

Article

Extending a Vision–Language Model with Audio Understanding: Introducing Qolda-AVL for the Kazakh Language

Batyr Arystanbekov ^{1,2} , Akylbek Maxutov ^{1,2,*} , Aspandiyar Nurimanov ^{1,2} and Huseyin Atakan Varol ^{1,2} 

¹ Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Astana 010000, Kazakhstan; batyr.arystanbekov@nu.edu.kz (B.A.); aspandiyar.nurimanov@gmail.com (A.N.); ahvarol@nu.edu.kz (H.A.V.)

² Department of AI & Big Data, Faculty of Information Technologies and Artificial Intelligence, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan

* Correspondence: akylbek.maxutov@nu.edu.kz

Abstract

Recent advances in multi-modal large language models have enabled systems to jointly process text, images, and audio. However, these developments have primarily benefited high-resource languages, leaving many low-resource communities underserved. In response, we introduce Qolda-AVL, a compact five-billion-parameter audio–vision–language model tailored for Kazakh. Qolda-AVL extends our previous Qolda vision–language model by adding a dedicated audio perception branch while maintaining strong visual and linguistic performance. Built on the Qwen3-VL-Thinking backbone, we incorporate Audio DeepStack, which transfers features from three intermediate Whisper encoder layers into the first three layers of the language model using dedicated projections and residual connections. The model is trained through a four-stage pipeline: adapting the Whisper encoder and language model to Kazakh, aligning the new audio branch to the language backbone, and jointly fine-tuning all modules on chain-of-thought reasoning tasks across audio, image, and text. All audio, vision, and language capabilities are evaluated using the model’s native reasoning mode, and a chain-of-thought trace is generated before each final answer during the performance assessment. To facilitate further research, we open-source the model along with the adapted Kazakh versions of four audio benchmarks, covering spoken attribute reasoning, spoken mathematical question answering, and audio captioning with question answering.

Keywords: audio–vision–language models; large language models; Kazakh natural language processing



Academic Editor: Giuseppe Maria Luigi Sarnè

Received: 30 April 2026

Revised: 4 June 2026

Accepted: 10 June 2026

Published: 15 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In recent years, large language models (LLMs) have evolved from text-only systems into architectures capable of jointly reasoning over text, images, and video. Early open-source models, such as LLaVA [1], InternVL [2], Molmo [3], and Qwen-VL [4], demonstrated that aligning visual encoders with LLMs yields powerful cross-modal reasoning capabilities. In parallel, similar advances have been made in integrating the audio modality with LLMs, and more recently, these lines of work have converged into omni-modal systems that span text, image, audio understanding, and, in some cases, speech generation, including the proprietary GPT-4o [5] and open-source flagship models such as the Qwen-Omni series [6,7].

The importance of incorporating the audio modality into an LLM backbone extends beyond simply unifying audio, vision, and text within a single model. Tasks, such as automatic speech recognition (ASR), audio classification, and speech-to-text translation (S2TT), are efficiently handled by relatively compact specialized models, most notably OpenAI’s Whisper [8], which is widely adopted and fine-tuned by the community for diverse speech-related tasks. However, tasks that require true reasoning over audio input, such as answering questions about a recording, performing multi-step inference that combines acoustic cues with world knowledge, or following complex instructions, cannot be fully addressed by specialized audio models alone. A transcription-then-text pipeline is similarly limited: it discards paralinguistic and environmental information at the ASR stage, leaving the downstream LLM to reason over incomplete input. This has motivated the research community to develop audio–language models that integrate a robust audio encoder directly with an LLM backbone, mirroring the encoder-plus-LLM design pattern established in vision–language models [9–11].

Despite this rapid progress in the multi-modal domain, LLMs are predominantly trained on data from high-resource languages such as English and Chinese, even multilingual ones. Many low-resource languages remain underserved as a result. Kazakh is a representative example: it is the official language of Kazakhstan, spoken by over 13 million people worldwide, yet it is substantially underrepresented in the pretraining corpora of frontier language systems. As a Turkic language with agglutinative morphology, in which words are constructed by stacking suffixes onto word stems, Kazakh additionally poses challenges for tokenization and various NLP tasks [12]. Several major initiatives in Kazakh-language resources and modeling, further discussed in detail in Section 2.4, have begun to close this gap, including our prior work on Qolda [13], the first vision–language model (VLM) designed specifically for Kazakh while preserving its multilingual capabilities. However, no open-source system to date integrates audio alongside vision and text capabilities tailored to Kazakh, leaving a clear gap in the multi-modal landscape for this language.

Our central research question is whether a compact, Kazakh-specialized tri-modal model can be built despite the limited or entirely absent training data, and whether it can remain competitive with much larger general-purpose systems. Addressing this requires a method for generating the missing multi-modal reasoning data, which is the main focus of our work. As a secondary, architectural question, we also explore how an audio modality can be injected into a pretrained vision–language model at multiple levels of abstraction without degrading its existing capabilities.

To address these questions, we present Qolda-AVL, a 5B-parameter tri-modal model that jointly processes audio, image, and text, with a particular focus on Kazakh. The name builds on our prior Qolda model [13], where *qolda* means “at hand” or “in support” in Kazakh, with AVL denoting the audio–vision–language modality scope. The model is built on the Qwen3-VL [14] backbone, and extended with a Whisper-based audio branch [8] and an Audio DeepStack mechanism that injects features from multiple encoder depths into the early layers of the language model decoder. To preserve the pretrained capabilities of each component while integrating audio into a Kazakh-centric model, we adopt a staged training pipeline that adapts individual modalities in isolation before joint tri-modal fine-tuning. Despite its compact size, Qolda-AVL establishes a strong baseline for Kazakh audio understanding, while preserving and, in several cases, improving upon the vision and language capabilities of its backbone. The main contributions of this work are summarized as follows:

- **Qolda-AVL**, the first open-source tri-modal (audio, vision, language) model designed specifically for Kazakh, released together with model weights, training code (<https://doi.org/10.3390/bdcc10060192>).

[//github.com/IS2AI/ms-swift-Qolda-AVL](https://github.com/IS2AI/ms-swift-Qolda-AVL) (accessed on 1 June 2026)), and inference code (<https://huggingface.co/issai/Qolda-AVL-5B> (accessed on 30 April 2026)).

- **Integration of Audio DeepStack into a vision–language backbone.** We adapt the hierarchical multi-level injection mechanism, originally proposed for vision in DeepStack [15] and Qwen3-VL [14], to the audio modality, routing features from three intermediate Whisper encoder layers into the early LLM decoder layers.
- **A staged training pipeline** for integrating an audio branch into a pretrained vision–language model, together with a synthetic chain-of-thought (CoT) generation methodology that constructs audio reasoning data for Kazakh through two complementary approaches: direct generation with a reasoning audio model, and refinement of intermediate-checkpoint outputs by a stronger external LLM.
- **A publicly released Kazakh audio understanding benchmark suite**, covering spoken attribute reasoning, spoken mathematical question answering, and audio captioning with QA. The suite is the first multi-task audio benchmark collection targeted at Kazakh and is released as a Hugging Face collection (<https://huggingface.co/collections/issai/qolda-avl-audio-benchmarks> (accessed on 29 April 2026)).

The remainder of this paper is organized as follows. Section 2 reviews related work on vision–language and audio–language models, omni-modal systems, and multi-modal resources for Kazakh. Section 3 describes the Qolda-AVL architecture, our staged training pipeline, and the evaluation protocol. Section 4 presents the benchmark results across the three modalities, along with an ablation study. Section 5 discusses our findings and their implications, and Section 6 concludes the paper.

2. Related Work

Our work builds on progress in vision–language modeling, audio–language integration, omni-modal architectures, and low-resource language adaptation. We structure this section accordingly, beginning with vision–language models, as many architectural patterns now used in audio–language systems originated in the vision domain. We then review audio–language models and omni-modal systems, and conclude with the landscape of NLP and multi-modal research for Kazakh and related Turkic languages.

2.1. Vision–Language Models

The LLaVA family [1] popularized a simple and scalable paradigm: connecting a pretrained CLIP [16] vision encoder to an instruction-tuned LLM through a lightweight projection layer, trained on synthetic instruction-following data. Subsequent refinements introduced dynamic high-resolution handling and unified single-image, multi-image, and video understanding, while preserving the efficiency of the projection-based design. Other lines of work explored deeper fusion strategies, including trainable visual experts integrated into every LLM layer (CogVLM [17]), scaled vision backbones with progressive alignment strategy (InternVL [2]), and explored hybrid encoders combining semantic and fine-details features (DeepSeek-VL [18]). Despite these variations, the projection-based design has remained the prevailing pattern in open VLMs.

The Qwen-VL series represents another influential line of development within this paradigm. Model iterations introduced a position-aware vision–language adapter, dynamic-resolution encoding, and Multi-modal Rotary Position Embedding (M-RoPE) as a unified positional scheme for text, images, and video [4,19,20]. Most recently, Qwen3-VL [14] introduced DeepStack integration for multi-level vision–language fusion, which is central to our work, alongside interleaved M-RoPE for balanced spatial–temporal modeling and text-based timestamp alignment for video. The DeepStack mechanism, originally proposed

in [15], departs from the standard practice of feeding all visual tokens into the first LLM layer. Instead, features are extracted from multiple depths of the vision encoder and routed to corresponding early layers of the LLM. This allows the language model to process visual information hierarchically across its depth with minimal additional computational cost. Qwen3-VL adapts this idea by extracting features from three distinct depths of its SigLIP-2 vision encoder, projecting each through a dedicated multi-layer perceptron (MLP) merger, and integrating them into the hidden states of the first three LLM layers via residual connections. Our work directly builds on the Qwen3-VL backbone and extends its DeepStack mechanism from the vision modality to audio.

2.2. Audio–Language Models

Modern audio–language models build on robust pretrained audio encoders. Self-supervised speech models such as wav2vec 2.0 [21] and HuBERT [22] demonstrated that pretraining on unlabeled speech produces transferable representations for a wide range of downstream speech tasks [23], while general (non-speech) audio understanding has been targeted by encoders such as BEATs [24] and CLAP variants [25,26]. Whisper [8], a Transformer encoder–decoder trained on 680K hours of weakly labeled multilingual speech, demonstrated strong zero-shot transfer for ASR and speech translation. The model has become a widely adopted audio backbone for downstream multi-modal systems [11,27], as its representations encode non-trivial information about non-speech sounds and acoustic events despite being supervised only with text transcriptions [28].

Early efforts to integrate audio with LLMs adopted modular, pipeline-based architectures, either chaining ASR, a text LLM, and text-to-speech (TTS) into sequential cascades, or positioning the LLM as a controller invoking specialized audio models on demand [29]. Such systems, however, suffer from error propagation, loss of paralinguistic information through intermediate text representations, and increased latency, motivating the development of end-to-end audio–language models.

Latent representation-based methods bypass the text bottleneck by projecting audio encoder outputs directly into the LLM’s embedding space through a modality adapter, mirroring the encoder–projector–LLM pattern established in the vision–language domain. Representative systems include SALMONN [11], which combines Whisper and BEATs encoders described earlier with a window-level Q-Former adapter to support speech, sound events, and music. Qwen-Audio [10] and its successor Qwen2-Audio [27] pair a Whisper-based encoder with an LLM through a multi-task pretraining framework covering over 30 audio tasks.

A critical design decision in latent representation-based models is modality adaptation: how to compress long, frame-level audio sequences into a token-level representation compatible with the LLM. Common strategies include convolutional downsampling, CTC-based compression, and Q-Former-based query mechanisms [9]. Most approaches extract features only from the final layer of the audio encoder and inject them at a single point in the LLM, typically the input embedding layer. This discards the hierarchical representations that emerge across encoder layers, where lower layers capture fine-grained acoustic–phonetic features and higher layers encode more abstract, semantic information. Different designs make different use of this hierarchy. GAMA [30] aggregates features from multiple Audio Spectrogram Transformer layers into a single fused representation injected at one point in the LLM, exploiting multi-layer information through feature fusion. An alternative is depth-aligned injection, in which features from distinct encoder depths are routed to distinct LLM layers, preserving the encoder’s layer-wise progression rather than collapsing it. The latter strategy has proven effective in the vision domain through DeepStack [15] and Qwen3-VL [14].

Our Audio DeepStack mechanism adapts this idea to audio by extracting features from three distinct Whisper encoder layers and injecting each into a corresponding early LLM layer through dedicated projectors and residual addition, directly mirroring Qwen's vision-side pipeline. Concurrent and independent work by MOSS-Audio [31] applies a conceptually similar DeepStack-inspired cross-layer injection for audio understanding. Our approach differs in two important respects. First, we build on a pretrained Whisper encoder with light domain fine-tuning, whereas MOSS-Audio trains a dedicated audio encoder from scratch. Second, our model extends Qwen3-VL rather than a text-only backbone, yielding a tri-modal system that jointly handles vision, audio, and text within a single LLM, whereas MOSS-Audio is a robust audio-only model.

2.3. Omni-Modal Systems

A growing line of work moves beyond bi-modal architectures toward models that jointly process text, images, video, and audio within a single framework. On the proprietary side, GPT-4o [5] and some variants of Gemini 3 [32] support native real-time audio–visual interaction, though their architectures remain undisclosed. Among open-source efforts, the Qwen-Omni series has been particularly influential. Qwen2.5-Omni [6] introduced the Thinker–Talker architecture for joint understanding and streaming speech generation. Qwen3-Omni [7] upgraded both components to mixture-of-experts (MoE) designs and replaced the Whisper audio encoder with a custom Audio Transformer trained from scratch on 20 million hours of supervised data. Other open-source systems, including VITA [33], Baichuan-Omni [34], OpenOmni [35], and Ola [36], have explored complementary strategies for omni-modal integration, typically through multi-stage training pipelines that progressively align modalities to mitigate cross-modal conflicts. Our training strategy follows a similar progressive, modular approach, sequentially adapting individual components before joint multi-modal training, as detailed in Section 3.

Our work does not aim to build a full omni-modal system with speech generation. Instead, we focus on the understanding side, adding an audio perception branch to an existing VLM while preserving its visual capabilities through modular design, with particular emphasis on adapting the resulting tri-modal model to the low-resource Kazakh language. In this sense, our contribution can be viewed as an intermediate step toward fully omni-modal systems for underserved linguistic communities.

2.4. Language and Multi-Modal Resources for Kazakh

Research on Kazakh NLP has advanced primarily through systematic resource construction. Foundational text datasets cover the main downstream tasks, including named entity recognition, sentiment analysis, open-domain question answering, and machine translation [12]. Subsequently, the first generation of Kazakh-centric large language models has emerged. KazLLM [37] and Sherkala [38] are both 8B-parameter models adapted from Llama 3.1 and trained on multilingual corpora covering Kazakh, English, Russian, and Turkish. However, beyond these Kazakh-specialized efforts, evaluation studies of general-purpose multilingual LLMs on Kazakh consistently report a performance gap relative to high-resource languages, particularly for smaller open-source models [39].

On the speech side, multiple open-source corpora are available for Kazakh, unified at industrial scale in KSC2 [40], a 1128h corpus covering television, radio, parliamentary sessions, and podcasts. On the modeling side, Whisper [8] includes Kazakh among its 99 supported languages, but its zero-shot performance remains limited, motivating several fine-tuning efforts. For instance, pseudo-labelling combined with a GPT-based language model has been shown to yield over 10% absolute WER reduction on Whisper for Kazakh [41]. Cross-lingual transfer has also been explored in TurkicASR [42], a multilingual

model spanning ten Turkic languages, and in Soyle [43], which fine-tunes Whisper jointly on eleven Turkic languages and the six official UN languages.

In contrast, Kazakh multi-modal research is only just emerging and has so far been limited to vision–language models. Qolda [13], our prior work, was among the first VLMs specifically adapted for Kazakh, built as a compact, parameter-efficient 4B model on top of InternVL3.5 and fine-tuned on trilingual (Kazakh, English, Russian) image–text data. A parallel effort, HordeVision [44], adapted the 8B Qwen3-VL model for Kazakh visual understanding with a relatively small amount of training data. No open-source system, however, currently integrates audio alongside vision and text for Kazakh. We emphasize that the bottleneck for Kazakh is not raw speech data or models, since KSC2 [40] alone provides more than a thousand hours and fine-tuned Whisper variants already perform competitively [41]. What is missing is joint audio–text reasoning data, which we synthesize (Section 3.2.5). The present work addresses this gap with Qolda-AVL, which extends Qolda into the tri-modal setting while remaining in the same compact, parameter-efficient design suitable for low-resource deployment.

3. Methodology

This section describes the design, training and evaluation of Qolda-AVL. We first present the model architecture, detailing the vision branch, the newly introduced audio branch, and their integration into the shared language backbone (Section 3.1). We then describe our staged training pipeline, which progressively integrates the audio modality into the pretrained vision–language model (Section 3.2). Finally, we outline the evaluation protocol used to assess language, vision, and audio capabilities (Section 3.3).

3.1. Model Architecture

Qolda-AVL extends Qwen3-VL from a bi-modal vision–language model into a tri-modal architecture that jointly processes audio, vision, and text within a single backbone. As illustrated in Figure 1, the model consists of three components: (i) a vision branch, comprising the SigLIP-2 encoder, its vision–language merger, and the associated Vision DeepStack modules, all inherited from Qwen3-VL; (ii) an audio branch, which is the primary contribution of this work and combines a Whisper-based encoder with a dedicated audio projection and the introduced Audio DeepStack mechanism; and (iii) a Qwen3 language model decoder that serves as the shared reasoning backbone across all modalities.

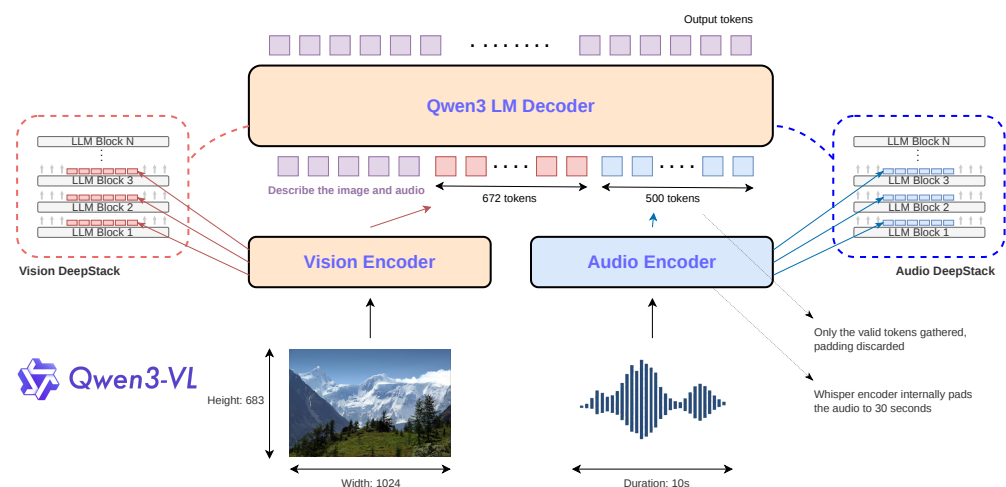


Figure 1. Overview of the Qolda-AVL architecture. The Qwen3 LM decoder processes a single sequence of interleaved text, visual, and audio tokens, while DeepStack modules inject multi-level features from both the vision encoder and the Whisper-based audio encoder into the first three decoder layers through residual connections.

3.1.1. Vision Branch

The vision branch follows the Qwen3-VL architecture without modification. It employs the SigLIP-2-Large encoder, a 24-layer Vision Transformer with a hidden dimension of 1024, that processes images at dynamic native resolutions using 2D Rotary Position Embeddings (2D-RoPEs). After encoding, a two-layer MLP merger compresses each 2×2 spatial patch group into a single visual token aligned with the LLM's hidden dimension of 2560.

In addition to the main merger that feeds into the first LLM layer, Qwen3-VL integrates a Vision DeepStack mechanism that extracts features from three intermediate ViT layers (layers 5, 11, and 17) and injects them into the hidden states of the first three LLM decoder layers via dedicated merger modules and residual addition. This multi-level injection preserves hierarchical visual information across the language model, mirroring the progression from lower-level to higher-level features that naturally emerges across ViT depth. All vision branch parameters, including the DeepStack mergers, are initialized from the pretrained Qwen3-VL-4B-Thinking checkpoint.

3.1.2. Audio Branch

The audio branch is the primary architectural contribution of Qolda-AVL. It mirrors the design principles of the vision branch, adapting the same encoder–projection–DeepStack pattern to the audio modality.

Audio Encoder. We use the encoder component of Whisper-large-v3-turbo [8], a 32-layer Transformer with a hidden dimension of 1280. Following standard Whisper preprocessing, input audio is resampled to 16 kHz and converted to a 128-channel log-Mel spectrogram, which is internally padded to 30 s (3000 mel frames). The encoder's convolutional frontend downsamples this by a factor of two along the time axis, producing up to 1500 frame-level representations. After encoding, only the tokens corresponding to the valid (non-padded) portion of the input are retained; for a typical 10 s utterance, this yields approximately 500 audio tokens.

Audio Projection. The final-layer features from the Whisper encoder are projected into the LLM's embedding space through a two-layer MLP with a GELU activation, preceded by a LayerNorm that stabilizes the encoder's output distribution. The MLP maps each audio frame from the encoder's 1280-dimensional space into the 2560-dimensional embedding space of the LLM decoder. Unlike the vision merger, which both compresses 2×2 spatial neighborhoods and projects them into the LLM hidden space, the audio module performs only the projection step, as audio frames form a one-dimensional temporal sequence with no spatial structure to merge. For this reason, we refer to it as a *projection* rather than a *merger*. The projected tokens are then inserted into the input sequence of the LLM decoder alongside text and visual tokens.

Audio DeepStack. We incorporate Audio DeepStack, which extends the DeepStack mechanism, previously applied to vision in Qwen3-VL, to the audio modality. Analogous to how the Vision DeepStack extracts features from three intermediate ViT layers, Audio DeepStack extracts hidden states from three intermediate Whisper encoder layers, specifically layers 8, 16, and 24 of the 32-layer encoder. These layers are evenly spaced across the encoder's depth (at approximately its quarter, half, and three-quarter points), directly mirroring the uniform-depth sampling used by the Vision DeepStack, which draws from ViT layers 5, 11, and 17 of its 24-layer encoder. This heuristic spreads the three injection points across low-, mid-, and high-level representations. However, we did not tune these indices, and a systematic study of the optimal injection depths is left for future work. Each set of extracted features is then processed by a dedicated MLP projection with the same architecture as the main audio projection, and the resulting representations are added

to the hidden states of the first three LLM decoder layers (0, 1, and 2) through residual connections, mirroring the injection points of the vision-side DeepStack.

An important design choice is the initialization of the Audio DeepStack modules. While the Vision DeepStack mergers carry pretrained weights from Qwen3-VL, the Audio DeepStack is newly introduced and must not disrupt the pretrained model's behavior at initialization. We therefore apply zero initialization to the last linear layer of each audio MLP projector. This ensures that at the start of training, the Audio DeepStack outputs are identically zero, and the model's behavior on text and vision inputs is exactly preserved. As training progresses, the projections gradually learn to inject meaningful audio features into the appropriate LLM layers.

This hierarchical injection design is motivated by the observation that, like vision encoders, audio encoders encode distinct kinds of information across their depth, with different layers being most informative for different audio tasks [28]. By routing features from different encoder depths to corresponding early LLM layers, Audio DeepStack allows the language model to process these representations at matching levels of abstraction, rather than compressing all audio information into a single injection point.

3.1.3. Multi-Modal Integration

The Qwen3 LM decoder serves as the shared backbone for reasoning over all three modalities. Text tokens are processed through the standard embedding layer, while visual tokens (from the vision merger) and audio tokens (from the audio projection) are inserted into the input sequence using modality-specific delimiter tokens: `<|vision_start|>/<|vision_end|>` for images and `<|audio_start|>/<|audio_end|>` for audio. This allows the model to handle arbitrary combinations of text, images, and audio within a single input sequence.

Positional information is encoded using Interleaved MRoPE [14], which distributes temporal, horizontal, and vertical position components across the embedding dimensions in an interleaved fashion to produce a balanced frequency spectrum. Visual tokens receive three-dimensional position IDs reflecting their spatial and temporal layout, while audio tokens are assigned one-dimensional sequential position IDs, with the same value replicated across all three MRoPE axes to reflect the sequential nature of audio signals. The DeepStack injection follows the same pattern for both modalities: at LLM decoder layers 0, 1, and 2, the corresponding DeepStack features from the vision and audio encoders (when present) are added to the hidden states via residual connections, using separate vision and audio projection modules.

3.2. Training Pipeline

Training Qolda-AVL requires integrating a new audio modality into a pretrained vision–language model without degrading its existing visual and linguistic capabilities. Following the progressive, modular paradigm established by prior models (e.g., VITA-1.5 [33] and Ola [36]), we adopt a stage-wise strategy that first prepares the individual components in isolation and then gradually connects them, progressing from uni-modal adaptation to joint tri-modal training. The full pipeline (see Figure 2) comprises four stages: two preparatory stages that independently adapt the audio encoder and the language model to the target languages and domain (Stages 1.1 and 1.2), an audio alignment stage that connects the adapted encoder to the LLM backbone (Stage 2), and a joint multi-modal fine-tuning stage that trains the complete tri-modal system end-to-end (Stage 3).

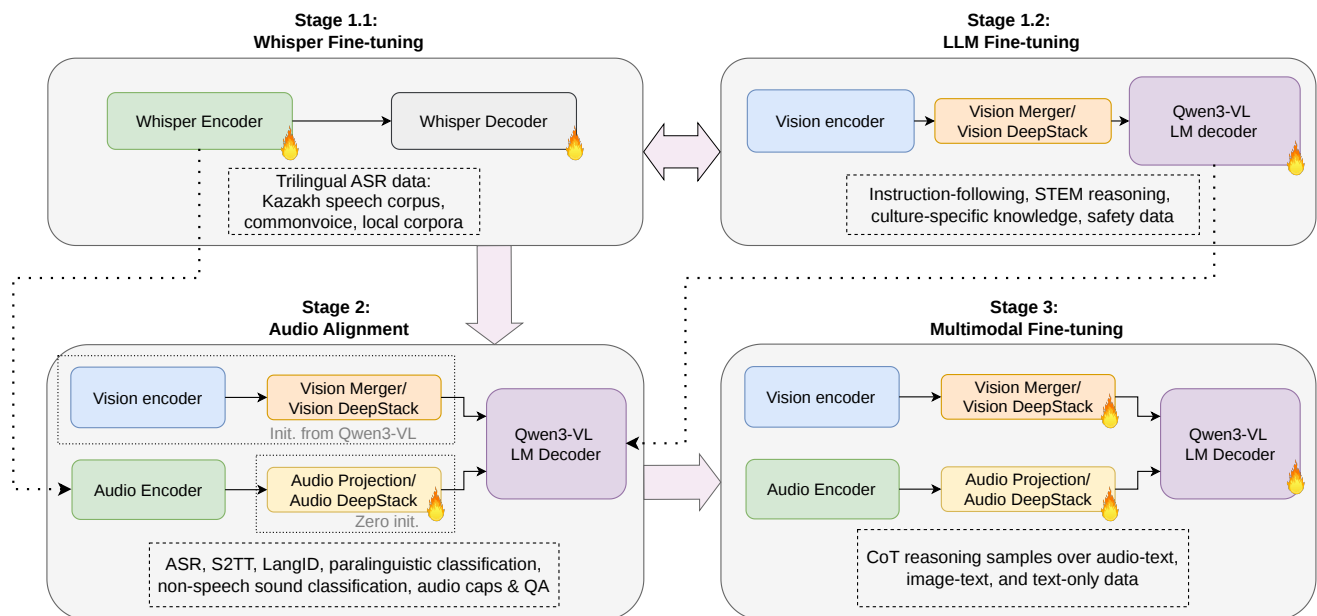


Figure 2. Overview of the Qolda-AVL training pipeline. Stages 1.1 and 1.2 are preparatory: Stage 1.1 fine-tunes the Whisper encoder–decoder, while Stage 1.2 adapts the Qwen3-VL LM decoder with the vision branch kept frozen, both on trilingual data with a focus on Kazakh. In Stage 2, only the zero-initialized audio projection and Audio DeepStack modules are trained on a diverse mixture of audio understanding tasks. Stage 3 performs joint multi-modal fine-tuning on CoT reasoning samples drawn from audio–text, image–text, and text-only data, updating the audio modules, the Vision Merger and Vision DeepStack, and the LM decoder. Flame icons indicate trainable components at each stage; dotted arrows indicate component initialization from prior stages.

3.2.1. Stage 1.1: Whisper Fine-Tuning

The goal of this stage is to strengthen the Whisper encoder’s representations for Kazakh, which is underrepresented in the original Whisper pretraining data, while preserving its multilingual capabilities for English and Russian. This stage operates on the standalone Whisper encoder–decoder model, independently of the multi-modal architecture. The training corpus combines trilingual ASR data, which includes KSC2 [40], Mozilla Common Voice [45] samples, and, additionally, collected speech recordings. In total, the data comprises 1128 h of Kazakh, 580 h of Russian, and 1000 h of English speech. Training follows the standard ASR objective: given a log-Mel spectrogram input, the model is trained to autoregressively produce the corresponding text transcript, with a cross-entropy loss computed over the decoder output tokens. We fine-tune both the Whisper encoder and decoder end-to-end; no other components of the Qolda-AVL architecture are involved at this stage. The decoder is required to provide the ASR training signal: Whisper is an encoder–decoder model, and the encoder is adapted by backpropagating the decoder’s transcription loss through the full model. After this stage, the decoder is discarded, as its role is taken over by the LLM backbone in the multi-modal architecture.

3.2.2. Stage 1.2: LLM Fine-Tuning

In parallel with Stage 1.1, we fine-tune the Qwen3-VL-4B-Thinking model on Kazakh-focused text-only data. This stage is dedicated to language adaptation of the LLM decoder that will subsequently serve as the shared backbone of the multi-modal architecture. The training corpus spans instruction-following examples in Kazakh, English, and Russian; STEM reasoning problems; culture-specific knowledge about Kazakhstan sourced from specialized Kazakh datasets and related materials; and safety-oriented samples. We rely

heavily on the text-only data collection introduced in our previous work on the Qolda model [13]. In total, the training corpus comprises over 5M multilingual samples. We keep the Qwen3-VL tokenizer unchanged. Our Kazakh adaptation modifies data rather than vocabulary, so the encoder vocabulary remains stable throughout training. We conduct supervised fine-tuning (SFT), updating only the LLM decoder of the Qwen3-VL checkpoint while keeping the vision encoder, vision merger, and Vision DeepStack modules frozen. Although fine-tuning the decoder on text-only data can, in principle, induce representation drift at the vision–language interface, we observe no noticeable degradation in visual capabilities after this stage. Moreover, any residual misalignment is further mitigated by the inclusion of vision–language data in Stage 3.

3.2.3. Model Assembly

Following the Stages 1.1 and 1.2, the full Qolda-AVL architecture is assembled by combining components from both initial stages. The LLM decoder, together with the vision encoder, vision merger, and Vision DeepStack mergers, is initialized from the Stage 1.2 checkpoint, while the Whisper encoder is initialized from the Stage 1.1 checkpoint. The audio projection and Audio DeepStack projections are newly instantiated with zero-initialized final layers, as described in Section 3.1.2, so that the assembled model is functionally identical to the Stage 1.2 checkpoint at initialization and its behavior on text-only and vision–language inputs is fully preserved. This assembly step involves no parameter updates and serves solely as the starting point for the subsequent multi-modal training stages.

The model contains approximately 5B parameters in total, distributed across a 4B-parameter Qwen-VL language backbone, a 400M-parameter ViT vision encoder, a 600M-parameter Whisper-large audio encoder, and lightweight projection and DeepStack modules for audio and vision. Input token embeddings and the output language modeling head are tied.

3.2.4. Stage 2: Audio Alignment

Once the full Qolda-AVL architecture has been assembled, the next stage trains the newly introduced components that bridge audio perception and language reasoning. Specifically, we update only the audio projection and the Audio DeepStack modules, so that the Whisper-encoded representations are mapped into a form over which the LLM backbone can reason effectively. At this stage, we expose the model to a diverse set of audio understanding tasks that span speech content, speaker characteristics, and environmental sound. The training mixture covers the following task categories:

- *Automatic Speech Recognition (ASR)*: The model receives a speech input and is asked to produce the corresponding text transcript. The ASR mixture is predominantly Kazakh (~3000 h), complemented by English (~1000 h) and Russian (~600 h), with a smaller residual portion (under 200 h) covering other languages.
- *Speech-to-Text Translation (S2TT)*: The model is given speech in a source language and is asked to generate a written translation in a target language. Training covers directional pairs among English, Russian, Kazakh, and Turkish, thereby exposing the model to cross-lingual understanding of spoken input.
- *Language Identification (LangID)*: Given a short speech clip, the model is asked to predict the spoken language. The training set spans 17 languages drawn from a curated subset of VoxLingua107 [46], with a deliberate emphasis on the Turkic family (Kazakh, Azerbaijani, Turkish, Turkmen, Tatar, Uzbek), alongside a selection of major world languages (Arabic, German, English, Spanish, Persian, French, Italian, Japanese, Korean, Russian, Chinese).

- *Audio Question Answering (audio QA)*: Given an audio clip and a natural-language question, the model is asked to produce a short factual answer. The data is drawn from the MECAT dataset [47], whose questions probe different aspects of the provided audio, including direct perception of events, sound characteristics, acoustic environment, and contextual reasoning. The text annotations are in English.
- *Audio Captioning*: The model is asked to produce a natural-language description of an audio clip. The data is again based on MECAT, and the audio clips are shared with the audio QA subset. Six prompt variants target complementary aspects of the recording: a long-form description of the overall content, a short one-sentence summary, the acoustic environment, the speech content, the musical content, and the discrete sound events. This ensures that a single clip can be described from multiple perspectives.
- *Gender and Emotion Recognition*: The model is asked to classify the speaker's gender and/or expressed emotion from a single utterance. The data for this task is entirely in English.
- *Audio Classification*: The model is asked to assign a categorical label to a non-speech audio clip, for example, in the classification of animal sounds.

Table 1 reports summary statistics for each category, including the number of training samples, the total and unique audio hours, and the underlying datasets. The role of this data mixture is to establish alignment between the pretrained audio encoder and the LLM backbone, teaching the model to interpret a wide range of audio inputs spanning both speech and non-speech content.

We note, however, that some categories, in particular audio QA, audio captioning, and audio classification, have not yet been diversified in terms of prompt variants or languages and remain predominantly English at this stage. We defer the enrichment of these categories to the subsequent training stage, in which the model is further trained on a smaller but more linguistically and structurally diverse set of reasoning-oriented data. We also note that this Stage 2 mixture is strongly imbalanced toward speech: ASR alone accounts for 4750 row-h, whereas non-speech audio classification contributes only 3 h. In this work, we use the available public corpora as is and do not rebalance the mixture, neither by upsampling the underrepresented task categories nor by collecting additional non-speech audio. As we discuss in Section 5, this imbalance is the primary factor limiting performance on the most data-scarce categories, particularly non-speech animal sound recognition, and we leave its mitigation to future work.

At this training stage, only the audio projection and Audio DeepStack modules are updated, while all other components, namely the LLM backbone, the audio and vision encoders, the vision merger, and the Vision DeepStack modules, remain frozen.

Table 1. Stage 2 training data distribution across tasks. *Row-wise* hours sum clip length over every training row, while *unique hours* deduplicate clips that carry multiple annotations: paraphrased prompts, re-structured tasks (e.g., joint ASR+emotion on KazEmoTTS), or translated targets.

Task	Samples	Row-Wise Hours	Unique Hours	Primary Audio Sources
ASR	2780K	4750	3620	KSC2 [40], CommonVoice [45], KazEmoTTS [48], Golos [49], YouTube transcripts
S2TT	833K	1078	1078	CommonVoice [45], KazEmoTTS [48], VCTK [50], Hi-Fi TTS [51]
LangID	362K	978	978	VoxLingua107 [46] (17-language subset of major world and Turkic languages)
Audio QA	100K	279	56	MECAT-QA [47] (train subset)
Audio Captioning	89K	249	56	MECAT-Caption [47] (shares audio clips with MECAT-QA)
Gender/Emotion	37K	28	10	IEMOCAP [52], in-house gender + emotion set
Audio Classification	2K	3	3	ESC-50 [53]
Total	4.2M	7365	5800	

3.2.5. Stage 3: Multi-Modal Fine-Tuning

After audio alignment, we next aim to improve the model's reasoning ability over audio input. The standard procedure is to unfreeze the LLM backbone and continue training the full model on high-quality data. However, training on a single modality, such as audio alone, can noticeably degrade performance on the other modalities, i.e., text and vision. For this reason, in this stage we unfreeze the audio and vision branches together with the LLM backbone and train the model jointly on mixed-modality data.

Since the original Qwen3-VL is a reasoning model, i.e., it produces an internal thinking trace before the final answer, we rely on CoT samples to preserve this behavior. For image and text data, we reuse the exact reasoning image-text and text-only collection that we constructed for our previous Qolda model; we refer the reader to the Qolda paper [13] (Stage 3, Joint Fine-tuning) for a detailed description of how this data was assembled.

However, the situation is different for audio. Datasets pairing audio clips with reasoning annotations are still rare, and no corpus of this kind exists for Kazakh. Our methodology therefore relies on synthesizing such data from the audio material already used in Stage 2. The underlying idea is simple: given an audio clip, its prompt, and the corresponding ground-truth (GT) answer, a CoT trace that leads to that answer can be reconstructed. Depending on the task domain and the target language, we adopt two complementary approaches: (i) direct generation with a reasoning audio model, and (ii) refinement of Qolda-AVL's own output with a capable language model.

Approach 1: Generation with a Reasoning Audio Model. In the first approach, we use the recent Step-Audio-R1 model [54] from StepFun Team. The authors explore reasoning in the audio domain in depth and propose Modality-Grounded Reasoning Distillation (MGRD), an iterative training framework that progressively shifts the model's reasoning basis from text-based patterns toward genuine acoustic analysis. Since running Step-Audio-R1 at scale is computationally expensive, we process only a sampled subset of the Stage 2 data. For each sampled audio clip, we feed both the clip and its corresponding prompt to Step-Audio-R1 and collect the reasoning output. We then use the GT answers as a filtering signal, retaining only the samples whose generated answer matches the reference. Step-Audio-R1 was predominantly trained on English and Chinese, so we apply it only to audio clips paired with English prompts and answers. For a limited number of Kazakh and Russian samples, we additionally query the proprietary Gemini-3-Flash [32] model to obtain reasoning traces.

Approach 2: Refining Qolda-AVL Output. The second approach is designed as an alternative for Kazakh samples, for which Step-Audio-R1 performs suboptimally. After the audio alignment stage, Qolda-AVL already exhibits a degree of reasoning ability over audio inputs, even though it has not yet been explicitly trained on audio CoT data. This emerges from the fact that the LLM backbone remains frozen during Stage 2: the model does not lose its reasoning ability but rather generalizes it to the newly aligned audio modality. In our experiments, this behavior is consistent. While not highly robust, the model's quality is sufficient for producing reasonable outputs given an audio input. However, a refinement step is still necessary to correct inaccurate CoT traces and incorrect answers from this intermediate, under-trained checkpoint. For this, we use the Qwen3.5-35B-A3B MoE model [55], which offers strong instruction following and multilingual capabilities, and apply it in text-only mode. The model receives the original prompt (in text form), the corresponding Qolda-AVL output, and the GT answer, and rewrites the output so that the reasoning trace leads coherently to the GT answer. Importantly, we do not regenerate the output from scratch; doing so would replace Qolda-AVL's reasoning style with a foreign one, which could undermine, rather than reinforce, the model's reasoning behavior during subsequent training.

We note one caveat regarding synthetic CoT generation. Neither teacher model in Approach 1 fully discloses its training corpus, so we cannot certify exact non-overlap with our benchmarks. Two design choices, however, limit any contamination risk. First, all synthetic traces are generated from Stage 2 training material, with evaluation splits held out entirely. Second, in Approach 1 we retain only samples whose answer matches the source-provided ground truth, so the teacher supplies a reasoning trace toward a pre-existing answer rather than label information. The Kazakh path (Approach 2) uses no external audio model at all.

We additionally diversify the training data by augmenting prompts with paraphrases, translations, and auxiliary formatting instructions (for instance, requests to return the answer in a particular style or format). Following these two approaches, we collect approximately two million samples covering around 3K unique audio hours. Roughly 40% of the data is in English, while the remaining 60% is predominantly Kazakh and Russian, with a small share of other languages. In terms of modality balance, audio samples account for about 50% of the mixture, image–text samples for 40%, and text-only samples for the remaining 10%, with every sample consisting of a CoT reasoning trace followed by the final answer.

During this stage, the audio and vision branches are trained jointly with the LLM backbone on mixed-modality data, with only the two encoders kept frozen. This design choice follows standard practice in both modalities: LLaVA-1.5 [56] and the Qwen-VL series [19,20] freeze the vision encoder during instruction tuning, while SALMONN [11] and Step-Audio-R1 [54] keep the Whisper-based audio encoder frozen throughout all training stages. We do this for two reasons. The pretrained encoders have been exposed to corpora orders of magnitude larger than our fine-tuning mixture, so unfreezing them on a smaller, narrower corpus risks catastrophic forgetting of the general-purpose representations that make them valuable in the first place. In addition, the Audio and Vision DeepStack mechanisms already give the LLM direct access to hidden states from multiple encoder depths, allowing it to select the appropriate level of abstraction per task without reshaping the encoders themselves.

3.2.6. Training Configuration

Stage 1.1 fine-tunes the Whisper encoder–decoder using fp16 mixed precision. We train for five epochs with AdamW at an initial learning rate of 1×10^{-5} , followed by a linear decay schedule with a warm-up ratio of 0.01. The generation maximum length is set to 225 tokens, consistent with the standard Whisper decoding budget for ASR. We use a per-device batch size of 20 with no gradient accumulation.

Stage 1.2 is implemented with the MS-Swift [57] framework using bfloat16 mixed precision and FlashAttention. We train for one epoch with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1) at an initial learning rate of 2×10^{-5} , followed by a cosine decay schedule with a warm-up ratio of 0.05 and gradient clipping at a maximum norm of 1.0. The maximum sequence length is set to 8192 tokens, and samples are concatenated through sequence packing to improve throughput. We conduct the training across 8× NVIDIA H100 80 GB GPUs via DeepSpeed ZeRO-1, with a per-device batch size of two and gradient accumulation over four steps, resulting in a global batch size of 64. The stage completes in approximately 30 h of wall-clock time.

For Stage 2, i.e., the audio alignment process, we train the audio projection modules with a higher initial learning rate of 1×10^{-4} , as these modules are trained from scratch. The maximum sequence length is set to 4096 tokens, which sufficiently accommodates the Whisper audio tokens together with the associated text prompt and target, given the absence of long reasoning traces in the training data of this stage. Most of the remaining

hyperparameters follow the previous stage. We conduct the training across 8 H100 GPUs, with a per-device batch size of 16 and gradient accumulation over four steps, resulting in a global batch size of 512. The training completes in 14 h.

For Stage 3, i.e., the joint multi-modal fine-tuning, we set the initial learning rate to 2×10^{-5} , lower than in Stage 2 since the LLM decoder is again unfrozen and updated together with the audio and vision branches. The maximum sequence length is set back to 8192 tokens to accommodate the longer CoT traces that characterize this stage's training data. Most of the remaining hyperparameters, again, follow the previous stages. Here, we conduct the training across 16 H100 GPUs, with a per-device batch size of two and gradient accumulation over eight steps, resulting in a global batch size of 256. The training completes in 38 h.

Although the audio module applies projection without temporal compression, unlike the 2×2 spatial merging on the vision side, this does not create sequence-length pressure. Table 2 reports the average sequence composition per modality across the training mixture. Because the audio clips are short (mean ≈ 7 –8 s), audio inputs contribute ≈ 310 tokens on average, comparable to the ≈ 367 tokens contributed by downscaled images, so projection-only audio does not yield heavier inputs than the vision branch in practice. For all three modalities, the full sequence is dominated by the CoT completion rather than the prompt. We filter out the rare over-length samples prior to training, so that all training data fit within the predefined 4096 and 8192 token windows in Stages 2 and 3, respectively. The average total sequence length did not exceed 1900 tokens. The remaining length-related constraint applies to audio specifically: Whisper internally pads audio to 30 s, capping audio-specific tokens at 1500 and bounding the duration of an individual clip rather than the LLM sequence budget.

Table 2. Average sequence composition per modality in the training mixture, reported in tokens. Prompt media is the audio or image token count; prompt total includes the text instruction. Completion is the CoT trace plus final answer. Full sequence is prompt plus completion. Values are corpus-weighted by dataset size; audio and image token counts are computed with the model's actual processors.

Modality	Prompt Text	Prompt Media	Prompt Total	Completion	Full Seq.
Text-only	139	0	≈ 139	1674	1813
Audio-text	25	310 (audio)	≈ 336	881	1217
Image-text	90	367 (image)	≈ 459	1410	1869

3.3. Evaluation

We assess our extended model across three dimensions: language-only performance, vision-language tasks, and audio understanding. For each, we use established English benchmarks to measure general capability retention, and Kazakh- and Russian-specific benchmarks to evaluate our adaptation approach. The evaluation methodology for language and vision follows our previous work on Qolda [13], using the same datasets and protocols. Qolda-AVL is evaluated across all three dimensions using its native reasoning mode, inherited from the Qwen3-VL-4B-Thinking backbone. The model generates an internal chain-of-thought trace before each final answer, and we report the final-answer metric throughout.

3.3.1. Language Evaluation

We assess linguistic, cognitive, and cultural proficiency using benchmarks organized by target capability. For general world knowledge and scientific reasoning, we use MMLU [58], MMLU-Pro [59], and GPQA [60]. We evaluate common-sense reasoning with ARC [61] (easy and challenge sets) and mathematical problem-solving with GSM8K [62].

For multilingual reading comprehension, we use Belebele [63] (English, Kazakh, and Russian splits). Machine translation is evaluated with FLORES-200 [64] across six directions involving Kazakh, Russian, and English.

For benchmarks without native Kazakh and Russian support, we use adapted versions from our previous work [13]. To assess specialized knowledge of Kazakhstan, we use KazMMLU [65], KazCulture [66], and KazQAD [67]. KazCulture is evaluated in two paradigms: question-only (Q) and passage question (PQ).

Accuracy is the primary metric for most tasks. For translation, while our previous work [13] reported BLEU scores, we now use xCOMET-XXL [68] as the main translation quality metric. This metric is evaluated in reference-based mode, with both the source and reference provided, and the results are reported in the $[0, 1]$ range. BLEU measures surface-level n -gram overlap with a reference translation, but it often correlates poorly with human judgments. This limitation is especially significant for morphologically rich, low-resource languages such as Kazakh, where accurate translations may differ from the reference at the lexical level. xCOMET is a neural metric that achieves state-of-the-art correlation with expert Multidimensional Quality Metrics (MQMs) annotations at the sentence, system, and error-span levels. In addition to providing a single quality score, it predicts fine-grained error spans with severity labels (minor, major, and critical), offering a more reliable quality signal and greater interpretability than n -gram-based metrics.

We report xCOMET-XXL scores using analytical confidence intervals based on each evaluation set's sample size and the empirical segment-level standard deviation. We estimate this standard deviation from a held-out set of $N = 4018$ machine translation outputs scored with xCOMET-XXL under the same protocol, yielding $au_{extseg} = 0.186$. We use this value as a working estimate across language pairs, noting that au_{extseg} varies modestly by system and translation direction. As a result, thresholds should be considered approximate. For FLORES-200 ($N = 2009$ segments per direction), the per-system standard error is approximately 0.004, with a 95% confidence interval half-width of about ± 0.008 . Pair-wise differences are significant at $p < 0.05$ when $|\Delta| > 0.011$ under a conservative two-sample approximation that does not use per-segment pairing. Differences below this threshold fall within measurement variance and should not be interpreted as evidence of system superiority. We apply the same procedure to xCOMET-XXL scores for speech-to-text translation on FLEURS, with thresholds adjusted for the smaller per-language test sets.

All benchmarks are evaluated in a zero-shot setting, except GSM8K, which uses a three-shot CoT approach. Models are prompted to output structured JSON for automated answer extraction.

3.3.2. Vision Evaluation

We evaluate multi-modal visual capabilities using AI2D [69] for diagrammatic reasoning, MMStar [70] and RealWorldQA [71] for general visual question answering (VQA), MathVista [72] for visual mathematical reasoning, and OCRBench [73] for optical character recognition (OCR).

For Kazakh and Russian, we use localized vision benchmarks introduced in [13]. AI2D and MathVista are adapted to Kazakh by embedding localized text in images. MMStar and RealWorldQA are adapted for both Kazakh and Russian. For Kazakh OCR, we use OCRBench-Kazakh [13], which includes regular text recognition, handwritten text recognition, and charts VQA. For Russian OCR, we use MWS-Vision-Bench [74].

Accuracy is the primary metric for all vision benchmarks. For OCR tasks requiring long-form responses, GPT-4o serves as the LLM-as-a-Judge. Evaluations are conducted in a zero-shot setting as in the language evaluation.

3.3.3. Audio Evaluation

We assess our model's audio understanding through five tasks: ASR, S2TT, spoken attribute reasoning, spoken mathematical question answering, and audio captioning with question answering. Due to the lack of Kazakh-language audio evaluation benchmarks, we adapt established English benchmarks to Kazakh and Russian using machine translation. All adapted and newly collected benchmarks are publicly available on ISSAI's Hugging Face (<https://huggingface.co/collections/issai/qolda-avl-audio-benchmarks> (accessed on 29 April 2026)).

For ASR, we evaluate transcription using the FLEURS [75] benchmark in English, Kazakh, and Russian. The primary metric is the Word Error Rate (WER), reported for both unnormalized and normalized settings, with normalization including lowercase conversion and punctuation removal. For each setting, we report the mean and interquartile range (IQR) of the cleaned mean, which excludes outliers to provide a more robust performance estimate. Using FLEURS, we also evaluate S2TT, which directly translates spoken audio into text in the target language. We assess six translation directions: Kazakh to English, English to Kazakh, Kazakh to Russian, Russian to Kazakh, Russian to English, and English to Russian. xCOMET-XXL [68] is the evaluation metric, the same as for text-to-text machine translation in language evaluation. The average xCOMET-XXL score for each model calculated as the mean across all six pairs.

For spoken attribute reasoning, we adapt the SAKURA benchmark [76] to Kazakh and Russian. This benchmark covers four tasks: gender identification, language identification, emotion recognition, and animal sound recognition. Each task is evaluated in two settings: single-hop, where the model extracts an attribute directly from the audio, and multi-hop, where further reasoning is required. Since the audio remains the same across languages, we adapt the benchmark by machine-translating instructions and answer options. The dataset uses a multiple-choice format, and accuracy is reported separately for each task and setting.

For spoken mathematical question answering, we adapt a Spoken Math Question Answering (MQA) dataset [77] to Kazakh. The dataset includes four formats: short-digit recognition, long-digit recognition, single-step reasoning, and multi-step reasoning. We created the Kazakh version by machine-translating transcriptions and synthesizing speech with the Mangisoz TTS [78] service. The exact match accuracy is reported for each format and as an overall average.

For audio captioning and question answering, we adapt the WavCaps [79] and WavCaps-QA [80] datasets to Kazakh and Russian using machine translation. These datasets feature recordings of natural and environmental sounds. In captioning, the model generates a textual description of the audio. In question answering, the model answers questions about the sound events in the recording. For both tasks, we use GPT-4o as an LLM-as-a-Judge to compare model outputs with ground-truth references and assign a binary semantic consistency score.

All baselines and Qolda-AVL receive the same task-specific instructions, formatted according to each model card's recommendations. For ASR, the instruction directs the model to perform automatic speech recognition on the input audio. For S2TT, the model is instructed to transcribe and translate the audio into the target language. For SAKURA, Spoken MQA, WavCaps, and WavCaps-QA, each model receives the benchmark's question text as the prompt. No baseline is evaluated in default chat mode.

Table 3 summarizes our audio evaluation benchmarks by the audio source (natural recordings or synthesized) and by textual reference source (native or machine-translated into Kazakh and Russian). Only the Kazakh split of Spoken MQA uses synthesized audio. All other benchmarks use original natural recordings, with machine translation applied only to the textual references where relevant.

Table 3. Audio evaluation benchmarks are categorized by both audio and text sources. Only the Kazakh split of Spoken MQA uses synthesized audio, generated by Mangisoz TTS from machine-translated transcripts. All other benchmarks use natural audio from their source datasets, and any machine translation is applied only to the text.

Task	Benchmark	Audio Source	Text Source
ASR	FLEURS [75]	Natural	Native
S2TT	FLEURS [75]	Natural	Native
Spoken attribute reasoning	SAKURA [76]	Natural	MT into KK/RU
Spoken mathematical QA	Spoken MQA [77]	Natural	Native for EN
		Synth. (KK via TTS)	MT into KK
Audio captioning	WavCaps [79]	Natural	MT into KK/RU
Audio captioning QA	WavCaps-QA [80]	Natural	MT into KK/RU

3.3.4. Model Selection and Comparison

Baseline Selection. We compare Qolda-AVL to open-source baselines selected by architectural lineage and modality coverage. For lineage, we include Qolda [13], the direct predecessor of Qolda-AVL, evaluated in both thinking and non-thinking modes, as well as Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Thinking [14], which serve as the shared upstream backbone. These baselines help isolate the impact of our multi-stage fine-tuning and assess the addition of the audio modality to existing text and vision capabilities. For modality coverage, we compare against Qwen2.5-Omni-3B and Qwen2.5-Omni-7B [6], which are similar in scale and support all three input modalities, and Qwen3-Omni-30B-A3B-Instruct and Qwen3-Omni-30B-A3B-Thinking [7], which represent the open-source state of the art and serve as an upper-bound reference. We note an important capability-class asymmetry between Qolda-AVL and these baselines in key ways. The Qwen-Omni family is omni-modal, supporting speech generation and, in the 30B variants, real-time audio–visual interaction. In contrast, Qolda-AVL is designed solely for understanding and does not generate speech. The Qwen3-Omni-30B variants serve as upper-bound references rather than direct comparators, as they are about six times larger than Qolda-AVL. To address baseline selection asymmetries and isolate the effect of multi-modal training, we include two specialized non-omni references. For language-only comparisons, we use Llama-3.1-KazLLM-1.0-8B [37], a text-only 8B LLM adapted for Kazakh. This allows us to evaluate whether adding multi-modal capabilities affects text-only Kazakh performance. For Kazakh ASR, we include Sybyrla [81], a specialized Kazakh ASR system, as a dedicated speech reference to measure the impact of joint multi-modal training on ASR quality. Qolda-AVL is a thinking-only model built on the Qwen3-VL-4B-Thinking backbone, and all evaluations are conducted in reasoning mode.

Generation Settings. To ensure fair and reproducible comparisons, all models are evaluated using the generation settings recommended in their model cards, applied consistently across text, vision, and audio benchmarks. For Qolda-AVL, which operates only in thinking mode, we set temperature to 1.0, top-p to 0.95, top-k to 20, repetition penalty to 1.0, and presence penalty to 0.0 for text and vision benchmarks, aligning with the Qwen3-VL-4B-Thinking backbone. For audio benchmarks, we reduce the temperature to 0.1 while maintaining top-p at 0.95 and top-k at 20 to minimize variance in transcription and speech translation outputs.

4. Results

We evaluate Qolda-AVL on language, vision, and audio benchmarks in Kazakh, English, and Russian, comparing it to open-source baselines in vision–language and omni-modal architectures. Our results show that adding the audio modality maintains and, in some cases, enhances the model’s existing language and vision capabilities. The main

contribution of this work is the audio evaluation, which includes ASR, S2TT, spoken attribute reasoning, audio captioning with QA, and Spoken MQA. We also present ablation studies to assess the effects of Audio DeepStack and the preparatory Whisper fine-tuning stage on audio performance.

4.1. Language Results

General Language Benchmarks. Qolda-AVL delivers the strongest performance on Kazakh among all models in its parameter class, outperforming both its predecessor Qolda and the Qwen3-VL-4B-Thinking across nearly every benchmark (see Table 4). Compared to Qolda-think, a strong Kazakh-focused baseline, Qolda-AVL improves average accuracy by about four points, with consistent gains in knowledge-intensive tasks (MMLU-Pro, MMLU), reasoning (GPQA, GSM8K), and reading comprehension (Belebele). It achieves the highest score on Belebele among all evaluated systems, including the larger Qwen3-Omni-30B-A3B variants. Qolda-AVL also outperforms Qwen3-VL-4B-Thinking on average and on most individual benchmarks, demonstrating that our multi-stage fine-tuning enhances Kazakh-language capability without compromising general reasoning. Only Qwen3-Omni-30B-A3B-Thinking, with approximately six times as many parameters, surpasses Qolda-AVL on Kazakh.

Table 4. Evaluation results of language benchmarks. All values are reported as accuracy (%). Bold values indicate the best result for each benchmark–language pair.

Lang.	Model	Average	MMLU-Pro	MMLU	GPQA	ARC-easy	ARC-chg.	GSM8K	Belebele
Kazakh	KazLLM-8B	54.39	25.98	47.65	28.61	79.92	65.19	59.14	74.22
	Qolda-nothink (4.3B)	62.95	31.60	55.81	30.87	88.09	77.30	77.66	79.35
	Qolda-think (4.3B)	73.45	59.49	68.63	39.41	94.82	87.29	84.60	79.93
	Qolda-AVL-5B (Ours)	77.44	63.39	74.45	43.95	95.75	90.87	87.79	85.88
	Qwen3-VL-4B-Instruct	59.18	39.11	52.70	36.10	75.93	66.16	75.59	68.67
	Qwen3-VL-4B-Thinking	74.60	62.64	67.81	56.62	89.66	81.82	85.43	78.21
	Qwen2.5-Omni-7B	38.06	23.57	39.54	31.16	46.78	41.75	30.96	52.67
	Qwen2.5-Omni-3B	26.66	16.75	30.94	23.80	32.54	30.98	11.38	40.22
	Qwen3-Omni-30B-Instruct	73.20	54.55	67.26	47.36	92.71	84.85	85.43	80.22
	Qwen3-Omni-30B-Thinking	83.24	72.66	79.41	66.37	95.51	93.27	90.59	84.86
English	KazLLM-8B	67.64	37.24	63.75	29.36	93.31	82.51	77.10	90.22
	Qolda-nothink (4.3B)	72.59	43.92	67.77	31.41	96.14	87.32	88.21	93.33
	Qolda-think (4.3B)	80.83	66.88	76.75	45.43	96.27	90.24	93.71	96.50
	Qolda-AVL-5B (Ours)	83.14	70.60	79.41	49.54	98.15	94.11	94.84	95.31
	Qwen3-VL-4B-Instruct	80.61	65.74	74.59	46.94	96.53	92.59	94.77	93.11
	Qwen3-VL-4B-Thinking	87.95	75.87	82.75	71.59	98.22	95.26	95.30	96.65
	Qwen2.5-Omni-7B	69.48	43.56	67.22	33.03	96.27	89.56	64.04	92.65
	Qwen2.5-Omni-3B	62.42	29.87	59.94	27.62	91.53	80.47	60.68	86.86
	Qwen3-Omni-30B-Instruct	84.53	69.27	81.77	54.53	98.39	96.46	95.30	96.00
	Qwen3-Omni-30B-Thinking	91.31	81.90	87.50	79.61	98.73	96.63	97.27	97.54
Russian	KazLLM-8B	62.21	27.96	50.47	27.18	88.95	76.95	75.99	88.00
	Qolda-nothink (4.3B)	69.58	38.13	59.90	29.96	93.09	86.36	90.27	89.33
	Qolda-think (4.3B)	77.78	61.94	70.27	43.43	95.77	90.00	93.62	89.40
	Qolda-AVL-5B (Ours)	80.70	67.17	76.95	46.33	96.84	94.89	92.65	90.07
	Qwen3-VL-4B-Instruct	76.21	57.83	67.05	46.65	94.49	88.38	89.53	89.56
	Qwen3-VL-4B-Thinking	82.83	68.48	74.98	61.46	97.88	92.59	92.56	91.78
	Qwen2.5-Omni-7B	63.18	36.45	59.28	29.96	90.08	78.96	60.39	87.11
	Qwen2.5-Omni-3B	55.34	24.33	49.68	28.89	82.63	71.04	51.59	79.19
	Qwen3-Omni-30B-Instruct	81.23	64.35	75.73	49.73	97.97	94.44	94.39	92.00
	Qwen3-Omni-30B-Thinking	88.33	78.02	83.81	71.75	98.81	96.30	95.90	93.75

Qolda-AVL outperforms Llama-3.1-KazLLM-1.0-8B, a specialized text-only 8B Kazakh LLM, by over 23 points on the Kazakh average (77.44 vs. 54.39), despite having fewer parameters and supporting vision and audio modalities. This demonstrates that adding multi-modal capabilities does not reduce text-only Kazakh performance, and a multi-modal Kazakh-focused model can exceed specialized text-only models on language benchmarks.

Qolda-AVL also shows improvement over its predecessor Qolda in English and Russian, gaining about three points on average in both languages. This is notable, as extending a vision–language model to include audio often risks regression in original modalities and languages. Qolda-AVL not only avoids regression but also improves on every English and Russian benchmark compared to Qolda. However, it trails Qwen3-VL-4B-Thinking by a small margin in English (83.14% vs. 87.95% average) and Russian (80.70% vs. 82.83% average), mainly due to lower performance on GPQA, the PhD-level reasoning benchmark. On English GPQA, the gap exceeds 20 points (49.54 vs. 71.59), indicating a trade-off in the LLM decoder toward Kazakh specialization. In contrast, the gap on ARC, GSM8K, Belebele, and MMLU variants is minimal or absent. Qwen3-VL-4B-Thinking benefits from broader general-purpose reasoning training, while Qolda-AVL is specialized for Kazakh and audio understanding. On most other benchmarks (ARC, GSM8K, Belebele), the difference is minimal, and Qolda-AVL remains competitive or slightly ahead.

The Qwen2.5-Omni baselines perform poorly across all three languages, especially in Kazakh, with averages of 38.06% and 26.66% for the 7B and 3B variants, respectively. This reflects their limited exposure to low-resource languages during pretraining. In contrast, the Qwen3-Omni-30B-A3B family sets the upper bound across all benchmarks and languages, with the thinking variant achieving the highest average in each language. This outcome is expected given its scale and highlights how closely a 5B-parameter Kazakh-specialized model can approach a frontier-scale general-purpose system.

Text-to-Text Machine Translation. As shown in Table 5, Qolda-AVL achieves an average xCOMET-XXL score of 0.7783, outperforming all comparably sized open multi-modal baselines, including Qwen3-VL-4B-Thinking (0.7105), Qwen3-VL-4B-Instruct (0.5933), Qwen2.5-Omni-7B (0.4667), and Qwen2.5-Omni-3B (0.3963). It also surpasses the much larger Qwen3-Omni-30B-Instruct (0.7625). Within the Qolda family, Qolda-AVL achieves the highest score on four of six directions (kk–ru, en–ru, ru–en, and nearly kk–en), but trails the text-only Qolda-think on the into-Kazakh directions (en–kk and ru–kk). This result suggests a modest trade-off in Kazakh-generation quality for the benefit of integrated vision and audio capabilities. Among all systems evaluated, only the reasoning-enabled Qwen3-Omni-30B-Thinking achieves a higher overall average (0.8371), despite being much larger than Qolda-AVL.

Table 5. Model performance on text-to-text machine translation benchmark. The results are xCOMET-XXL scores in the range of [0, 1]. Higher is better. All xCOMET-XXL scores carry an approximate 95% confidence interval of ± 0.008 ($\sigma_{\text{seg}} = 0.186$, $N = 2009$); pair-wise differences greater than 0.011 are statistically significant at $p < 0.05$. Bold values indicate the best result for each language pair.

Model	Average	en–kk	kk–en	ru–kk	kk–ru	en–ru	ru–en
KazLLM-8B	0.7156	0.5542	0.7426	0.4945	0.7139	0.8612	0.9272
Qolda-nothink (4.3B)	0.7858	0.7230	0.7590	0.7081	0.7529	0.8509	0.9210
Qolda-think (4.3B)	0.8040	0.7406	0.7851	0.7111	0.7843	0.8714	0.9313
Qolda-AVL-5B (Ours)	0.7783	0.6534	0.7845	0.6146	0.7926	0.8875	0.9370
Qwen3-VL-4B-Instruct	0.5933	0.2342	0.6474	0.2400	0.6330	0.8637	0.9412
Qwen3-VL-4B-Thinking	0.7105	0.4895	0.6784	0.5042	0.7250	0.9199	0.9461
Qwen2.5-Omni-7B	0.4667	0.1387	0.4086	0.1408	0.3817	0.7965	0.9341
Qwen2.5-Omni-3B	0.3963	0.1270	0.2526	0.1425	0.2545	0.6907	0.9103
Qwen3-Omni-30B-Instruct	0.7625	0.6007	0.7647	0.5601	0.7712	0.9284	0.9499
Qwen3-Omni-30B-Thinking	0.8371	0.7505	0.7933	0.7354	0.8440	0.9424	0.9572

Kazakh-Specific Benchmarks. Table 6 presents the results on the Kazakh-specific evaluation suite. Qolda-AVL remains strong but does not lead in all areas. It records the third-highest KazMMLU score (69.47%), just behind Qwen3-Omni-30B-A3B-Thinking (77.80%)

and Qwen3-Omni-30B-A3B-Instruct (70.38%), and performs competitively on KazQAD (76.19%) and KazCulture-PQ (97.68%). However, its KazCulture-Q score drops to 45.28%, below both Qolda-think (48.87%) and Qolda-nothink (49.13%). Notably, Qolda-AVL still outperforms its upstream backbone Qwen3-VL-4B-Thinking on KazMMLU, matches it on KazQAD, and substantially exceeds it on both KazCulture splits. This demonstrates that Kazakh linguistic and cultural proficiency is preserved in our model compared to the base architecture.

Table 6. Model performance on Kazakh-specific benchmarks. The results are given as accuracy (%). Bold values indicate the best result for each benchmark.

Model	Average	KazMMLU	KazQAD	KazCulture	
				PQ	Q
KazLLM-8B	64.72	51.32	73.30	96.85	37.41
Qolda-nothink (4.3B)	68.64	52.95	74.19	98.27	49.13
Qolda-think (4.3B)	72.62	66.12	77.58	97.90	48.87
Qolda-AVL-5B (Ours)	72.16	69.47	76.19	97.68	45.28
Qwen3-VL-4B-Instruct	68.10	61.73	76.47	97.15	37.03
Qwen3-VL-4B-Thinking	68.42	68.44	76.47	97.15	31.63
Qwen2.5-Omni-7B	55.14	49.12	52.65	93.40	25.38
Qwen2.5-Omni-3B	49.29	41.11	41.96	87.71	26.39
Qwen3-Omni-30B-Instruct	72.44	70.38	78.44	98.35	42.58
Qwen3-Omni-30B-Thinking	73.56	77.80	80.46	98.50	37.48

4.2. Vision Results

As shown in Table 7, Qolda-AVL delivers the highest performance among models in its parameter class on Kazakh vision benchmarks, surpassing both Qolda and Qwen3-VL-4B-Thinking. Qolda-AVL improves average accuracy by over nine points compared to Qolda-think (62.73% vs. 53.20%), with gains on AI2D, MMStar, and MathVista. Qolda-AVL also outperforms Qwen3-VL-4B-Thinking on average across all benchmarks, demonstrating that our multi-stage fine-tuning enhances the base model’s visual reasoning in the Kazakh context. Additionally, Qolda-AVL achieves the highest score on RealWorldQA (55.95%), outperforming the much larger Qwen3-Omni-30B-A3B-Thinking on this benchmark. Overall, Qolda-AVL nearly matches Qwen3-Omni-30B-A3B-Thinking (62.73% vs. 64.48%), despite the latter having approximately 6 times as many parameters.

On English benchmarks, Qolda-AVL again shows improvement over its predecessor (Qolda-think), increasing the average accuracy by more than ten points (75.45% vs. 64.23%) and delivering consistent gains across all tasks. This demonstrates that adding audio capabilities and using multi-stage training does not reduce performance on English vision tasks. Compared to Qwen3-VL-4B-Thinking, Qolda-AVL trails only slightly on average (75.45% vs. 76.39%).

On Russian vision benchmarks, Qolda-AVL achieves the highest average among comparably sized models (56.52%), exceeding Qolda-think by over eight points and outperforming both Qwen3-VL-4B variants. Qolda-AVL also leads Qwen3-VL-4B-Thinking (56.52% vs. 54.68%).

4.3. Audio Results

Automatic Speech Recognition. As shown in Table 8, Qolda-AVL achieves the highest Kazakh ASR performance among all evaluated models, with a normalized WER mean of 0.1874 and a cleaned value of 0.1648. The next-best model, Qwen3-Omni-30B-A3B-Thinking, has a normalized mean WER of 0.7175, despite its much larger size. Qwen2.5-Omni baselines perform poorly on Kazakh ASR, with WER values above 1.0, indicating more errors than reference words and limited Kazakh transcription capability. These results

support the decision to specialize Qolda-AVL for Kazakh audio, as no general-purpose open-source model in this parameter range matches its Kazakh speech recognition performance. For English and Russian ASR, Qolda-AVL remains competitive but does not lead. In English, Qwen3-Omni-30B-A3B-Thinking achieves the lowest normalized mean WER (0.0540) and also leads on the cleaned metric (0.0532 normalized). Qolda-AVL ranks third overall (0.0982 normalized mean), outperforming both Qwen2.5-Omni variants. In Russian, the two Qwen3-Omni-30B-A3B models lead, with Qolda-AVL close behind (0.1293 normalized mean WER versus 0.0944 to 0.0953). The performance gap to the 30B-parameter models is small, and Qolda-AVL consistently outperforms Qwen2.5-Omni baselines across all languages and metrics.

Table 7. Accuracy (%) of evaluated models on vision-based benchmarks. A dash (–) indicates that the benchmark does not exist in the corresponding language. Bold values indicate the best result for each benchmark–language pair.

Lang.	Model	Average	AI2D	MMStar	RealWorldQA	MathVista	OCRBench
Kazakh	Qolda-nothink (4.3B)	44.85	46.73	37.30	46.01	37.54	56.67
	Qolda-think (4.3B)	53.20	55.18	59.37	49.02	45.09	57.36
	Qolda-AVL-5B (Ours)	62.73	72.49	67.59	55.95	67.24	50.39
	Qwen3-VL-4B-Instruct	54.02	61.32	50.60	51.24	50.25	56.69
	Qwen3-VL-4B-Thinking	60.62	67.16	65.70	55.03	66.70	48.53
	Qwen2.5-Omni-7B	44.31	52.03	45.46	35.42	44.90	43.76
	Qwen2.5-Omni-3B	35.29	39.57	33.80	21.31	37.80	43.99
	Qwen3-Omni-30B-Instruct	60.50	73.15	61.06	54.90	54.42	58.96
	Qwen3-Omni-30B-Thinking	64.48	75.37	68.85	54.77	71.01	52.38
English	Qolda-nothink (4.3B)	52.16	69.14	36.96	53.33	40.20	61.16
	Qolda-think (4.3B)	64.23	73.07	60.88	60.00	68.10	59.08
	Qolda-AVL-5B (Ours)	75.45	81.41	70.62	68.63	74.57	82.00
	Qwen3-VL-4B-Instruct	73.75	81.41	65.12	71.11	65.90	85.20
	Qwen3-VL-4B-Thinking	76.39	83.47	73.88	67.97	75.72	80.90
	Qwen2.5-Omni-7B	69.78	80.94	61.17	65.10	58.60	83.10
	Qwen2.5-Omni-3B	64.46	76.23	53.13	56.21	51.95	84.80
	Qwen3-Omni-30B-Instruct	76.98	86.59	70.57	73.33	69.10	85.30
	Qwen3-Omni-30B-Thinking	80.91	88.64	74.48	74.77	79.76	86.90
Russian	Qolda-nothink (4.3B)	37.98	–	36.87	48.10	–	28.96
	Qolda-think (4.3B)	47.91	–	54.95	56.07	–	32.72
	Qolda-AVL-5B (Ours)	56.52	–	67.23	64.18	–	38.14
	Qwen3-VL-4B-Instruct	52.54	–	57.68	61.70	–	38.25
	Qwen3-VL-4B-Thinking	54.68	–	71.33	60.92	–	31.80
	Qwen2.5-Omni-7B	49.09	–	58.87	46.01	–	42.40
	Qwen2.5-Omni-3B	42.52	–	50.13	42.48	–	34.95
	Qwen3-Omni-30B-Instruct	60.30	–	68.11	65.49	–	47.31
	Qwen3-Omni-30B-Thinking	61.88	–	71.83	65.36	–	48.46

Qolda-AVL matches or slightly outperforms Sybyrla, a specialized Kazakh ASR system, on cleaned Kazakh WER (0.1801 vs. 0.1968 unnormalized; 0.1648 vs. 0.1681 normalized). This demonstrates that Qolda-AVL’s joint multi-modal training does not reduce speech recognition performance in Kazakh. The model gains additional vision and language capabilities without compromising ASR quality.

Speech-to-Text Translation. Table 9 shows results for the FLEURS S2TT benchmark. Qolda-AVL achieves the highest average xCOMET-XXL score (0.7349) among all systems, outperforming the much larger Qwen3-Omni-30B-Thinking (0.6067) and Qwen3-Omni-30B-Instruct (0.5718). This advantage is especially clear when Kazakh is the source language: on kk-en (0.6924) and kk-ru (0.6870), Qolda-AVL exceeds the strongest 30B baseline by more than 2.5 times, demonstrating the value of language-specialized multi-modal training for low-resource speech translation. Qolda-AVL also

leads on ru–kk (0.6461), surpassing both Qwen3-Omni-30B variants, and performs on par with Qwen3-Omni-30B-Thinking (0.6377) on en–kk (0.6364), while maintaining a clear lead over Qwen3-Omni-30B-Instruct (0.6024). For high-resource pairs en–ru and ru–en, the Qwen3-Omni-30B models hold a slight advantage (e.g., 0.9095 vs. 0.8420 on en–ru) due to their scale and broader pretraining, though Qolda-AVL remains competitive. In contrast, the Qwen2.5-Omni baselines perform poorly in all directions with Kazakh as the source (kk–en and kk–ru at approximately 0.14 xCOMET), confirming their lack of Kazakh speech understanding.

Table 8. Model performance on ASR. The results are reported as WER; lower is better. Each cell reports *unnorm* | *norm*, i.e., WER computed on unnormalized and normalized text, respectively. *Cleaned* denotes the mean WER computed over the IQR-cleaned subset of utterances (with outliers removed), providing a more robust estimate of central performance. Bold values indicate the best result for each language-denotation pair.

Model	Kazakh		English		Russian	
	Mean	Cleaned	Mean	Cleaned	Mean	Cleaned
Sybyrla	0.2244 0.1951	0.1968 0.1681	0.1555 0.1423	0.1140 0.0991	0.1253 0.1142	0.1131 0.0999
Qolda-AVL-5B (Ours)	0.2027 0.1874	0.1801 0.1648	0.1130 0.0982	0.0995 0.0876	0.1388 0.1293	0.1222 0.1129
Qwen2.5-Omni-7B	3.6722 3.6671	1.0413 1.0362	0.1407 0.1160	0.1317 0.1033	0.3014 0.2395	0.2755 0.2103
Qwen2.5-Omni-3B	7.5563 7.5534	1.0828 1.0796	0.7357 0.7096	0.1626 0.1418	0.2892 0.2537	0.2653 0.2277
Qwen3-Omni-30B-Instruct	0.8350 0.8173	0.6217 0.6037	0.1118 0.0963	0.0996 0.0804	0.1074 0.0944	0.0928 0.0811
Qwen3-Omni-30B-Thinking	0.7300 0.7175	0.7025 0.6890	0.1105 0.0540	0.0964 0.0532	0.1048 0.0953	0.0939 0.0821

Table 9. Model performance on S2TT. The results are xCOMET-XXL scores in the range [0, 1]. xCOMET-XXL scores on FLEURS carry an approximate 95% confidence interval of ± 0.010 ($\sigma_{\text{seg}} = 0.186$, $N = 1225$ on the Kazakh FLEURS test set); pair-wise differences greater than 0.015 are statistically significant at $p < 0.05$. Bold values indicate the best result for each language pair.

Model	Average	kk–en	kk–ru	en–kk	en–ru	ru–kk	ru–en
Qolda-AVL-5B (Ours)	0.7349	0.6924	0.6870	0.6364	0.8420	0.6461	0.9053
Qwen2.5-Omni-7B	0.4457	0.1430	0.1440	0.2508	0.8018	0.4447	0.8898
Qwen2.5-Omni-3B	0.4090	0.1411	0.1449	0.2046	0.7103	0.3938	0.8595
Qwen3-Omni-30B-Instruct	0.5718	0.1924	0.2061	0.6024	0.9095	0.5888	0.9315
Qwen3-Omni-30B-Thinking	0.6067	0.2728	0.2707	0.6377	0.9066	0.6199	0.9324

Spoken Attribute Reasoning. Table 10 shows the SAKURA QA benchmark results on audio understanding in four categories: speaker gender, animal sounds, spoken language identification, and emotion recognition. Each category is tested in both single-hop and multi-hop reasoning settings.

For Kazakh, Qwen3-Omni-30B-A3B-Thinking achieves the highest average accuracy (72.44%), followed by Qolda-AVL (64.02%). Although Qolda-AVL ranks second overall, it leads in gender multi-hop (78.80%) and emotion multi-hop (37.80%), while Qwen3-Omni-30B-A3B variants perform best in animal sound recognition and language identification. Qwen2.5-Omni baselines perform poorly across all categories, with averages below 50%.

A similar pattern appears for English and Russian. Qwen3-Omni-30B-A3B-Thinking achieves the highest averages (78.70% for English, 76.85% for Russian). Qolda-AVL is third in English (68.04%, behind Qwen3-Omni-30B-A3B-Instruct at 69.33%) and second in Russian (66.34%), while leading in gender multi-hop (84.60% for English, 80.80% for Russian). The Qwen3-Omni-30B-A3B family excels in categories that benefit from broad pretraining, especially animal sounds and language identification.

Qwen2.5-Omni baselines remain weak, except for a single strong result in Russian animal single-hop (96.20% by Qwen2.5-Omni-7B).

Table 10. Model performance on the SAKURA QA audio benchmark. The results are given as accuracy (%). Bold values indicate the best result for each benchmark–language pair.

Lang.	Model	Average	Gender		Animal		Language		Emotion	
			Single	Multi	Single	Multi	Single	Multi	Single	Multi
Kazakh	Qolda-AVL-5B (Ours)	64.02	81.20	78.80	60.00	52.40	86.14	79.36	36.47	37.80
	Qwen2.5-Omni-7B	46.38	52.20	52.60	46.00	33.80	92.60	43.40	24.80	25.60
	Qwen2.5-Omni-3B	32.80	50.40	48.80	25.80	28.20	33.20	25.00	27.20	23.80
	Qwen3-Omni-30B-Instruct	62.58	79.00	52.80	95.00	50.60	97.00	56.80	36.80	32.60
	Qwen3-Omni-30B-Thinking	72.44	87.00	66.80	93.40	71.54	96.40	90.60	40.80	33.00
English	Qolda-AVL-5B (Ours)	68.04	88.80	84.60	57.20	69.40	86.75	88.20	34.80	34.60
	Qwen2.5-Omni-7B	61.18	64.60	51.00	94.60	63.00	94.20	58.80	35.00	28.20
	Qwen2.5-Omni-3B	49.15	51.80	49.80	94.40	50.40	62.00	32.60	27.00	25.20
	Qwen3-Omni-30B-Instruct	69.33	88.20	54.60	94.40	65.40	97.80	74.80	44.60	34.80
	Qwen3-Omni-30B-Thinking	78.70	91.80	76.20	93.00	85.60	98.40	96.40	50.00	38.20
Russian	Qolda-AVL-5B (Ours)	66.34	82.80	80.80	60.32	59.40	90.98	84.80	35.40	36.20
	Qwen2.5-Omni-7B	58.45	57.60	46.80	96.20	59.60	92.60	51.80	34.60	28.40
	Qwen2.5-Omni-3B	45.10	47.20	50.80	87.00	44.80	47.80	30.80	28.20	24.20
	Qwen3-Omni-30B-Instruct	65.70	76.80	46.20	94.80	67.20	99.20	63.60	43.40	34.40
	Qwen3-Omni-30B-Thinking	76.85	88.00	65.80	94.00	85.80	99.00	93.60	48.00	40.60

Audio Captioning and Question Answering. Table 11 presents results for audio captioning and caption-based question answering benchmarks, which assess how effectively models describe audio content and answer related questions. For Kazakh, Qolda-AVL achieves the highest Audio Caption QA score (38.82%) and ranks first in average score (25.89%), ahead of Qwen3-Omni-30B-A3B-Thinking (25.20%). The Qwen2.5-Omni baselines perform poorly, with averages of below five, reflecting limited Kazakh audio capability. While the Qwen3-Omni-30B-A3B variants are competitive, they lag behind Qolda-AVL on the QA component, where understanding Kazakh questions is essential. For English and Russian, Qolda-AVL is competitive but does not lead. In English, Qwen3-Omni-30B-A3B-Thinking achieves the highest average (37.69%), followed by Qwen2.5-Omni-7B (37.15%), Qwen3-Omni-30B-A3B-Instruct (36.18%), and Qwen2.5-Omni-3B (30.24%). Qolda-AVL ranks last (26.47%), with the gap mainly in raw audio captioning, where general-purpose audio pretraining offers an advantage. In Russian, Qolda-AVL ranks third on average (27.22%), behind Qwen3-Omni-30B-A3B-Thinking (33.71%) and Qwen3-Omni-30B-A3B-Instruct (29.89%), while outperforming both Qwen2.5-Omni baselines. Its strength is concentrated in the QA component: the Russian Audio Caption QA score (40.79%) slightly exceeds even the much larger Qwen3-Omni-30B-A3B-Instruct (40.46%).

Spoken Mathematical Question Answering. Table 12 demonstrates results for the Spoken MQA benchmark, which evaluates numerical and reasoning tasks from spoken audio inputs across four difficulty levels: short digit recognition, long digit recognition, single-step reasoning, and multi-step reasoning.

For Kazakh, Qolda-AVL leads the benchmark, achieving the highest scores across all four subtasks and the average (87.24%). The margin over the next-best model, Qwen3-Omni-30B-A3B-Thinking (19.10% average), is more than four times that. The Qwen2.5-Omni baselines perform poorly (averages below 2%), and even the larger Qwen3-Omni-30B-A3B variants struggle, with Qwen3-Omni-30B-A3B-Instruct averaging only 14.72%. This significant gap demonstrates that language specialization is essential for

spoken numerical reasoning in low-resource languages: a 5B-parameter Kazakh-focused model outperforms a 30B-parameter general-purpose model by an order of magnitude.

Table 11. Evaluation results on audio captioning and audio caption QA benchmarks. Scores reflect accuracy (%) as determined by an LLM-as-a-Judge. Bold values indicate the best result for each benchmark–language pair.

Lang.	Model	Average	Audio Cap.	Audio Cap. QA
Kazakh	Qolda-AVL-5B (Ours)	25.89	12.95	38.82
	Qwen2.5-Omni-7B	4.30	1.68	6.91
	Qwen2.5-Omni-3B	3.55	4.80	2.30
	Qwen3-Omni-30B-Instruct	18.83	11.33	26.32
	Qwen3-Omni-30B-Thinking	25.20	16.85	33.55
English	Qolda-AVL-5B (Ours)	26.47	16.76	36.18
	Qwen2.5-Omni-7B	37.15	22.97	51.32
	Qwen2.5-Omni-3B	30.24	19.02	41.45
	Qwen3-Omni-30B-Instruct	36.18	23.99	48.36
	Qwen3-Omni-30B-Thinking	37.69	24.06	51.32
Russian	Qolda-AVL-5B (Ours)	27.22	13.64	40.79
	Qwen2.5-Omni-7B	26.29	21.33	31.25
	Qwen2.5-Omni-3B	17.78	12.20	23.36
	Qwen3-Omni-30B-Instruct	29.89	19.31	40.46
	Qwen3-Omni-30B-Thinking	33.71	23.01	44.41

Table 12. Evaluation results on the spoken math QA (MQA) benchmark. Scores are reported as accuracy (%). Bold values indicate the best result for each benchmark–language pair.

Lang.	Model	Average	Short Digit	Long Digit	Single-Step Reason.	Multi-Step Reason.
Kazakh	Qolda-AVL-5B (Ours)	87.24	90.00	88.37	93.58	77.01
	Qwen2.5-Omni-7B	1.31	2.00	0.00	2.36	0.86
	Qwen2.5-Omni-3B	1.24	1.00	0.00	3.21	0.74
	Qwen3-Omni-30B-Instruct	14.72	14.00	4.65	26.35	13.86
	Qwen3-Omni-30B-Thinking	19.10	12.12	2.40	36.99	24.89
English	Qolda-AVL-5B (Ours)	88.65	88.00	84.88	94.93	86.78
	Qwen2.5-Omni-7B	71.63	86.00	48.26	89.19	63.07
	Qwen2.5-Omni-3B	65.41	75.00	43.02	83.28	60.34
	Qwen3-Omni-30B-Instruct	83.44	93.00	84.30	96.11	60.34
	Qwen3-Omni-30B-Thinking	94.34	93.00	94.77	95.44	94.53

For English, Qolda-AVL is highly competitive with frontier-scale baselines, achieving the second-highest average (88.65%), behind Qwen3-Omni-30B-A3B-Thinking (94.34%) and ahead of Qwen3-Omni-30B-A3B-Instruct (83.44%). Qolda-AVL ranks second on the single-step reasoning subtask (94.93%), just behind Qwen3-Omni-30B-A3B-Instruct (96.11%). Across the remaining subtasks, Qolda-AVL trails Qwen3-Omni-30B-A3B-Thinking by small margins, while substantially outperforming both Qwen2.5-Omni baselines. This strong English performance shows that Kazakh-focused training does not compromise English-speaking proficiency in math. Our 5B model remains close to a 30B-parameter frontier model on this challenging task.

4.4. Ablation Study

Effect of Audio DeepStack. We conduct an ablation study to demonstrate that hierarchical audio injection via Audio DeepStack improves audio reasoning performance over a standard single-layer projection baseline. Figure 3 presents the training and validation losses of two Qolda-AVL checkpoints trained during the audio alignment stage under

identical settings. The Audio DeepStack-based model exhibits faster loss convergence and a relative validation loss reduction of over 5%.

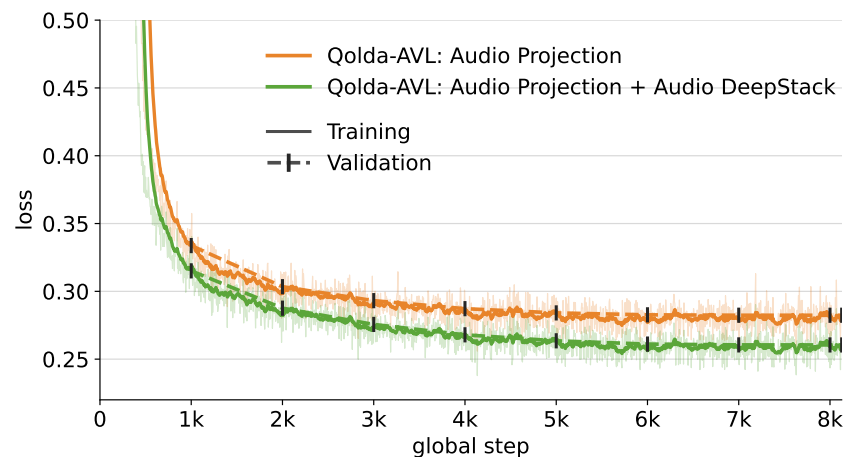


Figure 3. Effect of Audio DeepStack during audio alignment (Stage 2) training. Training and validation loss curves comparing two Qolda-AVL variants trained under identical Stage 2 conditions (see Section 3.2.6), differing only in the presence of Audio DeepStack.

While the loss curves provide a useful quantitative signal for comparing the two models, they are insufficient on their own for a formal assessment of downstream performance. We therefore further evaluate both checkpoints on a set of benchmark tasks following our evaluation protocol, and omitting the image and language benchmarks since the vision branch and the LLM remain frozen during Stage 2 and are not updated. The results show that hierarchical audio injection substantially improves performance on the animal sound classification task, with six wins out of six subtasks (two task modes $\times$ three languages), and on the emotion recognition QA task, again, with all wins. So, it reflects a consistent effect rather than variance from a single small evaluation. We can observe that these gains are most pronounced on tasks that rely on multi-level acoustic features, where both fine-grained spectral cues and higher-level semantic context matter. On S2TT, the Audio DeepStack-based model achieves consistent improvements on five of six language directions; however, a 5% drop on the en $\rightarrow$ kk direction alone offsets these gains, leaving the overall average roughly unchanged. In total, the Audio DeepStack-based model outperforms the baseline on 25 of the 41 evaluated tasks and ties on one. Table 13 (left) reports metric scores across task groups, averaged over languages and subtasks. The average score decreases where the baseline wins are negligible, with the exception of the gender recognition task, which exhibits a 2% drop in accuracy.

Crucially, the performance gain is obtained at negligible cost. The three Audio DeepStack projectors add only 29.5M parameter (0.58% of the 5B model). Because they inject features by residual addition onto existing audio-token positions rather than appending tokens, they leave the LLM's sequence length, attention, and KV-cache untouched. Measured on an H100, the entire Audio DeepStack path adds ~ 1.5 ms at prefill (a one-time cost over the audio tokens) and is statistically indistinguishable from zero over full CoT generation ($< 0.2\%$ end-to-end).

Effect of Whisper Fine-tuning. Our second ablation isolates the contribution of the preparatory Whisper fine-tuning (Stage 1.1) to downstream audio understanding. Since the ASR portion of the Stage 2 mixture partially overlaps with the data used in Stage 1.1, a natural question is whether Stage 1.1 can be skipped and the original Whisper encoder used directly when training the audio projection modules in Stage 2. To answer this, we train a second Qolda-AVL variant under identical Stage 2 conditions, but with the original

(non-fine-tuned) Whisper encoder in place of our Stage 1.1 checkpoint, and evaluate both variants on the audio subset of our benchmark suite.

The results show that the model variant initialized with the Stage 1.1 Whisper encoder checkpoint outperforms its counterpart on 26 of 41 evaluated tasks. As expected, the largest ASR gain is observed for Kazakh, with smaller but still positive improvements for English and Russian. Beyond ASR, the benefits of fine-tuning extend to higher-level audio understanding: language identification improves by 6.53 points on average, and reasoning accuracy on Kazakh increases by 3.81 points. The remaining task groups show small or negligible differences, with the only meaningful regression being a 2.87-point drop on emotion QA. Table 13 (right) summarizes the results across all tasks. Overall, these results indicate that adapting the Whisper encoder to the target languages prior to audio alignment benefits the subsequent audio adaptation process across a broader range of tasks than ASR alone, justifying the inclusion of Stage 1.1 in our training pipeline.

We note that this ablation isolates the value of using the Stage 1.1 checkpoint as the frozen Stage 2 encoder, rather than the more general question of whether a dedicated preparatory ASR phase outperforms an equivalent amount of ASR data folded into Stage 2. We leave the latter comparison to future work.

Table 13. Ablation studies on the audio subset of our benchmark suite. **Left:** Effect of the Audio DeepStack projection compared to a single-projection baseline. **Right:** Effect of Stage 1.1 fine-tuning of the Whisper encoder versus the original encoder. All checkpoints are Stage 2, trained under identical conditions; scores are averaged across languages and subtasks within each group. Arrows indicate the desired direction of each metric: ↓ means lower values are better, and ↑ means higher values are better. Deltas (Δ) shown in green denote a positive gain with respect to the metric direction; for ASR (WER ↓), a negative delta corresponds to an improvement.

Task	Audio DeepStack Ablation			Whisper Fine-Tuning Ablation		
	Baseline	Audio DS	Δ	Whisper-Orig.	Whisper-FT	Δ
ASR (WER ↓)	12.10%	11.98%	−0.12%	13.86%	11.98%	−1.88%
S2TT (xCOMET ↑)	0.7159	0.7152	−0.0008	0.7054	0.7152	+0.0098
QA Animal (Acc ↑)	45.90%	57.90%	+12.00%	55.82%	57.90%	+2.08%
QA Language (Acc ↑)	90.13%	89.63%	−0.50%	83.10%	89.63%	+6.53%
QA Emotion (Acc ↑)	35.43%	38.73%	+3.30%	41.60%	38.73%	−2.87%
QA Gender (Acc ↑)	42.00%	39.93%	−2.07%	39.87%	39.93%	+0.07%
Reasoning (Acc ↑)	79.12%	79.00%	−0.12%	75.19%	79.00%	+3.81%

5. Discussion

Evaluation results show that adding an audio modality and a second round of multi-modal fine-tuning did not reduce Qolda-AVL’s textual or visual performance. The model consistently outperforms both its predecessor, Qolda, and the Qwen3-VL-Thinking backbone on Kazakh, while remaining competitive in English and Russian. Some of the improvements over Qolda are due to the switch from InternVL3.5 to the stronger Qwen3-VL-4B-Thinking backbone. Therefore, Qwen3-VL-Thinking is the more relevant baseline for assessing our pipeline’s impact. Against this baseline, Qolda-AVL still demonstrates consistent gains in Kazakh that cannot be explained solely by the backbone change. These gains are most evident on knowledge-intensive and reading-comprehension benchmarks, indicating that additional Kazakh-focused text and joint reasoning data strengthened the model’s Kazakh linguistic abilities. The remaining small gap between English and Russian is mainly in graduate-level scientific reasoning, where a Kazakh-centric corpus cannot match the backbone’s broader pretraining. There is a modest trade-off in text-to-text translation: Qolda-AVL underperforms the text-only Qolda-think in both directions into Kazakh, but improves across all others. This suggests that moving from text-only to tri-

modal fine-tuning slightly reduces generative fluency in the most morphologically complex output language, while enhancing comprehension. The drop in the question-only split of KazCulture supports this, as this split emphasizes factual recall without context and is sensitive to how the model allocates its capacity across modalities.

Kazakh’s agglutinative morphology directly affects how multilingual models encode and generate text. The fertility of tokenizers, defined as the average number of tokens of sub-words per orthographic word [82], is a recognized measure of the efficiency with which a model represents a language [83]. Agglutinative languages like Kazakh typically have higher fertility because suffix stacking creates many inflected forms not present in the vocabulary. On the FLEURS validation set, the Qwen3-VL tokenizer produces 4.78 tokens per Kazakh word, compared to 1.26 for English and 2.46 for Russian (Table 14), resulting in a 3.79× ratio between Kazakh and English. Generating Kazakh text, therefore, requires nearly four times as many tokens as English for the same message, thereby extending the autoregressive horizon and increasing sensitivity to per-token errors. Surface-form metrics such as BLEU further penalize valid inflected variants, so we report translation quality using xCOMET-XXL throughout. Our adaptation addresses this bottleneck at the data level: we retain the Qwen3-VL tokenizer and oversample Kazakh text in Stage 1.2, enabling existing subwords to form well-structured Kazakh outputs across diverse inflected forms. This approach can be applied to other agglutinative low-resource languages, including other Turkic families.

Table 14. The fertility of the Qwen3-VL tokenizer was evaluated using the FLEURS [75] validation set in English, Russian, and Kazakh. *Tokens/word* represents the average number of subword tokens generated per orthographic word. Higher values indicate that more tokens are needed to encode the same content, increasing the input sequence length and extending the autoregressive decoding process.

Language	Words	Tokens	Tokens/Word	Ratio vs. English
English	22,503	28,303	1.26	1.00×
Russian	21,503	52,873	2.46	1.95×
Kazakh	20,693	98,961	4.78	3.79×

On the vision side, Qolda-AVL achieves clear improvements across all three languages, even though vision was not the main focus of the new training stages. The joint image–text–audio mixture in Stage 3 provides stronger cross-modal grounding than previous vision-only fine-tuning, with the largest improvements in diagrammatic reasoning, general VQA, and visual mathematical reasoning. Russian vision performance benefits are most likely due to the trilingual design of the Stage 1.2 and Stage 3 data. The remaining small gap in English reflects the lower proportion of English-only visual data and the allocation of some fine-tuning capacity to audio. Importantly, these vision improvements do not compromise Kazakh textual capabilities. The minor declines in English and Russian high-level reasoning (mainly GPQA) result from the Kazakh-focused specialization trade-off, not from vision-induced regression. This suggests the effectiveness of our training strategy. Preparing the audio branch in isolation before joint fine-tuning prevents the cross-modal conflicts, which are often observed when adding a new modality to a pretrained vision–language model.

The audio evaluation highlights the main contribution of this work. For Kazakh, Qolda-AVL achieves the strongest results in its parameter class across automatic speech recognition, speech-to-text translation, spoken mathematical question answering, and audio-based caption question answering. It also remains competitive with much larger models on spoken attribute reasoning. The performance gap with frontier-scale systems narrows

significantly for Kazakh, indicating that Kazakh audio remains an open challenge in multilingual speech and that a compact, specialized model can close this gap when the training data and adaptation process are well aligned with the target language. The Spoken MQA results for Kazakh highlight a significant performance gap. All evaluated systems, including the 30B-parameter Qwen3-Omni variants, achieve near-minimum accuracy, while Qolda-AVL matches its English performance. This gap is most likely due to the benchmark itself, not the audio reasoning capabilities of the competing models. The Kazakh benchmark was created using the Mangisoz TTS service, so the evaluation audio reflects the unique acoustic profile of this synthesizer rather than natural Kazakh speech. Qolda-AVL's training data includes Kazakh speech from diverse sources, which likely improves its robustness to domain shift, including the synthetic Mangisoz domain. Mangisoz-synthesized audio is not included in Stage 1.1 or Stage 2 training data. The Mangisoz TTS service is used only to create the Kazakh Spoken MQA evaluation benchmark. Therefore, the comparison in Table 12 is zero-shot for this synthesizer's acoustic domain. In contrast, general-purpose systems show low accuracy due to limited Kazakh coverage and minimal exposure to Mangisoz-generated speech. While this does not undermine the main conclusion, it suggests caution when interpreting the absolute size of the gap and highlights the need for future evaluation on naturally recorded Kazakh spoken math data. Overall, these results show that effective spoken reasoning in low-resource languages currently depends on language-specific training rather than model size. For English and Russian, the model is competitive but not dominant, as expected given its scale and the intentional focus on Kazakh in training.

Animal sound recognition results should be interpreted with caution, as this category is the least represented in our Stage 2 training mixture. As shown in Table 1, the audio classification task includes only three hours of training audio, which is significantly less than the ASR, S2TT, and language identification subsets that dominate the audio alignment stage. This data scarcity is further compounded by the speech-focused nature of the mixture and our choice of audio backbone. While frontier omni-modal models are pretrained on more non-speech audio, the Whisper encoder we use is supervised mainly with textual transcriptions. Although it encodes some non-speech information, it is not intended for general audio event tagging. Addressing this limitation will require expanding the non-speech audio component of the training mixture, which we see as a logical next step for the pipeline. Audio captioning and caption-based question answering yield lower scores across all evaluated systems, including the largest ones. This suggests that the main limitations are task formulation and data, rather than model capacity. Both tasks require describing multiple overlapping sound events in natural languages and are evaluated by an LLM-as-a-Judge using references that, for Kazakh and Russian, are machine-translated. Any noise from this translation step directly affects the scores, and the detailed semantic agreement required is more challenging than that required for closed-set answers in classification benchmarks. In our case, this challenge is exacerbated by the limited captioning data in Stage 2, which is mainly in English, sourced through MECAT, and by the lack of a dedicated non-speech reasoning stage. Qolda-AVL learns the general structure of the captioning task but does not match systems trained on larger, specialized audio captioning datasets. Our leading performance on Kazakh benchmarks reflects the lack of Kazakh audio understanding in the open ecosystem rather than a definitive claim about the quality of our captioning. This points to the need for a more diverse, multilingual captioning stage as a next step.

The two failure modes described above stem from the same bottleneck: non-speech audio in Stage 2 accounts for only 0.04% of the total row-hour mixture (3 h out of 7365 h; see Table 1). We propose three complementary mitigation strategies for future itera-

tions. First, incorporate large-scale general-purpose audio corpora, such as AudioSet [84], FSD50K [85], and AudioCaps [86], to expand the ESC-50 and MECAT components significantly in both clip count and class coverage. Second, upsample underrepresented Stage 2 task categories (classification, captioning, audio QA) to ensure ASR does not dominate the gradient signal, regardless of raw row-hour counts. Third, enhance the Whisper backbone with a non-speech-pretrained encoder, such as BEATs [24] or CLAP [25], routed through the DeepStack mechanism to provide a dedicated injection path for fine-grained acoustic-event features into the LLM.

6. Conclusions

In this work, we extended a vision–language model with audio understanding to build the first open-source tri-modal (audio, vision, language) model for Kazakh. Qolda-AVL integrates a Whisper-based audio branch with a Qwen3-VL backbone using Audio DeepStack, a hierarchical multi-level injection mechanism adapted from DeepStack and Qwen3-VL. The model is trained through a four-stage progressive pipeline that incrementally unifies independently adapted modalities into a single tri-modal system.

Qolda-AVL is a specialized Kazakh audio adapter built on the Qwen3-VL-Thinking backbone, focusing on Kazakh-centric tasks. While English and Russian capabilities are maintained or slightly improved, the model establishes a strong baseline for Kazakh audio understanding across language, vision, and audio benchmarks. It achieves the highest performance in its parameter class on all evaluated Kazakh audio tasks, significantly reducing the gap with larger frontier-scale systems. For speech-to-text translation and spoken mathematical reasoning, it outperforms these larger models. We also introduce the first multi-task Kazakh audio understanding benchmark suite, which includes spoken attribute reasoning, spoken mathematical question answering, and audio captioning with question answering. This suite is intended as a valuable resource for future Kazakh multi-modal research. Our ablation studies show that Audio DeepStack improves convergence and downstream performance on tasks that require multi-level acoustic features. Additionally, the preparatory Whisper fine-tuning stage benefits both Kazakh ASR and higher-level audio understanding tasks, such as language identification and spoken reasoning, supporting its inclusion in the pipeline.

Several directions remain for future work. The audio captioning and animal sound recognition results are mainly limited by the scale and diversity of non-speech audio data used during alignment. Adding a dedicated multilingual non-speech captioning stage is a logical next step. The Spoken MQA evaluation highlights the need for naturally recorded Kazakh spoken math benchmarks to supplement the synthesized data used here. To enhance the ecological validity of the audio evaluation suite, it is needed to reduce synthetic audio and machine translation noise. In future work, we will collect natural Kazakh recordings for this limitation. We are releasing the model and benchmark suite to support research in these areas and to advance inclusive, resource-efficient multi-modal language technology for underserved linguistic communities.

Author Contributions: Conceptualization, B.A., A.M., A.N. and H.A.V.; methodology, B.A., A.M. and A.N.; software, B.A., A.M. and A.N.; validation, B.A., A.M. and A.N.; investigation, B.A., A.M., A.N. and H.A.V.; resources, H.A.V.; data curation, B.A., A.M. and A.N.; writing—original draft preparation, B.A. and A.M.; writing—review and editing, B.A., A.M. and H.A.V.; visualization, B.A.; supervision, H.A.V.; project administration, H.A.V.; funding acquisition, H.A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study are publicly available open-source resources, distributed under their respective licenses by the original authors and cited throughout the paper. The Kazakh audio benchmarks introduced in this work, covering spoken attribute reasoning, spoken mathematical question answering, and audio captioning with question answering, are publicly released through the ISSAI Hugging Face collection at <https://huggingface.co/collections/issai/qolda-avl-audio-benchmarks> (accessed on 29 April 2026). The curated training mixture and the synthetic chain-of-thought data generated in Stage 3 are not publicly released at this stage.

Acknowledgments: The authors would like to thank Madina Mansurova for supervising the grant that facilitated this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 34892–34916.
2. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 17–21 June 2024; pp. 24185–24198.
3. Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J.S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 11–15 June 2025; pp. 91–104.
4. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv* **2023**, arXiv:2308.12966.
5. OpenAI. GPT-4o System Card. *arXiv* **2024**, arXiv:2410.21276. [[CrossRef](#)]
6. Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. Qwen2.5-Omni Technical Report. *arXiv* **2025**, arXiv:2503.20215. [[CrossRef](#)]
7. Xu, J.; Guo, Z.; Hu, H.; Chu, Y.; Wang, X.; He, J.; Wang, Y.; Shi, X.; He, T.; Zhu, X.; et al. Qwen3-Omni Technical Report. *arXiv* **2025**, arXiv:2509.17765. [[CrossRef](#)]
8. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; Mcleavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning, 23–29 July 2023*; *Proceedings of Machine Learning Research*; PMLR: New York, NY, USA, 2023; Volume 202, pp. 28492–28518. Available online: <https://proceedings.mlr.press/v202/radford23a.html> (accessed on 30 April 2026).
9. Yang, Z.; Shimizu, S.; Yu, Y.; Chu, C. When Large Language Models Meet Speech: A Survey on Integration Approaches. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, 27 July–1 August 2025; pp. 20298–20315. [[CrossRef](#)]
10. Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; Zhou, J. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv* **2023**, arXiv:2311.07919.
11. Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; Zhang, C. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria, 7–11 May 2024.
12. Veitsman, Y.; Hartmann, M. Recent Advancements and Challenges of Turkic Central Asian Language Processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, Abu Dhabi, United Arab Emirates, 20 January 2025; pp. 309–324.
13. Arystanbekov, B.; Nurimanov, A.; Maxutov, A.; Albrekht, V.; Kuzdeuov, A.; Varol, H.A. Qolda: A Small Vision-Language Model for the Kazakh Language. *IEEE Access* **2026**, *14*, 46392–46414. [[CrossRef](#)]
14. Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. Qwen3-VL Technical Report. *arXiv* **2025**, arXiv:2511.21631. [[CrossRef](#)]
15. Meng, L.; Yang, J.; Tian, R.; Dai, X.; Wu, Z.; Gao, J.; Jiang, Y.G. DeepStack: Deeply stacking visual tokens is surprisingly simple and effective for LMMs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 10–15 December 2024; NIPS '24.

16. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, 18–24 July 2021*; Proceedings of Machine Learning Research; PMLR: New York, NY, USA, 2021; Volume 139, pp. 8748–8763.
17. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. CogVLM: Visual expert for pretrained language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems NIPS '24*, Vancouver, BC, Canada, 10–15 December 2024.
18. Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv* **2024**, arXiv:2403.05525.
19. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv* **2024**, arXiv:2409.12191.
20. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-VL Technical Report. *arXiv* **2025**, arXiv:2502.13923. [[CrossRef](#)]
21. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems NIPS '20*, Red Hook, NY, USA, 6–12 December 2020.
22. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhota, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2021**, *29*, 3451–3460. [[CrossRef](#)]
23. Yang, S.W.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhota, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. SUPERB: Speech Processing Universal PERFORMANCE Benchmark. In *Proceedings of the Interspeech 2021*, Brno, Czech Republic, 30 August–3 September 2021; pp. 1194–1198. [[CrossRef](#)]
24. Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; Che, W.; Yu, X.; Wei, F. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning ICML'23*, Honolulu, HI, USA, 23–29 July 2023.
25. Elizalde, B.; Deshmukh, S.; Ismail, M.A.; Wang, H. CLAP Learning Audio Concepts from Natural Language Supervision. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [[CrossRef](#)]
26. Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; Dubnov, S. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Rhodes Island, Greece, 4–10 June 2023.
27. Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. Qwen2-Audio Technical Report. *arXiv* **2024**, arXiv:2407.10759 [[CrossRef](#)]
28. Gong, Y.; Khurana, S.; Karlinsky, L.; Glass, J. Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers. In *Proceedings of the Interspeech 2023*, Dublin, Ireland, 20–24 August 2023; pp. 2798–2802. [[CrossRef](#)]
29. Huang, R.; Li, M.; Yang, D.; Shi, J.; Chang, X.; Ye, Z.; Wu, Y.; Hong, Z.; Huang, J.; Liu, J.; et al. AudioGPT: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence; AAAI'24/IAAI'24/EAAI'24*; AAAI Press: Washington, DC, USA, 2024. [[CrossRef](#)]
30. Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C.K.R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; Manocha, D. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, FL, USA, 12–16 November 2024; pp. 6288–6313. Available online: <https://aclanthology.org/2024.emnlp-main.361> (accessed on 29 April 2026).
31. OpenMOSS Team. MOSS-Audio Technical Report. GitHub Repository. 2026. Available online: <https://github.com/OpenMOSS/MOSS-Audio> (accessed on 20 April 2026).
32. Google DeepMind. Gemini 3 Pro Model Card. 2025. Available online: <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf> (accessed on 6 April 2026).
33. Fu, C.; Lin, H.; Long, Z.; Shen, Y.; Dai, Y.; Zhao, M.; Zhang, Y.F.; Dong, S.; Li, Y.; Wang, X.; et al. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv* **2025**, arXiv:2408.05211. [[CrossRef](#)]
34. Li, Y.; Sun, H.; Lin, M.; Li, T.; Dong, G.; Zhang, T.; Ding, B.; Song, W.; Cheng, Z.; Huo, Y.; et al. Baichuan-Omni Technical Report. *arXiv* **2024**, arXiv:2410.08565. [[CrossRef](#)]
35. Luo, R.; Lin, T.E.; Zhang, H.; Wu, Y.; Liu, X.; Li, Y.; Chen, L.; Li, J.; Zhang, L.; Xia, X.; et al. OpenOmni: Advancing Open-Source Omnimodal Large Language Models with Progressive Multimodal Alignment and Real-time Emotional Speech Synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*; Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2025; Volume 38, pp. 158925–158953.
36. Liu, Z.; Dong, Y.; Wang, J.; Liu, Z.; Hu, W.; Lu, J.; Rao, Y. Ola: Pushing the Frontiers of Omni-Modal Language Model. *arXiv* **2025**, arXiv:2502.04328. [[CrossRef](#)]

37. Institute of Smart Systems and Artificial Intelligence. Kazakh Large Language Model (ISSAI KAZ-LLM). 2024. Available online: <https://huggingface.co/collections/issai/issai-kazllm-10-6732d58c81bc8177442c362> (accessed on 29 April 2026).
38. Koto, F.; Joshi, R.; Mukhituly, N.; Wang, Y.; Xie, Z.; Pal, R.; Orel, D.; Mullah, P.; Turmakhan, D.; Goloburda, M.; et al. Sherkala-Chat: Building a State-of-the-Art LLM for Kazakh in a Moderately Resourced Setting. In Proceedings of the Second Conference on Language Modeling, Montreal, QC, Canada, 7–10 October 2025.
39. Maxutov, A.; Myrzakhmet, A.; Braslavski, P. Do LLMs Speak Kazakh? A Pilot Evaluation of Seven Models. In Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), Bangkok, Thailand, 15 August 2024; pp. 81–91. Available online: <https://aclanthology.org/2024.sigturk-1.8/> (accessed on 30 April 2026).
40. Mussakhoyeva, S.; Khassanov, Y.; Atakan Varol, H. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 1367–1371. [\[CrossRef\]](#)
41. Li, J.; Pu, Y.; Sun, Q.; Zhang, W.Q. Improving Whisper’s Recognition Performance for Under-Represented Language Kazakh Leveraging Unpaired Speech and Text. In Proceedings of the Interspeech 2024, Kos, Greece, 1–5 September 2024; pp. 2514–2518. [\[CrossRef\]](#)
42. Mussakhoyeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. *Information* **2023**, *14*, 74. [\[CrossRef\]](#)
43. Mussakhoyeva, S.; Gilmullin, R.; Khakimov, B.; Galimov, M.; Orel, D.; Abilbekov, A.; Varol, H.A. Noise-Robust Multilingual Speech Recognition and the Tatar Speech Corpus. In Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), Osaka, Japan, 19–22 February 2024; pp. 732–737. [\[CrossRef\]](#)
44. Zubitskii, P.; Morozov, V.; Murzakhmetov, S.; Sagyndyk, B.; Umbet, S. HordeVision: An Open-Source Kazakh Vision–Language Model. *TechRxiv* **2026**. [\[CrossRef\]](#)
45. Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 4218–4222. Available online: <https://aclanthology.org/2020.lrec-1.520/> (accessed on 29 April 2026).
46. Valk, J.; Alumäe, T. VOXLINGUA107: A Dataset for Spoken Language Recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*; IEEE: New York, NY, USA, 2020; pp. 652–658. Available online: <https://api.semanticscholar.org/CorpusID:227209238> (accessed on 25 April 2026).
47. Niu, Y.; Wang, T.; Dinkel, H.; Sun, X.; Zhou, J.; Li, G.; Liu, J.; Liu, X.; Zhang, J.; Luan, J. MECAT: A Multi-Experts Constructed Benchmark for Fine-Grained Audio Understanding Tasks. *arXiv* **2025**, arXiv:2507.23511.
48. Abilbekov, A.; Mussakhoyeva, S.; Yeshpanov, R.; Varol, H.A. KazEmoTTS: A Dataset for Kazakh Emotional Text-to-Speech Synthesis. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 9626–9632. Available online: <https://aclanthology.org/2024.lrec-main.841/> (accessed on 26 April 2026).
49. Karpov, N.; Denisenko, A.; Minkin, F. Golos: Russian Dataset for Speech Research. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 1419–1423. [\[CrossRef\]](#)
50. Yamagishi, J.; Veaux, C.; MacDonald, K. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*; University of Edinburgh, The Centre for Speech Technology Research (CSTR): Edinburgh, UK, 2019. [\[CrossRef\]](#)
51. Bakhturina, E.; Lavrukhin, V.; Ginsburg, B.; Zhang, Y. Hi-Fi Multi-Speaker English TTS Dataset. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 2776–2780. [\[CrossRef\]](#)
52. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *J. Lang. Resour. Eval.* **2008**, *42*, 335–359. [\[CrossRef\]](#)
53. Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM International Conference on Multimedia MM ’15, Brisbane, Australia, 26–30 October 2015; pp. 1015–1018. [\[CrossRef\]](#)
54. Tian, F.; Zhang, X.T.; Zhang, Y.; Zhang, H.; Li, Y.; Liu, D.; Deng, Y.; Wu, D.; Chen, J.; Zhao, L.; et al. Step-Audio-R1 Technical Report. *arXiv* **2025**, arXiv:2511.15848.
55. Qwen Team. Qwen3.5: Towards Native Multimodal Agents. 2026. Available online: <https://qwen.ai/blog?id=qwen3.5> (accessed on 30 April 2026).
56. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 26286–26296. [\[CrossRef\]](#)
57. Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; et al. SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning. *arXiv* **2024**, arXiv:2408.05517. [\[CrossRef\]](#)
58. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Vienna, Austria, 4 May 2021.

59. Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 95266–95290. [\[CrossRef\]](#)
60. Rein, D.; Hou, B.L.; Stickland, A.C.; Petty, J.; Pang, R.Y.; Dirani, J.; Michael, J.; Bowman, S.R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *Proceedings of the First Conference on Language Modeling*, Philadelphia, PA, USA, 7–9 October 2024.
61. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457. [\[CrossRef\]](#)
62. Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. Training Verifiers to Solve Math Word Problems. *arXiv* **2021**, arXiv:2110.14168. [\[CrossRef\]](#)
63. Bandarkar, L.; Liang, D.; Muller, B.; Artetxe, M.; Shukla, S.N.; Husa, D.; Goyal, N.; Krishnan, A.; Zettlemoyer, L.; Khabsa, M. The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, 11–16 August 2024; pp. 749–775. [\[CrossRef\]](#)
64. Team, N.; Costa-jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv* **2022**, arXiv:2207.04672. [\[CrossRef\]](#)
65. Togmanov, M.; Mukhituly, N.; Turmakhan, D.; Mansurov, J.; Goloburda, M.; Sakip, A.; Xie, Z.; Wang, Y.; Syzdykov, B.; Laiyk, N.; et al. KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 27 July–1 August 2025; pp. 14403–14416. [\[CrossRef\]](#)
66. Maxutov, A.; Arystanbekov, B.; Makhataeva, Z.; Yergen, A.; Taizhanov, N.; Nauryzbaikyzy, G.; Varol, H.A. Introducing Cultural Knowledge in Language Models: KazCulture Dataset for Kazakh Culture. *IEEE Access* **2026**, *14*, 44027–44042. [\[CrossRef\]](#)
67. Yeshpanov, R.; Efimov, P.; Boytsov, L.; Shalkarbayuli, A.; Braslavski, P. KazQAD: Kazakh Open-Domain Question Answering Dataset. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, 20–25 May 2024; pp. 9645–9656. Available online: <https://aclanthology.org/2024.lrec-main.843> (accessed on 28 April 2026).
68. Guerreiro, N.M.; Rei, R.; Stigt, D.v.; Coheur, L.; Colombo, P.; Martins, A.F.T. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 979–995. [\[CrossRef\]](#)
69. Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; Farhadi, A. A Diagram is Worth a Dozen Images. In *Proceedings of the Computer Vision—ECCV 2016*, Amsterdam, The Netherlands, 11–14 October 2016; pp. 235–251.
70. Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 27056–27087. [\[CrossRef\]](#)
71. xAI. Grok-1.5 Vision Preview. 2024. Available online: <https://x.ai/news/grok-1.5v> (accessed on 20 May 2024).
72. Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.W.; Galley, M.; Gao, J. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria, 7–11 May 2024.
73. Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.C.; Liu, C.L.; Jin, L.; Bai, X. OCRBench: On the hidden mystery of OCR in large multi-modal models. *Sci. China Inf. Sci.* **2024**, *67*, 220102. [\[CrossRef\]](#)
74. MTS AI Research. MWS-Vision-Bench: Russian Multimodal OCR Benchmark. 2025. Available online: <https://huggingface.co/datasets/MTSAIR/MWS-Vision-Bench> (accessed on 29 April 2026).
75. Conneau, A.; Ma, M.; Khanuja, S.; Zhang, Y.; Axelrod, V.; Dalmia, S.; Riesa, J.; Rivera, C.; Bapna, A. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. *arXiv* **2022**, arXiv:2205.12446. [\[CrossRef\]](#)
76. Yang, C.K.; Ho, N.; Piao, Y.T.; Lee, H.-Y. SAKURA: On the Multi-hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information. In *Proceedings of the Interspeech 2025*, Rotterdam, The Netherlands, 17–21 August 2025; pp. 1788–1792. [\[CrossRef\]](#)
77. Wei, C.; Wang, B.; Kim, J.J.; Chen, N.F. Towards Spoken Mathematical Reasoning: Benchmarking Speech-based Models over Multi-faceted Math Problems. *arXiv* **2025**, arXiv:2505.15000.
78. Institute of Smart Systems and Artificial Intelligence. Mangisoz. 2026. Available online: <https://mangisoz.nu.edu.kz/soyle> (accessed on 22 March 2026).
79. Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M.D.; Zou, Y.; Wang, W. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 3339–3354. [\[CrossRef\]](#)

80. Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; Chen, N.F. AudioBench: A Universal Benchmark for Audio Large Language Models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 4297–4316. [\[CrossRef\]](#)
81. InflexionLab. Sybyrla. 2026. Available online: <https://huggingface.co/InflexionLab/sybyrla> (accessed on 28 May 2026).
82. Rust, P.; Pfeiffer, J.; Vulić, I.; Ruder, S.; Gurevych, I. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 3118–3135. [\[CrossRef\]](#)
83. Petrov, A.; Malfa, E.L.; Torr, P.; Bibi, A. Language Model Tokenizers Introduce Unfairness Between Languages. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
84. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780. [\[CrossRef\]](#)
85. Fonseca, E.; Favory, X.; Pons, J.; Font, F.; Serra, X. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2021**, *30*, 829–852. [\[CrossRef\]](#)
86. Kim, C.D.; Kim, B.; Lee, H.; Kim, G. AudioCaps: Generating Captions for Audios in The Wild. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 119–132. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.