Diagnosing Modality Confounds in Cross-Lingual Multimodal Retrieval for Formosan Languages

Anonymous ACL submission

Abstract

Multimodal corpora for low-resource languages are increasingly available, yet corpus-query systems remain largely limited to keyword search and metadata filtering. We present a diagnostic study of cross-lingual multimodal retrieval of interactional-pragmatic events on a pilot corpus of 30 Taiwan Indigenous Television (TITV) news clips spanning three Formosan languages: Amis, Paiwan, and Truku. We extract four complementary modalities: semantic text embeddings from on-screen Chinese subtitles using XLM-RoBERTa, acoustic representations (MFCC baselines with Whisper-encoder ablation), prosodic statistics derived from phone-level segmentation, and body-kinematic features from pose estimation. We evaluate all single- and multi-modal fusion combinations under query-by-example retrieval. Pragmatic-event labels are generated using a vision-language model annotator and validated against two human annotators ($\kappa_{\text{inter}} = 0.747$; $\kappa_{\text{avg(VLM)}} = 0.765$). Across modalities, text embeddings achieve the strongest overall retrieval performance ($P@1 = 0.633$), substantially outperforming TF-IDF baselines. Prosodic and kinematic features also exceed majority-class baselines, suggesting that interactional-pragmatic information is partially recoverable beyond lexical semantics. Most notably, body-kinematic features show candidate cross-lingual signal for Amis queries ($P@1 = 0.500$), matching text-based retrieval despite entirely different spoken languages. A central finding is that apparent modality failure is partly representation-dependent rather than modality-intrinsic. MFCC-based audio fusion consistently degrades retrieval performance, but replacing MFCC statistics with Whisper-encoder hidden states partially recovers cross-modal retrieval quality and reduces evidence of language-identity confounds. At the same time, audio fusion continues to underperform audio-free combinations, suggesting that both feature representation and unlearned fusion contribute to retrieval failure. Rather than proposing a state-of-the-art retrieval architecture, this work aims to diagnose how modality representations behave under low-resource cross-lingual multimodal retrieval. The resulting pipeline provides a reproducible diagnostic foundation for future utterance-level and large-scale multimodal pragmatics research on Formosan languages.

1 Introduction

Corpus linguistics has undergone a decisive shift from *symbolic* to *geometric* query paradigms. Classic concordancers such as CQP (Evert and Hardie, 2011) match formal patterns; modern dense-retrieval systems (Karpukhin et al., 2020; Reimers and Gurevych, 2019) embed queries and documents into a shared vector space where semantic similarity is measured by cosine distance. Text corpora have benefited enormously from this transition. Multimodal corpora, however, remain largely stranded in the symbolic era, where the dominant retrieval interface is still metadata filtering or keyword search. This gap is consequential for pragmatic research. Pragmatic phenomena—comforting a grieving interlocutor, urging collective action, explaining a cultural practice—are inherently multimodal: words alone may be ambiguous, but prosodic contour, gesture dynamics, and body posture together constitute the communicative act (Austin, 1962; Kendon, 2004; McNeill, 1992). Encoding such events in a unified vector space would allow a researcher to query "celebrating" and retrieve relevant instances

across modalities and languages.

We pursue this in the context of Taiwanese indigenous language documentation. Taiwan Indigenous Television (TITV) broadcasts daily news in at least 16 Formosan languages, producing a naturalistic corpus of oral-visual interaction across typologically diverse languages; each clip pairs indigenous-language speech with on-screen Chinese subtitles, providing a built-in parallel structure across thousands of hours of discourse. This resource has been explored for speech representation learning and phylogenetic inference (Hsieh et al., 2024), but not yet for pragmatic retrieval. A natural question is whether vision-language models could simply be prompted to fill this gap. State-of-the-art VLMs can describe visual content and read Chinese subtitles accurately (Appendix A), but they cannot identify Formosan-language speech, distinguish indigenous groups from visual cues, expose per-modality contributions, or support corpus-level retrieval. VLMs are description systems, not retrieval systems; embedding-based retrieval is complementary, enabling search-by-example, per-modality ablation, and corpus-level pattern discovery in ways that prompting cannot.

Building on the MultiMoco project's multimodal corpus infrastructure for Taiwan's languages, we reframe the challenge from corpus *construction* to corpus *query*. With 30 clips and four interactional-pragmatic categories, our goal is not to build a state-of-the-art retrieval system, but to diagnose how different modality representations contribute to cross-lingual multimodal retrieval under low-resource conditions. In particular, we ask whether apparent modality failure reflects intrinsic limits of the modality itself or confounds introduced by feature representation and fusion design. To this end, we examine which modalities carry pragmatic signal, which fusion combinations help versus hurt, and where apparent failures trace back to feature choices rather than modality limits. Our central ablation, comparing MFCC global statistics against Whisper-encoder representations, illustrates this concern directly: what initially appears to be "modality failure for interactional-pragmatic events retrieval" turns out to be specific to the feature representation, not to the modality itself.

Our contributions are as follows:

1. A **four-modality feature pipeline** (text, acoustic, prosodic, kinematic) with a BoW/TF-IDF reference baseline, evaluated via query-by-example retrieval with bootstrap 95% CI.
2. A **VLM-as-annotator** pipeline for silver interactional-pragmatic labels, validated against two human annotators ($\kappa_{\text{inter}} = 0.747$, $\kappa_{\text{avg(VLM)}} = 0.765$).
3. The **first cross-lingual multimodal interactional-pragmatic retrieval baseline** on Formosan languages, with ablations identifying a language-identity confound in MFCC features and candidate cross-lingual signal in body-kinematic features.

2 Related Work

Large-scale multimodal corpora such as IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Zadeh et al., 2018) have driven progress in emotion recognition, but both target acted or scripted speech rather than naturalistic interaction, and neither addresses pragmatic function in multilingual settings. Retrieval-oriented multimodal work has similarly focused on video-text alignment (Miech et al., 2019) and audio-visual correspondence (Arandjelović and Zisserman, 2017), not interactional pragmatics. Karpukhin et al. (2020) show that a dual-encoder framework can replace sparse BM25 retrieval with dense vector search; our query-by-example evaluation is a direct instantiation of this approach, using mean-pooled XLM-RoBERTa vectors (Conneau et al., 2020) whose cosine distance tracks semantic similarity by design (Reimers and Gurevych, 2019).

Our label taxonomy draws on speech act theory (Austin, 1962; Searle, 1969; Levinson, 1983). Kendon (2004) and McNeill (1992) establish the integral role of gesture in interactional-pragmatic meaning, motivating our inclusion of kinematic features; Kita and Ide (2007) further suggest that gestural pragmatic functions are partially but not fully shared across languages, supporting the cross-lingual hypothesis we test here. The motivation for treating body-kinematic fea-

tures as a pragmatic signal comes partly from Kucherenko et al. (2020), who show that gesture timing and semantic content can be jointly learned from audio and text, supporting the converse inference that kinematic features may carry recoverable pragmatic information.

Our kinematic pipeline builds on DWPose (Yang et al., 2023), which refines hand and facial keypoint precision over OpenPose (Cao et al., 2019) via two-stage distillation; Allosaurus (Li et al., 2020) enables language-agnostic phone-level segmentation for prosodic feature extraction. For Formosan languages, the NTU Corpus (Su et al., 2008) provides transcribed spoken data; the MultiMoco corpus (Wang et al., 2023) and Hsieh et al. (2024) establish the TITV-based multimodal infrastructure our work extends from corpus *construction* to corpus *query*. Within low-resource speech technology, Pratap et al. (2024) scale speech models to 1,400+ languages via weakly supervised training—-their features are a natural candidate for replacing MFCC statistics in future work—and Shi et al. (2023) find that recording-condition variation disrupts cross-lingual representations in ML-SUPERB, a confound relevant to any low-resource multimodal setting. Recent work shows that LLMs can match crowd-worker annotation quality (Gilardi et al., 2023; Alizadeh et al., 2023); we extend this to multimodal interactional-pragmatic events annotation via a vision-language model.

3 Data, Annotation, and Features

3.1 Corpus

We work with 30 broadcast clips from TITV's programme 《看見族語》 (*Seeing the Indigenous Language*), a daily news show produced in multiple Formosan languages with on-screen Chinese subtitles. Our pilot sample covers three languages (see Appendix B): Amis (10 clips, 2026-03-02), Paiwan (10 clips, 2026-03-07), and Truku (10 clips, 2026-03-05), all Austronesian / Formosan. Each clip pairs indigenous-language speech with OCR-detectable Chinese subtitles; pre-processed outputs include OCR-corrected subtitle JSON, phone-level TextGrid alignments (Allosaurus), and per-frame pose estimation. In this pilot we treat each clip as one *document-level event*; utterance-level segmentation via the aligned TextGrids is deferred to future work. The available pre-processed data streams per clip are shown in Appendix C.

3.2 Pragmatic Event Labels

We adopt a four-category taxonomy derived from speech act theory, selected for (a) observable multimodal correlates, (b) natural occurrence in news discourse, and (c) cross-linguistic comparability: Explaining (conveying information or demonstrating cultural practice), Urging (advocacy or policy pressure), Celebrating (marking an achievement or cultural event), and Reporting (neutral broadcast; default when no dominant function applies). A fifth category, Comforting, was included in our initial taxonomy but absent from the sampled clips.

An initial attempt using keyword-based heuristics on the Chinese subtitles proved unreliable: Urging medical-policy segments were misclassified as Explaining due to shared health-related vocabulary—motivating a **VLM annotation pipeline** using Claude Haiku (`claude-haiku-4-5-20251001`) (Anthropic, 2025): five uniformly-spaced keyframes and an aggregated OCR subtitle summary (confidence $\geq$ 0.8, non-logo only) are submitted per clip with a structured prompt (Appendix D) requesting JSON output `label, confidence, rationale`. All 30 clips received `confidence: high`; label distribution is shown in Table 1.

The **random-baseline** $P@1$ under this distribution is $\sum_c p_c^2 \approx 0.47^2 + 0.27^2 + 0.17^2 + 0.10^2 \approx 0.33$. Since Explaining dominates the distribution, we also report a stricter **majority-class baseline**: always predicting Explaining achieves $P@1 \approx 0.47$ under leave-one-out evaluation (when an Explaining query is held out, 13 of the remaining 29 clips share that label).

Human validation. Two annotators independently labeled all 30 clips using the same conditions as the automated annotator. Inter-annotator agreement reached $\kappa = 0.747$ (substantial), and average human–VLM agreement was $\kappa_{\text{avg}} = 0.765$ (substantial), supporting the

reliability of the silver labels. Two clips where both annotators agreed on a label differing from the VLM assignment were corrected accordingly.

3.3 Feature Extraction

Text (768 dimensions). Chinese subtitle text is aggregated per clip from the OCR stream (duplicates and strings shorter than two characters or labelled logo are discarded), then encoded with **XLM-RoBERTa base** (Conneau et al., 2020) via mean-pooling over final hidden states. As a non-neural reference baseline, we additionally evaluate **TF-IDF character n-grams** (1–3 grams, sublinear TF weighting) on the same subtitle text.

Audio (86 / 1280 / 2560 dimensions). As the primary acoustic baseline we extract global statistics over five feature sets via `librosa` (McFee et al., 2015): 20 MFCC coefficients, first-order delta, RMS energy, zero-crossing rate, and spectral centroid (86 dims). To assess whether audio's poor retrieval performance reflects a modality-level property or a feature-level confound, we additionally extract encoder hidden states from Whisper large-v3 (Radford et al., 2023) (30 s windows, 5 s overlap), evaluating mean-pool (1,280 dims) and mean+std-pool (2,560 dims) variants. Unlike MFCC statistics, Whisper's encoder abstracts substantially from language-specific phonetics. Results are compared as an ablation in Table 6.

Prosody (8 dimensions). Phone-level annotations are obtained from Allosaurus TextGrids (Li et al., 2020); `<blk>` intervals are treated as silence. Since Allosaurus was not evaluated on Formosan languages, derived features, particularly **phone_entropy** and **avg_phone_dur**, may carry systematic recognition errors. We compute eight clip-level statistics: speech rate, silence ratio, mean and std phone duration, phone entropy (Shannon entropy over phone labels), burst rate (silence-to-speech transitions per second), longest silence interval, and unique phone count.

Kinematic (18 dimensions). Per-frame pose estimation uses YOLO-based detection followed by DWPose (Yang et al., 2023). Since keypoint count varies across frames due to occlusion, we apply centroid-based aggregation over all visible keypoints per speaker rather than fixed-index statistics; absent speakers default to neutral values. Per speaker we compute body centroid mean and std (x, y), hand presence ratios, and face centroid std—16 dims total—plus two global aggregates (mean body movement energy and hand asymmetry), yielding 18 dims; details are shown in Appendix E.

3.4 Fusion and Evaluation Protocol

Each modality vector is ℓ_2-normalised; multimodal combinations concatenate and renormalise. We intentionally use unweighted concatenation rather than learned fusion to maximise diagnostic transparency: any retrieval change is attributable directly to feature properties rather than fusion model interactions. We evaluate all $2^4 - 1 = 15$ non-empty subsets; the Whisper ablation is reported separately in §4.1.

Query-by-example retrieval. Each clip serves in turn as the query against the remaining 29 clips. Metrics are $P@k$ (fraction of top-k results sharing the query label) and MAP (mean average precision over the full ranking).

Cross-lingual retrieval. For each query clip, candidates are restricted to clips from the other two languages; $P@k$ is computed on this restricted ranking.

4 Diagnostic Experiments

Table 2 reports retrieval performance for each modality in isolation.

All configurations exceed the random baseline for $P@1$. **Text** achieves the highest $P@1$ (0.633, +0.30 above random and +0.17 above the majority-class baseline), confirming that XLM-RoBERTa subtitle embeddings capture pragmatic-event distinctions beyond lexical overlap. As a non-neural reference, the **BoW (TF-IDF)** baseline achieves $P@1 = 0.467$, roughly equal to the majority-class baseline; the gap of 0.167 points over BoW demonstrates that neural semantic embeddings provide genuine added value. **Prosody** is the strongest non-text modality (0.533), followed by Kinematic (0.500) and Audio-MFCC (0.433); both Prosody and Kinematic exceed the majority-class baseline, indicating that

Label	Amis	Paiwan	Truku	Total	Proportion
Explaining	7	5	2	14	47%
Urging	0	3	5	8	27%
Celebrating	3	1	1	5	17%
Reporting	0	1	2	3	10%
Total	10	10	10	30	

Table 1: Label distribution after VLM annotation and human validation.

	P@1 [95% CI]	P@3	MAP	Δ vs. random
Text (OCR subtitles)				
BoW/TF-IDF†	0.467 [0.30,0.63]	0.489	0.506	+0.14
XLM-RoBERTa	0.633 [0.47,0.80]	0.522	0.494	+0.30
Audio				
MFCC	0.433 [0.27,0.63]	0.244	0.351	+0.10
Whisper-encoder	0.400 [0.23,0.57]	0.444	0.429	+0.07
Prosody	0.533 [0.37,0.70]	0.456	0.455	+0.20
Kinematic	0.500 [0.33,0.67]	0.456	0.457	+0.17
Majority-class baseline	≈0.47	—	—	—
Random baseline	≈0.33	—	—	—

Bootstrap 95% CI, $n = 2000$ resamples. All CI bands overlap; rankings are tendencies, not confirmed orderings.

Table 2: Single-modality retrieval performance (30 clips, 4 modalities, 6 configurations).

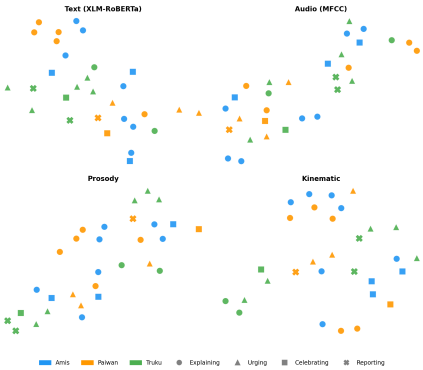

Figure 1: t-SNE projections of the four modality embeddings, coloured by language and shaped by pragmatic-event label. Text embeddings form the most coherent label clusters, while audio embeddings show strong language-based separation.

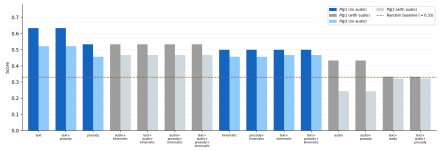

Figure 2: P@1 and P@3 for all 15 modality combinations, sorted by P@1. `audio` = MFCC (86d). The dashed horizontal line marks the random baseline (≈0.33).

these features carry interactional-pragmatic signal independent of label frequency.

Multi-modal fusion. Table 3 summarises all 15 combination scores.

Three patterns emerge:

(1) Text + Prosody is the best combination. Adding prosodic statistics to text leaves P@1 unchanged but marginally improves MAP (0.494 → 0.495), suggesting prosody contributes complementary ranking signal without diluting text semantics.

(2) Audio consistently hurts. Any combination containing `audio` yields lower P@1 than the same combination without it, most starkly for `text+audio` (P@1 = 0.333, well below the text-only baseline of 0.633). MFCC global statistics appear to encode language identity rather than interactional-pragmatic function, introducing misleading similarity signals. Table 6 shows that replacing MFCC with Whisper-encoder features partially, but not fully, addresses this pattern.

(3) Non-text modalities show limited fusion gains. `audio+kinematic` (P@1 = 0.533) marginally outperforms both `kinematic` alone (0.500) and `audio` alone (0.433), but the gain lies within bootstrap CI width and the combination still falls well below text (0.633).

Per-language analysis. Table 4 reports within-language retrieval for the text modal-

Modality Combination	P@1	P@3	MAP
`text`	**0.633**	**0.522**	0.494
`text+prosody`	**0.633**	**0.522**	**0.495**
`prosody`	0.533	0.456	0.455
`audio+kinematic`	0.533	0.467	0.453
`text+audio+kinematic`	0.533	0.467	0.454
`audio+prosody+kinematic`	0.533	0.467	0.453
`text+audio+prosody+kinematic`	0.533	0.467	0.454
`kinematic`	0.500	0.456	0.457
`prosody+kinematic`	0.500	0.456	0.457
`text+kinematic`	0.500	0.467	0.460
`text+prosody+kinematic`	0.500	0.467	0.460
`audio`	0.433	0.244	0.351
`audio+prosody`	0.433	0.244	0.351
`text+audio`	0.333	0.322	0.370
`text+audio+prosody`	0.333	0.322	0.370
Random baseline	≈0.33	—	—

Table 3: All 15 modality-combination retrieval scores, sorted by P@1. `audio` denotes MFCC-based features; see Table 6 for Whisper-encoder comparison.

Language	P@1	P@3	MAP
Amis	**0.600**	**0.700**	**0.690**
Paiwan	**0.600**	0.467	0.531
Truku	0.400	0.400	0.435

Table 4: Within-language retrieval (text modality)

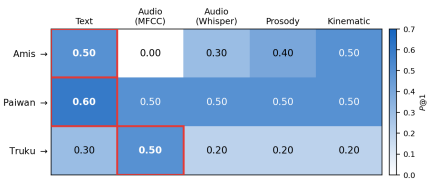

Figure 3: Cross-lingual P@1 heatmap: query language (rows) vs. modality (columns), with candidates restricted to the other two languages. The best cell per row is boxed.

ity (query and candidate pool restricted to the same language). Amis and Paiwan are tied for the highest within-language P@1 (0.600), while Truku scores lower (0.400). Amis benefits from 7 of its 10 clips sharing the Explaining label (cultural arts education content), forming a tight semantic cluster; Paiwan's performance reflects its diverse label mix (Urging and Explaining) that XLM-RoBERTa can distinguish via policy-specific vocabulary. Both effects reflect corpus composition rather than enhanced model capability.

Cross-lingual retrieval. Table 5 reports P@1 when queries are from one language and candidates are drawn exclusively from the other two.

The most notable finding is the **Amis-kinematic** result: kinematic matches text (P@1 = 0.500 each) for Amis cross-lingual queries, while audio collapses to near-chance under MFCC features and 0.300 under Whisper-encoder features. This parity suggests that body-movement dynamics generalise across language boundaries comparably to semantic text features for Amis content, where Explaining clips involve hands-on demonstration (weaving, plant-dyeing) with distinctive kinematic signatures that transfer to similar clips in other languages.

Paiwan shows the opposite pattern: text lead (P@1 = 0.600) reflects formulaic policy-advocacy language that XLM-RoBERTa maps cross-lingually; Truku's elevated MFCC cross-lingual score (P@1 = 0.500) drops to 0.200 with Whisper-encoder features, consistent with a recording-condition artefact.

4.1 Audio representation ablation

Motivation. The most counterintuitive finding in Table 3 is that any combination containing `audio` yields lower P@1 than the corresponding audio-free combination, with `text+audio` collapsing to P@1 = 0.333. Two interpretations are available. The *strong interpretation* holds that acoustic signal carries no recoverable interactional-pragmatic information beyond what text and kinematic features already encode. The *weak interpretation* holds that global MFCC statistics encode properties orthogonal to interactional-

Query Language	Text	Audio (MFCC)	Audio (Whisper)	Prosody	Kinematic
Amis → other	**0.500**	0.000	0.300	0.400	**0.500**
Paiwan → other	**0.600**	0.500	0.500	0.500	0.500
Truku → other	**0.300**	0.500	0.200	0.200	0.200

Table 5: Cross-lingual $P@1$ per source language and modality. Audio (Whisper) uses Whisper-meanstd pooling.

pragmatic function—most plausibly speaker identity, recording channel, and language identity—which act as confounds in cross-lingual retrieval. To quantify this confound, we compute the nearest-neighbour language purity of MFCC representations: 70% of clips (21/30) retrieve a same-language clip as their nearest neighbour, more than twice the random baseline of 33%. By contrast, MFCC nearest-neighbour pragmatic-label purity is only 43%, barely above the random baseline of 33%, confirming that MFCC similarity is driven by language (or recording-session) identity rather than interactional-pragmatic function. The two interpretations have opposite implications for future multimodal interactional-pragmatic work.

Method. We disambiguate by replacing MFCC global statistics with encoder hidden states from Whisper large-v3 (Radford et al., 2023), evaluating two pooling strategies: mean-pool (**Whisper-mean**, 1,280 dims) and mean + std-pool (**Whisper-meanstd**, 2,560 dims).

Results and interpretation. Three patterns emerge. **(1) Whisper does not improve single-modality P@1.** Under overlapping bootstrap CIs, neither Whisper variant clearly outperforms MFCC at the overall level, contrary to our pre-registered expectation. **(2) Audio fusion remains partially destructive but less so.** `text+audio_whisper_meanstd` recovers from $P@1 = 0.333$ (MFCC) to 0.433, consistent with the weak interpretation that part of the audio penalty is feature-specific. Fusion nonetheless still falls below the text-only baseline (0.633), indicating that the problem is not solely a MFCC artefact—unlearned fusion is also a factor. **(3) The Truku cross-lingual anomaly is resolved.** Truku's MFCC cross-lingual score (0.500) drops to 0.200 with Whisper features, supporting the hypothesis that the MFCC result reflected recording-condition similarity rather than genuine interactional-pragmatic generalisation. Amis cross-lingual $P@1$ improves from 0.000 (MFCC) to 0.300–0.400 (Whisper), further consistent with MFCC encoding language identity as a confound.

Taken together, the ablation supports the *weak interpretation*: MFCC global statistics encode a language-identity confound that explains part—but not all—of the audio penalty. Unlearned fusion and the small sample size are additional limiting factors that preclude a definitive conclusion about whether semantically-informed acoustic features could eventually contribute positively to interactional-pragmatic event retrieval.

5 Discussion

5.1 Modality contributions are representation-dependent

Under MFCC-based representations, text features tend to outperform other modalities ($P@1$: 0.633, 0.533, 0.500, 0.433 for text, prosody, kinematic, and audio respectively). However, the rankings should be read as tendencies rather than confirmed orderings. The ablation in §4.1 shows that the bottom position of audio is specific to the MFCC feature representation, not an intrinsic property of the modality. MFCC global statistics encode *language identity*, the three Formosan languages have distinct phonological inventories, which acts as a confound that overrides pragmatic similarity in cross-lingual retrieval.

Replacing MFCC with Whisper-encoder hidden states partially addresses this confound: `text+audio` fusion recovers from $P@1 = 0.333$ to 0.433, and Truku's cross-lingual score drops from 0.500 to 0.200. Audio fusion nonetheless continues to underperform audio-free combinations, indicating that unlearned concatenation is an additional lim-

Config	P@1 [95% CI]	P@3	MAP	Amis CL
`audio_mfcc`	0.433 [0.27, 0.63]	0.244	0.351	0.000
`audio_whisper_mean`	0.367 [0.20, 0.53]	0.422	0.427	0.400
`audio_whisper_meanstd`	0.400 [0.23, 0.57]	0.444	0.429	0.300
`text+audio_mfcc`	0.333 [0.17, 0.50]	0.322	0.370	—
`text+audio_whisper_meanstd`	0.433 [0.27, 0.60]	0.444	0.437	—
`kinematic+audio_mfcc`	0.533 [0.37, 0.70]	0.467	0.453	—
`kinematic+audio_whisper_meanstd`	0.467 [0.30, 0.63]	0.433	0.459	—

Table 6: Audio representation ablation. "Amis CL" = cross-lingual P@1 for Amis queries against other languages. CI = bootstrap 95% CI.

iting factor independent of feature choice.

5.2 Body-kinematic features carry candidate cross-lingual signal

The finding that kinematic features match text for Amis cross-lingual queries has theoretical implications. Kendon (2004) argues that gesture is not merely illustrative but constitutive of interactional-pragmatic acts. If body-kinematic patterns encode pragmatic function independently of language, they could serve as a language-agnostic bridge for cross-cultural retrieval.

Concretely, Amis Explaining clips involve hands-on cultural demonstrations (e.g., weaving, plant-dyeing, and bamboo crafts) with distinctive kinematic signatures (high hand presence ratios, large body centroid variance) that transfer across language boundaries, successfully matching Paiwan textile-weaving segments despite entirely different subtitle content.

5.3 Limitations

This diagnostic study has several limitations that should be acknowledged:

1. **Scale and imbalance.** With 30 clips and Explaining at 47%, bootstrap CI bands overlap across all comparisons (Table 2); modality rankings are tendencies, not confirmed orderings. Utterance-level segmentation is the critical next step.
2. **Feature representations.** MFCC statistics encode language identity; Allosaurus prosodic features are unvalidated on Formosan languages; all fusion is unweighted concatenation. Each of these is an addressable engineering choice.
3. **Silver labels.** VLM annotations were validated against two human annotators ($\kappa_{\text{avg}} = 0.765$) but not by indigenous community members; labels should not be treated as ground truth for cultural analysis. Accordingly, the present study should be interpreted as exploratory and hypothesis-generating rather than confirmatory.

6 Conclusion

We have presented a pilot framework for cross-lingual retrieval of interactional-pragmatic events, applied to Taiwanese indigenous language television news. By combining semantic text embeddings, acoustic features, prosodic statistics, and body-kinematic aggregates across 30 clips from Amis, Paiwan, and Truku, and annotating pragmatic labels via a VLM pipeline, we establish a first retrieval baseline. Text embeddings achieve P@1 = 0.633 overall; kinematic features match text (P@1 = 0.500 each) for cross-lingual Amis queries while substantially outperforming both audio representations, suggesting body movement as a *candidate* cross-cultural pragmatic signal. Audio fusion consistently underperforms audio-free combinations under both MFCC and Whisper-encoder features, motivating learned modality weighting as a priority for future work. The full pipeline—corpus, feature extraction, VLM annotation, and retrieval evaluation—provides a foundation for scaling to concept-based retrieval over utterance-level events and additional Formosan languages. Promising directions include utterance-level segmentation via the existing TextGrid alignments, self-supervised speech features (Hsu et al., 2021; Pratap et al., 2024), and contrastive cross-modal fusion (Radford et al., 2021).

Ethical Considerations

This study uses publicly broadcast television news from TITV, a state-funded public media platform for Taiwan's indigenous peoples. All footage was downloaded from TITV's official YouTube channel and used solely for non-commercial research purposes. Pragmatic-event labels were generated by an LLM (Claude Haiku) and have not been validated by indigenous community members; we caution against interpreting these silver labels as ground-truth characterisations of indigenous communicative norms. Future work will involve consultation with Formosan-language speakers and cultural advisors.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Anthropic. 2025. Claude Haiku 4.5. https://www.anthropic.com/claude. Large language model.

Relja Arandjelović and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617.

John Langshaw Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: Real-time multi-person 2d pose estimation using part affinity fields. *Preprint*, arXiv:1812.08008.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Shu-Kai Hsieh, Yu-Hsiang Tseng, Da-Chen Lian, and Chi-Wei Wang. 2024. Self-supervised learning for Formosan speech representation and linguistic phylogeny. *Frontiers in Language Sciences*, 3:1338684.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge.

Sotaro Kita and Sachiko Ide. 2007. Nodding, aizuchi, and final particles in Japanese conversation: How conversation reflects the ideology of communication and social relationships. *Journal of Pragmatics*, 39(7):1242–1254.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, pages 242–250. ACM.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and Shinji Watanabe. 2023. ML-SUPERB: Multilingual speech universal PERformance benchmark. In *Proceedings of Interspeech 2023*, pages 884–888.

Lily I-wen Su, Li-May Sung, Shuping Huang, Fuhui Hsieh, and Zhemin Lin. 2008. NTU corpus of Formosan languages: A state-of-the-art report. *Corpus Linguistics and Linguistic Theory*, 4(2):291–294.

Po-Ya Angela Wang, Pin-Er Chen, Hsin-Yu Chou, Yu-Hsiang Tseng, and Shu-Kai Hsieh. 2023. Lexical retrieval hypothesis in multimodal context. In *Proceedings of the 4th Conference on Language, Data and Knowledge (LDK 2023)*, pages 455–465, Vienna, Austria. NOVA CLUNL, Portugal.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4210–4220.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246. Association for Computational Linguistics.

Appendices

A VLM Baseline Analysis

To motivate the embedding-based approach, we tested state-of-the-art VLMs (Claude Sonnet 4.5, via keyframe extraction) on TITV Paiwan-language news footage. We evaluated five tasks: open-ended description, language identification, cultural recognition, subtitle OCR, and pragmatic event classification.

The VLM achieved near-perfect scores on tasks solvable via Chinese subtitles and general visual understanding (scene description, subtitle OCR), but exhibited fundamental limitations:

- **Language identification:** The VLM correctly identified the language as Paiwan *only* because the on-screen label "Paiwan" was visible. When asked to identify the language from audio or visual cues alone, it could not distinguish Paiwan from other Formosan or Austronesian languages.

- **Cultural specificity:** The VLM identified "Taiwanese indigenous culture" but could not distinguish Paiwan from Amis or Truku based on visual elements such as clothing or floral headwear.
- **Modality decomposition:** The VLM could not report which modality (subtitle text, facial expression, scene context) contributed most to its interactional-pragmatic judgement—it operates as a black box.
- **Cross-lingual retrieval:** VLMs are not designed for corpus-level retrieval. They cannot take one clip and return a ranked list of pragmatically similar clips from a corpus.

These findings confirm that VLMs and embedding-based retrieval are complementary: VLMs excel at per-instance description, while our framework addresses corpus-level search and modality-decomposed analysis.

B Pilot Corpus

Language	Clips	Date	Language Family
Amis	10	2026-03-02	Austronesian
Paiwan	10	2026-03-07	Austronesian
Truku	10	2026-03-05	Austronesian

Table 7: Pilot corpus overview. Each clip is approximately 5 minutes.

C Pre-processed Data Stream

Stream	Description
`h264/`	MP4 video (720p)
`wav/`	16 kHz mono WAV
`ocr_corrected/`	Frame-level OCR bounding boxes, text, confidence, semantic label (subtitle / title / logo / ticker)
`emit2/`	Allosaurus phone-level TextGrid (IPA labels with start/end times)
`pose/`	Per-frame pose JSON: body skeleton (variable keypoints), bilateral hand (21 keypoints each), facial landmarks; tracked for two speakers (left / right screen position)
`aligned_trees/`	Utterance-aligned parse trees (for future work)

Table 8: Pre-processed data streams available per clip.

D LLM-as-Annotator Prompt

The following prompt was used with Claude Haiku (`claude-haiku-4-5-20251001`) for pragmatic-event label annotation. Each API call included 5 keyframe images and the OCR subtitle summary as additional input.

> 你正在協助一個多模態語料庫研究計畫標注原住民族電視台（TITV）的新聞影片片段。
>
> 以下是這支影片的：
>
> 1. 五張均勻抽樣的畫面截圖
>
> 2. OCR 擷取的中文字幕摘要
>
> 請判斷這支影片中，**主要互動者之間的互動性質**屬於哪一類：
>
> - **Comforting**：慰問、關懷、安慰當事人（例如：災後探視、表達同情）
> - **Celebrating**：慶祝、歡迎、歡慶（例如：豐年祭、頒獎、交車典禮）
> - **Explaining**：解說、教導、文化傳承、分享知識（例如：耆老講解傳統、老師示範）
> - **Urging**：呼籲、勸告、政策訴求（例如：呼籲民眾參與、政策說明施壓）
> - **Reporting**：純粹的新聞播報，無明顯的上述互動性質
>
> 注意：
>
> - 主播在攝影棚播報算 Reporting，除非明顯帶有上述互動功能
> - 若影片含多個新聞，以**最主要**或**最具語用特色**的段落爲準
> - 若無法判斷，選 Reporting
>
> 請用以下 JSON 格式回答（只輸出 JSON，不要其他文字）：

```
{"label": "Explaining",
"confidence": "high",
 "rationale": "一句話說明判斷依據"}

confidence 值：high / medium / low
```

E Kinematic Dimensions

For each speaker $s \in \{\text{left}, \text{right}\}$ we compute: (1) body centroid mean (x, y) — average spatial position (2 dims); (2) body centroid std (x, y) — movement range (2 dims); (3) left/right hand presence ratio — fraction of frames in which the respective hand tracker fires (2 dims); (4) face centroid std (x, y) — head movement dynamics (2 dims). This gives $8 \times 2 = 16$ dimensions per clip, plus two global aggregates: (5) mean body movement energy across both speakers, and (6) hand asymmetry (absolute difference in hand presence ratios), yielding 18 dimensions in total.