UNIVERSE: BENCHMARKING AND ENHANCING LALMS ON CULTURALLY INCLUSIVE LOW-RESOURCE MUSIC UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large audio-language models (LALMs) have unlocked impressive capabilities in music understanding and reasoning. However, current LALMs exhibit severe performance degradation when applied to diverse global musical traditions, particularly low-resource folk music rooted in rich cultural contexts. Due to significant data imbalance and the absence of dedicated evaluation protocols, existing models often fail to capture the nuanced structural, acoustic, and stylistic characteristics of local genres. To bridge this gap, we present UNIVERSE, a comprehensive, reproducible framework for culturally-aware low-resource music understanding. UNIVERSE consists of two core components: (1) UNIVERSEBENCH, an expert-guided benchmark featuring 5,042 meticulously curated Q&A pairs across 38 distinct cultural entities to evaluate fine-grained musical comprehension; and (2) UNIVERSESET, a fully automated multi-turn dialogue dataset synthesized to facilitate multimodal alignment. Leveraging UNIVERSE, we present the first systematic study investigating multimodal imbalance learning strategies across both Dense and Mixture-of-Experts (MoE) LALM architectures. Extensive experiments demonstrate that our automated curation paired with latent representation alignment yields significant gains on dense backbones (+14.9 pp), while standard SFT remains a resilient baseline on sparse MoE architectures. Nevertheless, diagnostic evaluations reveal a remaining gap between surface-level textual alignment and deep acoustic comprehension, highlighting key challenges of building general LALMs for culturally inclusive music. Our anonymous demo page is available here.

1 INTRODUCTION

Recent advancements in multimodal tokenization and Large Language Models (LLMs) have significantly boosted the performance of Large Audio-Language Models (LALMs) in music reasoning and captioning. However, existing LALMs exhibit a pervasive systematic bias toward high-resource popular genres while consistently failing on regional modal/tuning systems, microtonal scales, and idiomatic rhythms inherent to global folk traditions. Folk music is usually low-resource—characterized by heterogeneous recording formats, sparse documentation, and scarce expert annotations. Even when embedded within massive pre-training corpora, these recordings lack structured supervision, preventing LALMs from capturing their fine-grained acoustic and cultural nuances. Although recent benchmarks such as CMI-bench (Ma et al., 2025a) have begun to quantify this cultural gap, they remain limited in scale, language coverage, and automated evaluation frameworks.

To empirically probe these vulnerabilities, we conducted qualitative diagnoses on representative folk music tracks using state-of-the-art LALMs (e.g., Gemini 3 Pro), revealing two critical failure modes:

(1) **Misidentification and Hallucination of Acoustic Features:** Models exhibit severe fragility when recognizing low-level acoustic patterns. For instance, given an audio recording of an Indian classical raga, the model misclassifies *Raag Abhogi* as *Todi*. Similarly, for *Ze wei lao mo* (a traditional Yi folk track from Yunnan, China), the model mistakenly forces a rigid 4/4 meter, failing to perceive its additive and mixed rhythmic structures. Such failures may stem from three intrinsic challenges: unconventional non-standard tuning systems, noisy field-recording environments, and complex stylistic ornaments (e.g., microtonal pitch bending and region-specific rhythmic/metric idiosyncrasies).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

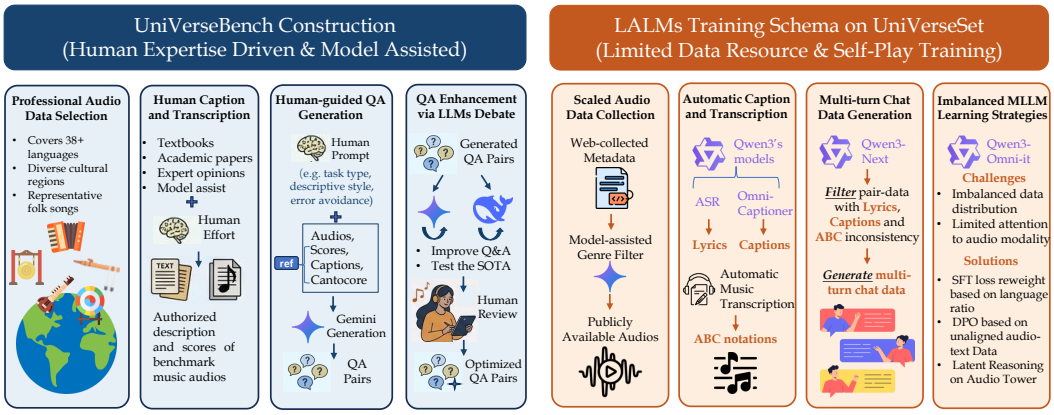


Figure 1: The construction pipeline of UNIVERSEBENCH and the post-training schema for LALMs on UNIVERSESET.

(2) **Textual Bias at the Expense of Acoustic Evidence:** Models demonstrate an over-reliance on metadata, leading to severe text-driven hallucination and cross-cultural interference. For example, explicitly adding the text prompt *"Chinese music"* misleads the model into categorizing an Indian *Raga* audio as the Chinese *Gong Pentatonic* scale. Furthermore, descriptive metadata can distract the model from objective audio analysis—e.g., revealing the geographic metadata *"Honghe Hani and Yi Autonomous Prefecture"* causes a Yi music track to be misclassified as Hani music.

While such fragility is known in general LALMs (Kuan & Lee, 2024; Sun et al., 2025; Cheng et al., 2026), it becomes exceptionally acute in low-resource music traditions where data scarcity and text-dominant bias amplify model failure. To address these limitations, we present UNIVERSE, a holistic framework encompassing three major contributions:

- **UNIVERSEBENCH for Heterogeneous Contexts:** We construct UNIVERSEBENCH, an evaluation benchmark comprising 5,042 Q&A pairs across 372 audio recordings spanning over 38 languages and cultural entities. To balance musicological rigor with evaluation scale, the benchmark is anchored by a core set of 1,873 expert-audited items, complemented by automated consensus-filtered extensions. It incorporates multimodal representations including symbolic sheet music scores, computational ethnomusicological features via CantoCore (Savage et al., 2012), and aligned metadata.

- **Human-AI Collaborative Curation Pipeline:** We establish a reproducible, scalable human-AI collaborative pipeline that balances expert rigor with automated scalability. We systematically delineate the boundary of LLM-driven knowledge integration and introduce a human-in-the-loop verification protocol to ensure annotation reliability and authenticity.

- **Systematic Investigation of Post-Training Schemas:** Using UNIVERSESET—a large-scale, automatically synthesized training dataset—we conduct the first systematic exploration of multimodal imbalance learning for low-resource music across Dense and Mixture-of-Experts (MoE) LALM architectures. We evaluate language-aware loss reweighting, Direct Preference Optimization (DPO) across text/audio towers, and latent reasoning strategies, offering actionable insights for culturally-aware audio intelligence.

2 RELATED WORK

2.1 MUSIC UNDERSTANDING ACROSS CULTURALLY DIVERSE TRADITIONS

Datasets and Benchmarks. Early computational musicology initiatives, such as CompMusic (Serra, 2014), laid the foundation for multicultural music analysis, paving the way for regional datasets covering diverse folk traditions (Zhang et al., 2023; Zhou et al., 2025; Krishnan et al., 2024; Papaioannou et al., 2022; Han et al., 2024). These primary resources have subsequently enabled instruction-following benchmarks, including Audio-Flan (Xue et al., 2025), MusiQAl (Christodoulou et al., 2025), GlobalMood (Lee et al., 2025), TAU (Lin et al., 2025), and ArtistMus (Kwon et al.,

2025). While these benchmarks evaluate scattered musical attributes like emotion, artist, and timbre, they suffer from two major limitations: (1) *Taxonomic Fragility*: they predominantly rely on rigid, culturally bounded taxonomies that fail to scale to broader regional genres; and (2) *Ad-hoc Curation*: their non-reproducible construction pipelines prevent adaptation to other low-resource musical domains. Consequently, existing benchmarks offer limited utility for evaluating how LALMs generalize to unseen or underrepresented global folk traditions.

Training and Evaluation. To endow models with cross-cultural auditory perception, recent studies have explored continual pre-training on multicultural corpora (Kanatas et al., 2025), contrastive or generative representation alignment (Wu et al., 2025), and domain-specific supervised fine-tuning (SFT) (Papaioannou et al., 2025). However, existing evaluation protocols remain confined to high-level semantic tags (e.g., broad genre, mood), leaving nuanced structural features—such as regional modal systems, microtonal pitch organization, and additive rhythms—largely unexamined. As a result, current LALMs frequently reduce cross-cultural reasoning to superficial text-label matching, failing to cultivate genuine auditory comprehension of diverse musical heritage.

2.2 Post-Training and Alignment in MLLMs

Multimodal Perception and Hallucination Mitigation. Recent vision-language frameworks increasingly decouple perception from reasoning to mitigate cultural bias and factual hallucination (Qiao et al., 2024; Romero et al., 2024; Liu et al., 2025). In the audio domain, Large Audio-Language Models (LALMs) exhibit severe acoustic hallucinations regarding object existence, temporal order, and fine-grained acoustic properties (Kuan & Lee, 2024; Sun et al., 2025; Cheng et al., 2026). This fragility is further exacerbated in low-resource music traditions due to text-dominance bias, where rich textual prompts override actual acoustic evidence (Zang et al., 2025). To suppress text-driven hallucination and enforce genuine acoustic grounding, prior works have investigated preference optimization and negative sampling strategies (Chaubey et al., 2026; Wang et al., 2024a). Extending these insights, UniVerse minimizes textual prompt complexity to directly stress-test the model's intrinsic auditory perception.

Imbalanced Learning in Low-Resource Scenarios. Considering diverse cultural traditions and different accessibility of digital sources, music understanding easily suffers from long-tailed distribution and cross-modal imbalance. To bridge this gap during post-training, we synthesize insights across three emerging paradigms: (1) *Language-aware Loss Reweighting*, which dynamically rebalances long-tailed supervision without requiring explicit data resampling (Jung & Plas, 2024); (2) *Audio-Grounded Direct Preference Optimization (DPO)*, which reinforces fine-grained acoustic alignment over text priors by contrasting matched versus mismatched audio-text pairs (Rafailov et al., 2023; Chaubey et al., 2026); and (3) *Latent Representation Alignment (REPA)*, which aligns continuous latent states rather than discrete text tokens, preserving delicate acoustic structures across both Dense and Mixture-of-Experts (MoE) backbones (Dai et al., 2026; Xu et al., 2025b).

3 Methodology

Figure 1 illustrates the unified framework of UniVerse, encompassing the curation of the UniVerseBench evaluation suite and the scalable training schema applied to UniVerseSet. To minimize potential evaluation bias, we enforce a rigorous zero-overlap protocol between the data sources and generation pipelines of the two processes (§3.1 and §3.2).

3.1 UniVerseBench Construction: From Local Traditions to Global Scales

3.1.1 Multi-Source Content Curation and Annotation

Figure 2 details the composition of UniVerseBench. The global subset extends the Voices of Civilizations (VoC) benchmark (Wu et al., 2026). To balance musicological rigor, public relevance, and cultural authenticity, we formulate a two-pronged curation strategy:

- **Expert-Led Selection**: A panel of ethnomusicologists curated the core corpus by synthesizing authoritative folk music anthologies, regional government archives, high-fidelity field recordings, and streaming platforms. This preserves both historical artifacts and "living heritage".
- **Multimodal Annotation**: Experts provided foundational textual analyses based on scholarly literature, which were subsequently expanded by Gemini 3 Pro for linguistic diversity. Professional transcribers manually converted performance audio into symbolic ABC notation. We acknowledge that staff notation is ill-suited for expressing the nuanced characteristics of diverse musical tradi-

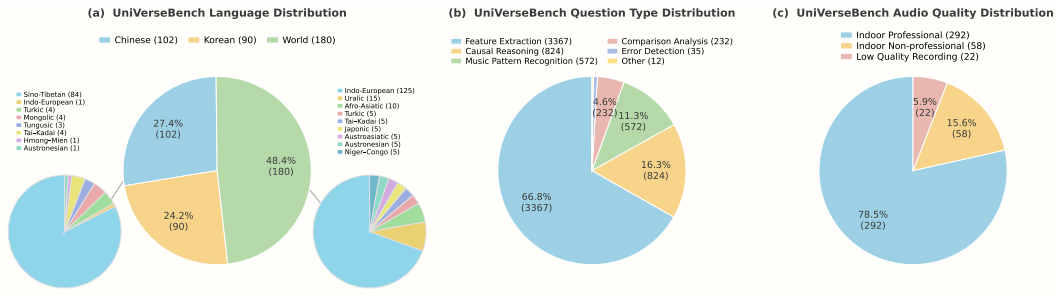


Figure 2: Composition of UNIVERSEBENCH. **(a) Language distribution**: Chinese and World slices partitioned by macro-language families based on metadata ISO codes. **(b) Question-type distribution**: spanning feature extraction, causal reasoning, pattern recognition, comparative analysis, error detection, and music completion. **(c) Audio Recording Quality**: expert audit distribution categorizing tracks into indoor professional, indoor non-professional, and low-fidelity field recordings.

tions. However, intentionally benchmarking its ceiling serves to quantify what standard, generic frameworks can achieve—thereby highlighting the essential need for tailored, culture-specific knowledge. To explicitly extract melodic contours in culturally diverse musical traditions, we integrated CantoCore (Savage et al., 2012), a computational ethnomusicology framework, whose strengths and limitations will be discussed in Appendix C.1. Cross-language acoustic cluster visualizations are provided in Figure 6.

3.1.2 INSTRUCTION DESIGN AND QA FORMULATION

To systematically probe the reasoning boundaries of LALMs, we formulate a structured QA paradigm. Given a question Q, options O, audio clip X, and background knowledge K, a benchmark tuple is defined as:

$$Q_{\text{pair}} = (Q, O, X, K, A_{\text{ns}}, \tau, \ell, \delta) \tag{1}$$

where A_{ns} denotes the ground-truth answer, τ specifies the target task taxonomy (e.g., rhythmic analysis, modal identification), ℓ denotes the regional language/tradition identifier, and $\delta \in \{0, 1\}$ is an explicit text-dependency indicator. When $\delta = 0$, the item is strictly unanswerable without processing the auditory signal X, effectively isolating pure acoustic perception from textual shortcutting.

3.1.3 QUALITY OPTIMIZATION AND CONSENSUS VERIFICATION

Option Refinement via Perceptual Index To prevent language models from exploiting textual priors, we calculate a Perceptual Index (PI) using a text-only LLM (Zang et al., 2025). Items exhibiting low PI (i.e., answerable purely via textual reasoning) are flagged, and their distractor options O are iteratively regenerated to be acoustically plausible yet textually indistinguishable.

Hybrid Human-AI Verification Protocol We implement a randomized human-AI consensus protocol. Expert evaluators review candidate QA tuples alongside ABC scores, textual captions, and CantoCore feature maps. An item is accepted if at least two initial experts reach consensus. In cases of disagreement, an independent senior musicologist reviews the item assisted by Gemini 3 Pro. The item is retained if the senior expert's judgment aligns with either Gemini's final inference or the initial expert consensus; otherwise, it is permanently discarded. Appendix A outlines the prompt templates and automated verification workflows.

3.2 POST-TRAINING FRAMEWORK FOR LOW-RESOURCE MUSIC UNDERSTANDING

3.2.1 UNIVERSESET: AUTOMATED DATA CONSTRUCTION PIPELINE

To enable scalable post-training without human annotation bottlenecks, we construct UNIVERSESET via a three-stage automated synthesis pipeline:

1. Data Curation and Alignment. We extract track metadata (titles, artists, regional tags) from global streaming platforms, align them with YouTube identifiers, and download high-fidelity audio streams. Tracks exceeding 10 minutes are filtered out to maintain computational efficiency.

2. Multi-Aspect Feature Annotation. To bridge raw audio waveforms with LLM text spaces, we extract structured symbolic, linguistic, and acoustic representations:

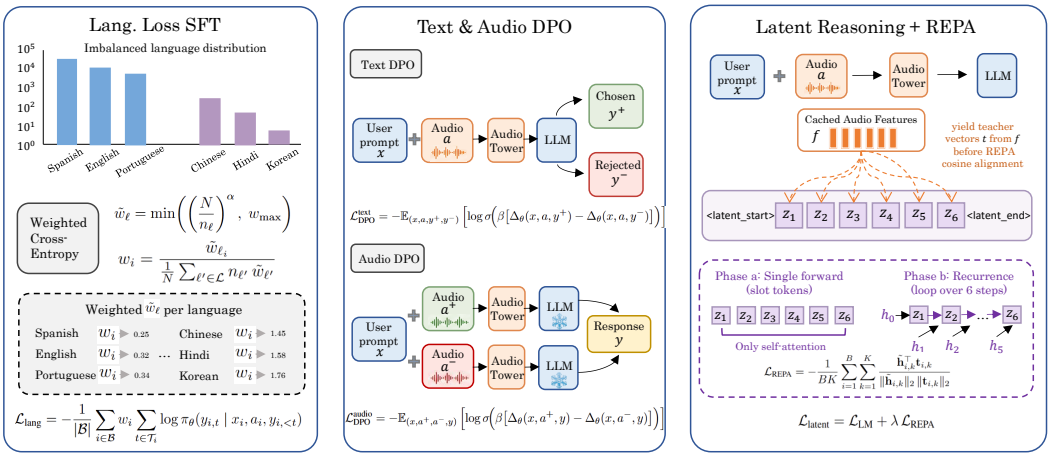


Figure 3: Overview of the post-training objectives. **Lang. Loss SFT** scales token-level loss by inverse language frequency to boost low-resource traditions. **Text/Audio DPO** optimizes textual preference under fixed audio and audio preference under fixed responses. **Latent Reasoning + REPA** imposes continuous representation alignment via continuous slot reasoning (dense Qwen2.5-Omni-7B-Instruct Xu et al. (2025a)) or encoder-side latent pooling (MoE Qwen3-Omni-30B-A3B-Instruct Xu et al. (2025b)).

- *Symbolic Scores*: Melody contours and pitch progressions are transcribed into ABC notation using an in-house fine-tuned SheetSage model (Donahue et al., 2022).
- *Linguistic Lyrics*: Lyrics and language IDs are transcribed using Qwen3-ASR (Shi et al., 2026).
- *Acoustic Captions*: Fine-grained acoustic properties (timbre, instrumentation, ornamentation) are captioned via Qwen3-Omni-Captioner (Ma et al., 2025b).

Extracted modalities are cross-validated against initial metadata using Qwen3-Next-80B-A3B-Instruct[1]. Inconsistent entries are automatically pruned.

3. Multi-Turn Dialogue Synthesis. We prompt Qwen3-Next-80B-A3B-Instruct to simulate interactive listening sessions by assigning structured persona profiles (demographic, regional, and task-oriented). Both user and assistant are modeled as audio-only listeners. The assistant is instructed to generate explicit step-by-step reasoning traces prior to outputting answers.

The resulting UNIVERSESET contains 113,023 multi-turn dialogues encompassing 510,078 QA pairs across 36 languages, mirroring a realistic long-tailed global distribution (Appendix A.1–A.3).

3.2.2 LOW-RESOURCE MULTIMODAL LEARNING STRATEGIES

As illustrated in Figure 3, we introduce a three-pillar post-training objective to mitigate multilingual long-tail imbalance and enforce strict acoustic grounding.

Language-Reweighted SFT To prevent dominant languages from overwhelming long-tailed regional traditions during SFT, we apply a smooth inverse-frequency loss weight w_i to each example belonging to language ℓ_i:

$$w_i \propto \min\left(\left(\frac{N}{n_{\ell_i}}\right)^{\alpha}, w_{\max}\right), \quad \text{s.t.} \quad \mathbb{E}_i[w_i] = 1 \tag{2}$$

where n_{ℓ_i} is the corpus count for language ℓ_i, N is the total count, and α controls smoothing. The reweighted supervised objective scaling the gradient contribution of underrepresented languages is:

$$\mathcal{L}_{\text{lang}} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} w_i \sum_{t\in\mathcal{T}_i} \log \pi_\theta(y_{i,t} \mid x_i, a_i, y_{i,<t}) \tag{3}$$

[1]https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct

Complementary Text and Audio-Tower DPO Following SFT, we apply Direct Preference Optimization (Rafailov et al., 2023; Chaubey et al., 2026) to reinforce acoustic reliance over language priors:

- *Text DPO*: Keeping audio a_i fixed, we contrast high-quality responses y_i^+ against text-only or hallucinated responses y_i^-, jointly updating the LLM and audio tower across both dense and MoE backbones.
- *Audio-Tower DPO*: Applied to dense backbones, we fix the response y_i and contrast the true audio a_i^+ against a hard-negative swapped audio track a_i^-. By freezing the LLM and updating only the audio encoder, the model is penalized for generating responses invariant to acoustic changes (Baid et al., 2026).

REPA-Grounded Latent Reasoning To prevent extended chain-of-thought reasoning from drifting away from raw acoustic context, we integrate Representation Alignment (REPA) (Yu et al., 2025) with continuous latent reasoning:

$$\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{LM}} + \lambda \, \mathcal{L}_{\text{REPA}} \tag{4}$$

For dense architectures (Qwen2.5-Omni), latent reasoning vectors $z_1 \cdots z_K$ are evolved directly within the decoder and aligned with audio-encoder features via cosine distance. For MoE architectures (Qwen3-Omni), to maintain single-pass vLLM serving efficiency, we freeze the MoE decoder and restrict REPA alignment to K-step pooled audio activations within the encoder and routing gates. Detailed mathematical formulations and MoE stability techniques are provided in Appendix B.1.

4 Experiments

4.1 Experimental Settings

All experiments are implemented using `ms-swift` and Megatron-LM (Shoeybi et al., 2020) on eight 80GB NVIDIA GPUs. The training pipeline employs `bf16` mixed precision, FlashAttention, activation recomputation, and a maximum sequence length of 16,384. We evaluate our framework across two baseline backbones with audio-conditioned thinking dialogues: the dense Qwen2.5-Omni (configured with tensor parallelism) and the MoE Qwen3-Omni (configured with tensor parallelism with expert parameters shared across ranks). Hyperparameters and data configurations are detailed in Appendix B.2.

4.2 Main Results

4.2.1 Overall Performance Comparison

Table 1 presents the performance breakdown on UniVerseBench. Proprietary baselines—Gemini 3 Flash and Qwen3.5-Omni-plus (QwenTeam, 2026)—are evaluated via direct option generation. In contrast, open-resource baselines (KimiTeam et al., 2025; Dinkel et al., 2025; Ghosh et al., 2025), Qwen2.5-Omni and Qwen3-Omni are strictly benchmarked under both direct-option and thinking modes. Under thinking mode, a prediction is considered correct only if both the reasoning chain and the final symbolic answer are valid. Although direct-option mode generally yields higher accuracy, we anchor our subsequent analyses to thinking mode to perform a fine-grained evaluation of internal reasoning capabilities.

Qwen3.5-Omni-plus yields the highest overall accuracy (74.4%), serving as an empirical performance ceiling. Crucially, thinking-format SFT consistently elevates baseline capabilities, boosting Qwen2.5-Omni by 6.7 percentage points (and up to 14.9 points with latent representation alignment) and Qwen3-Omni by 5.9 points. Most models follow a clear difficulty gradient: *LoC* $>$ *L-oth.* $>$ *ko*, reflecting the higher structural diversity and acoustic heterogeneity within the LoC stratum. Across skill dimensions, Qwen3.5-Omni-plus dominates in *feature extraction*, *causal reasoning*, *pattern recognition*, and *error detection*, whereas Gemini 3 Flash leads in *comparison analysis*. While Gemini 3 Flash achieves benchmark-leading performance on comparison analysis (84.9%), fine-grained cross-stream acoustic alignment remains an active challenge for other architectures.

4.2.2 Imbalanced Learning Strategies

Table 2 compares the efficacy of data reweighting, preference optimization, and representation-aligned objectives across both Qwen backbones.

For **Qwen2.5-Omni**, language loss reweighting, its combination with Text DPO, and single-phase latent reasoning (Phase a) yield only marginal gains over standard thinking-format SFT. In contrast,

Table 1: Overall accuracy (%) on UNIVERSEBENCH across language and skill strata. Language subsets include *LoC* (Languages of China), *ko* (Korean), and *L-oth.* (other languages). Skill strata comprise *FE* (feature extraction), *CR* (causal reasoning), *PR* (pattern recognition), *CA* (comparison analysis), and *ED* (error detection). "Post-trained" denotes models fine-tuned on our thinking-format post-training data.

Model	*Overall*	*LoC*	*ko*	*L-oth.*	*FE*	*CR*	*PR*	*CA*	*ED*
Kimi-audio (only options)	46.2	37.6	56.4	46.8	44.7	47.6	48.6	57.3	51.4
Kimi-audio (w/ think)	34.9	32.9	38.2	34.7	33.7	39.3	34.6	36.6	42.9
MiDashengLM (only options)	42.5	38.1	48.2	42.6	41.1	46.6	41.3	52.2	34.3
MiDashengLM (w/ think)	31.0	27.8	35.6	31.0	29.4	37.6	33.2	25.4	40.0
Music Flamingo (only options)	56.0	47.0	64.2	57.5	55.3	59.8	55.1	54.3	57.1
Music Flamingo (w/ think)	42.3	36.0	53.6	41.3	42.2	46.0	40.9	35.3	40.0
Gemini 3 Flash	70.8	67.6	**77.6**	69.5	68.7	73.3	71.3	**84.9**	55.6
Qwen3.5-Omni-plus	**74.4**	**71.0**	75.5	**75.7**	**73.6**	**80.1**	**72.6**	71.4	**77.1**
Qwen2.5-Omni (options)	46.4	40.9	52.9	46.7	44.1	**54.7**	50.2	43.1	37.1
Qwen2.5-Omni (Post-trained, options)	46.7	41.9	53.1	46.7	45.8	50.2	46.5	**47.4**	**54.3**
Qwen2.5-Omni (w/ think)	33.9	29.2	45.2	32.0	32.5	38.3	35.7	32.8	37.1
Qwen2.5-Omni (Post-trained, w/ think)	**48.8**	**44.4**	**57.6**	**47.8**	**47.4**	52.4	**54.4**	42.2	48.6
Qwen3-Omni (options)	**59.7**	**55.7**	**65.5**	**59.5**	56.3	**68.4**	**65.0**	**61.6**	**74.3**
Qwen3-Omni (Post-trained, options)	58.0	54.1	65.0	57.3	**56.8**	64.1	57.5	52.6	68.6
Qwen3-Omni (w/ think)	47.5	44.3	52.6	47.3	44.9	58.3	51.4	37.5	51.4
Qwen3-Omni (Post-trained, w/ think)	53.4	46.4	62.6	53.5	50.8	61.4	55.1	56.0	65.7

Table 2: Ablation study of imbalanced learning and alignment strategies on UNIVERSEBENCH. Values report strict parseable accuracy (%) presented as Qwen2.5-Omni | Qwen3-Omni. Underlined values highlight the best accuracy per backbone.

Method	Qwen2.5	Qwen3	Method	Qwen2.5	Qwen3
SFT (w/ think)	40.6	<u>53.4</u>	Audio DPO / Enc.-side REPA	45.7	52.9
Lang. Loss	41.8	51.9	LR. + REPA (Phase a)	42.9	—
Lang. Loss + Text DPO	42.6	50.9	LR. + REPA (Phase b) / Route-b ΔCE	<u>48.8</u>	52.9

recurrent latent alignment (Phase b) achieves optimal dense performance (48.8%). Audio DPO improves upon Phase a but falls short of Phase b.

For **Qwen3-Omni**, standard thinking-format SFT achieves peak accuracy (53.4%). Conversely, language loss reweighting and Text DPO induce slight performance regressions. Encoder-side REPA and its joint variant with router-gate ΔCE weighting offer robust domain adaptation, matching SFT performance while preserving single-forward serving efficiency.

Furthermore, correlation analysis reveals that performance gains are fundamentally decoupled from training volume: across all post-training strategies and backbones, neither Pearson's r nor Spearman's ρ exhibits a statistically significant positive correlation with per-language utterance counts ($p > 0.05$; see Appendix B.3.2 and Table 6).

4.2.3 CHAIN-OF-THOUGHT ANALYSIS

To investigate how post-training influences the model's internal reasoning, we partition model outputs on UNIVERSEBENCH into four mutually exclusive categories based on option correctness and explicit grounded token verbalization (Figure 5): *Correct w/ mention*, *Correct w/o mention*, *Wrong w/ mention*, and *Wrong w/o mention*. Following post-training, models articulate the ground-truth concepts more frequently in their CoT traces; however, the proportion of *Wrong w/ mention* instances also rises.

Qualitative inspection reveals a structural shift in model behavior. On one hand, post-training reinforces domain-specific reasoning, allowing models to articulate music theory concepts, aesthetic

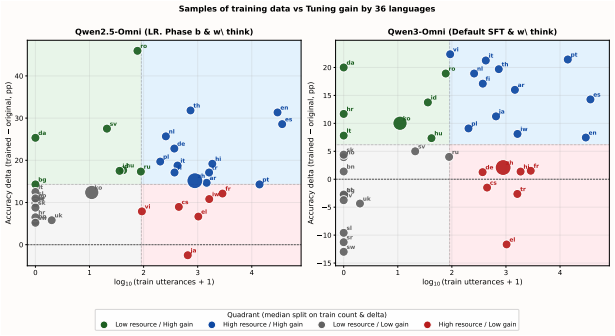


Figure 4: Per-language training count versus accuracy delta (trained − original, percentage points) across 36 languages.

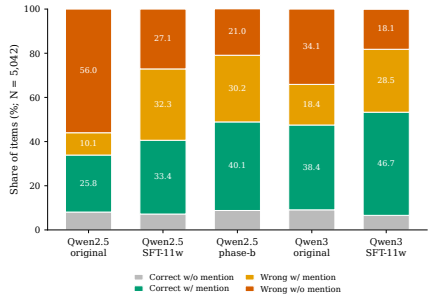


Figure 5: Distribution of reasoning-selection settings across original and post-trained checkpoints.

nuances, and explicit acoustic cues more accurately. On the other hand, this analytical strength introduces a key trade-off termed the *reasoning–label disconnect*: as internal chain-of-thought traces grow longer and more elaborate, models occasionally suffer from an *answer–option mismatch*. In these cases, sound semantic logic fails to map onto the correct option key due to attention decay and format alignment bias (Wang et al., 2024b; Turpin et al., 2023). Furthermore, LALMs still exhibit a representation bottleneck on low-resource regional musical elements (e.g., highly specialized folk terminology and regional timbres), over-indexing on surface-level textual or acoustic anchors rather than forming holistic, structural representations. Extended case studies are provided in Appendix C.2, and Figure 9 visualizes the diverse non-Western musical entities and instruments articulated during model reasoning traces.

4.3 Cross-cultural Generalization

We evaluate cross-cultural transfer across three granular dimensions: language exposure volume, Chinese regional traditions, and Korean geographic/melodic modes (*tori*). Unless otherwise noted, evaluations use the optimal checkpoint for each backbone.

Language Transfer. The grouping of languages across training volume and accuracy gain, as visualized in Figure 4 (with the corresponding numerical values detailed in Table 4), reveals a structural decoupling between pre-training volume and post-training efficacy. In the *low-resource/high-gain* quadrant, Qwen2.5-Omni achieves notable zero- and few-shot improvements on minimally exposed languages (e.g., Romanian and Danish). Conversely, the *high-resource/low-gain* quadrant exhibits diminishing returns, where high-volume languages yield modest gains or slight regressions under Qwen3-Omni (e.g., French and Greek). This indicates that post-training is not universally Pareto-improving: latent reasoning in dense models effectively propagates structural musical priors to underrepresented languages, whereas SFT on MoE models risks disrupting pre-existing representations in high-resource regimes.

Chinese Regional Variation. Aggregating macro-regional metrics obscures significant sub-regional variance (Table 3). Qwen2.5-Omni demonstrates marked accuracy boosts in previously low-performing regions (e.g., Jiangsu and Shanghai), though regions such as Qinghai and Hainan remain persistent bottlenecks. Qwen3-Omni achieves top-tier regional performance in Chongqing and Tianjin, but experiences slight regressions in areas where the base model was already strong (e.g., Jiangsu and Beijing). This emphasizes that culturally grounded failure modes are highly sensitive to fine-grained regional variations within a single language.

Korean Geography and Melodic Mode. Evaluating the Korean subset across provincial styles and melodic modes (*tori*) indicates consistent cross-regional gains (e.g., Pyeongan; see Appendix B.3.1, Table 5). Stratification by *tori* highlights broad performance improvements, particularly in *Susimga* and *Nanbongga*. Nonetheless, persistent music-theoretic bottlenecks remain in specific modes such as *Sin-gyeong*.

In summary, cross-cultural evaluation exhibits multi-scale behavior: language-level exposure is insufficient to capture regional or music-theoretic dynamics. Latent reasoning in Qwen2.5-Omni

Table 3: Best post-training performance versus original performance on the Chinese regional subset (sorted by mean trained accuracy). Guangdong* includes Hong Kong and Macau. Regions highlighted in violet indicate those with official regional or indigenous languages. Format: original | trained.

Region	Qwen2.5-Omni	Qwen3-Omni	Region	Qwen2.5-Omni	Qwen3-Omni
Chongqing	31.2 | 75.0	62.5 | 75.0	Ningxia	8.7 | 43.5	56.5 | 43.5
Hebei	31.0 | 56.9	37.9 | 48.3	Guizhou	31.1 | 42.6	39.3 | 42.6
Shanghai	34.5 | 56.4	40.0 | 47.3	Taiwan	29.1 | 41.8	41.8 | 48.1
Liaoning	18.5 | 55.6	55.6 | 59.3	Shandong	26.1 | 41.3	52.2 | 52.2
Jiangsu	34.4 | 54.1	59.0 | 49.2	Gansu	28.3 | 41.3	30.4 | 43.5
Henan	39.1 | 52.2	52.2 | 52.2	Jiangxi	26.2 | 41.0	47.5 | 50.8
Inner Mongolia	35.0 | 50.0	30.0 | 48.3	Shaanxi	18.2 | 40.9	56.8 | 47.7
Heilongjiang	27.3 | 50.0	27.3 | 45.5	Anhui	39.0 | 40.7	50.8 | 45.8
Xizang	40.7 | 48.1	51.9 | 55.6	Tianjin	33.3 | 40.0	53.3 | 60.0
Xinjiang	30.4 | 47.8	53.6 | 44.9	Jilin	32.3 | 38.7	32.3 | 35.5
Hubei	31.2 | 46.9	35.9 | 45.3	Guangdong*	18.5 | 38.5	27.7 | 38.5
Yunnan	30.2 | 46.5	27.9 | 44.2	Sichuan	10.3 | 33.3	56.4 | 48.7
Beijing	18.2 | 45.5	72.7 | 63.6	Shanxi	40.9 | 29.5	47.7 | 38.6
Zhejiang	33.3 | 45.2	54.8 | 50.0	Hainan	16.7 | 26.7	40.0 | 33.3
Hunan	40.5 | 45.2	42.9 | 35.7	Qinghai	38.6 | 26.3	42.1 | 45.6
Fujian	26.5 | 44.1	32.4 | 44.1	Guangxi	26.8 | 43.9	31.7 | 43.9

preferentially lifts lower baselines, whereas SFT in Qwen3-Omni preserves higher peak accuracy at the expense of potential regressions in high-resource strata. Achieving robust cross-cultural alignment thus necessitates fine-grained geographic and musicological stratification beyond language-averaged metrics. Detailed case studies are provided in Appendix C.1.

5 Conclusion and Future Work

We introduced UniVerse, an open and reproducible framework for culturally grounded, low-resource music understanding, comprising the multilingual benchmark UniVerseBench and the post-training dataset UniVerseSet. Benchmarking and training across dense (Qwen2.5-Omni) and MoE (Qwen3-Omni) architectures reveals that neither scale-driven language exposure nor a single global post-training objective guarantees competence in culture-specific musical reasoning. Empirically, three core insights emerge:

1. Audio language models lack fine-grained acoustic perception. While semi-automated consensus pipelines accelerate data synthesis, current LALMs struggle to discern nuanced acoustic commonalities and stylistic variations across musical traditions. AI-driven benchmark generation frequently fails under intricate musicological constraints, showing that current models rely heavily on text-surface associations rather than true acoustic reasoning and highlighting the necessity of expert-in-the-loop verification.

2. Cultural competence is decoupled from data volume and driven by representation alignment. Utterance counts show minimal correlation with downstream performance gain, allowing low-resource traditions to match high-volume counterparts and rendering simple loss reweighting ineffective. Instead, representation-aligned objectives—such as recurrent latent alignment in dense backbones—fundamentally reshape reasoning traces and enforce acoustic grounding, whereas adapting preference optimization to sparse MoE routers remains an open challenge.

3. Cross-cultural transfer is multi-scale and non-monotone. While latent reasoning in dense models systematically lifts lower baseline performance, full-parameter SFT on MoE architectures can inadvertently disrupt pre-existing representations in high-resource regimes. Crucially, broad language-level metrics mask substantial variance across fine-grained sub-regional traditions and specific music-theoretic dimensions.

To move toward culturally authentic audio intelligence, we highlight three future directions: (1) expert-in-the-loop frameworks translating ethnomusicological knowledge into verifiable alignment signals; (2) controllable music generation constrained by structural and regional priors; and (3) interactive human–AI co-creation environments paired with dynamic benchmarks to evaluate real-world workflows.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and proofreading (e.g., refining grammar, clarity, and readability of the written text). We have reviewed all AI-assisted work. Specifically, more than three authors independently compared the AI-suggested edits against the original drafts, verified that no factual content, numerical data, or technical claims were altered, and ensured that the final wording accurately reflects our intended meaning and scientific rigor. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All audio materials used in this work are sourced from publicly accessible URLs, and we do not claim ownership over any original content. The accompanying datasets—including sheet music, textual descriptions, lyrics, and any associated metadata—are collected exclusively for non-commercial academic research purposes. These data are not intended for commercial exploitation, redistribution, or any for-profit use. We acknowledge that all copyrights, intellectual property rights, and other proprietary rights remain with their respective owners. Any use of these data beyond the scope of academic research is strictly prohibited without explicit prior permission from the rights holders.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our findings, we provide complete details regarding our datasets, code implementations, training configurations, and evaluation benchmarks:

- **Datasets and Benchmarks:** The complete UNIVERSEBENCH evaluation suite and the UNIVERSESET post-training dataset—including audio source metadata, multilingual transcripts, expert-verified reasoning chains, and task category split definitions—will be made publicly available upon publication.
- **Codebase and Checkpoints:** All model training and evaluation scripts are built upon open-source frameworks. The complete codebase, evaluation pipelines, and post-trained model weights will be open-sourced under an Apache 2.0 license.
- **Training Details and Environment:** Full experimental configurations are explicitly specified in Section 4 and Appendix B.2.

REFERENCES

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Ami Baid, Zihui Xue, and Kristen Grauman. Don't let the video speak: Audio-contrastive preference optimization for audio-visual language models. *arXiv preprint arXiv:2604.14129*, 2026. URL https://arxiv.org/abs/2604.14129.

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. MoD-DPO: Towards mitigating cross-modal hallucinations in omni LLMs using modality decoupled preference optimization. *arXiv preprint arXiv:2603.03192*, 2026. URL https://arxiv.org/abs/2603.03192.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, and Lijie Wen. Omnidpo: A preference optimization framework to address omni-modal hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20172–20180, 2026.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. Aha-bench: Benchmarking audio hallucinations in large audio-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Anna-Maria Christodoulou, Kyrre Glette, Olivier Lartillot, and Alexander Refsum Jensenius. MusiQAl: A Dataset for Music Question and Answering through Audio and Video Fusion. *Transactions of the International Society for Music Information Retrieval*, 8(1), 2025. ISSN 2514-3298. doi: 10.5334/tismir.222.

Yifan Dai, Zhenhua Wu, Bohan Zeng, Daili Hua, Jialing Liu, Bozhou Li, Yuran Wang, Chengzhuo Tong, Hao Liang, Xiaochen Ma, et al. Latentomni: Rethinking omni-modal understanding via unified audio-visual latent reasoning. *arXiv preprint arXiv:2605.22012*, 2026.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. MiDashengLM: Efficient Audio Understanding with General Audio Captions, 2025.

Chris Donahue, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. In *ISMIR*, 2022.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Music flamingo: Scaling music understanding in audio language models, 2025. URL https://arxiv.org/abs/2511.10289.

Danbinaerin Han, Mark Gotham, Dongmin Kim, Hannah Park, Sihun Lee, and Dasaem Jeong. Six Dragons Fly Again: Reviving 15th-Century Korean Court Music with Transformers and Novel Encoding, 2024.

Vincent Jung and Lonneke van der Plas. Understanding the effects of language-specific class imbalance in multilingual fine-tuning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 2368–2376, 2024. doi: 10.18653/v1/2024.findings-eacl.157. URL https://aclanthology.org/2024.findings-eacl.157/.

Angelos-Nikolaos Kanatas, Charilaos Papaioannou, and Alexandros Potamianos. CultureMERT: Continual Pre-Training for Cross-Cultural Music Representation Learning, 2025.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Venkatakrishnan Vaidyanathapuram Krishnan, Noel Alben, Anish Nair, and Nathaniel Condit-Schultz. Sanidha: A Studio Quality Multi-Modal Dataset for Carnatic Music. 2024. doi: 10.5281/zenodo.14877433.

Chun-Yi Kuan and Hung-yi Lee. Can Large Audio-Language Models Truly Hear? Tackling Hallucinations with Multi-Task Assessment and Stepwise Audio Reasoning, 2024.

Daeyong Kwon, SeungHeon Doh, and Juhan Nam. ArtistMus: A Globally Diverse, Artist-Centric Benchmark for Retrieval-Augmented Music Question Answering, 2025.

Harin Lee, Elif Çelen, Peter Harrison, Manuel Anglada-Tort, Pol van Rijn, Minsu Park, Marc Schönwiesner, and Nori Jacoby. GlobalMood: A cross-cultural benchmark for music emotion recognition, 2025.

Yi-Cheng Lin, Yu-Hua Chen, Jia-Kai Dong, Yueh-Hsuan Huang, Szu-Chi Chen, Yu-Chen Chen, Chih-Yao Chen, Yu-Jung Lin, Yu-Ling Chen, Zih-Yu Chen, I.-Ning Tsai, Hsiu-Hsuan Wang, Ho-Lam Chung, Ke-Han Lu, and Hung-yi Lee. TAU: A Benchmark for Cultural Sound Understanding Beyond Semantics, 2025.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. CultureVLM: Characterizing and Improving Cultural Understanding of Vision-Language Models for over 100 Countries, 2025.

Yinghao Ma, Siyou Li, Juntao Yu, Emmanouil Benetos, and Akira Maezawa. Cmi-bench: A comprehensive benchmark for evaluating music instruction following. *arXiv preprint arXiv:2506.12285*, 2025a.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025b.

Richard Y Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637, 2024.

Charilaos Papaioannou, Ioannis Valiantzas, Theodoros Giannakopoulos, Maximos Kaliakatsos-Papakostas, and Alexandros Potamianos. A Dataset for Greek Traditional and Folk Music: Lyra, 2022.

Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. Universal Music Representations? Evaluating Foundation Models on World Music Corpora, 2025.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A Framework for Decoupling and Assessing the Capabilities of VLMs. abs/2406.14544, 2024. URL https://arxiv.org/abs/2406.14544.

QwenTeam. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark, 2024.

Patrick Savage, Emily Merritt, Tom Rzeszutek, and Steven Brown. Cantocore: A new cross-cultural song classification scheme. 2012.

Xavier Serra. Creating research corpora for the computational study of music: The case of the compmusic project. In *AES 53rd International Conference: Semantic Audio; 2014 Jan 27-29; London, UK. New York: Audio Engineering Society; 2014. Article number 1-1 [9 p.].* Audio Engineering Society, 2014.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

Shiliang Sun, Zhilin Lin, and Xuhan Wu. Hallucinations of large multimodal models: Problem and countermeasures. *Information Fusion*, 118:102970, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a. URL https://arxiv.org/abs/2406.11839.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024b.

Yuhui Wang, Changjiang Li, Guangke Chen, Jiacheng Liang, and Ting Wang. Reasoning or retrieval? a study of answer attribution on large reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 102235–102258, 2026.

Shangda Wu, Yashan Wang, Ruibin Yuan, Guo Zhancheng, Xu Tan, Ge Zhang, Monan Zhou, Jing Chen, Xuefeng Mu, Yuejie Gao, Yuanliang Dong, Jiafeng Liu, Xiaobing Li, Feng Yu, and Maosong Sun. CLaMP 2: Multimodal Music Information Retrieval Across 101 Languages Using Large Language Models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 435–451, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.27.

Shangda Wu, Ziya Zhou, Yongyi Zang, Yutong Zheng, Dafang Liang, Ruibin Yuan, and Qiuqiang Kong. Voices of civilizations: A multilingual qa benchmark for global music understanding, 2026. URL https://arxiv.org/abs/2603.00533.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Liumeng Xue, Ziya Zhou, Jiahao Pan, Zixuan Li, Shuai Fan, Yinghao Ma, Sitong Cheng, Dongchao Yang, Haohan Guo, Yujia Xiao, et al. Audio-flan: An instruction-following dataset for unified audio understanding and generation of speech, music, and sound. *arXiv preprint arXiv:2502.16584*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2025. URL https://arxiv.org/abs/2410.06940.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. Are you really listening? Boosting Perceptual Awareness in Music-QA Benchmarks, 2025.

Yu Zhang, Ziya Zhou, Xiaobing Li, Feng Yu, and Maosong Sun. CCOM-HuQin: An Annotated Multimodal Chinese Fiddle Performance Dataset. *Transactions of the International Society for Music Information Retrieval*, 6(1):60–74, 2023. ISSN 2514-3298. doi: 10.5334/tismir.146.

Monan Zhou, Shenyang Xu, Zhaorui Liu, Zhaowen Wang, Feng Yu, Wei Li, and Baoqiang Han. CCMusic: An Open and Diverse Database for Chinese Music Information Retrieval Research. *Transactions of the International Society for Music Information Retrieval*, 8(1):22–38, 2025. ISSN 2514-3298. doi: 10.5334/tismir.194.

APPENDIX

A UNIVERSE BENCHMARK VERIFICATION

To ensure the precision, musical authenticity, and logical consistency of the evaluation items, we establish a systematic, multi-stage hybrid verification protocol that integrates automated filtering, expert annotation, AI-assisted modification, adversarial selection, and a final manual audit into a continuous workflow.

First, an automated quality and adversarial filtering pass is conducted on the raw pool of generated QA candidates to select a refined subset of 1,873 high-quality candidate QA pairs designated for comprehensive human expert validation. This process integrates baseline-agnostic quality screening with a programmatic adversarial selection pipeline designed to maximize evaluation difficulty. It prioritizes boosted questions, specifically those where baseline models failed or where Gemini optimization flagged misleading acoustic cues. To guarantee uniform benchmark density across the dataset, we enforce a strict minimum of five questions per audio clip, dynamically backfilling the quota using the baseline models' incorrect items.

Next, a multi-category human expert annotation phase is executed, where each of the 1,873 selected candidate items is assigned to at least two ethnomusicology or music theory expert. The experts evaluate the audio-text alignment and categorize each QA pair into one of three statuses:

- **Directly Usable**: The question, options, and ground-truth answers are correct and require no modifications.
- **Requires Modification**: The item contains minor flaws such as ambiguous question phrasing, overlapping distractors, or slightly inaccurate musical terminology, and the expert provides explicit, granular feedback detailing the necessary corrections to the question stem or options.
- **Unusable**: The item suffers from fundamental conceptual or alignment errors.

To maintain the targeted scale of the 1,873-item human-audited core benchmark, unusable pairs trigger a rejection-replacement process where the expert reviews alternative candidate QA pairs generated from the same audio clip, ensuring all 1,873 items meet strict musicological standards.

Following the human annotation phase, we leverage DeepSeek-4.0-Pro for secondary quality control. DeepSeek-4.0-Pro consolidates the expert annotations, textual modification guidelines, and replacement candidates to reconstruct and refine the question stems, options, and annotations. After processing these updates, the model conducts a secondary automated check over the entire consolidated dataset to verify formatting uniformity and logical consistency before exporting.

Finally, a senior panel of musicology experts conducts a final manual audit of this compiled subset. This final pass manually verifies that all human-requested modifications and replacement pairs were correctly integrated, establishing the 1,873-item core expert-audited split. To expand evaluation coverage, this core is augmented with 3,169 automated consensus-filtered QA pairs (incorporating verified global items extending VOC), yielding the complete 5,042-item UNIVERSEBENCH suite.

A.1 DATA COLLECTION AND ALIGNMENT

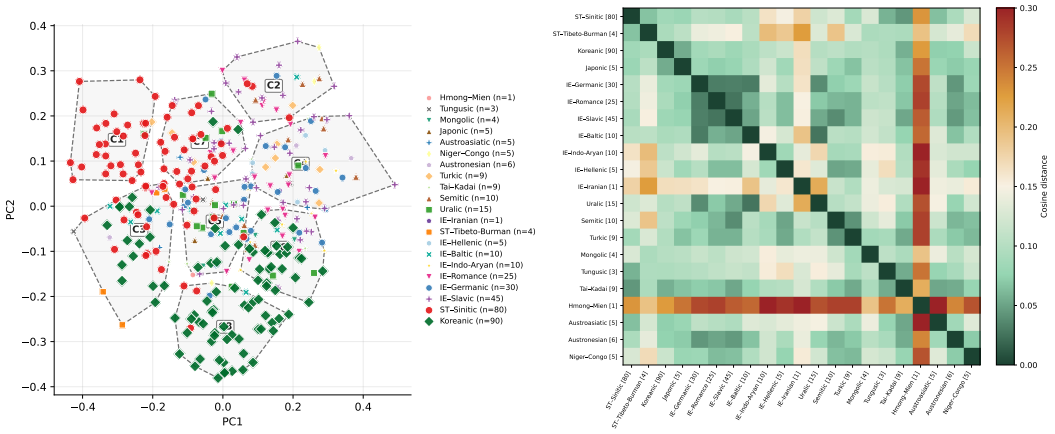
To ensure the cultural diversity of our low-resource music corpus, our metadata curation specifically targeted traditional and regional music. We queried a global music streaming service using a carefully compiled list of **770 traditional and regional genre tags**. Once the track titles and corresponding artist names were extracted, we utilized the SoundCharts platform[2] to bridge the textual metadata with raw audio. By querying these title-artist pairs via SoundCharts, we located their linked YouTube identifiers, which were subsequently used to download raw audio.

A.2 CROSS-VALIDATION HEURISTICS

Since the feature annotations (scores, lyrics, captions) are generated by independent models, we employed Qwen3-Next-80B-A3B-Instruct to conduct rigorous cross-modal verification. A track is excluded from the final training corpus if the LLM identifies any of the following fatal inconsistencies:

- **Symbolic-Acoustic Mismatch:** Conflicts in musical attributes, such as key signatures or time signatures, between the ABC notation (from SheetSage) and the acoustic captions (from Qwen3-Omni-Captioner).

[2]https://soundcharts.com/



(a) Audio embeddings by language family.

(b) Cosine distance matrix between language family centroids.

Figure 6: Acoustic structure of the benchmark audio excerpts in Qwen2.5-Omni embedding space. Points in panel (a) are colored by language family and each cluster's convex hull is outlined by dashed polygons; Sino-Tibetan branches are labeled `ST-Sinitic` / `ST-Tibeto-Burman` (parallel to `IE-`$\star$)). Panel (b) shows cosine distances between language family centroids.

- **Linguistic Hallucination:** Instances where the caption explicitly quotes or references specific lyrics that are entirely absent from the Qwen3-ASR transcription.

A.3 Dialogue Synthesis Rules and Constraints

To simulate realistic, multi-turn listening sessions while preventing data leakage, we enforced a strict set of rule-based constraints during the LLM dialogue generation.

A.3.1 User Profile and Task Sampling

Before generating a dialogue, a structured user profile is instantiated for each track:

- **Language Alignment:** The conversation language is dynamically matched to the lyric metadata. If the language field is missing, empty, or labeled as unknown, the pipeline robustly defaults to English.
- **Conditional Persona Mapping:** To prevent logical contradictions, user personas are generated via a two-stage sequential sampling process. We first select a demographic style group (*Gen Z, Middle-aged, Elderly,* or *Professional Academic*). We then sample a music expertise level restricted to that demographic's permitted subset. For instance, a Gen Z profile is restricted to *Lay listener* or *Amateur musician*, whereas a Professional Academic is restricted to *Professional composer* or *Cultural historian*.
- **Turn and Intent Constraints:** The dialogue length T is randomized between 1 and 9 turns. We sample I unique intents ($1 \leq I \leq \min(|Pool|, T)$) from a pool of 10 tasks. These include six general inquiries (*Factual Inquiry, Structural Analysis, Subjective Appreciation, Creative/Cultural Inquiry, Technical Verification,* and *Ethnomusicological Comparison*) and four academic tasks (*Multiple-Choice Question (MCQ), Fill-in-the-blanks, Short-answer Analysis,* and *True/False Inquiry*).
- **Mode Assignment:** For academic tasks, we assign either a *Direct Answer* mode (optimized for zero-shot quantitative benchmark evaluation) or an *Analytical* mode (optimized for reasoning-heavy instruction tuning), with a 50% probability each.

A.3.2 Query Constraints for Blind Simulation

Both the user and the assistant are framed as blind listeners with no prior access to the text annotations. User queries are heavily constrained to prevent answer leakage:

- **Feature Anchoring:** Questions are strictly limited to the verified features (score, lyrics, or captions) actually available for that specific track.

- **Information Isolation:** Queries must be concise and free of contextual hints, descriptive adjectives, or lyric quotes that could reveal the ground-truth answer (e.g., prohibiting leading phrases like "Since this is a G-minor song...").
- **Task-Specific Formatting:** For fill-in-the-blank tasks, the user provides only the template sentence containing the blank placeholder. For MCQs, all options (A, B, C, D) are constrained to have similar lengths, vocabulary, and tone to prevent heuristic guessing.

A.3.3 Assistant Reasoning and Output Rules

To reduce factual hallucinations, the assistant's behavior is explicitly restricted in both its latent reasoning and final response:

- **Verbatim Grounding:** Each assistant turn must feature a reference block citing the exact source keys (e.g., `caption`) and verbatim substrings from the extracted annotations to anchor deductions in verified facts.
- **Auditory Simulation:** The assistant's step-by-step thinking trace must be articulated as direct listening experiences (describing timbre, rhythm, and phonetic features) instead of explicitly mentioning the underlying JSON data sources.
- **Conditional Feedback Modes:** In *Direct Answer Mode*, the final output must contain only the exact option or term (e.g., "A" or "True") without greetings. In *Analytical Mode*, the assistant must provide a comprehensive, structured musicological explanation.

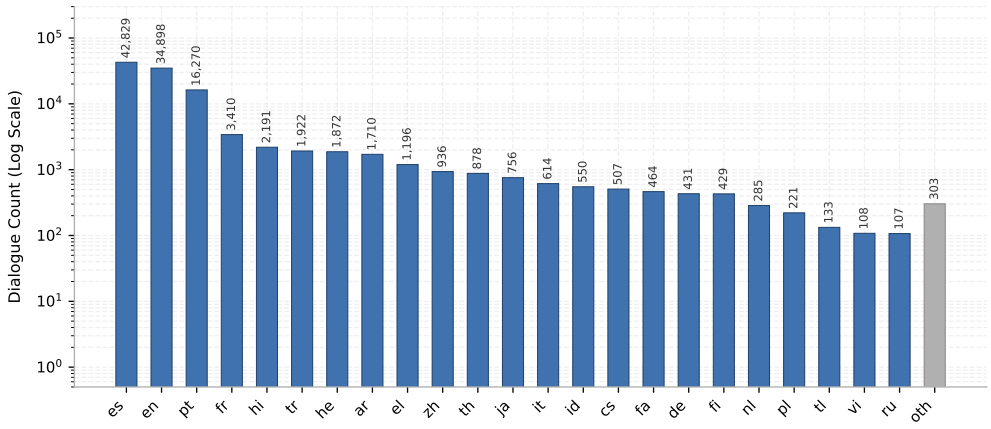


Figure 7: Language distribution of the synthesized corpus plotted on a logarithmic scale. The label `oth` aggregates all remaining low-frequency languages with fewer than 100 dialogues.

B UniVerSet Training Strategies and Experiments

This section states the detailed objectives and implementations of three training strategies illustrated in Section 3.2.2.

B.1 Methodologies

B.1.1 Language-reweighted SFT

Let example i have language label $\ell_i \in \mathcal{L}$, with counts $n_\ell = \sum_i [\ell_i = \ell]$ and $N = \sum_\ell n_\ell$. The language weight $\tilde{w}_\ell$ is

$$\tilde{w}_\ell = \min\left(\left(\frac{N}{n_\ell}\right)^{\alpha},\ w_{\max}\right), \tag{5}$$

where α is a smoothing factor and $w_{\max}$ is a cap. Weights are normalized so that $\mathbb{E}_i[w_i] = 1$:

$$w_i = \frac{\tilde{w}_{\ell_i}}{\frac{1}{N}\sum_{\ell' \in \mathcal{L}} n_{\ell'}\, \tilde{w}_{\ell'}}. \tag{6}$$

On audio-conditioned dialogues (x_i, a_i, y_i), we minimize language-reweighted cross-entropy over target tokens $\mathcal{T}_i$:

$$\mathcal{L}_{\text{lang}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \sum_{t \in \mathcal{T}_i} \log \pi_\theta(y_{i,t} \mid x_i, a_i, y_{i,<t}), \tag{7}$$

where $\mathcal{B}$ is the mini-batch.

B.1.2 TEXT AND AUDIO-TOWER DPO

After Language-reweighted SFT, we optimize preferences against a frozen reference π_{ref}. For a response y,

$$\Delta_\theta(x, a, y) = \log \pi_\theta(y \mid x, a) - \log \pi_{\text{ref}}(y \mid x, a). \tag{8}$$

Text DPO. Given pairs (x_i, a_i, y_i^+, y_i^-) with fixed audio a_i and a dispreferred response y_i^- (e.g., text-only or incorrect answers), we minimize

$$\mathcal{L}_{\text{DPO}}^{\text{text}} = -\mathbb{E} \log \sigma\Big(\beta\big[\Delta_\theta(x, a, y^+) - \Delta_\theta(x, a, y^-)\big]\Big). \tag{9}$$

The LLM and audio tower are updated jointly. This stage is used for both dense and MoE backbones.

Audio-tower DPO. Audio-tower DPO is used only for the dense backbone. Following modality-contrast preference learning (Wang et al., 2024a; Chen et al., 2026) with an RPO-style chosen-side likelihood term (Pang et al., 2024), we keep y_i fixed, contrast aligned audio a_i^+ with a hard-negative swap a_i^- (silence at most sparsely), freeze the LLM, and update only the audio tower:

$$\mathcal{L}_{\text{DPO}}^{\text{audio}} = -\mathbb{E} \log \sigma\Big(\beta\big[\Delta_\theta(x, a^+, y) - \Delta_\theta(x, a^-, y)\big]\Big) + \alpha\, \mathcal{L}_{\text{NLL}}(y \mid x, a^+), \tag{10}$$

where $\alpha > 0$.

MoE Restriction. On MoE models the same procedure is unstable. Without RPO, audio-only DPO collapses (near-zero loss and large negative rewards on both branches). RPO, smaller learning rates or β, gradient clipping, or unfreezing router gates instead cause early gradient-norm explosion (NaNs). MoE acoustic adaptation therefore uses supervised cross-entropy ($\mathcal{L}_{\text{CE}}$ and ΔCE); Audio-tower DPO is used only for dense models.

B.1.3 REPA-GROUNDED LATENT LEARNING

We add K continuous latent reasoning steps and align them to offline audio-encoder features with REPA. For each example (x_i, a_i, y_i),

$$\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{LM}} + \lambda\, \mathcal{L}_{\text{REPA}}, \qquad \lambda > 0, \tag{11}$$

where $\mathcal{L}_{\text{LM}}$ is next-token cross-entropy on y_i. At step $k \in \{1, \dots, K\}$, a student $\mathbf{s}_{i,k} \in \mathbb{R}^{d_s}$ is matched to a frozen teacher $\mathbf{t}_{i,k} \in \mathbb{R}^{d_a}$. Teachers are computed offline: frame-level features $\mathbf{f}_i \in \mathbb{R}^{T_i \times d_a}$ are cached and pooled as $\mathbf{t}_{i,k} = \text{Pool}_K(\mathbf{f}_i)_k$. Students $\{\mathbf{s}_{i,k}\}$ are computed in training and depend on the architecture. If $d_s \neq d_a$, $\tilde{\mathbf{s}}_{i,k} = \mathbf{W}\mathbf{s}_{i,k}$; otherwise $\tilde{\mathbf{s}}_{i,k} = \mathbf{s}_{i,k}$. With ℓ_2 normalization,

$$\mathcal{L}_{\text{REPA}} = -\frac{1}{BK} \sum_{i=1}^{B} \sum_{k=1}^{K} \frac{\tilde{\mathbf{s}}_{i,k}^\top \mathbf{t}_{i,k}}{\|\tilde{\mathbf{s}}_{i,k}\|_2 \, \|\mathbf{t}_{i,k}\|_2}, \tag{12}$$

where B is the batch size. No gradient flows through $\mathbf{f}_i$.

Dense Omni Models. For Qwen2.5-Omni, joint LLM–audio updates and multi-pass decoding are feasible. The target is

$$\hat{y}_i = \texttt{<latent_start>}\, z_1 \cdots z_K\, \texttt{<latent_end>}\, y_i, \tag{13}$$

where z_k are label-masked slots and $\mathbf{s}_{i,k} \equiv \mathbf{h}_{i,k} \in \mathbb{R}^{d_h}$ are final-layer decoder states. *Phase a* inserts learnable <latent_slot> embeddings in one pass. *Phase b* replaces them with <latent_rec> and feeds $\mathbf{h}_{z_{k-1}}$ as the input at z_k ($z_0 \equiv$ <latent_start>), which requires an extra pass. Latents are computed in the LLM and aligned to audio by REPA.

MoE Omni Models. For Qwen3-Omni, recurrent latent decoding is costly and incompatible with single-forward serving (e.g., vLLM); slot REPA on routed states is also unstable. We therefore freeze the MoE decoder and update only the audio tower and router gates under $\mathcal{L}_{\text{LM}} + \lambda\mathcal{L}_{\text{REPA}}$. Students are pooled from audio activations $\mathbf{g}_i$ as $\mathbf{s}_{i,k} = \text{Pool}_K(\mathbf{g}_i)_k$. $\mathcal{L}_{\text{LM}}$ is backpropagated through the frozen decoder to update routing without changing the compute graph. A frozen probe scores $\delta_i = \mathcal{L}_{\text{text-only}}(x_i, y_i) - \mathcal{L}_{\text{audio}}(x_i, a_i, y_i)$ and sets $w_i = \text{clip}(1 + \beta\delta_i,\, w_{\min},\, w_{\max})$, giving

$$\mathcal{L}_{\text{MoE}} = \sum_i w_i\, \mathcal{L}_{\text{CE}}(x_i, a_i, y_i) + \lambda\, \mathcal{L}_{\text{REPA}}. \tag{14}$$

Layer-wise routing indices are frozen from a reference pass. REPA is used only in training, so inference is standard single-forward decoding, with no extra tokens, multi-pass decoding, or routing metadata.

B.2 Detailed Parameters

All completed runs are trained with Megatron-LM through `ms-swift` on eight 80 GB GPUs, using FlashAttention, full activation recomputation, maximum sequence length 16384, one epoch, and a 1% validation split. All supervised and preference stages use the **full** post-training corpus of (thinking-format jsonl, or the corresponding full preference-pair export); we do not use reduced-scale subsets. Thinking-format data enable `add_non_thinking_prefix`. Vision towers remain frozen; audio generation is disabled during training.

Hyperparameters of Backbones. Qwen2.5-Omni is trained as a dense model with tensor parallel size 4 (sequence parallel on, except Phase b where it is disabled). Qwen3-Omni uses tensor parallel 8 (with expert parameters shared across data ranks), MoE auxiliary loss coefficient 10^{-6}, and expert capacity factor 2.0. Both use CPU optimizer offload (fraction 0.5; Text DPO on Qwen3-Omni uses 0.64) with the precision-aware optimizer. Unless noted, SFT uses $\text{lr} = 10^{-5}$, audio-tower $\text{vit_lr} = 10^{-6}$, warmup fraction 0.05, and $\text{min_lr} = 10^{-6}$, updating the full LLM together with `visual.audio_tower`.

Qwen2.5-Omni Special Setting. SFT (w/ think) fine-tunes on the full thinking corpus with micro-/global batch sizes 4/8. Lang. Loss uses the same recipe with language-level `sample_weight` on the full corpus. Latent Reasoning Phase a trains on the full lang-loss corpus with batch sizes 4/16, K=6 latent steps, and REPA weight λ=0.5; Phase b continues the Phase a checkpoint with batch sizes 1/4 under recurrent decoding (same K and λ). Text DPO is initialized from the lang-loss SFT checkpoint and trains the full LLM plus audio tower on full text-preference pairs with $\text{lr} = 10^{-7}$, $\text{vit_lr} = 5\times10^{-7}$, β=0.03, warmup 0.1, clip_grad=0.25, and batch sizes 1/4. Vanilla Audio DPO freezes the LLM, trains only the audio tower on full M2 audio-swap pairs with $\text{lr} = \text{vit_lr} = 10^{-6}$, β=0.03, warmup 0.1, clip_grad=0.25, and batch sizes 2/4 with a chosen-likelihood regularization weight $\alpha = 0.05$.

Qwen3-Omni Special Setting. SFT (w/ think) fine-tunes on the full thinking corpus with batch sizes 1/4. Lang. Loss matches this setup with language `sample_weight` on the full corpus. Encoder-side REPA initializes from the full-corpus think SFT checkpoint, freezes the MoE decoder, and updates the audio tower with projections on the full lang-loss corpus (batch sizes 2/32, K=6, λ_{REPA}=0.1, $\text{vit_lr}=10^{-5}$). Encoder-side REPA + ΔCE + gates uses the same initialization and batch sizes, but optimizes audio parameters together with MoE router gates on the full ΔCE-weighted corpus ($\text{lr} = \text{vit_lr}=10^{-5}$). Text DPO follows the Qwen2.5-Omni text-DPO schedule (full LLM + audio; β=0.03; batch 1/4) on full text-preference pairs, starting from the lang-loss SFT checkpoint.

B.3 Experimental Analysis

B.3.1 Cross-Cultural Generalization

Table 4 presents the exact numerical values corresponding to Figure 4. Table 5 shows the models' performances cross Korean provinces and melodic modes (tori).

B.3.2 Correlation Analysis

We compute Pearson's r and Spearman's ρ between post-training utterance counts (n_ℓ) and per-language accuracy across $N = 25$ languages ($n \geq 10$) on the full benchmark, as well as across four primary skills (detailed statistics in Appendix Table 6). Across all strategies and backbones, no statistically significant positive correlation emerges ($\alpha = 0.05$). Overall correlation coefficients remain near zero for both models. While weak negative correlations occasionally appear within

Table 4: Per-language accuracy (original | trained) for all benchmark languages. Rows are grouped by the Qwen2.5-Omni training-count vs. gain quadrant (median splits on $\log_{10}(n+1)$ and Δ=trained−original): green = low-resource/high-gain, blue = high-resource/high-gain, grey = low-resource/low-gain, red = high-resource/low-gain; within each block, sorted by n (high→low).

Language [#]	Qwen2.5-omni	Qwen3-omni	Language [#]	Qwen2.5-omni	Qwen3-omni
ru (Russian) [88]	30.67 | 48.00	50.67 | 54.67	ro (Romanian) [76]	12.16 | 58.11	48.65 | 67.57
hu (Hungarian) [41]	25.00 | 42.65	39.71 | 47.06	id (Indonesian) [35]	37.50 | 55.00	42.50 | 56.25
sv (Swedish) [20]	23.75 | 51.25	41.25 | 46.25	bg (Bulgarian) [0]	31.43 | 45.71	52.86 | 50.00
da (Danish) [0]	25.33 | 50.67	36.00 | 56.00	es (Spanish) [36476]	25.97 | 54.55	37.66 | 51.95
en (English) [29936]	20.90 | 52.24	58.21 | 65.67	pt (Portuguese) [13911]	33.33 | 47.62	41.67 | 63.10
hi (Hindi) [1864]	35.62 | 54.79	60.27 | 61.64	tr (Turkish) [1624]	26.32 | 43.42	52.63 | 50.00
ar (Arabic) [1463]	34.67 | 49.33	42.67 | 58.67	zh (Chinese) [886]	29.24 | 44.40	44.26 | 46.35
th (Thai) [736]	30.30 | 62.12	54.55 | 74.24	it (Italian) [423]	31.25 | 50.00	40.00 | 61.25
fi (Finnish) [372]	32.89 | 50.00	42.11 | 59.21	de (German) [369]	21.52 | 44.30	50.63 | 51.90
nl (Dutch) [257]	27.03 | 52.70	29.73 | 48.65	pl (Polish) [202]	33.33 | 53.03	53.03 | 62.12
ko (Korean) [10]	45.19 | 57.56	52.55 | 62.57	uk (Ukrainian) [1]	40.58 | 46.38	56.52 | 52.17
bn (Bengali) [0]	31.94 | 37.50	36.11 | 37.50	hr (Croatian) [0]	32.47 | 38.96	40.26 | 51.95
et (Estonian) [0]	31.51 | 41.10	49.32 | 46.58	lv (Latvian) [0]	38.75 | 48.75	46.25 | 42.50
lt (Lithuanian) [0]	28.12 | 40.62	40.62 | 48.44	no (Norwegian) [0]	26.32 | 36.84	50.00 | 53.95
sr (Serbian) [0]	32.39 | 42.25	54.93 | 43.66	sk (Slovak) [0]	31.87 | 40.66	43.96 | 48.35
sl (Slovenian) [0]	43.84 | 54.79	60.27 | 50.68	sw (Swahili) [0]	40.26 | 45.45	51.95 | 38.96
fr (French) [2874]	39.39 | 51.52	53.03 | 54.55	iw (Hebrew) [1634]	36.49 | 47.30	33.78 | 41.89
el (Greek) [1027]	45.00 | 51.67	68.33 | 56.67	ja (Japanese) [654]	35.00 | 32.50	43.75 | 55.00
cs (Czech) [447]	37.31 | 46.27	62.69 | 61.19	vi (Vietnamese) [92]	47.37 | 55.26	38.16 | 60.53

Table 5: Korean accuracy stratified by region (left) and melodic mode (*tori*, right). Format: original | trained.

Region	Qwen2.5-Omni	Qwen3-Omni	Mode	Qwen2.5-Omni	Qwen3-Omni
Gyeonggi	41.5 | 51.8	52.3 | 52.3	Yukjabaegi	50.2 | 56.5	55.9 | 66.1
Gangwon	58.7 | 61.9	50.8 | 60.3	Menari	46.7 | 54.4	50.9 | 56.2
Gyeongsang	39.7 | 53.8	47.4 | 60.3	Susimga	37.1 | 61.0	40.0 | 65.7
Jeolla	50.0 | 56.2	56.0 | 65.6	Jin-gyeong	42.2 | 54.1	55.0 | 55.0
Jeju	44.4 | 59.3	54.3 | 59.3	Gyeong	28.6 | 71.4	57.1 | 64.3
Hwanghae	45.2 | 65.8	50.0 | 69.2	Ban-gyeong	50.0 | 55.3	46.1 | 57.9
Pyeongan	30.6 | 62.9	54.8 | 71.0	Nanbongga	40.0 | 65.0	57.5 | 73.8
Hamgyeong	34.4 | 56.2	40.6 | 62.5	Sin-gyeong	30.8 | 53.8	30.8 | 50.0
Gyeonggi/Chungcheong	36.4 | 54.5	36.4 | 63.6	Jeju scale	46.6 | 63.8	55.2 | 60.3

specific skills (e.g., Spearman's ρ on CA), overall accuracy remains fundamentally decoupled from training volume.

These findings underscore two key conclusions: (1) **Benchmark characteristics:** UNIVERSEBENCH evaluates culturally grounded musical understanding rather than text-language frequency, ensuring low-resource languages are not systematically disadvantaged. (2) **Post-training dynamics:** Imbalanced-learning interventions enhance reasoning by reshaping internal representations, rather than merely amplifying high-frequency training data.

C CASE STUDY

C.1 KEYWORD CONCENTRATION IN MELODIC-CONTOUR QUESTIONS

We analyze melodic-contour questions to investigate pipeline limitations stemming from the overrepresentation of the label `undulating`. Specifically, we quantify this label bias, evaluate whether models exploit it as a shortcut, and trace its root cause across CantoCore, its automatic parser, and Gemini-based generation.

C.1.1 EXAMINATION SETUP

We restricted the audit to questions written in Chinese or English. Then, we programmatically and manually searched the benchmark for melodic-contour questions, which returned 191 questions, each paired with a distinct audio recording. Two examples from the retained subset are shown below.

Table 6: Correlation between post-training utterance count and per-language accuracy. LR. represents Latent Reasoning. REPA++ represents REPA+ΔCE+gates. Comparison Analysis merges plain + neutral variants. Category columns report Pearson's r and Spearman's ρ within each skill.

	Overall				FE		CR		PR		CA	
Strategy	r	p	ρ	p	r	ρ	r	ρ	r	ρ	r	ρ
Q2.5 SFT (think)	0.255	0.219	0.028	0.895	0.279	0.105	0.346	0.096	−0.014	0.204	−0.272	−0.531
Q2.5 Lang. Loss	0.270	0.193	0.095	0.651	0.350	0.140	0.099	−0.077	0.062	0.031	−0.130	−0.480
Q2.5 Text DPO	0.339	0.097	0.064	0.760	0.327	−0.108	0.150	0.017	0.105	0.164	0.252	−0.133
Q2.5 Audio DPO	0.366	0.072	0.130	0.536	0.355	0.051	0.222	0.103	0.209	0.312	−0.205	−0.186
Q2.5 LR Phase-a	0.312	0.129	−0.037	0.862	0.353	−0.121	0.207	0.113	0.049	0.211	−0.158	−0.442
Q2.5 LR Phase-b	0.124	0.556	−0.173	0.407	0.146	−0.157	0.018	−0.081	0.197	0.250	−0.157	−0.313
Q3 SFT (think)	0.091	0.666	0.046	0.827	0.173	0.038	−0.004	0.044	−0.166	−0.083	0.040	−0.240
Q3 Lang. Loss	0.038	0.858	−0.133	0.525	−0.001	−0.202	0.090	0.037	0.016	−0.042	0.056	−0.204
Q3 Text DPO	0.059	0.779	−0.081	0.701	0.193	−0.073	−0.169	−0.249	−0.187	0.050	−0.003	−0.547
Q3 Enc.-side REPA	0.123	0.559	0.120	0.568	0.199	0.125	−0.204	−0.090	0.165	−0.009	0.150	−0.238
Q3 Enc.-side REPA++	0.139	0.506	0.109	0.604	0.308	0.280	−0.299	−0.260	−0.071	−0.074	−0.119	−0.308

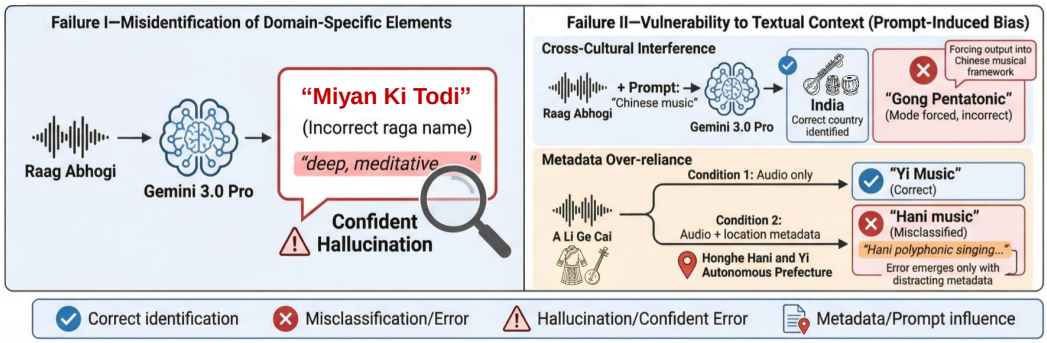


Figure 8: Two systematic failure modes of LALMs on Raag Abhogi and A Li Ge Cai: misidentification & prompt-induced bias.

Example 1. Which of the following best describes the melodic contour of the vocal line in this excerpt?
A. Subtle Segmented Descents
B. Staccato
C. Arpeggiated
D. Undulating
Correct answer: D. Undulating

Example 2. Based on what you hear, how would you characterize the melodic contour of this piece?
A. Arched, undulating, and wave-like
B. Linear and descending, with arched segments
C. pointillist smooth
D. syncopated jagged
Correct answer: A. Arched, undulating, and wave-like

We then extracted the melodic-contour feature produced by our CantoCore parser. Table 7 summarizes the audit results for this 191-question subset: it reports the proportion of correct options containing `Undulating` and, for the corresponding recordings, the distribution of parser-derived melodic-contour features that contain only `Undulating`, contain `Undulating` together with other labels, or do not contain `Undulating`. Note that `Undulating` occurs in 178/191 correct answers (93.19%), this is a downstream answer regularity and therefore motivates the shortcut analysis below. We call these 178 questions undulating-positive questions in later sections.

C.1.2 Model results and the limits of the shortcut claim

The concentration of the semantic label `Undulating` in every correct option creates a corpus-level cue that a model could potentially use without resolving the audio in full. Thus, we examine the results of every Qwen2.5-Omni and Qwen3-Omni configuration on 178 undulating-positive questions reported in Table 8. Overall accuracy treats invalid outputs as incorrect, whereas valid-only accuracy removes them from the denominator. This heterogeneous and generally moderate performance does

Table 7: Answer-label concentration and CantoCore alignment.

Observation	Count	Proportion (%)
Correct option contains `Undulating`	178/191	93.19
CantoCore contains only `Undulating`	95/178	53.37
CantoCore is multi-coded and includes `Undulating`	83/178	46.63
CantoCore does not include `Undulating`	0/178	0.00

Table 8: Qwen2.5-Omni and Qwen3-Omni results on 178 Undulating-positive questions. "Valid" denotes valid response count ($N = 178$). Accuracy is reported as Overall/Valid-only (%) for Full response, and Overall (%) for Options only. A dash denotes no evaluated setting.

Qwen2.5-Omni

	Full response		Options only
Configuration	Valid	Acc (%)	Acc (%)
Original	176	44.38 / 44.89	48.31
Default SFT	163	41.01 / 44.79	42.13
Lang. Loss	166	44.38 / 47.59	42.70
Text DPO	168	53.93 / 57.14	42.13
Audio DPO	161	42.70 / 47.20	39.89
LR Phase-a	158	45.51 / 51.27	–
LR Phase-b	178	54.49 / 54.49	44.38

Qwen3-Omni

	Full response		Options only
Configuration	Valid	Acc (%)	Acc (%)
Original	154	57.87 / 66.88	65.17
Default SFT	158	47.19 / 53.16	49.44
Lang. Loss	158	46.63 / 52.53	47.75
Text DPO	160	46.63 / 51.88	47.75
Enc.-side REPA	159	48.88 / 54.72	50.56
Enc.-side REPA++	158	43.26 / 48.73	49.44

not exhibit a uniformly high accuracy that would support a claim of systematic shortcut use. The more direct finding of this case study is the answer regularity itself and what it reveals about the construction pipeline.

C.1.3 Examining the automatic CantoCore parser

The CantoCore specification treats melodic contour as a phrase-level qualitative character and explicitly requires annotator discretion. Temporary direction changes that do not materially alter the dominant contour—particularly changes lasting only one or two notes—should be ignored; otherwise, the specification warns, too many contours will be classified as *undulating*, reducing the character's informativeness.

Our CantoCore parser implementation provides a deterministic operationalization of this guidance. The parser reads the human-transcribed XML score, segments phrases at rests or breath marks, and converts every remaining interval into an ascending or descending sign. A phrase is labeled `Undulating` whenever this sign sequence changes direction more than once; the recording-level output is the set union of all phrase labels. The following examples illustrate this rule:

$$\boxed{C \nearrow D \searrow C \nearrow D} \implies (+,-,+) \implies \texttt{Undulating},$$

$$\boxed{C \nearrow E \searrow D \searrow B \nearrow C} \implies (+,-,-,+) \implies \texttt{Undulating}.$$

This operationalization is reproducible but has four limitations. First, reversals caused by ornaments, passing tones, or phrase-final motion may satisfy the symbolic rule without shaping the perceived principal contour. Second, one qualifying phrase adds `Undulating` regardless of its duration or frequency, so the recording-level union establishes presence rather than predominance. Third, even human-produced scores can simplify glides, microtonal motion, and culturally specific vocal inflections. Finally, rest- and breath-based boundaries need not coincide with perceived musical syntax, and different segmentation can change the number of within-phrase reversals.

Thus, complete parser alignment for the 178 undulating-positive recordings demonstrates consistency with the automatic extraction rule, not universal agreement with listeners' holistic perception. This is a trade-off for scalable, reproducible annotation and motivates targeted perceptual validation. On the other hand, fully replacing this stage with human annotation under the CantoCore protocol would not remove all interpretive limitations, which will be discussed later, and would require considerably more time and labor. The relevant goal is therefore to understand and validate this trade-off, rather than to treat either automatic extraction or exhaustive human annotation as an unqualified reference.

C.1.4 Granularity considerations at the Gemini question-generation stage

Gemini faithfully transfers the supplied CantoCore label or complete multi-label set. The limitation instead arises from scope: the unordered recording-level set indicates that a contour occurs in at least one parser-defined phrase, but contains no location, duration, frequency, or prominence information. When Gemini converts this representation into a natural-language multiple-choice question, however, the wording often refers to the contour of the piece, song, vocal line, or main theme, and may use expressions such as *overall*, *primary*, or *predominant*. This can therefore expand local presence into a claim of global characterization.

Exact transfer guarantees traceability, but not alignment between question scope and evidence scope. A listener may recognize a local undulating phrase while perceiving another contour as dominant, so accuracy can conflate holistic audio judgment with recovery of a score-derived label. Global wording should therefore, if possible, be supported by richer phrase-level evidence or targeted perceptual review; otherwise, presence-based wording is more appropriate.

C.1.5 Examining CantoCore as a human reference

CantoCore provides a common vocabulary for structured comparison while leaving room for subjective perceptual interpretation: its reported mean inter-rater reliability is $\kappa = 0.47$, and $\kappa = 0.37$ for melodic contour. Phrase grouping, multi-coding, and prominence therefore require judgment even under manual annotation, which is also substantially more costly. This perspective also clarifies our design choice. Score-based automation improves scalability, consistency, and traceability relative to exhaustive feature-by-feature human coding, while the human-produced transcription preserves expert musical evidence.

C.2 Qualitative Analysis of Chain-of-Thought Alignment

In this section, we present twelve representative items (organized into six comparative pairs) by comparing the model's responses before and after post-training across six behavioral transition categories. For instance, **Example 1** and **2** show the original and SFT Qwen3-Omni's response with wrong and correct answer, respectively, but both contain the correct mention. We paint the correct answer with CoT contents in green, and the wrong ones in red.

Wrong w/ mention → Correct w/ mention.

Example 1. 这首歌曲中的"衬词"在艺术表现上起到了什么主要作用?
A. 以方言吟唱方式营造怀旧氛围，作为背景人声装饰音色
B. 通过方言乡音烘托正词，并以密集的句式展现轻松欢快的情趣
C. 复杂韵律创造的独特音响效果，凸显了方言衬词的表现力
D. 自由变化的句式结构展现了民间音乐的即兴活力
Correct answer: B.
Song title: 龙船调(Dragon Boat Tune)

Original (Qwen3-Omni original). Predicted: C.
Option B is a strong contender: the chorus's energetic, rhythmic interjections punctuate the main lyrical lines, and the mood is lively. Option C more directly addresses the artistic function of the fillers themselves—their rhythmic complexity and unique sonic effect. Option B's "烘托正词" is a secondary effect. Therefore C is the most accurate answer.

Post-trained (Qwen3-Omni SFT). Predicted: B.
衬词以密集、重复、节奏鲜明的方言短语构成，语义空转而成为情绪载体。选项A中衬词并非背景；选项C强调复杂韵律，但衬词结构简单重复；选项D的自由变化与高度固定的衬词句式不符。选项B指出通过方言乡音烘托正词，并以密集的句式展现轻松欢快的情趣，与听觉特征吻合。

Example 2. Based on the vocal timbre and melodic ornaments, which regional style of Korean Minyo does this belong to?
A. The deep, husky, and dramatic Namdo style
B. The bright, lyrical, and agile Gyeonggi style
C. The nasal and melancholic Seodo style
D. The simple and robust Dongbu style
Correct answer: B.
Song title: 발자랑손자랑(Boasting of Hands and Feet)

Original (Qwen3-Omni original). Predicted: D.

The vocal timbre is bright, clear, and high-pitched. Ornamentation is minimal, and the style is straightforward and energetic, characteristic of the Dongbu (East Coast) style. Namdo is deep and dramatic, Gyeonggi is lyrical and agile, and Seodo is nasal and melancholic. Therefore D.

Post-trained (Qwen3-Omni SFT). Predicted: B.

The vocal timbre is bright, clear, and slightly nasal—a hallmark of the Gyeonggi Minyo style, known for lyrical, agile, ornamented delivery. Melismatic runs and trills confirm this. Namdo is deep and dramatic; Seodo is nasal and melancholic but lacks the lyrical flow; Dongbu is simple and robust. The only option that aligns with the auditory evidence is B.

Correct w/ mention → Wrong w/ mention.

Example 3. Which combination of instruments reflects the unique orchestration style of this piece?

A. a qanun and frame drum combined with a full symphony orchestra
B. Qawwali singer and harmonium combined with a minimalist piano, prepared guitar, and looping electronic beats
C. Traditional Arabic instruments combined with a Western string section, electric guitar, and organ
D. A renowned female vocalist performing traditional Arabic qawwali with a large electronic dance music drop synthesized entirely

Correct answer: C.
Song title: Qareat El Fengan (The Coffee Cup Reader)
Original (Qwen2.5-Omni original). Predicted: C.

The unique orchestration style is best reflected by the combination of traditional Arabic instruments with a Western string section, electric guitar, and organ. This blend creates a fusion of Eastern and Western musical elements characteristic of the piece.

Post-trained (Qwen2.5-Omni SFT). Predicted: B.

The audio begins with a cinematic orchestral introduction featuring strings, brass, and percussion, then a male vocalist singing in Arabic. The combination of Western orchestral techniques with Arabic lyrical content creates a hybrid form. The presence of a full orchestra and cinematic reverb aligns with a modern Arabic pop ballad fusing traditional emotional expression with contemporary production.

Example 4. What is the primary narrative theme conveyed by the lyrics and melody of this piece?

A. A tale exploring metaphors of nature's reclaiming power against human encroachment
B. The nostalgic reconstruction of a peaceful, paradisiacal rural childhood, contrasting it with the perceived corruption and complexities of contemporary urban life
C. A personal reflection on the homeland as a sacred heirloom and a nurturing mother from birth to death
D. A tale of creation grounded in nature imagery, conveyed through lush woodwind interludes and rhythmic storytelling motifs

Correct answer: C.
Song title: Indonesia Pusaka

Original (Qwen2.5-Omni original). Predicted: C.

The lyrics and melody convey a personal reflection on the homeland as a sacred heirloom and a nurturing mother from birth to death. The imagery of Indonesia as a motherland, with references to its natural beauty and the protection it offers, suggests a deep emotional connection to the homeland.

Post-trained (Qwen2.5-Omni SFT). Predicted: B.

The lyrics, sung in Indonesian, frame Indonesia as a maternal figure and a sacred homeland. The core message is deeply personal and devotional: the homeland as a nurturing mother, worthy of worship and sacrifice. This aligns precisely with option B, which captures the nostalgic, sacred contrast between an idealized past and present devotion to the nation.

Wrong w/o mention → Correct w/o mention.

Example 5. Based on the rhythmic organization, how is the meter of this piece classified?

A. Rhythmic symmetry
B. Iso-metric
C. Non-metric
D. Compound Hemiola

Correct answer: B.
Song title: 尕 老汉 (The Little Old Man)

Original (Qwen2.5-Omni original). Predicted: D.

Based on the rhythmic organization, the meter of this piece is classified as:

Post-trained (Qwen2.5-Omni SFT). Predicted: B.

The audio is a monophonic, unaccompanied vocal performance with a steady, unvarying rhythmic pulse. The vocal line exhibits a consistent, non-syncopated beat that repeats without deviation. This is not free-form, nor does it exhibit polyrhythmic layering or metric modulation. The other options—rhythmic symmetry, non-metric, or compound hemiola—are contradicted by the unyielding single-meter pulse.

Example 6. Which of the following best describes the formal structure of each stanza?

A. Strophic form with octaves (eight-line stanzas)
B. Strophic form with repeated couplets (AABB)
C. Parnassian form with balanced stanzas and enjambment (ABAB) Ballad form with verses and a chorus (AABA)
D. Verse-refrain form with the verse following AABB and the refrain being repeated (AABB / R)

Correct answer: B.
Song title: Ach synku, synku (Oh, My Son, My Son)

Original (Qwen3-Omni original). Predicted: D.

Each stanza has a verse followed by a refrain. The verse is a repeating two-line pattern (AABB); the refrain is also AABB. This matches option D: verse-refrain form with AABB verse and a repeated refrain. Option B describes a single structure without a distinct refrain.

Post-trained (Qwen3-Omni SFT). Predicted: B.

The lyrics have four distinct lines per stanza, each repeated verbatim (XX YY ZZ WW). This is not eight-line octaves, nor alternating rhyme, nor a ballad with a chorus. Each line is doubled into pairs. Option D implies a chorus that is not present.

Correct w/o mention → Wrong w/o mention.

Example 7. 在这首民歌的扩充段落中，南方汉族民间对这种排比性的短句称为什么？

A. 散句
B. 垛字
C. 赶板（推进节奏的重复）
D. 赶句

Correct answer: D.
Song title: 郎在对门唱山歌*(The Lad Across the Way Sings a Mountain Song)*

Original (Qwen3-Omni original). Predicted: D.

The same phrase is repeated multiple times in quick succession, a technique used to build energy. This driving, repetitive structure is a form of rhythmic repetition that propels the music forward. “垛字” refers to stacking syllables; “散句”means scattered free-form phrases. Option D is the most accurate answer.

Post-trained (Qwen3-Omni SFT). Predicted: C.

扩充段落中短句机械性重复，语义不递进，仅通过节奏与音节叠加制造律动。这种结构在南方汉族民间音乐中被称作赶板，即通过重复短句推动节奏。选项A散句为无规律自由句式；选项B垛字多用于叠词或衬词；选项D并非专业术语、无学术依据。赶板是唯一符合音乐学定义的术语。

Example 8. Which instrument provides the primary rhythmic accompaniment in this audio?

A. Janggu (hourglass drum) provides the primary rhythmic accompaniment because it is a percussion instrument often used for keeping time in the same style
B. The primary rhythm is provided by the Gayageum, playing syncopated rhythmic patterns alongside the drums, but primarily shaping the rhythm structurally
C. Janggu provides the primary rhythm because Buk, being a larger barrel drum, is typically reserved for bass sounds and less intricate rhythms in this context
D. Buk (barrel drum)

Correct answer: D.
Song title: 새타령(Bird Song)

Original (Qwen3-Omni original). Predicted: D.

The audio features a prominent, deep, resonant drum that provides a steady, foundational beat. The hourglass drum is typically higher-pitched and used for more intricate, syncopated rhythms. In this piece the deep, grounding beat is the primary rhythmic foundation. The plucked zither does not provide the primary rhythmic accompaniment.

Post-trained (Qwen3-Omni SFT). Predicted: A.

Two percussion instruments are present. One produces a steady, syncopated rhythm and is treated as the primary rhythmic driver; the other provides deep, resonant beats that punctuate the rhythm in a

supportive role. The plucked zither provides melodic and harmonic support, not the primary rhythm. Option D is incomplete and does not name the correct instrument. The only accurate choice is A.

Wrong w/o mention → Wrong w/ mention.

Example 9. Which description best characterizes the melodic range and contour of this song?

A. Upward contour with a broad melodic range
B. Moderate melodic range with a jagged, fluctuating contour
C. Small melodic range with an undulating contour
D. Medium melodic range with a stepwise, mostly stable contour (like a low range with small, steady steps up and down a few times)

Correct answer: C.
Song title: Holka modrooká (Blue-Eyed Girl)

Original (Qwen2.5-Omni original). Predicted: D.

The song's melody is characterized by a broad range and an upward contour. The lyrics are in Czech, and the music is in the key of B major. The song is a traditional folk song from the Czech Republic.

Post-trained (Qwen2.5-Omni phase-b). Predicted: D.

A solo accordion plays a lively dance-like melody built from repeated stepwise motifs with minimal leaps. The contour is not dramatic or wide, but rather undulating, with gentle rises and falls. This is not expansive, not jagged, and not undulating in a flowing sense, but grounded and repetitive. Option D is the only one that reflects this restraint; A, B, and C imply larger scales or broader range.

Example 10. What is the central theme of the song's lyrics?

A. Lyrics suggesting a comfortable, established rhythm to a long-standing relationship
B. De-escalating a conflict over a misunderstanding
C. Nostalgically recalling simpler, happier times in a forgotten town
D. Praising the beauty and endurance of true, understanding love

Correct answer: B.
Song title: Sina Makosa (I Am Not To Blame)
Original (Qwen3-Omni original). Predicted: D.

The lyrics repeatedly stress an inseparable bond and a plea for unity, pointing to a celebration of profound and enduring love. The other options do not align with the lyrical content. Therefore D.

Post-trained (Qwen3-Omni Enc.-side REPA). Predicted: D.

The refrain is a declaration of inseparable unity, not a plea for reconciliation or a memory of the past. The line is not about resolving conflict, but affirming identity through union. Only D captures praise of love as an enduring, understanding union. The other options introduce conflict (B), nostalgia (C), or comfort as a settled state (A), which are absent from the text.

Wrong w/ mention → Wrong w/ mention.

Example 11. What is the typical tempo and character of this Spanish children's song?

A. Fast and lively (Allegretto)
B. Moderate and peaceful (Andante tranquillo)
C. Cheerful yet slightly waltzing (Andante grazioso)
D. Playful and skipping along (Allegro vivace)

Correct answer: A.
Song title: La Tarara

Original (Qwen2.5-Omni original). Predicted: D.

La Tarara is a lively and upbeat Spanish children's song with a cheerful melody. The tempo is fast and energetic, and the character is playful and skipping along, making it a fun listen for children.

Post-trained (Qwen2.5-Omni phase-b). Predicted: D.

The tempo is fast and propulsive, with a playful, almost skipping cadence, especially in the refrain. Among the options, Allegro vivace (D) matches both the fast tempo and the playful, skipping character. The others are too slow (Allegretto), too calm (Andante tranquillo), or too waltz-like (grazioso).

Example 12. Based on the instrumental arrangement, which group of instruments defines the 'folk-rock' transition in this version?

A. A double bass and mandolin duo joined by a distorted lead guitar and a driving drum pattern with heavy use of cymbals
B. A combination of traditional folk instruments like bagpipes and violin with rock elements like drums and electric guitars
C. An acoustic blues harmonica player improvising with a resonator guitar transitioning into a groove led by a fuzz bass, distortion pedals, and a pounding kick drum

D. A folk choir singing a cappella, accompanied by a six-string bass and minimal percussion, transitioning into a heavily distorted guitar riff

Correct answer: B.
Song title: Mezi horami (Between the Mountains)

Original (Qwen3-Omni original). Predicted: A.

The audio begins with violin and acoustic guitar, then adds a distorted electric guitar and a more driving rock rhythm section. This fusion of traditional folk instruments with rock elements is folk-rock. Option A describes that combination. Option B mentions bagpipes, which are not present. Therefore A.

Post-trained (Qwen3-Omni audio REPA). Predicted: A.

A mournful fiddle and acoustic guitar give way to a driving drum kit, fuzz bass, and distorted lead electric guitar. Option A is taken to identify that transition. Option B incorrectly includes bagpipes, which are not mentioned. Option C misidentifies a harmonica; option D describes an a cappella choir. Only A is treated as matching the instrumentation.

From the examples above, we observe several systematic strengths and limitations of the post-trained model:

- **Domain Knowledge Reasoning: Examples 1 and 2** focus on musicological knowledge, where the former addresses overall aesthetic comprehension and the latter evaluates song-level musical features. Post-training significantly enhances the model's capacity for reasoning over fundamental music theory and acoustic concepts.

- **Semantic–Label Mismatch:** Conversely, **Examples 3 and 4** illustrate a critical post-training failure mode: *answer–option mismatch*, wherein the model reasons correctly but selects the wrong option key. This issue arises when autoregressive language models decouple semantic inference from label mapping due to noisy training signals, attention decay, asymmetric loss weighting, and RLHF-induced format bias—collectively allowing content correctness to diverge from symbolic accuracy Wang et al. (2024b; 2026); Arcuschin et al. (2025); Turpin et al. (2023).

- **Robustness in Reasoning Format:** Furthermore, **Examples 5 and 6** demonstrate that the post-trained model gains stronger reasoning capability even without explicitly repeating the option text during inference.

- **Deficits in Low-Resource & Localized Elements:** Nevertheless, as shown in **Example 7** (where regional concepts like *Ganju* and *Ganban* are conflated) and **Example 12** (where iconic timbre like the *bagpipe* is missed), a lack of fine-grained knowledge of low-resource musical elements hinders deep and culturally diverse music understanding.

- **Superficial Grounding vs. Holistic Comprehension:** Finally, **Examples 9–12** demand complex, intrinsic comprehension of the entire piece. Over-reliance on local surface features (e.g., localized audio grounding or prominent lyrics) exposes the model's fundamental weakness in abstracting higher-level musical structures.

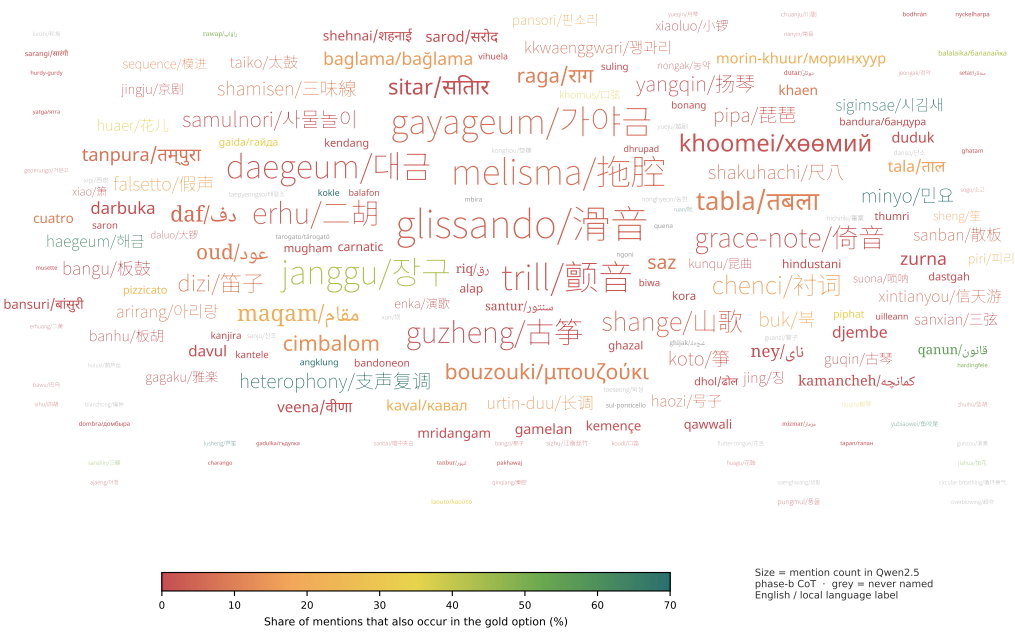


Figure 9: Named non-Western instruments, techniques, and genres in Qwen2.5-Omni phase-b thinking traces: type size is mention count, and colour is the share of those mentions that also occur in the gold option (grey: never named)