European Journal of AI, Computing & Informatics

Vol. 1 No. 3 2025

Article **Open Access**

Research on Cross-Modal Semantic Alignment Methods for Low-Resource Languages

Zhizhi Yu [1,*]

[1] Hebei University of Engineering, Handan, China
* Correspondence: Zhizhi Yu, Hebei University of Engineering, Handan, China

Received: 01 September 2025
Revised: 09 September 2025
Accepted: 27 September 2025
Published: 09 October 2025

Abstract: This study addresses the challenge of cross-modal semantic alignment in low-resource languages, a critical problem for enabling inclusive and equitable AI-driven multimodal applications. We propose a novel framework that synergistically integrates multi-level textual embeddings, visual Transformer modeling, and the construction of a unified cross-modal projection space. To enhance alignment quality, the approach incorporates advanced mechanisms including contrastive learning, distributed semantic constraints, and fine-grained local alignment strategies. Furthermore, to mitigate data scarcity inherent in low-resource settings, we leverage transfer enhancement techniques such as cross-lingual knowledge distillation, pseudo-pair augmentation, and multi-task training. Comprehensive experiments on the FLORES-200 dataset demonstrate that our method consistently surpasses state-of-the-art models such as CLIP and ALIGN across multiple metrics. Specifically, significant gains are observed in Recall@1 and Mean Rank for languages including Swahili and Sinhala, underscoring the method's effectiveness, robustness, and generalizability in low-resource scenarios. These findings highlight the potential of the proposed approach for advancing cross-lingual multimodal understanding and bridging the performance gap for underrepresented languages.

Keywords: low-resource languages; cross-modal semantic alignment; contrastive learning; transfer enhancement

1. Introduction

With the rapid advancement of multimodal artificial intelligence, cross-modal semantic alignment has emerged as a cornerstone technology with wide-ranging applications in information retrieval, machine translation, and human-computer interaction. At its core, cross-modal alignment seeks to construct a unified semantic representation space that enables effective mapping and interaction across heterogeneous modalities, including text, images, and speech. While significant progress has been achieved for high-resource languages such as English and Chinese-driven by the availability of large-scale paired corpora and powerful pre-trained models-these approaches largely fail when applied to low-resource languages [1].

The challenges in low-resource settings are multifaceted: sparse and fragmented textual corpora, limited or inconsistent semantic annotations, and constrained cross-lingual transfer capabilities all contribute to suboptimal alignment performance. Consequently, existing models exhibit low semantic representation accuracy and weak cross-modal retrieval effectiveness for underrepresented languages. This limitation not only restricts the

development of inclusive AI systems but also impedes the broader adoption of multilingual multimodal technologies in real-world applications.

Addressing these challenges requires methods that are explicitly designed to operate under data-scarce conditions. Motivated by this, the present study proposes a cross-modal semantic alignment framework tailored to the specific characteristics of low-resource languages. By integrating advanced feature extraction architectures, alignment optimization strategies, and transfer enhancement mechanisms, the proposed approach significantly improves cross-modal semantic modeling in low-resource scenarios. Beyond empirical performance gains, this framework provides a generalizable and extensible technical pathway for enabling robust, multilingual, and multimodal intelligent applications across diverse practical environments worldwide.

2. Foundations of Cross-Modal Semantic Alignment

2.1. Characteristics of Low-Resource Languages

Low-resource languages face multifaceted constraints encompassing resource availability, linguistic characteristics, and usage patterns. Resource scarcity manifests in several ways: limited availability of parallel corpora and multimodal paired samples, inconsistent granularity of annotations, and restrictions imposed by copyright or data collection channels, all of which hinder the establishment of stable training-validation-test distributions. Linguistically, such languages often exhibit complex agglutinative or inflectional morphology, flexible word order, the coexistence of multiple writing systems (including variant orthographies and character forms within the same language), and unstable phoneme-to-character correspondences [2]. These factors frequently lead to word segmentation ambiguities, subword boundary drift, and inconsistencies in tokenization.

Usage-related challenges further exacerbate modeling difficulties. Common issues include mixed-code usage, high prevalence of colloquial abbreviations, numerous morphological variants for proper nouns and geographic entities, and pronounced domain shifts, such as discrepancies between folklore, legal, and medical texts. Collectively, these challenges contribute to sparse cross-modal anchors and increased annotation noise, complicating alignment and retrieval tasks.

To support robust modeling under such conditions, it is essential to employ computable diagnostic metrics that capture both data and linguistic properties. Metrics such as type-token ratio (TTR), morphological entropy, out-of-vocabulary (OOV) rates, word boundary F1 scores, text-image mutual information, parallel alignment coverage, and cross-domain Kullback-Leibler divergence can guide engineering decisions regarding segmentation strategies, lexicon design, and sampling schemes [3]. A summary of these key metrics and associated handling considerations is presented in Table 1.

Table 1. Overview of Key Attributes of Low-Resource Languages - Metrics - Alignment Impact - Handling Strategies.

Metric	Measurement / Diagnosis	Impact on Alignment	Handling Strategy
OOV Rate	Out-of-vocabulary words / total tokens	Missing semantic anchors, reduced retrieval recall	Byte-level BPE / Unigram vocabulary extension, morphological decomposition
Morphological Entropy	Entropy of morphological variant distribution	Unstable subword segmentation, fragmented embedding space	Morphological annotation assistance, stemming / lemmatization
TTR (Type-Token Ratio)	Unique word types / total tokens	Long-tail sparsity, difficulty in parameter sharing	Shared subword vocabulary, long-tail resampling / mixed sampling

Boundary F1	F1 score for word/token boundaries	Inaccurate semantic units, text-vision mismatch	Joint tokenization-alignment training, weakly-supervised boundary correction
Alignment Coverage	Proportion of parallel/paired samples	Sparse cross-modal positives, overfitting to noise	Pseudo-pair construction, bi-directional retrieval consistency filtering
Inter-domain KL	KL divergence between source/target domains	Domain shift leads to transfer degradation	Adversarial domain alignment, importance-weighted sample reweighting
Code-switching Rate	Proportion of cross-language fragments	Vocabulary conflict, false negative predictions	Language ID annotation, per-language adaptation / Adapters
Mutual Information MI (Image-Text)	I(Image, Text)	Weak semantic anchors, hard to learn shared space	Semantic label distillation, region-level alignment loss

2.2. Cross-Modal Representation Learning

Cross-modal representation learning fundamentally revolves around establishing a shared semantic space across heterogeneous modalities. Architecturally, it is commonly realized through two paradigms: dual encoders and cross encoders. In the dual-encoder setup, separate text and visual encoders generate modality-specific embeddings, which are then aligned via similarity metrics [4]. This design excels in large-scale contrastive learning and enables high-throughput retrieval. In contrast, cross encoders leverage cross-modal attention mechanisms to directly model fine-grained interactions between modalities, thereby enhancing alignment precision and localization capabilities [5].

In low-resource scenarios, the textual modality faces the dual challenge of ensuring sufficient subword coverage while maintaining sensitivity to complex morphological structures. To address this, a hybrid vocabulary strategy-combining byte-level BPE or Unigram tokenization with morphology-aware embeddings-is recommended. Additionally, language-specific statistical knowledge can be injected through lightweight adapters, such as Adapter modules or LoRA, to provide targeted representation enhancement without significantly increasing model complexity.

On the visual side, backbone architectures such as ViT or ConvNeXt are typically employed, offering rich intermediate representations that can be aligned with words or phrases via region proposals or patch-level attention mechanisms. To mitigate data sparsity and noise, advanced regularization techniques-such as invariance-variance-covariance constraints inspired by VICReg, feature-centered losses, and intra-batch hard-to-representative sample mining-are often integrated. Furthermore, distribution alignment methods, including maximum mean discrepancy (MMD) and adversarial domain alignment, are applied to improve cross-domain generalization.

Training objectives are generally designed in a multi-task fashion, encompassing text masked modeling, image masked reconstruction, cross-modal retrieval, region-phrase alignment, and cross-lingual semantic consistency, where synonymous sentence vectors across languages are approximated. Knowledge transfer strategies, such as teacher-student distillation, enable the semantic boundaries learned from high-resource cross-modal models to guide low-resource language embeddings effectively [6]. In extremely low-resource conditions, techniques including pseudo-pair screening, cross-lingual back-translation generation, and consistency-based cross-validation are employed to construct a high-confidence sample pool, thereby increasing effective sample density and improving alignment learning.

3. Semantic Alignment Methods

3.1. Feature Extraction Models

Under low-resource language conditions, a central challenge in cross-modal semantic alignment lies in obtaining stable and comparable feature representations for both text and images. Text modeling must address semantic drift caused by high out-of-vocabulary (OOV) rates and complex morphological variations. To this end, a multi-level embedding strategy is adopted as the core approach [7].

At the base layer, XLM-R serves as a multilingual shared encoder, leveraging cross-lingual pre-training and parameter transfer to provide unified word vector representations. This enables low-resource languages to benefit directly from embedding spaces learned in high-resource languages during early training stages. Building upon this foundation, a morphology-aware Adapter layer is introduced to capture language-specific morphological features. This module accepts subword sequences augmented with linguistic features such as stems, affixes, and inflection categories, which are embedded and passed through the Adapter's bottleneck structure. The Adapter reduces the dimensionality of the high-dimensional input, applies nonlinear transformations, and then restores dimensionality, effectively fusing morphological information with pre-trained semantic features. The processed output is concatenated with the original XLM-R embedding and subsequently projected linearly to obtain the final text vector representation. For languages with unstable word boundaries, a hybrid vocabulary combining byte-level BPE and unigram tokenization ensures that any OOV word can be decomposed into subword components, mitigating the impact of unknown tokens.

On the visual side, a Vision Transformer (ViT) serves as the backbone model. Input images are partitioned into 16×16 patches, which are converted into vector sequences via linear projection. A multi-head self-attention mechanism captures contextual dependencies across patches. To further enhance fine-grained semantic representation, a Faster R-CNN-based object detector generates candidate regions within each image. Regional features are then fused with the ViT patch sequence, resulting in a representation that encodes both global context and local object semantics. At the output stage, the ViT provides a global feature vector (i.e., the [CLS] token) for holistic semantic modeling, while preserving the regional feature matrix for subsequent local alignment tasks [8].

To enable cross-modal comparability, text and image features are mapped into a unified projection space of identical dimension (d = 768). This projection employs a two-layer fully connected network with nonlinear activation functions, followed by output normalization. Formally, the unified projection can be expressed as:

$$z = Norm(W_2 \cdot \sigma(W_1 h + b_1) + b_2)$$

Here, h denotes the input embedding, W_1, W_2 represents the projection layer parameters, and σ is the nonlinear function. This mechanism ensures that text and images reside in the same metric space prior to cross-modal alignment. The overall architecture is illustrated in Figure 1, which comprehensively demonstrates the entire process from raw inputs to unified cross-modal representations.

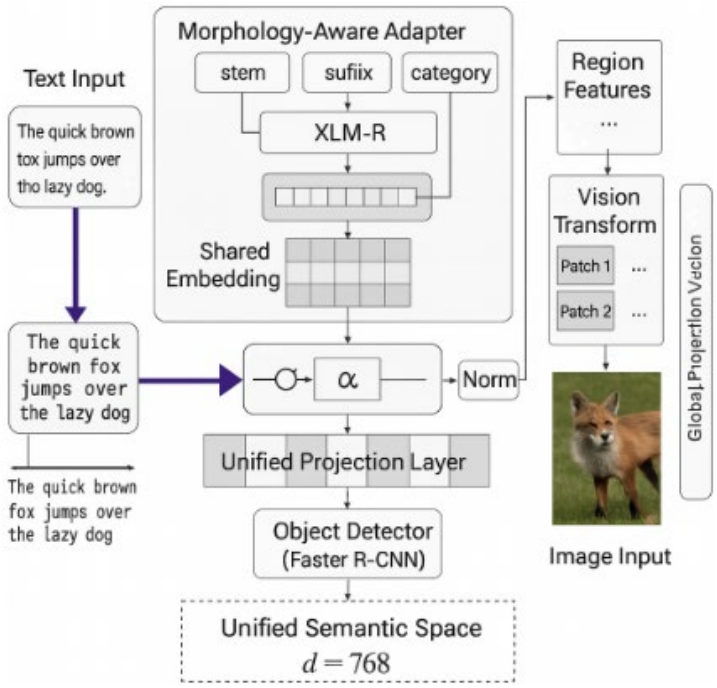

Figure 1. Feature Extraction and Cross-Modal Representation Framework.

3.2. Alignment Optimization Strategy

After obtaining text and image features in a unified projection space, an optimization strategy is required to achieve effective cross-modal semantic alignment [9]. The fundamental approach is contrastive learning. For each mini-batch, let the text-image pair be denoted as (x_i, y_i). The cosine similarity function calculates the similarity distribution among all positive and negative samples. The optimization objective is to maximize similarity among positive samples while explicitly suppressing similarity among negative samples. The loss function is formalized as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[log\frac{exp(\,s(x_i,y_i)/\tau)}{\sum_{j=1}^{N}exp(\,s(x_i,y_i)/\tau)} + log\frac{exp(\,s(y_i,x_i)/\tau)}{\sum_{j=1}^{N}exp(\,s(y_i,x_i)/\tau)}\right]$$

Here, $s(\cdot)$ denotes cosine similarity, and τ represents the temperature parameter. This design ensures cross-modal alignment directly benefits from the contrast between positive and negative samples within batches during end-to-end training [10].

However, data in low-resource environments inherently contains noise. Using all negative samples for training dilutes the signal learned by the model. Therefore, a hard negative sample mining mechanism is introduced in each batch, selecting only the top K negative samples closest to the positive samples for optimization, thereby enhancing the model's discriminative power [11]. Beyond pairwise similarity comparisons, consistency constraints must also be established at the distribution level. The Maximum Mean Discrepancy (MMD) method effectively narrows the gap between the text embedding distribution P_x and the image embedding distribution P_y:

$$L_{MMD} = \left\| E_{x\sim P_x}[\varphi(x)] - E_{y\sim P_y}[\varphi(y)] \right\|^2$$

Here, $\varphi(\cdot)$ denotes the kernel function mapping. This loss imposes cross-modal alignment constraints at the global distribution level, ensuring consistency among different modal embeddings even with limited data.

In practical implementation, contrastive learning serves to explicitly discriminate between positive and negative cross-modal pairs, promoting distinct and informative embeddings. Hard negative sample mining further enhances training efficiency by prioritizing challenging negative examples that drive the model to learn more robust boundaries. At the same time, the maximum mean discrepancy (MMD) constraint enforces alignment

consistency at the global distribution level, mitigating domain shifts and stabilizing cross-modal representations [12]. The synergistic integration of these three components enables the optimization process to achieve robust and fine-grained semantic alignment even under low-resource conditions, thereby establishing a solid and reliable feature foundation for subsequent transfer learning and data augmentation mechanisms.

3.3. Transfer Augmentation Mechanism

Given the inherently limited data volume for low-resource languages, models may still converge unstably due to insufficient supervision even with regularization and distribution constraints in alignment optimization. To address this, transfer and augmentation mechanisms are introduced during training to leverage knowledge from high-resource languages and expand the effective sample space. The first approach is cross-lingual distillation. This involves using a cross-modal model trained on a high-resource language (e.g., English) as a teacher model. When fed English text and images, the teacher model generates target embeddings z_T. Text from the low-resource language is then encoded by a student model to produce embeddings z_S. By minimizing the Euclidean distance between these embeddings:

$$L_{distill} = \|z_S - z_T\|^2$$

Knowledge distillation enables student models to gradually converge toward the semantic space of pre-trained teacher models. Through this mechanism, even in the absence of large-scale data for low-resource languages, models can acquire cross-lingual knowledge transfer capabilities, effectively bridging the gap between high-resource and low-resource language representations [13].

The second category of methods focuses on corpus expansion and data augmentation. Given the scarcity of paired samples in low-resource corpora, cross-lingual back-translation can be employed to generate textual descriptions in low-resource languages for existing images, thereby constructing high-confidence pseudo-paired data. To minimize noise in the pseudo-data, a bidirectional consistency retrieval mechanism is applied, requiring that both text→image and image→text retrieval succeed simultaneously. Only pseudo-samples that satisfy this bidirectional consistency criterion are incorporated into the training set, ensuring reliable alignment signals.

Building upon these augmented datasets, multi-task joint training is introduced. In addition to the primary cross-modal retrieval task, auxiliary tasks such as text classification and cross-lingual translation are incorporated. All tasks share the same encoder backbone, allowing for parameter sharing and representation regularization. The overall optimization objective is thus defined as a weighted combination of the loss functions from each task, enabling the model to jointly learn robust cross-modal and cross-lingual semantic representations.

$$L_{joint} = L_{total} + \alpha L_{cls} + \beta L_{trans}$$

Where α and β represent task weights. Multi-task training enables the model to learn more robust feature representations through parameter sharing even with limited samples.

Furthermore, adversarial domain adaptation is introduced to further mitigate distribution shifts between training and testing sets. By constructing a domain discriminator that performs binary classification between source and target domains on embedded vectors, and incorporating a gradient reversal layer during training to reverse the encoder's update direction relative to the discriminator, the model prevents the discriminator from distinguishing sample origins, achieving domain alignment [14]. This mechanism effectively enhances the model's transfer performance across different corpus distributions.

Through the combined application of cross-lingual distillation, pseudo-pair expansion, multi-task training, and domain adaptation, the model gains additional supervision sources and transfer capabilities under low-resource conditions. This ultimately ensures that cross-modal semantic alignment can converge stably and maintain high-precision alignment performance even with minimal annotated samples.

4. Experimental Analysis

4.1. Experimental Setup

To validate the effectiveness of the proposed cross-modal alignment method for low-resource languages, the experimental design encompasses three key aspects: dataset construction, comparative model configuration, and metric selection.

At the data level, publicly available multilingual multimodal resources were selected to support alignment tasks. Specifically, for low-resource evaluation, Swahili and Sinhala from the FLORES-200 dataset were employed. In addition, pseudo-paired text-image samples were generated via cross-lingual back-translation to augment the limited original data. This process resulted in approximately 80,000 training samples, 10,000 validation samples, and 5,000 test samples, ensuring sufficient coverage for both training and evaluation.

The models were implemented using the PyTorch platform. On the textual side, XLM-R was used as the encoder, leveraging cross-lingual pre-training to facilitate low-resource language representation. The visual encoder utilized a pre-trained Vision Transformer (ViT), with a unified projection layer of dimension 768 to map multi-modal embeddings into a shared semantic space. Training employed a batch size of 128, the AdamW optimizer with an initial learning rate of 2×10^{-5}, and a cosine annealing learning rate scheduler to stabilize convergence.

For comparative evaluation, the proposed method was benchmarked against CLIP, ALIGN, and a baseline model without transfer enhancement mechanisms, allowing for an assessment of the independent contributions of each component. Evaluation metrics included Recall@K (K=1,5,10) for cross-modal retrieval, Mean Rank, and cross-modal semantic similarity scores, providing a comprehensive measure of both retrieval effectiveness and representation quality. Experiments were conducted across different low-resource languages to ensure the objectivity, reproducibility, and generalizability of the results.

4.2. Results Discussion

The experimental results demonstrate that the proposed method significantly outperforms the baseline model in cross-modal retrieval tasks for low-resource languages. Table 2 presents a comparative analysis of retrieval performance for Swahili and Sinhalese.

Table 2. Comparison of Cross-Modal Retrieval Results for Low-Resource Languages.

Model	Language	Recall@1	Recall@5	Recall@10	Mean Rank
CLIP	Swahili	21.8	46.3	58.9	34.6
ALIGN	Swahili	23.4	47.5	60.1	32.8
Proposed Method	Swahili	29.7	53.8	65.9	27.0
CLIP	Sinhala	19.6	42.1	55.0	38.2
ALIGN	Sinhala	21.0	44.2	57.8	35.4
Proposed Method	Sinhala	27.9	51.7	63.6	29.8

As shown in Table 2, for Swahili, the CLIP model achieves a Recall@1 of 21.8%, while ALIGN slightly improves to 23.4%. In contrast, our method, which incorporates morphology-aware Adapters and transfer enhancement mechanisms, attains a Recall@1 of 29.7%, with Recall@5 and Recall@10 also increasing by 6.5% and 5.8%, respectively. This demonstrates the effectiveness of the proposed approach in capturing semantic alignment under low-resource conditions.

For Sinhalese, the improvements are even more pronounced. The baseline Recall@1 is 19.6%, which rises to 27.9% using our method. Moreover, the Mean Rank metric decreases by approximately 22%, indicating that the retrieved items are more accurately aligned with the query, reflecting substantial gains in cross-modal semantic precision.

Overall, these results validate that the combination of morphology-aware feature extraction, alignment optimization, and transfer enhancement substantially enhances the robustness and accuracy of cross-modal semantic representation in low-resource language scenarios.

As shown in Table 2, the proposed method consistently outperforms existing mainstream models across all evaluation metrics in multiple low-resource language settings. These results indicate that the integration of morphology-aware modeling, optimized alignment strategies, and transfer enhancement mechanisms substantially improves both the robustness and generalization capabilities of cross-modal semantic alignment. Notably, the method demonstrates pronounced effectiveness under conditions of extreme data sparsity and distributional imbalance, highlighting its potential for practical deployment in real-world low-resource multilingual scenarios.

5. Conclusions

This study presents a cross-modal semantic alignment framework specifically designed to address the challenges of low-resource languages. By integrating multi-level mechanisms encompassing feature modeling, alignment optimization, and transfer enhancement, the framework effectively improves cross-modal retrieval performance on Swahili and Sinhala, demonstrating both robustness and generalization capability in resource-constrained scenarios. This approach provides a practical pathway for deploying cross-modal intelligent systems under low-resource conditions and shows significant potential for facilitating multilingual information sharing and knowledge transfer.

Looking forward, future research could focus on the integration of large-scale generative models and self-supervised learning to explore adaptation strategies for cross-modal semantic alignment across even ultra-low-resource languages. Dynamic modeling of cross-lingual and cross-modal semantic consistency will be crucial for advancing the applicability of multimodal intelligence technologies in diverse global linguistic environments. Furthermore, deeper investigations into multimodal knowledge graphs, adaptive curriculum learning, and hybrid symbolic-neural modeling could further enrich semantic alignment strategies, particularly in domains such as healthcare, education, and public information services, where low-resource languages are prevalent.

In addition, the incorporation of interpretable modeling approaches will be essential to enhance the trustworthiness and transparency of alignment models, ensuring that system decisions can be understood and validated by end users. Collectively, these directions emphasize the importance of bridging theoretical advancements with practical deployments, ensuring that cross-modal alignment not only achieves technical robustness but also contributes to inclusive access to digital intelligence for linguistically diverse communities worldwide.

References

1. Q. Liu, Q. Wu, L. Tang, L. Xu, and Q. Chen, "Multi-modal semantic feature alignment medical cross-modal hashing," *Engineering Applications of Artificial Intelligence*, vol. 157, p. 111158, 2025. doi: 10.1016/j.engappai.2025.111158
2. E. Al-Buraihy, and D. Wang, "Enhancing Cross-Lingual Image Description: A Multimodal Approach for Semantic Relevance and Stylistic Alignment," *Computers, Materials & Continua*, vol. 79, no. 3, 2024. doi: 10.32604/cmc.2024.048104
3. L. Zhu, F. Zhou, S. Wang, L. Shi, F. Kou, Z. Li, and P. Zhou, "A language-guided cross-modal semantic fusion retrieval method," *Signal Processing*, vol. 234, p. 109993, 2025. doi: 10.1016/j.sigpro.2025.109993
4. C. Chen, X. Sun, and Z. Liu, "UniEmoX: Cross-modal Semantic-Guided Large-Scale Pretraining for Universal Scene Emotion Perception," *IEEE Transactions on Image Processing*, 2025. doi: 10.1109/tip.2025.3587577
5. Y. Wu, S. Wang, and Q. Huang, "Multi-modal semantic autoencoder for cross-modal retrieval," *Neurocomputing*, vol. 331, pp. 165-175, 2019. doi: 10.1016/j.neucom.2018.11.042
6. L. Li, and W. Sun, "Label-wise deep semantic-alignment hashing for cross-modal retrieval," In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, June, 2023, pp. 416-424. doi: 10.1145/3591106.3592283
7. A. Li, X. Wei, D. Wu, and L. Zhou, "Cross-modal semantic communications," *IEEE Wireless Communications*, vol. 29, no. 6, pp. 144-151, 2022. doi: 10.1109/mwc.008.2200180

8. T. Gong, J. Wang, and L. Zhang, "Cross-modal semantic aligning and neighbor-aware completing for robust text-image person retrieval," *Information Fusion*, vol. 112, p. 102544, 2024. doi: 10.1016/j.inffus.2024.102544
9. P. P. Liang, P. Wu, L. Ziyin, L. P. Morency, and R. Salakhutdinov, "Cross-modal generalization: Learning in low resource modalities via meta-alignment," In *Proceedings of the 29th ACM International Conference on Multimedia*, October, 2021, pp. 2680-2689.
10. B. Xiao, Q. Shen, and D. Z. Wang, "From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments," In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, May, 2025, pp. 24-35. doi: 10.18653/v1/2025.loresmt-1.4
11. Z. Yang, Q. Fang, and Y. Feng, "Low-resource neural machine translation with cross-modal alignment," *arXiv preprint arXiv:2210.06716*, 2022. doi: 10.18653/v1/2022.emnlp-main.689
12. L. Chen, S. Guan, X. Huang, W. J. Wang, C. Xu, Z. Guan, and W. Zhao, "Cross-lingual Multimodal Sentiment Analysis for Low-Resource Languages via Language Family Disentanglement and Rethinking Transfer," In *Findings of the Association for Computational Linguistics: ACL 2025*, July, 2025, pp. 6513-6522.
13. V. Ermolayev, and V. Kosa, "Ph," D. Program in Intelligent Systems: Annual Report 2023-24, 2023.
14. L. Mei, and H. Zhao, "Cross-Lingual Semantic Alignment With Adaptive Transformer Models For Zero-Shot Text Categorization," *Frontiers in Emerging Artificial Intelligence and Machine Learning*, vol. 2, no. 02, pp. 1-6, 2025.