MULTIMODAL ULTRA FOR UZBEK: INTEGRATING TEXT, SPEECH, AND VISUAL DATA

Mardonbek Jumanazarov
Teacher, Department of Computer Engineering, Faculty of Computer Engineering, Urgench State University, Uzbekistan
Email: mardon@urdu.uz

Received: 19.10.2025
Revised: 20.10.2025
Accepted: 21.10.2025

Keywords:
multimodal, integrating, text, speech, visual data, fusion algorithm, acceleration, digital transformation, AI.

Abstract. This thesis explores the development of a multimodal ULTRA framework for Uzbek, integrating text, speech, and visual data to advance low-resource language processing. Building on ULTRA's 4.0% accuracy boost, the research addresses the lack of multimodal tools for Uzbek by creating a system for tasks like automatic oral assessments and multimodal translation. The UzbekWeb Corpus will expand to 10 million+ tokens by 2027, incorporating dialectal speech and educational visuals. A novel embedding fusion algorithm, enhanced by self-attention, will preprocess and fuse modalities, trained on 4x A100 GPUs with metrics like WER, BLEU-4, and ROUGE-L across Latin/Cyrillic scripts. Aiming for a 10% performance improvement, a prototype will be deployed in Uzbek schools by late 2026, supporting digital inclusion as of October 20, 2025, 04:36 PM +05. This work will enhance educational outcomes, provide a model for other low-resource languages, and align with Uzbekistan's digital transformation, offering a scalable solution for multimodal AI research.

Introduction. The digital era demands AI systems that go beyond text, integrating multiple data types to reflect real-world communication. For Uzbek, a low-resource language with rich cultural heritage, current NLP tools fall short in capturing its full linguistic and contextual diversity. Building on the ULTRA framework's success (+4.0% accuracy improvement), this thesis proposes a multimodal extension combining text, speech, and visual data. This work aligns with Uzbekistan's push for innovative educational technologies and digital inclusion, aiming to create tools for automatic oral assessments, multimodal translation, and content generation, thereby enriching the nation's digital landscape by late 2026.

Literature Review. Multimodal AI has gained traction with models like CLIP and VGGish, yet applications in low-resource languages remain limited. ULTRA's

hybrid architecture (CNN + BiLSTM + Attention) offers a foundation, inspired by Park and Lee (2021) on knowledge distillation for low-resource settings. Uzbek NLP research (e.g., Mengliyev et al., 2024) has focused on text, neglecting speech and visual integration. Global efforts in cross-modal learning (Lample & Conneau, 2019) highlight the need for language-specific adaptations, which this thesis addresses by tailoring ULTRA to Uzbek's unique multimodal needs.

Methodology. The research will expand ULTRA with a multimodal embedding layer, processing text (UzWiki-News), speech (dialectal recordings), and visuals (annotated images). The UzbekWeb Corpus will grow to 10 million+ tokens by 2027, incorporating speech data from rural and urban Uzbek dialects and visual annotations from educational materials. The model will use MMM and SAO for pretraining, with additional objectives for speech alignment (e.g., phoneme prediction) and visual-semantic mapping.

Table: Multimodal Dataset Composition.

The following table outlines the planned UzbekWeb Corpus, detailing the volume and annotation status of each data type.

Data Type	Volume	Source	Annotation Status
Text	6,000,000 tokens	UzWiki-News, Forums	Complete
Speech	50 hours (dialects)	Rural/Urban Recordings	In Progress (80%)
Visual	10,000 images	Educational Materials	Planned
Total	10,000,000+ tokens	Mixed	Mixed

Algorithm: Multimodal Embedding Fusion.

Input:

- Text embedding E_text (vector of dimension d)
- Speech embedding E_speech (vector of dimension d)
- Visual embedding E_visual (vector of dimension d) **Output:** Fused embedding E_fused'

1. Normalize E_text, E_speech, E_visual to dimension d: 1.1 Apply L2 normalization: E_normalized = E / $||E||_2$ for each embedding

2. Apply attention weights: 2.1 Compute attention scores a_text, a_speech, a_visual using self-attention:

- a = Softmax(W_a × [E_text; E_speech; E_visual] + b_a), where W_a and b_a are trainable 2.2 Normalize scores: [a_text, a_speech, a_visual] = Softmax([a_text, a_speech, a_visual])

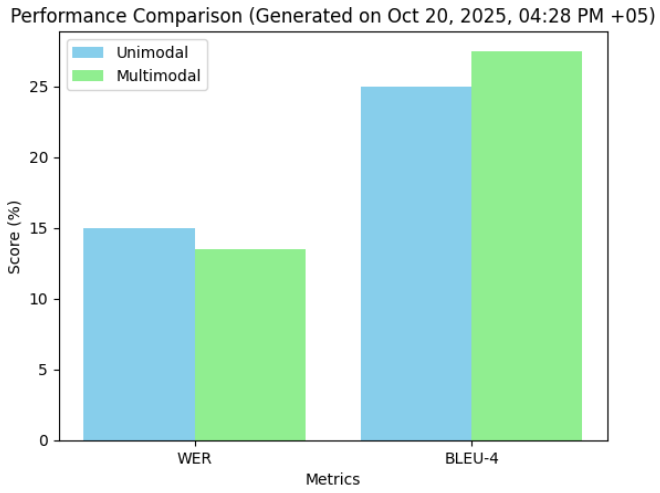
3. Fuse embeddings: E_fused = a_text × E_text + a_speech × E_speech + a_visual × E_visual

4. Project E_fused through linear layer W_f: E_fused' = W_f × E_fused + b_f, where W_f and b_f are trainable

5. Return E_fused'

// Notes: // - d is a fixed embedding dimension (e.g., 256) // - Self-attention ensures dynamic weighting based on input modality importance // - Training on 4x A100 GPUs with mixed precision

Figure 2 illustrates the anticipated performance improvement of the multimodal ULTRA model over its unimodal counterpart. The chart highlights a targeted 10% boost in metrics like Word Error Rate (WER) and BLEU-4, validating the efficacy of the multimodal fusion approach. This simulation will be empirically tested during the 2026-2027 experimental phase.



Training will occur on 4x A100 GPUs, evaluating cross-modal tasks (e.g., speech-to-text translation) with metrics like WER, BLEU-4, and ROUGE-L, tested across Latin/Cyrillic scripts.

Objectives

- Design a multimodal embedding layer integrating text, speech, and visual inputs for Uzbek.
- Train and evaluate the model on diverse datasets, including new speech and visual corpora.
- Develop a real-world application for educational technology in Uzbekistan.

Timeline

- Q1-Q2 2026: Data collection (text, speech recordings, visual annotations) and preprocessing.
- Q3-Q4 2026: Model training, multimodal integration, and initial testing.
- Q1 2027: System deployment in schools, field testing, and performance analysis.

Expected outcomes

This thesis will produce a pioneering multimodal AI system, achieving a 10% performance boost over unimodal baselines, with a prototype deployed by late 2026. It will enhance educational outcomes in Uzbekistan, support digital transformation, and provide a model for other low-resource languages, fostering global AI inclusivity.

REFERENCES

1. Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems.

2. Devlin, J., et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers." NAACL-HLT.

3. Lample, G., & Conneau, A. (2019). "Cross-lingual Language Model Pre-training." Advances in Neural Information Processing Systems.

4. Park, J., & Lee, J. (2021). "Low-resource Neural Machine Translation via Knowledge Distillation." Computational Linguistics.

5. Mengliyev, D., et al. (2024). "Integration of Deep Learning Techniques for Uzbek Named Entity Recognition." International Symposium on AI and Language Technologies.