work differs from previous surveys by specifically focusing on the intersection of multimodality and low-resource languages, while addressing all major techniques relevant to this domain. While several prior surveys have separately explored multimodality or low-resource languages, none has comprehensively examined both aspects across such a diverse range of languages and approaches.

In summary, our contribution is fourfold:

- We provide the first comprehensive analysis of LMMs specifically focused on LR languages, examining 117 studies across 96 languages.
- We develop a novel taxonomy (see Fig. 3) that categorizes existing approaches into six main categories: multimodal data creation, synthetic data generation, multimodal fusion techniques, visual enhancement techniques, cross-modal transfer learning and architectural innovations.
- We systematically organize the literature to enable a clear understanding of current approaches and remaining challenges in making LMMs more accessible to speakers of LR languages.
- We provide an open-source repository that includes implementation details, datasets, and benchmarks to facilitate future research in this emerging field.

Organization. The remainder of this survey is organized as follows. In Section 2, we present an overview of the constructed taxonomy and discuss research trends between 2018 and 2025. Sections 3 and 4 are dedicated to resource-oriented contributions. In Section 3, we identify and categorize common approaches for dataset creation for LR languages. In Section 4, we discuss automated data generation techniques. Sections 5–8 are dedicated to method-oriented contributions. In Section 5, we categorize and compare strategies used to fuse multiple modalities. In Section 6, we discuss techniques used to enhance machine translation by using visual information. In Section 7, we present methods that perform cross-modal transfer learning. In Section 8, we analyze architectural contributions. In Section 9, we discuss current evaluation challenges and propose several ways to address the identified challenges. In Section 10, we draw our conclusions, point out current research gaps, and propose ways to mitigate them in future work.

2. Taxonomy

To organize the diverse approaches in the rapidly evolving field of LMMs for low-resource languages, we develop a comprehensive taxonomy through a systematic analysis of the 117 papers in our survey. Our methodology involves initial coding of each paper's primary contributions and techniques, iterative refinement through thematic analysis to identify recurring patterns, and hierarchical organization of approaches based on their functional relationships and chronological development in the field. Our analysis reveals that researchers addressing the challenges of LMMs for low-resource languages typically follow a progression from resource development to architectural refinement. This progression is reflected in our taxonomy, which organizes approaches into six main categories that represent both the current state of the field and the primary research strategies for addressing challenges in the context of underrepresented languages.

In Fig. 3, we systematically organize LMMs for LR languages into six main categories. The first two categories focus on constructing high-quality resources. While the first category discusses multimodal data creation either from scratch or via extending existing datasets, the second approach centers on synthetic data generation, which automatically

Table 2
Comparison of our survey with related work on LLMs and LMMs across different focuses, modalities, languages, techniques, and additional coverage.

Survey	Focus	Multi-modality	Languages	Techniques					Additional Coverage
				Data Creation	Fusion	Visual Enh.	Transfer	Adaptation	
Gu et al. [18]	LLM-as-a-judge		High-resource						Evaluation, reliability, applications
Joshi et al. [9]	Language resources		Low-resource	✓					Digital divide
Paullada et al. [19]	Dataset development		Low-resource	✓					Data challenges
Ruder et al. [20]	Cross-lingual NLP		Low-resource				✓		Benchmarking
Zhao et al. [21]	LLMs		Both	✓			✓	✓	Pre-training, adaptation, utilization
Zhu et al. [22]	Multilingual LLM		Both	✓			✓		Cross-lingual transfer
Gan et al. [23]	Vision-language	✓	High-resource	✓	✓	✓			Pre-training objectives
Wang et al. [24]	Pre-training	✓	High-resource	✓				✓	Model architectures
Li et al. [25]	Large VLMs	✓	High-resource						Benchmark evaluations, challenges
Yin et al. [26]	LMM architectures	✓	High-resource	✓	✓			✓	Training strategies
Xie et al. [27]	Large multimodal agents	✓	High-resource						Agentic AI, evaluation methods
Xu et al. [28]	Resource-efficient models	✓	Both					✓	Efficient algorithms, system designs
Alam et al. [29]	LLMs for LR contexts	✓	Low-resource				✓	✓	Capabilities, prompting, evaluation
Mu et al. [30]	Mixture-of-Experts	✓	Both					✓	Algorithms, theory, applications
Our Survey	**LMMs for LR languages**	✓	**Low-resource**	✓	✓	✓	✓	✓	**Systematic taxonomy, 96 languages, 117 studies**

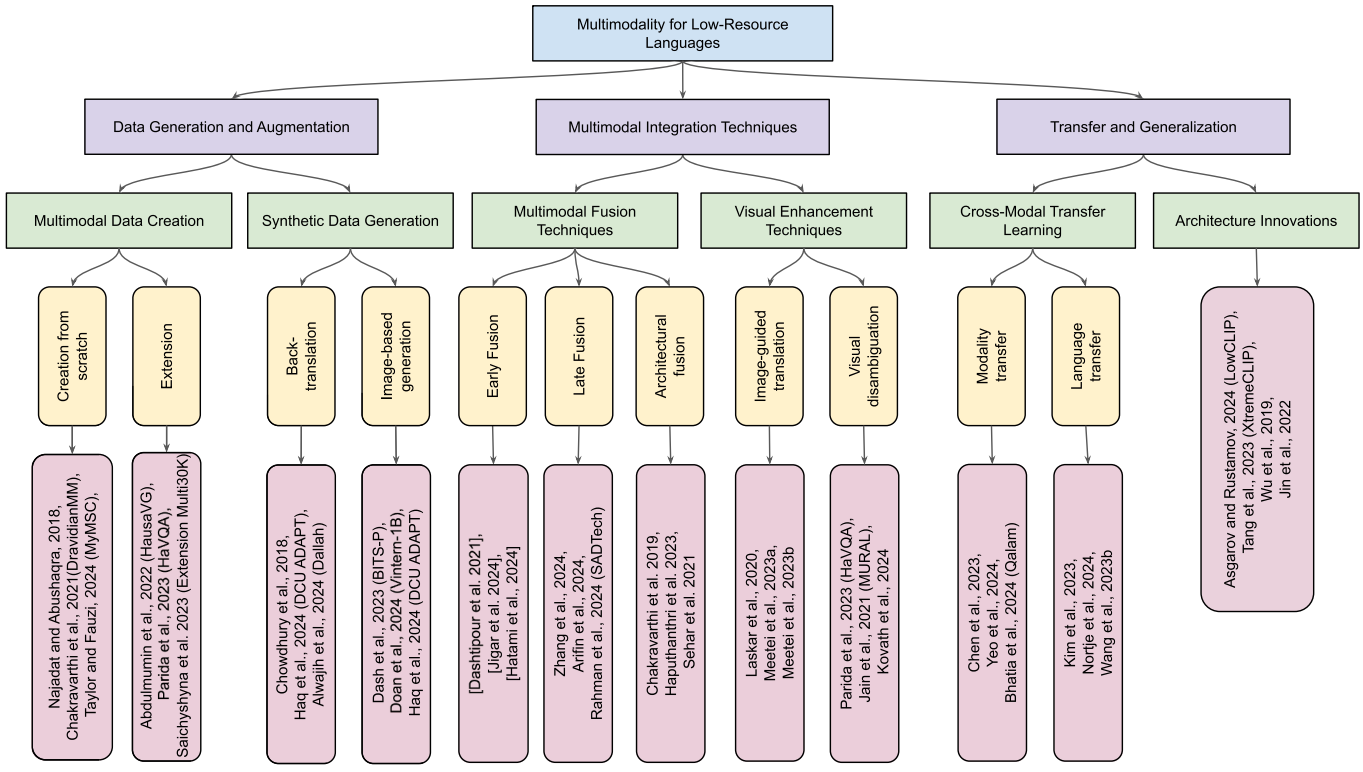

Fig. 3. High-level taxonomy of LMMs for low-resource languages. We depict six main categories (inside boxes with green background), which are further divided into subcategories, exemplified via a few representative studies.

expands available resources via back-translation and image-based generation. Building upon this work, we present several multimodal fusion techniques and provide various strategies for effectively combining this information, ranging from early and late fusion to more complex hybrid approaches. In the fourth category, we illustrate visual enhancement techniques that harness visual information through image-guided translation and visual disambiguation methods, highlighting their importance for improving translation quality and resolving ambiguities. Expanding from the single-modality solutions, the next category focuses on cross-modal transfer learning approaches that can facilitate knowledge sharing based on both modality transfer and language transfer. Finally, our last category comprises architectural innovations specifically tailored for multimodal tasks in the context of LR languages.

It is important to note that several studies naturally span multiple areas. For instance, some studies that fuse visual and textual features could also be viewed as cross-modal transfer when they leverage pre-trained vision-language models. Similarly, papers that introduce new datasets may incorporate synthetic augmentation, and architecture-focused contributions may sometimes rely on transfer-learning mechanisms. In such cases, we assign each study to the category that reflects its primary tech-

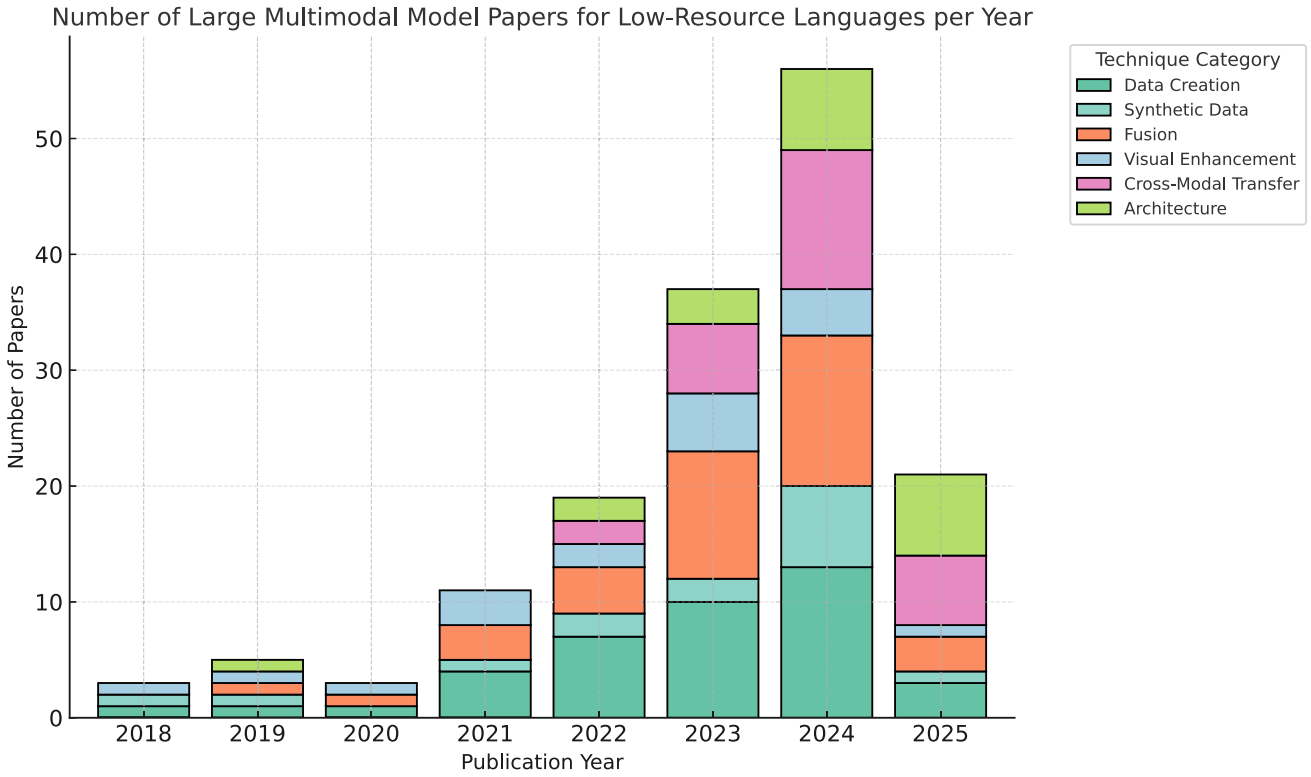

Fig. 4. Number of LMM papers for LR languages published per year (2018–2025), categorized by technique: Multimodal Data Creation, Synthetic Data Generation, Multimodal Fusion Techniques, Visual Enhancement Techniques, Cross-Modal Transfer Learning, and Architectural Innovations. Best viewed in color.

nical contribution. This principle helps maintain clear boundaries, while acknowledging the natural overlap across multimodal methods for low-resource languages.

Furthermore, to understand how these categories have evolved over time, we analyze the publication trends from 2018 to 2025. Fig. 4 shows the number of papers per year in each category, illustrating how early work focused primarily on data creation and synthetic augmentation, while recent years saw an increase in fusion strategies, cross-modal transfer, and architectural innovations, reflecting the shift toward foundation-model-based approaches.

Our taxonomic structure not only organizes existing research, but also highlights the inter-dependencies between different approaches and reveals gaps in current research, particularly in the exploration of complex multimodal combinations involving video and speech for low-resource languages. We structure the remainder of this article according to our novel taxonomy shown in Fig. 3.

3. Multimodal data creation

There are two main approaches to create multimodal datasets for LR languages. The first is based on multimodal dataset creation from scratch, while the second is based on using an existing resource as a starting point. We next discuss papers introducing novel datasets based on the two alternatives.

Dataset creation from scratch. Dataset creation from scratch has emerged as a crucial approach for enabling multimodal research in LR languages, particularly for sentiment analysis and specific language tasks. Multiple research teams have focused on creating specialized datasets through direct data collection and annotation, such as collecting Arabic videos with multimodal features for sentiment analysis [31], building comprehensive Tamil and Malayalam video review datasets [32], developing new corpora for languages such as Malay [33], creating speech translation resources for Fongbe [34] and compiling Arabic multimodal sentiment collections [35]. A significant trend has been the creation of meme-based datasets, with efforts focused on Bengali, through MemoSen and MUTE [36,37], and Romanian, through RoMemes [38], all incorporating multiple levels of annotation.

These dataset creation efforts have expanded beyond sentiment analysis to encompass other crucial applications, such as sign language recognition with ArabSign [39], and multi-purpose datasets like BIG-C for Bemba [40]. Additionally, the creation of a Manipuri-English parallel corpus with accompanying audio recordings for speech-to-text translation [41] provides an important resource for research in low-resource languages. More recently, Farsi et al. [42] introduced a comprehensive suite of multimodal datasets for Persian, covering tasks such as VQA, OCR, visual abstraction reasoning, and cultural knowledge grounding. These projects typically involve careful quality control by using multiple annotators, standardized recording environments, and expert validation, demonstrating a shift toward building comprehensive resources specifically designed for LR languages, rather than relying on translation or transfer from high-resource languages.

Dataset extension. In addition to building data from scratch in the context of LR language understanding, there have been several efforts for leveraging existing datasets of rich-resource languages and building upon them. Sen et al. [43] introduced the Bengali Visual Genome (BVG) dataset, which extends the Visual Genome dataset [44] with Bengali translations and annotations, enabling the development and evaluation of multimodal models for Bengali-English machine translation (MT) and image captioning. Similarly, Abdulmumin et al. [45] created the Hausa Visual Genome (HaVG) dataset by translating a subset of the Visual Genome dataset into Hausa, providing a valuable resource for English-to-Hausa multimodal MT. Building upon prior work and continuing the focus on the Hausa language, Parida et al. [46] introduced

the Hausa Visual Question Answering (HaVQA) dataset, which adapts question-answer pairs from the Visual Genome dataset to the Hausa language through manual translation, creating the first visual question-answering (VQA) dataset for Hausa. Extending this trend to Indian languages, Parida et al. [47] introduced OVQA, a multimodal dataset for the Odia language, by translating over 6000 question-answer pairs and associated captions from the Visual Genome dataset into Odia. Similarly, Anwar et al. [48] introduced MuAViC, a multilingual audio-visual corpus providing 1200 hours of audio-visual speech across 9 languages, establishing the first open benchmark for audio-visual speech-to-text translation.

Apart from the focus on African languages, Saichyshyna et al. [49] extended the Multi30K dataset [50] to include Ukrainian translations and captions, facilitating integrated vision and language research in Ukrainian. More recently, Lovenia et al. [51] presented SEACrowd, a comprehensive multilingual and multimodal data hub and benchmark suite for Southeast Asian languages, which covers 13 tasks across three modalities (text, image, and audio) and 38 Southeast Asian indigenous languages, while Lent et al. [52] introduced CreoleVal, an extensive collection of benchmarks for 28 Creole languages, addressing the significant resource gap for these historically marginalized language varieties.

4. Synthetic data generation

An alternative approach to efficiently create multimodal datasets for LR languages relies on synthetic data generation. Unlike traditional dataset creation, which typically involves intensive manual data collection, human annotation, and domain-specific curation, synthetic data generation leverages existing resources and automated techniques to produce new multimodal content with minimal human input. This distinction is critical, as synthetic methods offer a scalable alternative for low-resource settings, where manual annotation is often costly or infeasible.

Back-translation. A common approach for synthetic data generation relies on the usage of back-translation, which has proven to be an effective technique to enhance the data for multilingual MT (MMT) in LR language pairs. This technique works by translating text from an HR language into an LR language, and then back again, helping to generate additional aligned examples without requiring human involvement. Dutta Chowdhury et al. [53] demonstrated the effectiveness of this technique for training a neural MMT system in the context of LR language pairs by leveraging the Flickr30k dataset [54] and translating the source-language (English) captions to the target LR language (Hindi). Meetei et al. [55] extended this approach for low-resource multimodal neural machine translation in the news domain for English-Hindi. In the WMT24 English-to-Low-Resource Multi-Modal Translation task, Haq et al. [56] showcased the effectiveness of back-translation for Hindi. Another use case of back-translation was shown by Alwajih et al. [57], who, starting from English-based image-text pairs, employed translation to Arabic, as well as back-translation. This was necessary for evaluating the quality of the translation, before passing the data to humans for Arabic dialect translation and training a dialect-aware LMM, named Dallah. However, the consistency of back-translated synthetic data has been a concern. To address this issue, Wang et al. [58] proposed a framework to improve the robustness of models when adapting grounded VQA models to LR languages, aiming to improve the performance without relying on machine-translated data. Wang et al. [59] further explored this challenge by introducing noise-robust learning for cross-lingual cross-modal retrieval to handle translation noise in machine-translated sentences.

Image-based generation. Another mainstream approach for synthetic data generation uses images as a starting point. In the case of Indic language multimodal MT [60], synthetic images generated by diffusion models were deemed beneficial, their main goal being that of capturing the complexity of the target domain, and augmenting the existing image dataset. Similarly, Haq et al. [56] created exhaustive image descriptions in addition to the already existing short region-based descriptions. Doan et al. [61] utilized image-based generation for several purposes, such as description generation and relevant information extraction, to develop Vintern-1B, an efficient LMM for Vietnamese. Nath et al. [62] applied this approach for image caption generation in the low-resource Assamese language using an encoder-decoder framework that combines CNNs and RNNs to generate descriptions from images. Jiang et al. [63] expanded these approaches with multimodal seed data augmentation for the low-resource audio Latin Cuengh language, demonstrating how seed data can enhance intelligent recognition and comprehension of low-resource dialects. Collectively, these studies demonstrate the versatility and effectiveness of synthetic data for tackling a diverse set of multimodal tasks in the context of LR languages.

An emerging approach that avoids both traditional back-translation or image-based methods consists in leveraging the outputs of Vision-Language Models (VLMs). Qu et al. [64] generated multilingual responses for image-text inputs, translated them into English, and compared them with trusted references to detect hallucinations. These automatically mined hallucination-aware pairs are then used for direct preference optimization [65], enabling scalable fine-tuning without manual annotations, especially useful for low-resource languages.

Another innovative approach focuses on optimizing the composition of training data itself. Shukor et al. [66] developed systematic methods to determine optimal domain weights for multimodal pre-training using scaling laws, validating their approach across Large Language Models, Native Multimodal Models, and Large Vision Models. This methodology provides principled alternatives to costly trial-and-error approaches for data mixture optimization, particularly valuable in resource-constrained settings, typical for low-resource language development.

Data sovereignty concerns. Synthetic data generation introduces risks beyond technical quality. Back-translation and LLM-based augmentation propagate source-language biases to target languages, potentially encoding cultural assumptions misaligned with target communities [67]. More critically, the CARE principles [4] assert that Indigenous communities must retain authority over their linguistic data, a requirement that synthetic generation pipelines rarely accommodate. Evidence from Sámi language technology demonstrates the consequences: LLMs trained on available corpora without community oversight produce outputs that appear valid to non-speakers, but constitute nonsense to native speakers [68]. We thus recommend that synthetic data pipelines incorporate community validation protocols and explicit data governance agreements prior to deployment.

5. Multimodal fusion techniques

Multimodal fusion refers to the process of combining information from different modalities (such as text, images, audio) to make more informed predictions or generate better outputs. Fusion can be seen as the “meeting point” where information from separate sensory channels comes together, similar to how humans integrate what they see, hear and smell to understand their environment. The choice of fusion strategy significantly impacts performance, especially in low-resource settings, where each modality might provide crucial complementary information that others lack. Below, we describe the primary approaches to fusion that represent different philosophies about when and how this integration should occur.

We identify three distinct types of fusion approaches employed in multimodal learning, categorized into early fusion, late fusion, and architectural fusion approaches. An overview of the different fusion strategies is provided in Fig. 5. The diagram depicts the various ways in which textual, visual and auditory features can be combined at different stages to enable effective integration of multimodal information. In Table 3, we provide a summary of computational requirements and efficiency trade-offs for a series of representative fusion approaches. We further discuss each fusion strategy independently, referring to the computational requirements and trade-offs along the way.

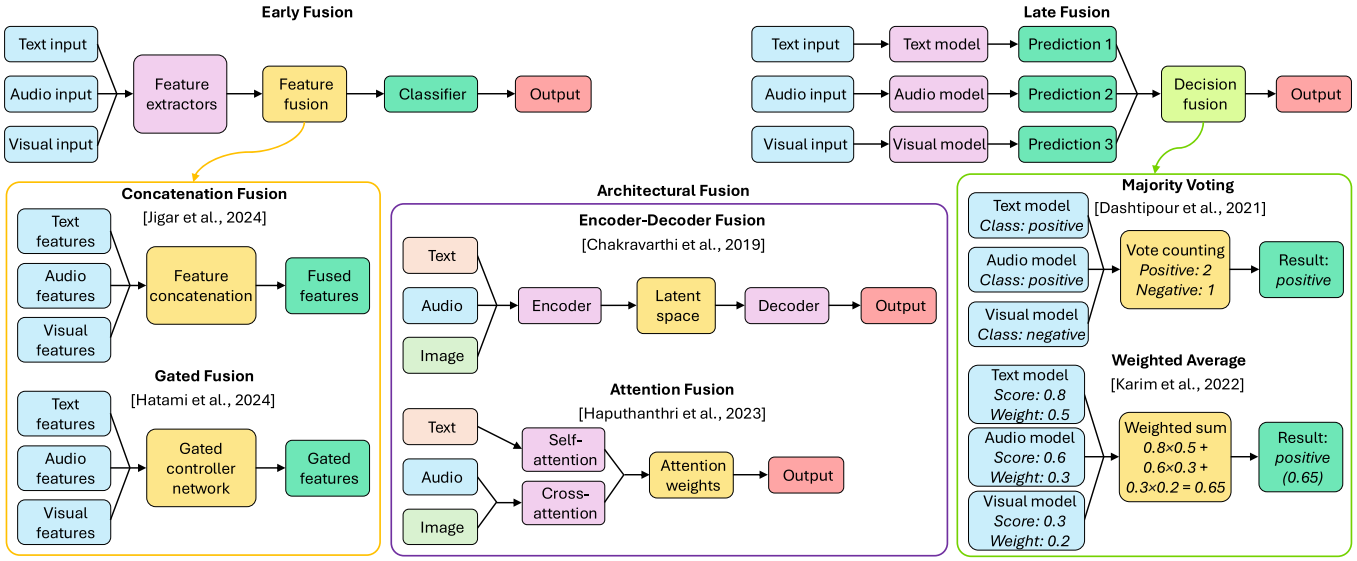

Fig. 5. An overview of various fusion strategies employed in LMMs, categorized into early fusion, late fusion, and architectural fusion approaches. Early fusion combines features from different modalities (text, audio, and visual) using feature extractors and fusion techniques, before passing them to a classifier for the final output. Concatenation fusion directly concatenates features from different modalities, while gated fusion employs a gate controller network to regulate information flow between modalities. Late fusion processes each modality using separate models, then combines their predictions using decision-level fusion methods, such as majority voting or weighted averaging. Architectural fusion approaches, such as attention fusion and encoder-decoder fusion, provide more sophisticated methods for multimodal integration. Attention fusion leverages self-attention layers and learned attention weights to selectively focus on relevant features across modalities.

Early fusion. Early fusion, also known as feature-level fusion, involves combining features from different modalities at the input level before passing them through a unified model [69,70]. Early fusion can be conceptualized as "combining ingredients before cooking", i.e. all modalities are mixed at the beginning of the processing pipeline. This allows the model to learn cross-modal interactions from the start, potentially capturing subtle relationships between modalities.

In Persian sentiment analysis, Dashtipour et al. [71] demonstrated the effectiveness of early fusion by combining acoustic, visual, and textual features through a context-aware framework, achieving 91.39% accuracy with A + V + T concatenation. Similarly, Al-Azani et al. [72] showed that early fusion of textual, auditory, and visual modalities achieved over 94% accuracy for Arabic sentiment analysis.

The shared task on Tamil and Malayalam multimodal sentiment analysis [73] also revealed that early fusion techniques are particularly effective for handling code-mixed content and cultural nuances specific to these languages [74].

For Amharic hate speech detection in memes, Jigar et al. [75] employed concatenation, directly combining visual features from memes with textual features, achieving 75% accuracy and demonstrating the effectiveness of this straightforward approach for LR languages [76]. The integration of multimodal features through gating mechanisms has shown particular promise in LR scenarios, as demonstrated in English-to-Low-Resource translation tasks for Hindi, Malayalam, Bengali, and Hausa, where Hatami et al. [77] used gated fusion to selectively combine visual and textual information. This approach was further validated by Alalem et al. [78] in their Audio-Text Fusion model for English and Egyptian Arabic, where they employed Group Gated Fusion to dynamically control the flow of information between modalities, achieving superior performance over traditional fusion methods.

From a computational perspective, early fusion approaches such as Multi-Representative Fusion (MRF) [79] demonstrate that competitive results can be achieved on consumer-grade hardware (GTX 1080Ti with 11 GB VRAM), reaching 84.1% accuracy on the ICT-MMMO dataset within 100 epochs. However, early fusion typically requires 2–3× more memory during training due to joint feature processing, and demands strict temporal alignment between modalities.

Late fusion. Late fusion, also known as decision-level fusion, combines predictions from separate modality-specific models at the decision stage rather than fusing features early in the pipeline [84,87]. Late fusion can be conceptualized as "requesting multiple expert opinions and then voting on a final decision" [83,86]. In this approach, each modality is processed by its own specialized model, which becomes an expert in that particular type of data. Only after these individual experts have made their predictions are the results combined. This is particularly valuable when certain modalities might be missing or corrupted in real-world applications [88].

Two popular late fusion strategies are weighted averaging and majority voting. In weighted averaging, the predictions from different modalities are combined using a weighted sum, with weights determining the contribution of each modality to the final decision [81,89]. The weights can be uniform or learned to optimize performance. Majority voting employs gating mechanisms to control information flow between modalities and determine which modality should be emphasized [90]. For example, Dashtipour et al. [71] used gating networks to adaptively combine predictions from audio, visual and textual models based on their estimated reliability for Persian sentiment analysis. Their results showed that intelligent fusion using gates improved performance compared with simple averaging, highlighting the benefits of selective information integration in multimodal systems.

Late fusion strategies are especially suitable for resource-constrained environments due to their flexibility and lower memory requirements. Since each modality is processed by independent models, the system can continue functioning when one modality is unavailable, enabling graceful degradation with missing inputs [87,88]. For Arabic rumor detection, Albalawi et al. [82] achieved 83.83% accuracy with late fusion (MARBERTv2 + VGG-19 ensemble), compared with 85.57% for early fusion, demonstrating that the 1.7% performance gap is often smaller than the computational cost savings. Late fusion also enables parallel training of modality-specific models, reducing wall-clock time by 25–40% compared with end-to-end early fusion training [84].

Architectural fusion. Architectural fusion comprises more sophisticated integration methods that go beyond simple concatenation or averaging of features. Encoder-decoder fusion can be understood as a "trans-

Table 3
Computational requirements and efficiency trade-offs for multimodal fusion techniques in low-resource settings. Bold values indicate configurations accessible for researchers with limited computational resources. A dash line indicates that the respective information is not specified in the original publication.

Method/Model	Fusion Type	GPU Req.	Training	Params	Performance	Key Trade-off
Early Fusion Approaches						
MRF [79]	Early	**1080Ti 11GB**	**100 ep.**	≈50M	84.1% Acc	Noise-robust; needs multiple representations
ViT + mBERT [80]	Early	–	**40 ep.**	≈200M	72.4% Acc	High #params; moderate accuracy
Swin + XLM-RoBERTa [80]	Early	–	**40 ep.**	≈280M	75.8% Acc	Best early fusion; heavier
A + V + T Concat [71]	Early	–	–	≈30M	91.4% Acc	Simple; sync-sensitive
BiLSTM Multimodal [75]	Early (Concat)	–	**32–64 ep.**	≈10M	75.0% Acc	Low #params; limited complexity
Late Fusion Approaches						
XLM-R + DenseNet [81]	Late	**GTX 1050**	**5-fold CV**	≈400M	83.0% F1	Best multimodal; high memory
MARBERTv2 + Ensemble [82]	Late	–	**100 ep.**	≈180M	85.6% Acc	Robust to missing modalities
Intermediate / Architectural Fusion						
SentimentFormer [80]	Intermediate	–	**30 ep.**	≈220M	79.0% Acc	Best overall; balanced cost
AVTF-TBN [83]	Attention	RTX 3090 24GB	300 ep.	≈100M	78.0% F1	High compute; medium accuracy
CNN-LSTM Tagalog [84]	Intermediate	–	**12 h**	≈20M	89.5% Acc	25% faster than A + V
Encoder–Decoder & Advanced Fusion						
URSA (3D-CNN + BLSTM) [85]	Feature-level	–	–	128 + 64 cells	95.4% Acc	Feature > decision fusion
Feature-Extract [86]	Sep. + Merge	**T4 16GB**	–	**8.48M**	93.3% Acc	Low #params; specialized pipeline

lation system" between modalities, i.e. information from each modality is first converted into a common "language" (shared representation space) by encoders, before being decoded into the final output [70,80]. This allows the model to find complex mappings between very different data types. For example, Chakravarthi et al. [91] employed an encoder-decoder framework with phonetic transcription to improve machine translation between Dravidian languages, while Sehar et al. [85] utilized an encoder-decoder architecture to fuse audio, video and text features for Urdu sentiment analysis. Similarly, Meetei et al. [92] showed that encoder-decoder fusion of correlated modalities can enhance translation quality for LR languages. The key advantage of encoder-decoder architectures is their ability to first encode input features from different modalities into a shared representation space before decoding them into the target output.

Attention-based fusion has also proven to be highly effective for multimodal integration [93]. As shown by Haputhanthri et al. [94] for Sinhala sign language recognition, attention mechanisms allow the model to dynamically focus on the most relevant features across modalities. Yang et al. [95] successfully employed attention fusion for Mongolian sentiment analysis by combining features from audio, text and visual inputs. Zhang et al. [83] demonstrated that attention-based fusion of multimodal data improves depression risk detection by allowing the model to attend to salient information across audio, video and text modalities. The ability of attention mechanisms to learn dynamic weights between modalities makes them particularly suitable for tasks requiring adaptive integration of complementary sources.

Intermediate and architectural fusion approaches offer a balance between performance and accessibility. SentimentFormer [80] achieves the highest Bangla meme accuracy (79.04%) with only 30 epochs, outperforming both early (75.83%) and late fusion (74.80%) on the same dataset. At the high-resource end, attention-based models such as AVTF-TBN [83] require an RTX 3090 (24 GB) and 300 epochs for clinical-grade depression detection accuracy.

Comparative analysis of fusion techniques. Each fusion approach presents distinct advantages and challenges in the context of LR languages [69,84]. Early fusion enables deep interaction between modalities from the start, but can be computationally expensive and may suffer when one modality is noisy. Late fusion offers flexibility and robustness when modalities are missing, but may miss important cross-modal interactions [87]. Architectural fusion approaches show promise in capturing complex relationships between modalities, but require careful tuning and substantial computational resources. A notable innovation in this space is the Multi-Representative Fusion (MRF) mechanism [79],

Table 4
Performance comparison of early, late and intermediate fusion for low-resource languages. Best score on each row is highlighted in bold.

Language/Task	Early	Late	Intermediate	Best Strategy
Bangla Memes [80]	75.83%	74.80%	**79.04%**	Intermediate
Arabic Rumors [82]	**85.57%**	83.83%	–	Early
Urdu Sentiment [85]	**95.35%**	91.23%	–	Early
Javanese Emotion [86]	71.15%	–	**93.32%**	Separate Processing
Amharic Memes [75]	**75.00%**	–	–	Early
Persian Video [71]	**91.39%**	90.32%	–	Early

which generates diverse representations for each modality and selectively chooses the best fusion via attention. This approach has shown particular promise in handling noisy inputs, achieving state-of-the-art performance on several LR sentiment analysis benchmarks.

Handling noisy modalities. A critical consideration for real-world deployment is robustness to corrupted modalities. The MRF mechanism [79] addresses this by generating multiple diverse representations for each modality and using attention to select the most informative fusion. When acoustic features are corrupted, MRF automatically reduces their contribution (from approximately 15% to <5% of the final prediction), maintaining robust performance. However, MRF fails when all three modalities are simultaneously noisy for utterances critical to prediction. For Javanese emotion recognition [86], separate modality processing achieves an accuracy of 93.32% compared with 71.15% for joint processing, specifically because independent processing minimizes interference when one channel contains noise.

Architectural complexity considerations. Our analysis suggests that the additional complexity of architectural fusion is justified in three cases: (1) when cross-modal interactions are semantically rich and task-critical, as in Sinhala sign language recognition [94] and Mongolian sentiment analysis [95], where temporal alignment between visual gestures and linguistic features requires learned attention weights; (2) when modalities have different noise characteristics or information densities, as demonstrated for Urdu sentiment analysis where feature-level fusion (95.35%) substantially outperformed decision-level fusion (91.23%) [85]; and (3) for clinical or safety-critical applications where prediction errors have serious consequences [83]. Conversely, for rapid prototyping or tasks where text modality dominates (e.g. in Bengali hate speech detection, where a text-only XLM-RoBERTa achieves $F1 = 0.82$ vs. $F1 = 0.83$ for the multimodal pipeline [81]), simpler approaches may be preferable.

Table 5
Decision guide for selecting the fusion strategy based on constraints and requirements. ✓✓ = Strongly recommended, ✓ = Suitable, ∼ = Acceptable, ✗ = Not recommended. Based on empirical findings from [79,80,83,87].

Requirement/Constraint	Early	Late	Architectural
Missing modality robustness	✗	✓✓	✓
Noisy input handling	✗	✓	✓✓ (MRF)
Low memory (<8GB VRAM)	✗	✓✓	∼
Fast training (<50 epochs)	∼	✓✓	∼
Maximum accuracy	✓✓	∼	✓✓
Cross-modal interactions	✓✓	✗	✓✓
Rapid prototyping	∼	✓✓	✗
Clinical/safety-critical	∼	✗	✓✓

In Table 4, we present the performance of fusion strategies across different low-resource languages and tasks, while in Table 5, we provide a decision guide based on specific constraints. Early fusion generally achieves the highest accuracy (e.g. 95.35% for Urdu, 91.39% for Persian video analytics), but the performance gap between strategies is often smaller than the computational cost difference. For researchers with limited computational resources (single GPU, <16GB VRAM), we recommend starting with lightweight early fusion models such as BiLSTM (≈10M parameters) to establish baselines, before progressing to intermediate fusion with efficient architectures for improved performance.

6. Visual enhancement techniques

Visual enhancement techniques aim to improve MT quality by leveraging visual information to provide additional context and resolve ambiguities in the source text. These techniques broadly fall into two main categories: image-guided translation, which uses visual features to enhance the overall translation process, and visual disambiguation, which specifically focuses on resolving ambiguous words/phrases via visual context.

Image-guided translation. A promising direction for improving translation quality for LR languages is the use of image-guided translation approaches. Dutta Chowdhury et al. [53] showed that augmenting neural MT systems with visual features extracted from a pre-trained CNN and integrated into an encoder-decoder architecture can improve translation quality, achieving a bilingual evaluation understudy (BLEU) score of 24.2 for Hindi to English translation. Building upon these ideas, Laskar et al. [96,97] developed a multimodal neural MT system with a bidirectional RNN encoder and a doubly-attentive decoder for English-Hindi translation. Their system, which combines visual and textual features, and employs pre-trained word embeddings from monolingual data, outperforms a text-only baseline, achieving a BLEU score of 33.57 versus 27.75 on the test set.

Subsequent studies [56,98–101] have demonstrated the effective use of visual information for improving MT in LR settings, particularly for the English-Hindi language pair. Meetei et al. [100] proposed a video-guided multimodal MT framework that incorporates spatio-temporal video features, showing improvements of up to +4.2 BLEU over text-only baselines for English to Hindi translation, while Meetei et al. [101] explored multimodal translation for news domain data, showing that ResNet-based image features outperform VGG-based features and improve BLEU scores by +1.8 points. Additionally, Shi et al. [99] explored different approaches for extracting and integrating image features using VGG and ResNet models, achieving a +3 BLEU improvement over text-only translation. Another contribution is presented by Gain et al. [98], who showed how visual context enhances translation robustness under noisy conditions (e.g. OCR errors), even when image relevance is reduced. Extending this line of work, Tayir et al. [102] demonstrated that visual context can bridge structural gaps in distant language pairs, such as English-Uyghur, by introducing a visual masked language modeling approach for unsupervised multimodal MT. Similarly, Tayir et al. [103] improved translation for the same language pairs by harnessing varying-granularity image features in low-resource settings.

More recently, Haq et al. [56] presented a context-aware transformer model that integrates visual features via BERT encoding, demonstrating consistent improvements over text-only baselines. In a related direction, Lekshmy et al. [104] developed an English-Malayalam vision-aided translation system for visually impaired users, employing multimodal machine learning techniques to perform object recognition and generate translated descriptions in real-time.

Across all studies, qualitative analyses confirmed that visual cues are particularly beneficial for handling rare words and domain-specific terms, with both image and video modalities helping to resolve ambiguity and improve translation quality in LR scenarios.

Visual disambiguation. While image-guided translation aims to enhance overall translation quality by integrating visual context, the visual disambiguation techniques focus on task-specific ambiguities by grounding them in visual information. In this regard, studies revolving around the creation of Visual Genome datasets for LR languages, such as Hindi [105], Bengali [43] and Hausa [45], have played a pivotal role in advancing visual disambiguation techniques. Building upon previous work, Parida et al. [106] explored this line of research by developing a multimodal NMT system for English-Bengali using object tags extracted from images as auxiliary input, while Nortje et al. [107] introduced an innovative few-shot learning approach for visually-prompted keyword localization in Yoruba.

Several studies have investigated the use of visual features for disambiguation [96,108–112]. For example, Jain et al. [108] highlighted the benefits of using visual features for disambiguation. Their model, called MURAL, shows strong performance on text-to-image retrieval tasks, where it manages to retrieve relevant images for ambiguous queries. This finding is also supported by the qualitative examples, where MURAL successfully disambiguates word senses based on visual context. In addition, Kovath et al. [110] proposed a co-attention mechanism for Malayalam VQA that allows the model to jointly learn attention over both textual and visual inputs, demonstrating improved performance over baselines using only textual features.

Comparative analysis of visual enhancement techniques. Image-guided translation consistently demonstrates performance improvements over text-only baselines for LR languages, though effectiveness varies with dataset size and translation direction. These approaches excel at handling semantic ambiguities and culturally-specific concepts, but their success depends heavily on the quality of extracted visual features. A key limitation is the reliance on high-quality image-text pairs, which are often scarce for LR languages. While these techniques improve translation quality, they also introduce computational overheads. Future work should focus on developing more efficient visual feature extraction methods and better approaches for leveraging visual information with limited paired data.

7. Cross-modal transfer learning

Cross-modal transfer learning represents a critical approach for LR languages, allowing models to harness knowledge from data-rich modalities or languages to improve performance in resource-constrained settings. Unlike traditional transfer learning, which operates within a single modality, cross-modal transfer must bridge the significant gap between different types of data representations. This is conceptually similar to how a person might use their understanding of written language to help learn a sign language, or how knowledge of one spoken language can facilitate learning another. In the context of low-resource languages, two primary transfer directions have emerged: modality transfer, which moves knowledge between different data types (e.g. from text to speech), and language transfer, which leverages high-resource languages to improve performance in low-resource ones.

Modality transfer. Modality transfer addresses the challenge of transferring knowledge between different modalities to improve performance

Table 6
Overview of architectural innovations for low-resource multimodal learning. V = Vision, T = Text, A = Audio. Although Cycle-Attn is evaluated on EN and DE (high-resource), it is included as a key methodological reference. The authors simulated a low-resource scenario using the limited Multi30K dataset to demonstrate the efficacy of knowledge transfer from a rich monolingual corpus (EN) via cycle consistency constraints.

Model/Method	Year	Languages	Modalities	Task	Approach Category	Key Innovation
Cycle-Attn [120]	2019	EN, DE	V, T	Image Captioning	Translation + Alignment	Cycle consistency constraint for cross-lingual alignment
Multi-task Adversarial [121]	2022	EN, HI	T, A	Sentiment Analysis	Adversarial Learning	Cross-lingual transfer via shared embeddings
FEWVLM [122]	2022	EN, HI	V, T	VL Understanding	Prompt Engineering	Few-shot prompting with moderate-size VLM
Amharic Captioning [123]	2023	Amharic	V, T	Image Captioning	Attention-based DNN	Visual attention + Bi-GRU decoder
Auxiliary CTC [113]	2023	102 langs	A, T	Multilingual ASR	CTC Conditioning	LID-conditioned auxiliary objectives
Sanskrit-Malayalam NMT [124]	2022	SA, ML	T, A	Machine Translation	Multimodal NMT	Morphology + WSD embedding fusion
XtremeCLIP [125]	2023	EN, HI	V, T	VL Understanding	Parameter-efficient	Prototype affinity matching (5-7K params)
LowCLIP [126]	2024	Azerbaijani	V, T	Image Retrieval	Efficiency-first	mBERT + lightweight image encoders
Yoruba ASR [127]	2024	EN, YO	T, A	Speech Recognition	Transfer Learning	MFCC-based acoustic modeling
Llama 3 [128]	2024	200 langs	V, T	General Multimodal	Foundation Model	Native multimodal MoE architecture
DeepSeek-V3 [129]	2024	14 langs	V, T	General Multimodal	MoE Architecture	MLA + FP8 training efficiency
Claude 4 [130]	2025	15 langs	V, T	General Multimodal	Foundation Model	Dual-mode reasoning operation
Apple AFM [131]	2024	16 langs	V, T	On-device/Server	Distillation + QAT	2-bit quantization for edge deployment
MMaDA [132]	2025	60+ langs	V, T	Multimodal Diffusion	Unified Diffusion	Discrete diffusion language modeling
MixLoRA [133]	2024	15 langs	V, T	Instruction Tuning	Dynamic PEFT	Conditional mixture routing for adaptation

on low-resource tasks. This approach is particularly valuable when certain modalities have more abundant data than others. For example, text data is often easier to collect than paired speech data for many languages. The fundamental challenge lies in bridging the representational gap between modalities, since text operates in a discrete symbolic space, while speech and vision exist in continuous signal spaces with very different statistical properties. Successful modality transfer requires finding meaningful mappings between these different representational spaces. A diversity of approaches has been used to achieve modality transfer. Chen et al. [113] proposed a progressive transfer learning strategy that leverages both general pre-training (Kinetics-400 for visual and CC25 for language) and domain-specific pre-training (sign-to-gloss translation) to bridge modalities for sign language translation. Amalas et al. [114] introduced a data-driven approach to select source languages and demonstrated that multilingual pre-training outperforms monolingual pre-training for text-to-speech systems. Wu et al. [115] developed a captioning approach via multi-objective optimization that addresses the challenge of utilizing both triplet datasets (image, HR language, LR language) and large-scale paired datasets during training. Yeo et al. [116] tackled LR visual speech recognition by using Whisper-based automatic transcriptions to generate training labels from unlabeled multilingual audio-visual data. For Arabic handwriting recognition, Bhatia et al. [117] employed modality transfer through an architecture combining SwinV2 for visual encoding and RoBERTa for text decoding, while Tran et al. [118] demonstrated successful modality transfer for Vietnamese through extensive pre-training of both vision and language components, combined with automated data curation methods. Notably, Onuoha et al. [119] challenged the assumptions about multimodal integration through their study of Igbo minimal pairs. Their findings show that native Igbo speakers can accurately distinguish minimal pairs through audio alone, suggesting that the benefits of cross-modal integration may be more relevant for non-native speakers.

Language transfer. Language transfer is an approach to harness knowledge from HR languages to improve model performance on LR languages. Recent work demonstrated several effective strategies. For instance, Wang et al. [58] adapted MDETR to new languages by using adapters and code-switching without relying on MT data. Cheema et al. [134] presented ViLanOCR, a novel approach that adapts multilingual vision-language transformers for low-resource Urdu optical character recognition by leveraging the Swin encoder and mBART-50 decoder. Kim et al. [135] focused on learning general speech knowledge from English for lip reading, and combining it with language-specific audio features. Aruna Gladys et al. [136] proposed a multimodal representation learning framework that uses cross-lingual transfer learning to analyze sentiment in LR language datasets, demonstrating significant performance improvements for Tamil language sentiment analysis. Chen et al. [137] improved multilingual ASR by conditioning models on language identity predictions from early layers to enhance performance across numerous languages. dos Santos et al. [138] proposed to use data augmentation and contrastive learning to improve multilingual contrastive language-image pre-training (CLIP) models for LR languages. Nortje et al. [139] showed that initializing a Yoruba few-shot word learning model with weights from an English speech-image model substantially improves performance. These approaches share the common theme of transferring learned representations and knowledge from HR languages (typically English), while developing techniques to adapt and fine-tune models for target LR languages.

The effectiveness of language transfer methods varies significantly based on linguistic similarity, writing systems, and cultural context. For instance, transfer between closely related languages (such as Spanish to Portuguese) typically outperforms transfer between distant language families (such as English to Tamil). The methods described above demonstrated different approaches to this challenge: Wang et al. [58] focused on architecture adaptation through adapters, while Kim et al. [135] emphasized feature-level knowledge transfer. Meanwhile, Nortje et al. [139] showed that even initialization from a different language can provide substantial benefits. For practitioners working with specific low-resource languages, the choice between these approaches should consider both linguistic factors and computational constraints.

8. Architectural innovations

Architectural innovations for low-resource multimodal learning focus on designing model structures that can effectively leverage limited data while maintaining reasonable computational requirements. The fundamental challenge lies in balancing model capacity (ability to learn complex patterns) with sample efficiency (ability to learn from limited examples). While simply scaling down large models designed for high-resource settings is one approach, the most successful innovations in this space incorporate architectural elements specifically designed to address the constraints of low-resource scenarios. These innovations generally fall into three categories: (1) efficiency-focused adaptations of existing architectures, (2) parameter-efficient fine-tuning methods, and (3) novel architectures designed specifically for low-resource multimodal learning. In Table 6, we provide a systematic overview of these architectural innovations, categorized by approach type, supported modalities, and target tasks. Tables 7 and 8 complement this overview with quantitative analyses of computational requirements and empirical performance, enabling direct comparison across methods with varying resource constraints.

Some recent architectural innovations in the context of LR languages have focused on adapting the

Table 7
Computational requirements for architectural innovations. A dash line indicates that the respective information was not reported in the original paper.

Model	Trainable Params	Total Params	Training Duration	Hardware (per paper)	Training Data
Parameter-Efficient Vision-Language Methods					
XtremeCLIP [125]	**5-7K**	149M	20–60 min	1× A100	2K-10K samples
LowCLIP [126]	192M	192M	37 hours	1× T4	500K+ captions
$FEWVLM_{base}$ [122]	224M	224M	30 epochs	–	Few-shot (16 ex.)
$FEWVLM_{large}$ [122]	740M	740M	30 epochs	–	Few-shot (16 ex.)
Language-Specific Architectures					
Amharic Caption [123]	–	–	35 epochs	–	8K images
Cycle-Attn [120]	–	–	50 epochs	–	30K pairs
Foundation Models (for reference)					
Llama 3 405B [128]	405B	405B	3.8×10^{25} FLOPs	16K× H100	15.6T tokens
DeepSeek-V3 [129]	37B active	671B	2.788M H800 hours	2048× H800	14.8T tokens

Table 8
Performance metrics for low-resource multimodal architectures. All values are extracted directly from source papers. Baseline methods and improvement calculations are specified for reproducibility. Full FT = full fine-tuning; Aug. = augmentation; – = not applicable or not reported.

Model	Task	Dataset	Metric	Score	Baseline	Δ
Parameter-Efficient Vision-Language Methods						
XtremeCLIP [125]	Visual Entailment	SNLI-VE (10K samples)	Accuracy	62.06%	51.10% (Full FT)	+21.4%
	Visual QA	VQA v2 (10K samples)	Accuracy	59.21%	54.10% (Full FT)	+9.4%
	Image Classification	FGVC (16-shot)	Accuracy	48.30%	28.14% (Full FT)	+71.6%
LowCLIP [126]	Image Retrieval	MSCOCO (AZ)	mAP	0.80	0.70 (Base Loss)	+14.3%
		Flickr30k (AZ)	mAP	0.87	0.84 (No Aug.)	+3.6%
$FEWVLM_{large}$ [122]	Visual QA	VQAv2 (few-shot)	Accuracy	51.1%	38.2% (Frozen 7B)	+33.8%
		OK-VQA (few-shot)	Accuracy	23.1%	12.6% (Frozen 7B)	+83.3%
Language-Specific Architectures						
Amharic Captioning [123]	Image Captioning	Flickr8k (AM)	4-gram BLEU	38.8	28.5 (CNN-GRU)	+36.1%
		BNATURE (AM)	4-gram BLEU	42.7	16.4 (CNN-GRU)	+160.4%
Cycle-Attn [120]	Image Captioning	Multi30K (DE)	CIDEr	43.78	42.91 (Dual-Attn+)	+2.0%
			BLEU-4	5.71	5.54 (Dual-Attn+)	+3.1%
Foundation Models (for reference)						
Llama 3 405B [128]	General	MMLU (5-shot)	Accuracy	87.3%	–	–
DeepSeek-V3 [129]	General	MMLU-Pro (5-shot CoT)	Accuracy	75.9%	–	–

CLIP architecture [140]. One such example is the LowCLIP model [126], which replaces the original text encoder trained primarily on English text with a multilingual BERT (mBERT). The authors evaluated various lightweight image encoders, such as EfficientNet-B0 and Tiny Swin Transformer, for a more computationally efficient approach, while also targeting LR languages like Azerbaijani. To compensate for the lighter architecture and the scarcity of image-text pairs in Azerbaijani, LowCLIP leveraged synthetic data generation via MT for text features, and image augmentation techniques, such as crop and rotation, for image features. In contrast, XtremeCLIP [125] took a different approach, in which the authors introduced a parameter-efficient method that only tunes a small prototype matrix, while keeping the visual and text encoders frozen. Their model also employs contrastive learning to provide additional supervisory signals in LR settings. Collectively, these efforts extend the applicability of CLIP to multimodal image retrieval tasks.

Approaches to adapting CLIP for LR settings illustrate different design philosophies. LowCLIP takes an efficiency-first approach, focusing on reducing both the model size and data requirements through lighter architectures and extensive data augmentation. In contrast, XtremeCLIP maintains most of the original model capacity, but introduces parameter-efficient tuning to learn a small set of adaptable weights. This trade-off between model capacity and training efficiency represents a key consideration for researchers working in low-resource settings, where both data and computational resources may be constrained. The choice between these approaches depends on the specific constraints of the application scenario, e.g. LowCLIP may be more suitable for deployment on edge devices or in settings with extremely limited data, while XtremeCLIP might be preferred when maintaining representation power for complex tasks is crucial. As shown in Table 9, EfficientNet-B0 achieves competitive retrieval performance (an mAP of 0.87 on Flickr30k), while requiring 16× fewer parameters and 45× fewer FLOPs than ViT-Base. The choice between these approaches depends on deployment constraints: LowCLIP suits scenarios requiring end-to-end retraining with domain-specific data, while XtremeCLIP is preferable when rapid adaptation with minimal computational overhead is essential.

Table 9
Image encoder performance comparison for low-resource image retrieval. Results are taken from LowCLIP [126].

Image Encoder	Params	GFLOPs	Size (MB)	mAP		
				COCO	Flickr8k	Flickr30k
ResNet-50	25.6M	4.09	97.8	0.80	0.76	0.73
EfficientNet-B0	5.29M	0.39	20.5	**0.81**	0.85	**0.87**
ViT-Base	86.6M	17.56	330.3	0.71	0.80	0.70
Swin-Tiny	28.3M	4.49	108.2	0.80	**0.84**	0.79

Another approach for multimodality in the context of LR languages is introduced by Wu et al. [120]. The approach combines two existing methods, a translation-based one and an alignment-based one, into a

Table 10
Comparison of parameter-efficient fine-tuning methods for low-resource vision-language understanding. VE = Visual Entailment, VQA = Visual Question Answering, IC = Image Classification. Results are taken from XtremeCLIP [125].

Method	Trainable Params	VE	VQA	IC	Avg.	Training Time
Zero-shot	0	33.74	52.03	39.17	42.89	–
Full fine-tuning	149M	51.10	54.10	28.14	51.12	hours
LLRD	149M	57.23	53.88	31.36	53.60	hours
BitFit	176-178K	59.56	54.72	41.61	55.66	minutes
BiNor	208-210K	59.54	54.75	41.73	55.67	minutes
CLIP-Adapter	131-262K	59.21	54.21	44.88	55.45	minutes
Tip-Adapter	5-10M	59.67	54.70	45.12	55.62	minutes
XtremeCLIP	**5-7K**	**62.06**	**59.21**	**48.30**	**57.73**	**20 minutes**
LoRA ($r = 4$)	≈4K	–	–	–	65.39	minutes
LoRA ($r = 16$)	≈16K	–	–	–	65.50	minutes
MixLoRA ($E = 16$)	≈8K/layer	–	–	–	**67.17**	hours

unified architecture to improve image captioning. The framework employs a model that first generates high-quality English captions, which are then used together with the images to produce captions in the LR language. The model achieves a fine-grained alignment between visual elements and captions in both languages via a cycle-consistency constraint, outperforming existing methods on standard metrics.

Jin et al. [122] introduced FEWVLM, showing that careful prompt engineering and efficient architectural design can achieve strong performance in the context of LMM usage with either little data or computational needs. They managed to develop a moderate-size VLM that combines a sequence-to-sequence transformer with prefix language modeling and masked language modeling, introducing effective prompt engineering approaches for visual-language tasks in the LR setting. Notably, FEWVLM outperforms Frozen [141], a model which is 31× larger. In turn, Frozen achieves comparable results with PICa [142], which is 246× larger. These results demonstrate that an effective design can compensate for model size. Building on parameter-efficient approaches, Shen et al. [133] introduced Conditional Mixture of LoRA (MixLoRA) for multimodal instruction tuning, which dynamically constructs adaptation matrices tailored to each input instance, addressing task interference challenges in multimodal scenarios. For specific language pairs, Laskar et al. [143] proposed a transliteration-based phrase augmentation approach for English-Assamese translation, which allows their model to share sub-word level information, and provides better word alignment through phrase pairs. In Table 10, we quantify the size vs. performance trade-off across these methods. XtremeCLIP achieves the highest average accuracy (57.73%) across visual entailment, VQA, and image classification benchmarks, while training only 5-7K parameters, compared with 149M for full fine-tuning. This demonstrates that task reformulation as prototype affinity matching can outperform conventional fine-tuning, while using less than 21,000× fewer trainable parameters. $\text{FEWVLM}_{\text{large}}$ (740M parameters) achieves 51.1% on VQAv2, surpassing the 7B-parameter Frozen model (38.2%) by 33.8%, validating the hypothesis that architectural efficiency can compensate for raw model scale. MixLoRA further improves upon standard LoRA by 8.3% on the MME benchmark through its conditional mixture routing mechanism, which dynamically selects expert combinations based on input characteristics.

Foundation models represent a qualitatively different design point, prioritizing broad capability over resource efficiency. We include them here to contextualize the computational differences that shape research accessibility. Dubey et al. [128] introduced the Llama 3 series with models ranging from 8B to 405B parameters, officially supporting 8 languages (English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai), with experimental multilingual capabilities on a broader set via the speech interface (34 languages). Llama 3 multimodal extensions for image, video, and speech understanding were described in their technical report, but remain under development and have not been publicly released along with the paper. In a similar endeavor, Liu et al. [129] presented DeepSeek-V3, a 671B-parameter MoE language model (37B active parameters per token) with multi-head latent attention (MLA) and FP8 mixed-precision training. It is important to note that DeepSeek-V3 is a text-only language model without native vision or audio capabilities. However, we include it to put efficient training strategies into perspective (DeepSeek-V3 requires 2.788M H800 GPU-hours, costing approximately $5.6M) and better inform future multimodal model development. For edge deployment scenarios, Gunter et al. [131] introduced Apple Intelligence Foundation Models with a novel Parallel-Track MoE architecture optimized for on-device processing, supporting 16 languages with 2-bit quantization-aware training. The prevalence of MoE architectures in these recent developments demonstrates the effectiveness of expert-based scaling for multimodal tasks, as also observed by Mu et al. [30]. Additionally, Yang et al. [132] proposed a unified diffusion architecture that combines multimodal understanding with generation capabilities, offering new perspectives on architectural design for LR contexts.

For specific language families and modality combinations, several innovative architectures have been proposed. Solomon et al. [123] developed a hybridized attention-based deep neural network for Amharic language image captioning, combining a CNN encoder with visual attention mechanisms and a bidirectional GRU decoder, achieving significant improvements in terms of BLEU. Rahul et al. [124] introduced a multimodal neural machine translation system between Sanskrit and Malayalam, which embeds morphology and word sense disambiguation awareness. It utilizes both textual and speech modalities via a two-level fusion approach of transform-based feature vectors. For African languages, Rahmon et al. [127] presented a speech recognition model for Yoruba that employs acoustic and language modeling with sequential MFCC features, achieving 83% accuracy in speech-to-text conversion. For sentiment analysis, Mamta et al. [121] explored multilingual, multi-task and adversarial learning approaches to transfer knowledge from HR languages to LR scenarios, leveraging shared semantic spaces through cross-lingual word embeddings. For Arabic, Alwajih et al. [144] introduced Peacock, a comprehensive family of LMMs with strong vision and language capabilities, alongside Henna, a benchmark for evaluating culturally-aware Arabic LMMs, further helping to bridge the gap between high-resource and low-resource languages, while addressing unique linguistic and cultural characteristics.

In Table 11, we synthesize the design trade-offs across architectural approaches, organized by methodological strategy. Three principal patterns emerge from our analysis. First, *parameter-efficient adaptation* methods (XtremeCLIP, LowCLIP, FEWVLM, MixLoRA) achieve competitive performance, while reducing trainable parameters by 3–5 orders of magnitude compared with full fine-tuning, making the corresponding models accessible to researchers with limited computational resources. Second, *cross-lingual transfer* approaches (Cycle-Attn, Amharic Captioning) effectively leverage high-resource language supervision, typically English, to bootstrap performance in target languages. However, this creates structural dependency on pivot language quality and availability. Third, foundation models occupy a distinct design regime. Since Llama 3 requires 3.8×10^{25} FLOPs and DeepSeek-V3 consumes 2.788M H800 GPU-hours for pre-training, these models remain inaccessible to most research groups focused on low-resource languages. The practical implication is that parameter-efficient methods currently offer the most viable path for researchers operating under resource constraints, while foundation models may serve as upstream components for transfer learning when API access or pre-trained weights are available.

The computational requirements documented in Table 7 reveal a structural divide with sociolinguistic implications. While parameter-efficient methods like XtremeCLIP (5-7K parameters, 20 minutes on one GPU) remain accessible, foundation models require resource-intensive infrastructure (Llama 3 consumes 3.8×10^{25} FLOPs across 16K H100 GPUs; DeepSeek-V3 requires 2.79M H800 GPU hours with an estimated

Table 11
Design strategies and trade-offs in multimodal architectures for low-resource settings. Core strategies are grouped by methodological approach.

Model	Core Strategy	Advantages	Constraints
Parameter-Efficient Adaptation			
XtremeCLIP [125]	Prototype affinity matching with frozen CLIP encoders; contrastive learning for supervision	21,000× less parameters vs. full fine-tuning; 20 min training on one A100; edge-deployable	Task performance bounded by frozen backbone capacity; requires labeled prototype examples
LowCLIP [126]	Lightweight image encoders (EfficientNet-B0) with mBERT; synthetic data via MT	Trainable on consumer GPU (T4); open-source; 37 hours total training	Performance depends on MT quality; cross-domain generalization gap observed
FEWVLM [122]	Seq2seq with PrefixLM + MaskedLM; prompt-based few-shot learning	Outperforms 31× larger Frozen model; comparable with 246× larger PICa	Zero-shot performance sensitive to prompt wording; task-specific prompt engineering required
MixLoRA [133]	Conditional mixture of LoRA experts; input-dependent routing	Reduces task interference in multi-task settings; 8.3% gain over standard LoRA on MME	Routing computation overhead; requires careful expert initialization
Cross-Lingual Transfer			
Cycle-Attn [120]	Translation + alignment hybrid with cycle consistency constraint	Fine-grained visual-textual alignment; leverages English captioning supervision	Requires pre-trained English captioner; limited to language pairs with English pivot
Amharic Captioning [123]	Inception-v3 encoder + Bi-GRU decoder with visual attention	End-to-end trainable; interpretable attention weights; significant BLEU increase on BNATURE	Requires translated Flickr8k data; architecture not tested on other LR languages
Multilingual Speech			
Auxiliary CTC [113]	LID-conditioned auxiliary objectives on Whisper encoder	Scales to 102 languages; 28% relative CER reduction on FLEURS	Requires pre-extracted Whisper features; multi-stage training pipeline
Foundation Models (for reference)			
Llama 3 [128]	Dense Transformer (405B params); multimodal extensions under development	Strong zero-shot; 8 officially supported languages; open weights	3.8×10^{25} FLOPs pre-training; multimodal capabilities not yet released
DeepSeek-V3 [129]	MoE with MLA (671B total, 37B active); FP8 mixed-precision training	2.788M H800 GPU hours ($5.6M); competitive with GPT-4o on benchmarks	Text-only model; no native vision/audio; requires 2048×H800 cluster
Apple AFM [131]	On-device (≈3B) with 2-bit QAT; server PT-MoE architecture	Edge-deployable; 16 languages; image understanding capability	Proprietary; Apple ecosystem only; version-specific adapters

cost of $5.6M). This asymmetry matters because LLMs exhibit systematic bias in knowledge acquisition. Indeed, new knowledge is learned less efficiently in LR languages, transfers less effectively to them, and is overwritten more easily by HR language information [145]. The implication is that scaling alone is not sufficient to achieve equity. Therefore, architectural innovations must explicitly counteract these biases.

Federated learning offers a technical framework aligned with data sovereignty principles, enabling collaborative training without data centralization [146]. Recent work demonstrates feasibility for multilingual LR settings. For example, federated prompt tuning achieves competitive performance while preserving data locality [147], and differential privacy integration protects against gradient inversion attacks [148]. For multimodal LR applications, federated approaches could enable geographically-distributed language communities to collaboratively improve models without ceding control over culturally-sensitive audiovisual data.

9. Evaluation challenges

Evaluation remains one of the most underdeveloped aspects of research on LMMs for LR languages. While the field has made significant strides in dataset creation, fusion strategies, and architectural innovations, the ways for measuring success have not kept pace. The lack of consistent and culturally-grounded evaluation protocols severely hampers the ability of researchers to compare models, reproduce findings, or interpret results in real-world contexts.

Limitations of standard metrics across cultural contexts. Most evaluation pipelines for LR multimodal models rely on automatic metrics originally designed for high-resource and predominantly Western-centric settings. Metrics such as BLEU, ROUGE, accuracy, and F1 implicitly assume that reference annotations reflect shared cultural, visual, and linguistic grounding. This assumption frequently fails in low-resource contexts.

One such case can be observed in multimodal tasks such as visual question answering, image captioning, and meme understanding, where the visual content itself often encodes culturally-specific assumptions regarding object salience, social roles, or everyday activities. For instance, a model trained primarily on Western image datasets may fail to recognize culturally significant objects (e.g. traditional clothing, local foods, religious symbols, etc.) that are common in LR language contexts. When benchmarks are translated or minimally adapted from high-resource languages, models may achieve high lexical overlap with reference answers while still producing outputs that are culturally inappropriate, semantically misleading, or pragmatically implausible for native speakers. These issues are further exacerbated in knowledge-intensive evaluations derived from English-centric benchmarks. For example, questions about local festivals, historical events, or social customs require cultural context that translation alone cannot provide. As a result, standard metrics

may overestimate progress or mask systematic failures that are only visible through culturally-grounded evaluation.

Dataset heterogeneity and comparability issues. A second major challenge concerns dataset heterogeneity, as existing studies evaluate multimodal models on datasets with widely distinct characteristics and assumptions. Many studies rely on translated versions of high-resource benchmarks, such as extensions of Multi30K for Ukrainian [49] or Visual Genome variants for Bengali [43], Hausa [45], and Hindi [105]. While translation-based approaches enable rapid benchmark construction, they often introduce Western cultural biases and may fail to reflect authentic language use or visual grounding in target communities. In contrast, newly introduced language-specific datasets, such as DravidianMultiModality [32], RoMemes [38], and ArabSign [39], better capture genuine linguistic and cultural phenomena, but typically suffer from limited coverage, non-standardized annotation protocols, and heterogeneous quality control practices, making cross-study comparison difficult. As a result, performance improvements reported across such heterogeneous evaluation settings are often not directly comparable.

Recommendations for fair evaluation practices. Based on our analysis, we propose the following recommendations for evaluation in LR multimodal research:

- *Report multiple metrics.* Studies should report multiple complementary metrics that capture different aspects of performance. In the context of machine translation tasks, researchers should report BLEU alongside COMET, or human evaluation scores. Another example in the context of VQA tasks, exact-match accuracy should be accompanied by relaxed matching that accounts for morphological variants and, when possible, human judgment of answer correctness.
- *Perform culturally-grounded human evaluation.* In addition to a diverse set of automated metrics, we believe that human evaluation conducted by native speakers from the target language community also plays a crucial role. Evaluators should assess whether outputs sound natural to native speakers, whether they are culturally appropriate, whether they convey the intended meaning accurately, and (for VQA) whether answers are semantically correct, even if worded differently from the reference.
- *Develop and use standardized benchmarks.* The field needs publicly available test sets for LR multimodal evaluation, following examples like SEACrowd [51] for Southeast Asian languages and CreoleVal [52] for Creole languages. Such benchmarks should cover different task types (VQA, captioning, translation, classification), include culturally-accurate content created together with language communities, provide multiple correct answers to account for natural variation, and document how data was labeled.
- *Compare to sensible baselines.* Rather than reporting absolute performance in isolation, studies should contextualize results relative to unimodal baselines (text-only or vision-only) to demonstrate the benefits of multimodal approaches, random and majority-class baselines to establish task difficulty, prior work on the same dataset when available, and performance on related HR languages to quantify the LR gap.

As shown above, evaluation challenges remain a major problem in LR multimodal research, but some steps have already been taken towards fixing this gap. Although standard metrics represent a great starting point for evaluation, they are designed for English and often miss what matters for LR languages and their cultural contexts. Solving these challenges requires creating culturally-appropriate benchmarks, using multiple and diverse evaluation metrics, as well as involving language communities in the evaluation process.

10. Conclusion and future work

Conclusion. Our survey has provided a comprehensive analysis of LMM-based approaches for LR languages, comprising 117 studies across 96 languages. Vision-language combinations dominate the current research landscape (65% of surveyed works), with an increasing trend toward incorporating video and speech in recent works. We observed a concentration of research in South Asian languages (including Hindi, Bengali, Malayalam), Southeast Asian languages (Vietnamese, Javanese, Malay), Middle Eastern languages (Persian, Arabic) and African languages (Hausa, Amharic), while 42 other languages appear in only one study each.

The landscape of LMMs for LR languages has shown remarkable progress across multiple dimensions, from data creation to fusion techniques and architectural innovations. Projects like HVG, SEACrowd, and BVG highlight growing attention to creating high-quality multimodal resources for understudied languages. Recent successes with models such as Qalam, LaVy, and Amharic LLaVA [149] demonstrate that carefully designed multimodal strategies can effectively leverage limited resources, while adapting large-scale architectures for low-resource contexts.

Challenges and gaps. Our analysis reveals several critical challenges in the current landscape of LMMs for LR languages. A significant modality imbalance exists, with text-image pairs dominating research (65% of studies), while audio and video modalities remain underexplored. This gap is particularly problematic for languages with strong oral traditions, where speech, tone and gesture carry essential linguistic information, with only 32% of studies incorporating audio, despite its crucial importance for predominantly oral languages, and only 8.5% of studies incorporating a video modality. We also identified persistent dataset scarcity and uneven language representation, with just three languages (Hindi, Arabic, Bengali) accounting for a disproportionate share of research attention. Technical limitations further constrain progress, as computational constraints limit the application of advanced fusion techniques in resource-constrained environments typical for LR contexts. Current cross-modal transfer methods struggle with catastrophic forgetting and inefficient knowledge transfer, particularly for languages that are structurally distant from high-resource counterparts. The field also lacks standardized evaluation frameworks for meaningful comparison across approaches, while recent work by Shen et al. [150] highlights significant safety challenges when deploying LLMs in multilingual contexts. Finally, sociolinguistic dimensions remain underexplored, including cultural representation, algorithmic bias, and potential impacts on language endangerment and revitalization efforts. These concerns are particularly acute given power imbalances between communities speaking low-resource languages and the primarily Western institutions developing these technologies.

Our study identifies three mechanisms through which LMMs may perpetuate digital inequalities. First, language model training inherently favors languages with larger training representation [67,145], introducing a bias towards modeling HR languages. Second, benchmarks derived from English (e.g. translated MMLU) embed Western cultural assumptions that disadvantage LR language speakers even when linguistic accuracy is achieved [151], introducing cultural biases in the evaluation. Third, computational requirements exclude researchers in LR language regions from model development, creating dependency on external institutions and biasing resource access. Addressing these biases requires community-centered approaches that prioritize local capacity building, alongside technical performance metrics.

Future work. Based on the challenges identified above, we propose several key directions for future research.

For short-term development, we propose the following actionable research directions for benchmark and dataset creation: (1) extend Visual Genome-style multimodal datasets to at least 20 additional LR languages, prioritizing the 42 languages currently represented by only a single study; (2) develop speech-image paired corpora for tonal languages (e.g. Yoruba, Igbo, Fongbe), where audio modality carries critical semantic distinctions absent in text; and (3) establish a standardized "LR-MMBench" evaluation suite with culturally-adapted visual question answering tasks, following SEACrowd's multilingual methodology, but

incorporating non-Western visual contexts and evaluation protocols validated by native speakers.

To develop and improve LMMs for LR language, several concrete directions emerge from our analysis: (1) develop catastrophic forgetting mitigation strategies maintaining over 95% source-language performance, while achieving over 80% target-language performance for language pairs with fewer than 1000 parallel sentences; (2) create language-agnostic visual encoders pre-trained on culturally-diverse image collections sourced from non-Western contexts, reducing the documented Western bias in current visual representations; and (3) establish explicit source-language selection guidelines based on typological similarity metrics (syntactic distance, shared writing systems, WALS features) to maximize positive transfer for specific target languages.

Several other research gaps require attention in future. Regarding the observed modality imbalance, researchers should prioritize incorporating audio and video for LR languages with limited writing traditions, enabling more robust applications that better reflect natural communication patterns, particularly for tonal languages and those where nonverbal communication is significant. For resource development, future work should advance synthetic data generation techniques (building on HVG, ELAICHI, Vintern-1B) and improve cross-lingual transfer methodologies (extending XtremeCLIP, LowCLIP) to accommodate greater linguistic diversity, while minimizing catastrophic forgetting. To overcome limitations in resource-constrained settings, researchers should investigate efficient fusion approaches including stacking-based late fusion, tensor fusion for complex interactions, and graphical fusion leveraging graph-based representations, all adapted for computational efficiency. Advancing adaptive integration through mechanisms that dynamically adjust the contribution of each modality based on input quality and task requirements will be crucial. Building on MRF [79], future work should explore hybrid approaches that combine strengths of different fusion strategies, while maintaining computational efficiency. Finally, adopting community-centered design approaches that address sociolinguistic dimensions alongside technical advances will ensure that developments benefit the intended language communities themselves.

Finally, we advocate for mandatory community engagement through: (i) participatory design frameworks requiring documented language community involvement in dataset creation, with explicit data governance and benefit-sharing agreements; (ii) open-source, mobile-first data collection libraries suitable for field conditions, where many LR languages are spoken; and (3) standardized model cards for LR multimodal systems, documenting limitations, cultural biases, and appropriate use cases, ensuring transparent communication with end-user communities.

CRediT authorship contribution statement

Marian Lupaşcu: Writing – original draft, Visualization, Methodology, Investigation, Conceptualization; **Ana-Cristina Rogoz:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization; **Mihai Sorin Stupariu:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization; **Radu Tudor Ionescu:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the project “Romanian Hub for Artificial Intelligence - HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416. The authors thank reviewers for the constructive feedback.

References

[1] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O.K. Mohammed, B. Patra, et al., Language is not all you need: aligning perception with language models, in: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeuIPS), 2023, pp. 72096–72109. https://dl.acm.org/doi/10.5555/3666122.3669277.

[2] D. Driess, F. Xia, M.S.M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., PaLM-E: an embodied multimodal language model, in: Proceedings of the 40th International Conference on Machine Learning (ICML), 2023, pp. 8469–8488. https://dl.acm.org/doi/10.5555/3618408.3618748.

[3] S. Khanna, X. Li, Invisible languages of the LLM universe, (2025). https://arxiv.org/abs/2510.11557.

[4] S.R. Carroll, I. Garba, O.L. Figueroa-Rodríguez, J.C. Holbrook, R. Lovett, S. Materechera, M.A. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe, et al., The CARE principles for indigenous data governance, Data Sci. J. 19 (2020) 43. https://doi.org/10.5334/dsj-2020-043

[5] S. Bird, Local languages, third spaces, and other high-resource scenarios, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, p. 7817–7829. https://doi.org/10.18653/v1/2022.acl-long.539

[6] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proc. Natl. Acad. Sci. 114 (13) (2017) 3521–3526. https://doi.org/10.1073/pnas.1611835114

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, et al., BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] M. Rungta, J. Singh, S.M. Mohammad, D. Yang, Geographic citation gaps in NLP research, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 1371–1383. https://doi.org/10.18653/v1/2022.emnlp-main.89

[9] P.M. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[10] S. Ranathunga, N. de Silva, Some languages are more equal than others: probing deeper into the linguistic disparity in the NLP world, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP), 2022, pp. 823–848. https://doi.org/10.18653/v1/2022.aacl-main.62

[11] A. Caines, M. Rei, The Geographic Diversity of NLP Conferences, 2019, (https://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/) Accessed: December 2025.

[12] K. Darwish, N. Habash, M. Abbas, H.S. Al-Khalifa, H.T. Al-Natsheh, S.R. El-Beltagy, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, W. El-Hajj, M. Jarrar, H. Mubarak, A panoramic survey of natural language processing in the arab world, Commun. ACM 64 (2020) 72–81. https://doi.org/10.1145/3447735

[13] O. Kanishcheva, CLARIN Knowledge Centre for Ukrainian NLP and Corpora (UkrNLP-Corpora), 2023, (https://www.clarin.eu/blog/introduction-clarin-knowledge-centre-ukrainian-nlp-and-corpora-ukrnlp-corpora) Accessed: December 2025.

[14] M. Romanyshyn (Ed.), Proceedings of the fourth Ukrainian natural language processing workshop (UNLP), Association for Computational Linguistics, 2025. https://doi.org/10.18653/v1/2025.unlp-1.0

[15] K. Akhynko, O. Kosovan, M. Trokhymovych, Hidden persuasion: detecting manipulative narratives on social media during the 2022 russian invasion of Ukraine, in: Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP), 2025, pp. 194–202. https://doi.org/10.18653/v1/2025.unlp-1.19

[16] M. Li, Top 50+ Chinese AI Investment Statistics [2025], 2025, (https://www.secondtalent.com/resources/chinese-ai-investment-statistics/) Accessed: December 2025.

[17] D. Normile, Chinese firm’s faster, cheaper AI language model makes a splash, Science 387 (6731) (2025) 238. https://www.science.org/content/article/chinese-firm-s-faster-cheaper-ai-language-model-makes-splash.

[18] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, Y. Wang, J. Guo, A survey on LLM-as-a-judge, (2024). https://arxiv.org/abs/2411.15594.

[19] A. Paullada, I.D. Raji, E.M. Bender, E. Denton, A. Hanna, Data and its (dis)contents: a survey of dataset development and use in machine learning research, Patterns 2 (11) (2021) 100336. https://www.sciencedirect.com/science/article/pii/S2666389921001847. https://doi.org/10.1016/j.patter.2021.100336

[20] S. Ruder, N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, D. Garrette, G. Neubig, M. Johnson, XTREME-R: towards more challenging and nuanced multilingual evaluation, in: Proceedings of the 2021 Conference on Empir-

ical Methods in Natural Language Processing (EMNLP), 2021, pp. 10215–10245. https://doi.org/10.18653/v1/2021.emnlp-main.802
[21] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, (2023). https://arxiv.org/abs/2303.18223.
[22] W. Zhu, Y. Lv, Q. Dong, F. Yuan, J. Xu, S. Huang, L. Kong, J. Chen, L. Li, Extrapolating Large Language Models to Non-English by Aligning Languages, (2023). https://arxiv.org/abs/2308.04948.
[23] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, Vision-language pre-training: basics, recent advances, and future trends, Found. Trends Compu. Graph. Vision 14 (3–4) (2022) 163–352. https://www.nowpublishers.com/article/Details/CGV-105. https://doi.org/10.1561/0600000105
[24] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, W. Gao, Large-scale multi-modal pre-trained models: a comprehensive survey, Mach. Intell. Res. 20 (2023) 447–482. https://doi.org/10.1007/s11633-022-1410-8
[25] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, G. Shi, A survey of state of the art large vision language models: benchmark evaluations and challenges, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025, pp. 1587–1606. https://openaccess.thecvf.com/content/CVPR2025W/TMM-OpenWorld/html/Li_A_Survey_of_State_of_the_Art_Large_Vision_Language_CVPRW_2025_paper.html.
[26] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, Natl. Sci. Rev. 11 (12) (2024) nwae403. https://academic.oup.com/nsr/article/11/12/nwae403/7896414. https://doi.org/10.1093/nsr/nwae403
[27] J. Xie, Z. Chen, R. Zhang, X. Wan, G. Li, Large multimodal agents: a survey, (2024). https://arxiv.org/abs/2402.15116.
[28] M. Xu, W. Yin, D. Cai, R. Yi, D. Xu, Q. Wang, B. Wu, Y. Zhao, C. Yang, S. Wang, et al., A survey of resource-efficient LLM and multimodal foundation models, (2024). https://arxiv.org/abs/2401.08092.
[29] F. Alam, S.A. Chowdhury, S. Boughorbel, M. Hasanain, LLMs for low resource languages in multilingual, multimodal and dialectal settings, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts (EACL), 2024, pp. 27–33. https://aclanthology.org/2024.eacl-tutorials.5/.
[30] S. Mu, S. Lin, A comprehensive survey of mixture-of-experts: algorithms, theory, and applications, (2025). https://doi.org/10.48550/ARXIV.2503.07137
[31] H. Najadat, F. Abushaqra, Multimodal sentiment analysis of arabic videos, J. Image Graph. 6 (1) (2018) 39–43. https://www.joig.net/index.php?m=content&c=index&a=show&catid=47&id=173.
[32] B.R. Chakravarthi, P. Jishnu, B. Premjith, K.P. Soman, R. Ponnusamy, P.K. Kumaresan, K.P. Thamburaj, J.P. McCrae, DravidianMultiModality: a dataset for multimodal sentiment analysis in Tamil and Malayalam, (2021). https://arxiv.org/abs/2106.04853.
[33] S. Taylor, F. Fauzi, Multimodal sentiment analysis for the malay language: new corpus using CNN-based framework, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 24 (2024) 1–26. https://doi.org/10.1145/3703445
[34] D.F. Kponou, F.A.A. Laleye, E.C. Ezin, FFSTC: Fongbe to French speech translation corpus, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024, pp. 7270–7276. https://aclanthology.org/2024.lrec-main.638/.
[35] A. Haouhat, S. Bellaouar, A. Nehar, H. Cherroun, Towards arabic multimodal dataset for sentiment analysis, in: Proceedings of Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA), 2023, pp. 126–133. https://doi.org/10.1109/IDSTA58916.2023.10317847
[36] E. Hossain, O. Sharif, M.M. Hoque, MemoSen: a multimodal dataset for sentiment analysis of memes, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), 2022, pp. 1542–1554. https://aclanthology.org/2022.lrec-1.165/.
[37] E. Hossain, O. Sharif, M.M. Hoque, MUTE: a multimodal dataset for detecting hateful memes, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12Th International Joint Conference on Natural Language Processing: Student Research Workshop (AACL-IJCNLP), 2022, pp. 32–39. https://doi.org/10.18653/v1/2022.aacl-srw.5
[38] V. Păiş, S. Niţă, A.-I. Jerpelea, L. Pană, E. Curea, RoMemes: a multimodal meme corpus for the Romanian language, (2024). https://arxiv.org/abs/2410.15497.
[39] H. Luqman, Arabsign: a multi-modality dataset and benchmark for continuous arabic sign language recognition, in: Proceedings of IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), 2023, pp. 1–8. https://doi.org/10.1109/FG57933.2023.10042720
[40] C. Sikasote, E. Mukonde, M.M.I. Alam, A. Anastasopoulos, BIG-C: a multimodal multi-purpose dataset for bemba, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 2062–2078. https://doi.org/10.18653/v1/2023.acl-long.115
[41] L. Sanayai Meetei, L. Rahul, A. Singh, S.M. Singh, T.D. Singh, S. Bandyopadhyay, An experiment on speech-to-text translation systems for manipuri to english on low resource setting, in: Proceedings of the 18th International Conference on Natural Language Processing (ICON), 2021, pp. 54–63. https://aclanthology.org/2021.icon-main.8/.
[42] F. Farsi, S. Shariati Motlagh, S. Bali, S. Sabouri, S. Momtazi, Persian in a court: benchmarking VLMs in persian multi-modal tasks, in: Proceedings of the First Workshop of Evaluation of Multi-Modal Generation (EvalMG), 2025, pp. 52–56. https://aclanthology.org/2025.evalmg-1.5/.
[43] S. Arghyadeep, S. Parida, K. Kotwal, S. Panda, O. Bojar, S.R. Dash, Bengali visual genome: a multimodal dataset for machine translation and image captioning, in: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA), 2021, pp. 63–70. https://doi.org/10.1007/978-981-16-6624-7_7
[44] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73. https://doi.org/10.1007/s11263-016-0981-7
[45] I. Abdulmumin, S.R. Dash, M.A. Dawud, S. Parida, S. Muhammad, I.S. Ahmad, S. Panda, O. Bojar, B.S. Galadanci, B.S. Bello, Hausa visual genome: a dataset for multi-modal english to hausa machine translation, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), 2022, pp. 6471–6479. https://aclanthology.org/2022.lrec-1.694/.
[46] S. Parida, I. Abdulmumin, S.H. Muhammad, A. Bose, G.S. Kohli, I.S. Ahmad, K. Kotwal, S. Deb Sarkar, O. Bojar, H. Kakudi, HaVQA: a dataset for visual question answering and multimodal research in hausa language, in: Findings of the

Association for Computational Linguistics (ACL), 2023, pp. 10162–10183. https://doi.org/10.18653/v1/2023.findings-acl.646

[47] S. Parida, S. Sahoo, S. Sekhar, K. Sahoo, K. Kotwal, S. Khosla, S.R. Dash, A. Bose, G.S. Kohli, S.S. Lenka, O. Bojar, OVQA: a dataset for visual question answering and multimodal research in odia language, in: Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP), 2025, pp. 58–66. https://aclanthology.org/2025.indonlp-1.7/.

[48] M. Anwar, B. Shi, V. Goswami, W.-N. Hsu, J. Pino, C. Wang, MuAViC: a multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation, in: Proceedings of Conference of the International Speech Communication Association (INTERSPEECH), 2023, pp. 4064–4068. https://www.isca-archive.org/interspeech_2023/anwar23_interspeech.html. https://doi.org/10.21437/Interspeech.2023-2279

[49] N. Saichyshyna, D. Maksymenko, O. Turuta, A. Yerokhin, A. Babii, O. Turuta, Extension multi30K: multimodal dataset for integrated vision and language research in ukrainian, in: Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), 2023, pp. 54–61. https://doi.org/10.18653/v1/2023.unlp-1.7

[50] D. Elliott, S. Frank, K. Sima'an, L. Specia, Multi30K: multilingual english-german image descriptions, in: Proceedings of the 5th Workshop on Vision and Language (VL'16), 2016, pp. 70–74. https://doi.org/10.18653/v1/W16-3210

[51] H. Lovenia, R. Mahendra, S.M. Akbar, L.J.V. Miranda, J. Santoso, E. Aco, A. Fadhilah, J. Mansurov, J.M. Imperial, O.P. Kampman, et al., SEACrowd: a multilingual multimodal data hub and benchmark suite for southeast asian languages, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 5155–5203. https://doi.org/10.18653/v1/2024.emnlp-main.296

[52] H. Lent, K. Tatariya, R. Dabre, Y. Chen, M. Fekete, E. Ploeger, L. Zhou, R.-A. Armstrong, A. Eijansantos, C. Malau, et al., CreoleVal: multilingual multitask benchmarks for creoles, Trans. Assoc. Comput. Linguist. 12 (2024) 950–978. https://aclanthology.org/2024.tacl-1.53/. https://doi.org/10.1162/tacl_a_00682

[53] K. Dutta Chowdhury, M. Hasanuzzaman, Q. Liu, Multimodal neural machine translation for low-resource language pairs using synthetic data, in: Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), 2018, pp. 33–42. https://aclanthology.org/W18-3405/. https://doi.org/10.18653/v1/W18-3405

[54] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Trans. Assoc. Computat. Linguist. 2 (2014) 67–78. https://aclanthology.org/Q14-1006/. https://doi.org/10.1162/tacl_a_00166

[55] L.S. Meetei, T.D. Singh, S. Bandyopadhyay, Low resource multimodal neural machine translation of english-hindi in news domain, in: Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL), 2021, pp. 20–29. https://aclanthology.org/2021.mmtlrl-1.4/.

[56] S. Haq, R. Huidrom, S. Castilho, DCU ADAPT at WMT24: english to low-resource multi-modal translation task, in: Proceedings of the Ninth Conference on Machine Translation (WMT), 2024, pp. 810–814. https://aclanthology.org/2024.wmt-1.75/. https://doi.org/10.18653/v1/2024.wmt-1.75

[57] F. Alwajih, G. Bhatia, M. Abdul-Mageed, Dallah: a dialect-Aware multimodal large language model for arabic, in: Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP), 2024, pp. 320–336. https://doi.org/10.18653/v1/2024.arabicnlp-1.27

[58] Y. Wang, J. Pfeiffer, N. Carion, Y. LeCun, A. Kamath, Adapting grounded visual question answering models to low resource languages, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 2596–2605. https://ieeexplore.ieee.org/document/10208296/. https://doi.org/10.1109/CVPRW59228.2023.00258

[59] Y. Wang, J. Dong, T. Liang, M. Zhang, R. Cai, X. Wang, Cross-lingual cross-modal retrieval with noise-robust learning, in: Proceedings of the 30th ACM International Conference on Multimedia (ACMMM), 2022, pp. 422–433. https://doi.org/10.1145/3503161.3548003

[60] A. Dash, H.R. Gupta, Y. Sharma, BITS-P at WAT 2023: improving indic language multimodal translation by image augmentation using diffusion models, in: Proceedings of the 10th Workshop on Asian Translation (WAT), 2023, pp. 41–45. https://aclanthology.org/2023.wat-1.3/.

[61] K.T. Doan, B.G. Huynh, D.T. Hoang, T.D. Pham, N.H. Pham, Q. Nguyen, B.Q. Vo, S.N. Hoang, Vintern-1B: an efficient multimodal large language model for Vietnamese, (2024). https://arxiv.org/abs/2408.12480.

[62] P. Nath, P.K. Adhikary, P. Dadure, P. Pakray, R. Manna, S. Bandyopadhyay, Image caption generation for low-resource assamese language, in: Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022), 2022, pp. 263–272. https://aclanthology.org/2022.rocling-1.33/.

[63] L. Jiang, J. Li, J. Zhang, Y. Shen, Multimodal seed data augmentation for low-resource audio latin cuengh language, Appl. Sci. 14 (20) (2024) 9533. https://www.mdpi.com/2076-3417/14/20/9533. https://doi.org/10.3390/app14209533

[64] X. Qu, M. Song, W. Wei, J. Dong, Y. Cheng, Mitigating multilingual hallucination in large vision-language models, (2024). https://arxiv.org/abs/2408.00550.

[65] R. Rafailov, A. Sharma, E. Mitchell, C.D. Manning, S. Ermon, C. Finn, Direct preference optimization: your language model is secretly a reward model, in: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS), 36, 2023, pp. 53728–53741. https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

[66] M. Shukor, L. Bethune, D. Busbridge, D. Grangier, E. Fini, A. El-Nouby, P. Ablin, Scaling laws for optimal data mixtures, (2025). https://arxiv.org/abs/2507.09404.

[67] R. Navigli, S. Conia, B. Ross, Biases in large language models: origins, inventory, and discussion, ACM J. Data Inf. Qual. 15 (2023) 1–21. https://doi.org/10.1145/3597307

[68] L. Wiechetek, F.A. Pirinen, B. Gaup, T. Trosterud, M. Kappfjell, S.N. Moshagen, The ethical question – use of indigenous corpora for large language models, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024, pp. 15922–15931. https://aclanthology.org/2024.lrec-main.1383/.

[69] F.Z. Youcef, F. Barigou, Arabic language investigation in the context of unimodal and multimodal sentiment analysis, in: Proceedings of 22nd International Arab Conference on Information Technology (ACIT), 2021, pp. 1–7. https://ieeexplore.ieee.org/document/9677274. https://doi.org/10.1109/ACIT53391.2021.9677274

[70] N. Al Roken, G. Barlas, Multimodal arabic emotion recognition using deep learning, Speech Commun. 155 (C) (2023) 103005. https://doi.org/10.1016/j.specom.2023.103005

[71] K. Dashtipour, M. Gogate, E. Cambria, A. Hussain, A novel context-aware multimodal framework for persian sentiment analysis, Neurocomputing 457 (C) (2021) 377–388. https://www.sciencedirect.com/science/article/abs/pii/S0925231221002666. https://doi.org/10.1016/j.neucom.2021.02.020

[72] S. Al-Azani, E.-S.M. El-Alfy, Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information, IEEE Access 8 (2020) 136843–136857. https://ieeexplore.ieee.org/document/9148603. https://doi.org/10.1109/ACCESS.2020.3011977

[73] B. Premjith, G. Jyothish Lal, V. Sowmya, B.R. Chakravarthi, R. Natarajan, K. Nandhini, A. Murugappan, B. Bharathi, M. Kaushik, S. Prasanth, R. Aswin Raj, S. Vijai Simmon, Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech), 2023, pp. 72–79. https://aclanthology.org/2023.dravidianlangtech-1.10/.

[74] R.G. Kodali, D.P. Manukonda, M. Pannakkaran, BytesizedLLM@dravidianlangtech 2025: abusive tamil and malayalam text targeting women on social media using XLM-RoBERTa and attention-BiLSTM, in: Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech), 2025, pp. 80–85. https://aclanthology.org/2025.dravidianlangtech-1.14/.

[75] M.A. Jigar, A.A. Ayele, S.M. Yimam, C. Biemann, Detecting hate speech in amharic using multimodal analysis of social media memes, in: Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying (TRAC), 2024, pp. 85–95. https://aclanthology.org/2024.trac-1.10/.

[76] A.G. Debele, M.M. Woldeyohannis, Multimodal amharic hate speech detection using deep learning, in: Proceedings of International Conference on Information and Communication Technology for Development for Africa (ICT4DA), 2022, pp. 102–107. https://ieeexplore.ieee.org/document/9971436/. https://doi.org/10.1109/ICT4DA56482.2022.9971436

[77] A. Hatami, S. Banerjee, M. Arcan, P. Buitelaar, J. Philip McCrae, English-to-low-resource translation: a multimodal approach for hindi, malayalam, bengali, and hausa, in: Proceedings of the Ninth Conference on Machine Translation (WMT), 2024, pp. 815–822. https://doi.org/10.18653/v1/2024.wmt-1.76

[78] S. Alalem, M.S. Zaghloul, O. Badawy, A novel deep learning multi-Modal sentiment analysis model for english and egyptian arabic dialects using audio and text, in: Proceedings of 24th International Arab Conference on Information Technology (ACIT), 2023, pp. 1–5. https://ieeexplore.ieee.org/document/10453875. https://doi.org/10.1109/ACIT58888.2023.10453875

[79] D.S. Chauhan, A. Ekbal, P. Bhattacharyya, An efficient fusion mechanism for multimodal low-resource setting, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2022, pp. 2583–2588. https://doi.org/10.1145/3477495.3531900

[80] F.T.J. Faria, L.H. Baniata, M.H. Baniata, M.A. Khair, A.I. Bani Ata, C. Bunterngchit, S. Kang, Sentimentformer: a transformer-based multimodal fusion framework for enhanced sentiment analysis of memes in under-resourced bangla language, Electronics 14 (4) (2025) 799. https://www.mdpi.com/2079-9292/14/4/799. https://doi.org/10.3390/electronics14040799

[81] M.R. Karim, S.K. Dey, T. Islam, M. Shajalal, B.R. Chakravarthi, Multimodal hate speech detection from bengali memes and texts, in: Proceedings of the International Conference on Speech and Language Technologies for Low-Resource Languages (SPELLL), 2022, pp. 293–308. https://link.springer.com/chapter/10.1007/978-3-031-33231-9_21.

[82] R.M. Albalawi, A.T. Jamal, A.O. Khadidos, A.M. Alhothali, Multimodal arabic rumors detection, IEEE Access 11 (2023) 9716–9730. https://ieeexplore.ieee.org/document/10026837. https://doi.org/10.1109/ACCESS.2023.3240373

[83] Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, G. Huang, Z. Liang, L. Zhang, L. Li, et al., Multimodal sensing for depression risk detection: integrating audio, video, and text data, Sensors 24 (12) (2024) 3714. https://www.mdpi.com/1424-8220/24/12/3714. https://doi.org/10.3390/s24123714

[84] N.J. Deocampo, M. Villarica, A. Vinluan, A lip-reading model for tagalog using multimodal deep learning approach, Int. J. Comput. Sci. Res. 8 (2024) 2796–2808. https://stepacademic.net/ijcsr/article/view/511.

[85] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, F. Khan, Urdu sentiment analysis via multimodal data mining based on deep learning algorithms, IEEE Access 9 (2021) 153072–153082. https://ieeexplore.ieee.org/document/9583225. https://doi.org/10.1109/ACCESS.2021.3122025

[86] F. Arifin, A. Nasuha, A. Priambodo, A. Winursito, T. Gunawan, Advanced multimodal emotion recognition for javanese language using deep learning, Indon. J.

Electr. Eng. Inform. 12 (3) (2024) 503–515. https://section.iaesonline.com/index.php/IJEEI/article/view/5662. https://doi.org/10.52549/ijeei.v12i3.5662
[87] O.Z. Mamyrbayev, K. Alimhan, B. Amirgaliyev, B. Zhumazhanov, D. Mussayeva, F. Gusmanova, Multimodal systems for speech recognition, Int. J. Mobile Commun. 18 (3) (2020) 314–326. https://doi.org/10.1504/ijmc.2020.107097
[88] K.T. Elahi, T.B. Rahman, S. Shahriar, S. Sarker, S.K.S. Joy, F.M. Shah, Explainable multimodal sentiment analysis on bengali memes, in: Proceedings of 26th International Conference on Computer and Information Technology (ICCIT), 2023, pp. 1–6. https://ieeexplore.ieee.org/document/10441342. https://doi.org/10.1109/ICCIT60459.2023.10441342
[89] M. Rahman, A. Raihan, T. Rahman, S. Ahsan, J. Hossain, A. Das, M.M. Hoque, Binary_beasts@dravidianlangtech-EACL 2024: multimodal abusive language detection in tamil based on integrated approach of machine learning and deep learning techniques, in: Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech), 2024, pp. 212–217. https://aclanthology.org/2024.dravidianlangtech-1.35/.
[90] R. Das, T.D. Singh, A multi-stage multimodal framework for sentiment analysis of assamese in low resource setting, Expert Syst. Appl. 204 (C) (2022) 117575. https://www.sciencedirect.com/science/article/abs/pii/S0957417422008879. https://doi.org/10.1016/j.eswa.2022.117575
[91] B.R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. Sridevy, M. Arcan, M. Zarrouk, J.P. McCrae, Multilingual multimodal machine translation for dravidian languages utilizing phonetic transcription, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages (LoResMT), 2019, pp. 56–63. https://aclanthology.org/W19-6809/.
[92] L. Meetei, T.D. Singh, S. Bandyopadhyay, Exploiting multiple correlated modalities can enhance low-resource machine translation quality, Multimed. Tools Appl. 83 (2024) 13137–13157. https://link.springer.com/article/10.1007/s11042-023-15721-2. https://doi.org/10.1007/s11042-023-15721-2
[93] N.-C. Ristea, R.T. Ionescu, Cascaded cross-modal transformer for request and complaint detection, in: Proceedings of the 31st ACM International Conference on Multimedia (ACMMM), 2023, pp. 9467–9471. https://doi.org/10.1145/3581783.3612846
[94] H.H.S.N. Haputhanthri, H.M.N. Tennakoon, M.A.S.M. Wijesekara, B.H.R. Pushpananda, H.N.D. Thilini, Multi-modal deep learning approach to improve sentence level sinhala sign language recognition, Int. J. Adv. ICT Emerging Reg. 16 (2) (2023) 21–30. https://icter.sljol.info/articles/10.4038/icter.v16i2.7264. https://doi.org/10.4038/icter.v16i2.7264
[95] Y. Yang, Q.-D.-E.-J. Ren, R.-F. He, Multi-modal sentiment analysis of mongolian language based on pre-trained models and high-resolution networks, in: Proceedings of International Conference on Asian Language Processing (IALP), 2024, pp. 291–296. https://ieeexplore.ieee.org/document/10661161/. https://doi.org/10.1109/IALP63756.2024.10661161
[96] S.R. Laskar, A.F. U.R. Khilji, P. Pakray, S. Bandyopadhyay, Multimodal neural machine translation for english to hindi, in: Proceedings of the 7th Workshop on Asian Translation (WAT), 2020, pp. 109–113. https://aclanthology.org/2020.wat-1.11/. https://doi.org/10.18653/v1/2020.wat-1.11
[97] S.R. Laskar, A.F. U.R. Khilji, D. Kaushik, P. Pakray, S. Bandyopadhyay, Improved english to hindi multimodal neural machine translation, in: Proceedings of the 8th Workshop on Asian Translation (WAT), 2021, pp. 155–160. https://aclanthology.org/2021.wat-1.17/. https://doi.org/10.18653/v1/2021.wat-1.17
[98] B. Gain, D. Bandyopadhyay, S. Mukherjee, C. Adak, A. Ekbal, Impact of visual context on noisy multimodal NMT: an empirical study for english to Indian languages, (2023). https://arxiv.org/abs/2308.16075.
[99] X. Shi, Z. Yu, Adding visual information to improve multimodal machine translation for low-resource language, Math. Prob. Eng. 2022 (1) (2022) 5483535. https://onlinelibrary.wiley.com/doi/10.1155/2022/5483535. https://doi.org/10.1155/2022/5483535
[100] L.S. Meetei, A. Singh, T.D. Singh, S. Bandyopadhyay, Do cues in a video help in handling rare words in a machine translation system under a low-resource setting?, Natur. Lang. Process. J. 3 (2023) 100016. https://www.sciencedirect.com/science/article/pii/S2949719123000134. https://doi.org/10.1016/j.nlp.2023.100016
[101] L.S. Meetei, S.M. Singh, A. Singh, R. Das, T.D. Singh, S. Bandyopadhyay, Hindi to english multimodal machine translation on news dataset in low resource setting, in: Proceedings of International Conference on Machine Learning and Data Engineering (ICMLDE), 218, 2023, pp. 2102–2109. https://www.sciencedirect.com/science/article/pii/S1877050923001862. https://doi.org/10.1016/j.procs.2023.01.186
[102] T. Tayir, L. Li, Unsupervised multimodal machine translation for low-resource distant language pairs, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23 (4) (2024) 1–22. https://dl.acm.org/doi/10.1145/3652161. https://doi.org/10.1145/3652161
[103] T. Tayir, L. Li, M. Maimaiti, Y. Muhtar, Low-resource machine translation with different granularity image features, in: Proceedings of Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2025, pp. 260–273. https://link.springer.com/chapter/10.1007/978-981-97-8620-6_18.
[104] H.O. Lekshmy, S. Jayaraman, English-malayalam vision aid with multi modal machine learning technologies, in: Proceedings of 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 2022, pp. 1469–1476. https://ieeexplore.ieee.org/document/9788187. https://doi.org/10.1109/ICICCS53718.2022.9788187
[105] S. Parida, O. Bojar, S.R. Dash, Hindi visual genome: a dataset for multimodal english to hindi machine translation, Computación y Sistemas 23 (4) (2019) 1499–1505. https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3294. https://doi.org/10.13053/cys-23-4-3294
[106] S. Parida, S. Panda, S.P. Biswal, K. Kotwal, S. Arghyadeep, S.R. Dash, P. Motlicek, Multimodal neural machine translation system for english to bengali, in: Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL), 2021, pp. 31–39. https://aclanthology.org/2021.mmtlrl-1.6/.
[107] L. Nortje, D. Oneaţă, G. Pirlogeanu, H. Kamper, Improved visually prompted keyword localisation in real low-resource settings, (2024). https://arxiv.org/abs/2409.06013.
[108] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, J. Baldridge, et al., MURAL: multimodal, multitask representations across languages, in: Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 3449–3463. https://aclanthology.org/2021.findings-emnlp.293/. https://doi.org/10.18653/v1/2021.findings-emnlp.293
[109] W. Jian, H. Hou, N. Wu, S. Sun, Z. Yang, Y. Wang, P. Wang, Multimodal neural machine translation for mongolian to chinese, in: Proceedings of 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8. https://ieeexplore.ieee.org/document/9892831. https://doi.org/10.1109/IJCNN55064.2022.9892831
[110] A.G. Kovath, A. Nayyar, O.K. Sikha, Multimodal attention-driven visual question answering for malayalam, Neural Comput. Appl. 36 (24) (2024) 14691–14708. https://link.springer.com/article/10.1007/s00521-024-09818-4. https://doi.org/10.1007/s00521-024-09818-4
[111] S.R. Laskar, R. Singh, M.F. Karim, R. Manna, P. Pakray, S. Bandyopadhyay, Investigation of english to hindi multimodal neural machine translation using transliteration-based phrase pairs augmentation, in: Proceedings of the 9th Workshop on Asian Translation (WAT), 2022, pp. 117–122. https://aclanthology.org/2022.wat-1.15/.
[112] S.R. Laskar, B. Paul, S. Paudwal, P. Gautam, N. Biswas, P. Pakray, Multimodal neural machine translation for english–assamese pair, in: Proceedings of International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 387–392. https://ieeexplore.ieee.org/document/9752181. https://doi.org/10.1109/ComPE53109.2021.9752181
[113] Y. Chen, F. Wei, X. Sun, Z. Wu, S. Lin, A simple multi-modality transfer learning baseline for sign language translation, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5110–5120. https://ieeexplore.ieee.org/document/9879103/. https://doi.org/10.1109/CVPR52688.2022.00506
[114] A. Amalas, M. Ghogho, M. Chetouani, R.O.H. Thami, A multilingual training strategy for low resource Text to Speech, (2024). https://arxiv.org/abs/2409.01217.
[115] Y. Wu, S. Zhao, Y. Zhang, X. Yuan, Z. Su, When pairs meet triplets: improving low-resource captioning via multi-objective optimization, ACM Trans. Multimedia Comput. Commun. Appl. 18 (3) (2022) 1–20. https://dl.acm.org/doi/10.1145/3492325. https://doi.org/10.1145/3492325
[116] J. Yeo, M. Kim, S. Watanabe, Y. Ro, Visual speech recognition for languages with limited labeled data using automatic labels from whisper, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 10471–10475. https://ieeexplore.ieee.org/document/10446720. https://doi.org/10.1109/ICASSP48485.2024.10446720
[117] G. Bhatia, E.M.B. Nagoudi, F. Alwajih, M. Abdul-Mageed, Qalam: a multimodal LLM for arabic optical character and handwriting recognition, in: Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP), 2024, pp. 210–224. https://aclanthology.org/2024.arabicnlp-1.19/. https://doi.org/10.18653/v1/2024.arabicnlp-1.19
[118] C. Tran, H.L. Thanh, LaVy: Vietnamese multimodal large language model, (2024). https://arxiv.org/abs/2404.07922.
[119] C.E. Onuoha, E. Uba, An analysis of minimal pairs in igbo using a multimodal approach to speech perception, Unizik J. Arts Human. 25 (2024) 31–50. https://www.ajol.info/index.php/ujah/article/view/272304. https://doi.org/10.4314/ujah.v25i1.2
[120] Y. Wu, S. Zhao, J. Chen, Y. Zhang, X. Yuan, Z. Su, Improving captioning for low-resource languages by cycle consistency, in: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 362–367. https://ieeexplore.ieee.org/document/8784910. https://doi.org/10.1109/ICME.2019.00070
[121] Mamta, A. Ekbal, P. Bhattacharyya, Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21 (5) (2022) 104. https://dl.acm.org/doi/10.1145/3514498. https://doi.org/10.1145/3514498
[122] W. Jin, Y. Cheng, Y. Shen, W. Chen, X. Ren, A good prompt is worth millions of parameters: low-resource prompt-based learning for vision-language models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 2763–2775. https://aclanthology.org/2022.acl-long.197/. https://doi.org/10.18653/v1/2022.acl-long.197
[123] R. Solomon, M. Abebe, Amharic language image captions generation using hybridized attention-based deep neural networks, Appl. Computat. Intell. Soft Comput. 2023 (1) (2023) 9397325. https://onlinelibrary.wiley.com/doi/10.1155/2023/9397325. https://doi.org/10.1155/2023/9397325
[124] C. Rahul, T. Arathi, L.S. Panicker, R. Gopikakumari, Morphology & word sense disambiguation embedded multimodal neural machine translation system between sanskrit and malayalam, Biomed. Signal Process. Control 85 (2023) 105051. https://www.sciencedirect.com/science/article/pii/S1746809423004846. https://doi.org/10.1016/j.bspc.2023.105051
[125] M. Tang, C. Wang, J. Wang, C. Tan, S. Huang, C. Chen, W. Qian, XtremeCLIP: extremely parameter-efficient tuning for low-resource vision language understanding, in: Findings of the Association for Computational Linguistics (ACL), 2023, pp.

6368–6376. https://aclanthology.org/2023.findings-acl.397/. https://doi.org/10.18653/v1/2023.findings-acl.397
[126] A. Asgarov, S. Rustamov, LowCLIP: adapting the CLIP model architecture for low-resource languages in multimodal image retrieval task, (2024). https://doi.org/10.48550/arXiv.2408.13909
[127] H.A. Rahmon, T.G. Jimoh, F.O. Madaiyese, Speech recognition model in yoruba language, Smartify J. Smart Educ. Pedag. 1 (1) (2024) 28–46. https://researchvision.us/index.php/smartify/article/view/5.
[128] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The Llama 3 herd of models, (2024). https://doi.org/10.48550/ARXIV.2407.21783
[129] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, et al., DeepSeek-V3 Technical Report, (2024). https://doi.org/10.48550/ARXIV.2412.19437
[130] Anthropic, Claude Opus 4 & Claude Sonnet 4 System Card, 2025. Accessed: December 2025, https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf.
[131] T. Gunter, Z. Wang, C. Wang, R. Pang, A. Narayanan, A. Zhang, B. Zhang, C. Chen, C. Chiu, D. Qiu, et al., Apple intelligence foundation language models, (2024).https://doi.org/10.48550/ARXIV.2407.21075
[132] L. Yang, Y. Tian, B. Li, X. Zhang, K. Shen, Y. Tong, M. Wang, MMaDA: multimodal large diffusion language models, (2025). https://doi.org/10.48550/ARXIV.2505.15809
[133] Y. Shen, Z. Xu, Q. Wang, Y. Cheng, W. Yin, L. Huang, Multimodal instruction tuning with conditional mixture of LoRA, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 637–648. https://aclanthology.org/2024.acl-long.38/. https://doi.org/10.18653/V1/2024.ACL-LONG.38
[134] M.D.A. Cheema, M.D. Shaiq, F. Mirza, A. Kamal, M.A. Naeem, Adapting multilingual vision language transformers for low-resource urdu optical character recognition (OCR), PeerJ Comput. Sci. 10 (2024) e1964. https://peerj.com/articles/cs-1964/.
[135] M. Kim, J.H. Yeo, J. Choi, Y.M. Ro, Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15359–15371. https://ieeexplore.ieee.org/document/10377080/. https://doi.org/10.1109/ICCV51070.2023.01409
[136] A. Aruna Gladys, V. Vetriselvi, Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning, Appl. Soft Comput. 157 (C) (2024) 111553. https://www.sciencedirect.com/science/article/abs/pii/S1568494624003272. https://doi.org/10.1016/j.asoc.2024.111553
[137] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti, S. Watanabe, Improving massively multilingual ASR with auxiliary CTC objectives, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. https://ieeexplore.ieee.org/document/10095326. https://doi.org/10.1109/ICASSP49357.2023.10095326
[138] G.O. dos Santos, D.A. Braga Moreira, A.I. Ferreira, J. Silva, L. Pereira, P. Bueno, T. Sousa, H. Maia, N. Da Silva, E. Colombini, H. Pedrini, S. Avila, CAPIVARA: cost-efficient approach for improving multilingual CLIP performance on low-resource languages, in: Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL), 2023, pp. 184–207. https://aclanthology.org/2023.mrl-1.15/. https://doi.org/10.18653/v1/2023.mrl-1.15
[139] L. Nortje, D. Oneaţă, H. Kamper, Visually grounded few-shot word learning in low-resource settings, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2024) 2544–2554. https://ieeexplore.ieee.org/document/10508479/. https://doi.org/10.1109/TASLP.2024.3393772
[140] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 139, 2021, pp. 8748–8763. http://proceedings.mlr.press/v139/radford21a.html.
[141] M. Tsimpoukelli, J. Menick, S. Cabi, S.M.A. Eslami, O. Vinyals, F. Hill, Multimodal few-shot learning with frozen language models, in: Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS), 2021, pp. 200–212. https://proceedings.neurips.cc/paper/2021/file/01b7575c38dac42f3cfb7d500438b875-Paper.pdf.
[142] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, An empirical study of GPT-3 for few-shot knowledge-based VQA, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 36, 2022, pp. 3081–3089. https://ojs.aaai.org/index.php/AAAI/article/view/20215. https://doi.org/10.1609/aaai.v36i3.20215
[143] S.R. Laskar, B. Paul, P. Pakray, S. Bandyopadhyay, English-assamese multimodal neural machine translation using transliteration-based phrase augmentation approach, in: Proceedings of International Conference on Machine Learning and Data Engineering (ICMLDE), 218, 2023, pp. 979–988. https://www.sciencedirect.com/science/article/pii/S1877050923000789. https://doi.org/10.1016/j.procs.2023.01.078
[144] F. Alwajih, E.M.B. Nagoudi, G. Bhatia, A. Mohamed, M. Abdul-Mageed, Peacock: a family of arabic multimodal large language models and benchmarks, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 12753–12776. https://aclanthology.org/2024.acl-long.689/. https://doi.org/10.18653/v1/2024.acl-long.689
[145] C. Wang, H. Tang, X. Yang, Y. Xie, J. Suh, S. Sitaram, J. Huang, Y. Xie, Z. Gong, X. Xie, F. Wu, Uncovering inequalities in new knowledge learning by large language models across different languages, (2025). https://arxiv.org/abs/2503.04064.
[146] B.Y. Lin, C. He, Z. Ze, H. Wang, Y. Hua, C. Dupuy, R. Gupta, M. Soltanolkotabi, X. Ren, S. Avestimehr, FedNLP: benchmarking federated learning methods for natural language processing tasks, in: Findings of the Association for Computational Linguistics (NAACL), 2022, pp. 157–175. https://aclanthology.org/2022.findings-naacl.13/. https://doi.org/10.18653/v1/2022.findings-naacl.13
[147] W. Zhao, Y. Chen, R. Lee, X. Qiu, Y. Gao, H. Fan, N.D. Lane, Breaking physical and linguistic borders: multilingual federated prompt tuning for low-resource languages, (2025). https://arxiv.org/abs/2507.03003. https://doi.org/10.48550/arXiv.2507.03003
[148] L. Tran, W. Sun, S. Patterson, A. Milanova, Privacy-preserving personalized federated prompt learning for multimodal large language models, (2025). https://doi.org/10.48550/arXiv.2501.13904
[149] M. Andersland, Amharic LLaMA and LLaVA: multimodal LLMs for low resource languages, (2024). https://doi.org/10.48550/arXiv.2403.06354
[150] L. Shen, W. Tan, S. Chen, Y. Chen, J. Zhang, H. Xu, B. Zheng, P. Koehn, D. Khashabi, The language barrier: dissecting safety challenges of LLMs in multilingual contexts, in: Findings of the Association for Computational Linguistics (ACL), 2024, pp. 2668–2680. https://aclanthology.org/2024.findings-acl.156/. https://doi.org/10.18653/v1/2024.findings-acl.156
[151] S. Singh, A. Romanou, C. Fourrier, D.I. Adelani, J.G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W.Q. Leong, Y. Susanto, et al., Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025, pp. 18761–18799. https://aclanthology.org/2025.acl-long.919/. https://doi.org/10.18653/v1/2025.acl-long.919