A PROGRESSIVE MULTI-STAGE MULTIMODAL DATASET AND BENCHMARK FOR REAL-WORLD CHINESE EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world emotion datasets must jointly balance acquisition control, ecological validity, cross-modal synchronization, label reliability, and reproducible evaluation. We construct a progressive multi-stage multimodal dataset for Chinese emotion research with 68,210 speech segments from 50 participants, nine perceived-emotion labels, synchronized timestamped photo sequences, heart-rate-variability (HRV) measurements, and inertial signals. A unified protocol links individual emotion elicitation, dyadic interaction, and daily-life recording while retaining the same participant cohort, label space, temporal reference, and quality controls. To address the subjectivity of perceived emotion, an auditable annotation pipeline combines independent ratings, absolute-majority aggregation, and adaptive annotator expansion. Across 65,456 segments with at least two alignable ratings, Fleiss' $\kappa = 0.4310$ and Krippendorff's $\alpha = 0.4188$; random resampling of 517 samples with ten complete ratings and fixed-model learnability tests further show that label agreement and learnability generally improve with more annotators. We establish unified, speaker-independent seven- and nine-class benchmarks; the complete seven-class benchmark covers 13 speech baselines, with matched-model comparisons on MELD and IEMOCAP. Chinese-HuBERT-Large achieves 78.42% Accuracy and 52.59% Macro-F1 on the seven-class task. A cumulative three-stage curriculum raises UAR from 51.63% to 54.07%, while yielding 76.61% Accuracy and 52.03% Macro-F1. The strict complete-modality intersection contains 37,043 segments (54.31%). On this subset, additional visual and sensor modalities mainly produce class-conditional performance changes: visual input improves accuracy for six non-neutral classes, while overall Macro-F1 remains close to the audio baseline. Together, the dataset, progressive acquisition framework, empirically validated consensus protocol, and systematic benchmark provide an auditable foundation for real-world Chinese emotion research.

1 INTRODUCTION

Emotion recognition is a central topic in affective computing, with applications in speech assistants, embodied agents, education, mental-health assessment, healthcare, and intelligent vehicles. As large-scale pretrained models and multimodal learning have advanced, emotion recognition has expanded from unimodal speech or image analysis to the joint modeling of speech, visual behavior, and physiological responses. Progress nevertheless depends on datasets that are sufficiently large, reliably labeled, and representative of real-world emotional expression.

Existing resources make different trade-offs among acquisition control, ecological validity, modality coverage, and annotation design. Controlled corpora such as EMO-DB and RAVDESS support stable recording and systematic class coverage; naturalistic resources such as MSP-Podcast and BIIC-Podcast broaden contextual diversity; and IEMOCAP, NNIME, CMU-MOSEI, and SEWA provide interaction or multimodal observations. The specific gap addressed here is the limited availability of public Chinese resources that combine, within one participant cohort and temporal reference, controlled elicitation, dyadic interaction, and daily-life collection, while also providing auditable multi-rater annotations, modality-coverage analysis, and a speaker-independent benchmark. This

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.

combination supports the study of changes across acquisition stages, label reliability, and cross-modal complementarity under a common protocol.

Constructing such a resource poses three linked challenges. Increasing ecological validity generally reduces experimental control; heterogeneous devices require a common temporal reference; and noise, multiple speakers, long recordings, and the subjectivity of perceived emotion substantially increase the cost of cleaning and annotation. A useful dataset must therefore jointly balance naturalness, controllability, synchronization, and label reliability.

Our central question is whether controlled elicitation, interactive tasks, and daily-life collection can be organized into a unified, auditable dataset and benchmark whose perceived-emotion labels support reproducible, leakage-free evaluation and reveal when additional modalities are complementary. We define perceived emotion as a stable categorical judgment made by external observers from audible and visible behavior. Labels and model predictions target this construct, while self-reports and physiological measurements provide complementary evidence for construct validity. We address the question through a unified acquisition and processing protocol, reliability and consensus analyses, and systematic baseline and diagnostic experiments.

We organize data collection into three progressive stages: individual elicitation provides relatively standardized emotional samples; dyadic, task-driven interaction introduces social feedback and spontaneous exchange; and daily-life collection broadens the temporal and contextual range of expression. The same participant identities, label space, temporal reference, and core quality-control criteria are retained across stages. Speech is the primary modality, accompanied by timestamped photo sequences and wearable physiological and inertial signals. Processing includes audio normalization, speech segmentation, target-speaker filtering, independent multi-rater annotation, consensus aggregation, and adaptive annotator expansion.

Our contributions are fourfold:

(1) **A multimodal dataset.** We construct and audit a real-world Chinese emotion dataset with 68,210 speech segments from 50 participants, nine original perceived-emotion labels, synchronized timestamped photo sequences, HRV measurements, and inertial signals. We report modality-specific coverage and define a strict multimodal intersection.

(2) **A progressive three-stage data framework.** Individual elicitation, dyadic interaction, and daily-life collection are integrated under a common participant cohort, label space, synchronization protocol, and quality-control framework, progressing from greater control toward greater ecological validity. A cumulative curriculum—individual, then individual plus dyadic, then all three stages—demonstrates that the framework supports curriculum-learning and class-balance diagnostics, with distinct effects on UAR, Accuracy, and Macro-F1.

(3) **Consensus annotation and empirical validation.** We provide an auditable pipeline based on independent ratings, absolute-majority aggregation, and adaptive annotator expansion. Agreement statistics, random resampling across annotator counts, and fixed-model learnability experiments validate inter-rater reliability, stability, and learnability of the perceived-emotion consensus labels.

(4) **A systematic benchmark.** We establish a unified, speaker-independent, and reproducible benchmark spanning self-supervised speech encoders, emotion-specific representations, generative speech models, and conventional acoustic features. On the strict multimodal intersection, aggregate metrics, class-wise results, and participant-level uncertainty characterize learnability, class difficulty, and modality complementarity.

2 Related Work

2.1 Emotion Datasets and Collection Paradigms

Emotion datasets trade off experimental control, ecological validity, label coverage, and annotation cost. Controlled corpora such as EMO-DB, CASIA, and RAVDESS provide stable recording conditions and systematic class coverage, while IEMOCAP introduces scripted and improvised dyadic interaction (Burkhardt et al., 2005; Chinese Linguistic Data Consortium, 2005; Livingstone & Russo, 2018; Busso et al., 2008). Naturalistic resources, including MSP-Podcast and BIIC-Podcast, broaden acoustic and contextual diversity through media-derived speech (Lotfian & Busso, 2019;

Table 1: Structural comparison with representative emotion datasets. Factual acquisition properties are reported instead of ranking dataset quality.

Dataset	Language/source	Scale	Setting	Modalities	Cross-setting cohort	Agreement/audit
IEMOCAP	English	10 actors	Scripted/improvised dyads	A,V,T	No	Multi-rater
MSP-Podcast	English media	>100 h	Natural podcast	A	No	Multi-rater
NNIME	Mandarin	44 participants	Dyadic interaction	A,V,ECG	No	49 raters
CHEAVD	Chinese media	2,381 clips	Film/web video	A,V	No	Multi-rater
M3ED	Chinese media	9,908 turns	Television dialogue	A,V,T	No	Multi-rater/context
BIIC-Podcast	Mandarin media	3,193 clips	Natural podcast	A	No	Consensus labels
Ours	Mandarin cohort	68,210 clips	Elicitation→dyad→daily life	A,photos,HRV,inertial	Yes	κ, α, growth audit

Upadhyay et al., 2023). Chinese audio-visual resources such as CHEAVD, M3ED, and MER2023 provide natural media, conversational context, or standardized challenge protocols (Li et al., 2017; Zhao et al., 2022; Lian et al., 2023). NNIME, CMU-MOSEI, and SEWA further demonstrate the value of interaction and multimodal observation (Chou et al., 2017; Zadeh et al., 2018; Kossaifi et al., 2021).

These paradigms offer complementary strengths. Controlled elicitation improves experimental repeatability, media-derived data increase contextual diversity, and interaction or multimodal recording adds behavioral context. Naturalistic collection additionally introduces background noise, class imbalance, uncertain speaker identity, missing modalities, and lower annotation agreement. Fair comparison therefore requires explicit split rules, leakage control, modality-coverage reporting, and metrics that are robust to class imbalance (Ma et al., 2024a; Wu et al., 2024). Recent shared tasks and unified benchmarks further emphasize noise robustness, open-vocabulary evaluation, and reproducible cross-corpus comparison (Lian et al., 2024; 2026).

2.2 Positioning of This Dataset

Table 1 summarizes the structural distinctions most relevant to this work. The central contribution is the integration of controlled elicitation, dyadic interaction, and daily-life collection within one participant cohort, label ontology, temporal reference, and audit framework. This design enables analyses of progression across settings, consensus-label reliability, modality missingness, and speaker-independent benchmarking under a shared protocol.

The remaining gap is thus a jointly auditable resource that combines several requirements: the same cohort spans progressively less controlled settings; heterogeneous devices share a temporal reference; long recordings are filtered for the target speaker; raw multi-rater decisions and aggregation rules remain auditable; and evaluation uses speaker-independent splits. Our dataset and benchmark are designed around this combination.

3 Overview of Datasets and Methods

3.1 Progressive Multi-Stage Collection Design

We organize individual elicitation, dyadic interaction, and daily-life collection as a continuum from controlled to naturalistic settings under one participant cohort, label space, synchronization protocol, and quality-control pipeline. "Progressive" has three operational dimensions: task constraints are gradually relaxed, social interaction increases from individual expression to dyadic exchange, and temporal and contextual openness expands from short controlled sessions to extended daily activities. Participant identity management, the label ontology, temporal references, and core quality criteria remain fixed across stages, reducing protocol variation unrelated to the acquisition setting. Emotional authenticity, annotation difficulty, and model behavior across these stages are evaluated empirically.

Speech is the primary modality, accompanied by timestamped photo sequences, HRV measurements, and inertial signals. All modalities are aligned to a shared temporal reference. Speech processing includes loudness normalization, activity detection, target-speaker filtering, and segmentation. Raw data remain read-only, while processed outputs and audit records are stored separately. Device and task settings, synchronization, and preprocessing are detailed in Appendix A; Figure 1 summarizes the full pipeline.

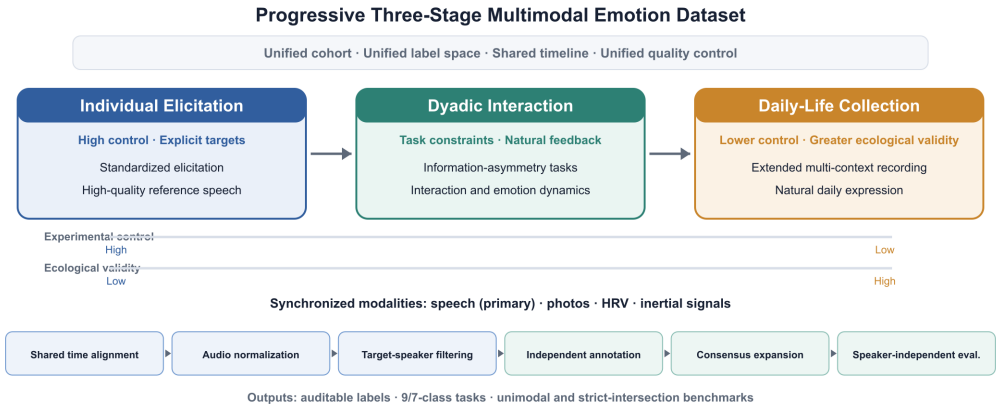


Figure 1: Overview of the progressive three-stage multimodal dataset construction and processing pipeline.

3.2 Consensus Annotation and Validation Framework

Each speech segment is independently assigned one of nine perceived-emotion labels based on its audible expression. Annotators are not shown the task target, self-report, other ratings, physiological signals, or visual data. A formal label is issued only when one category reaches an absolute majority. Samples without a majority are assigned to additional annotators; those still lacking a majority at the maximum annotator count are marked ambiguous and excluded. Raw votes, assignment order, and final status are retained for auditing. Section A.6 specifies the interface, annotator-count distribution, tie handling, and exclusion rules; Appendix B reports agreement and consensus-growth analyses.

Three complementary analyses evaluate the consensus labels. First, among 65,456 samples for which at least two raw ratings could be aligned, Fleiss' $\kappa = 0.4310$ and Krippendorff's $\alpha = 0.4188$, indicating a moderate shared structure beyond chance. Second, for 517 samples with ten complete ratings, we repeatedly draw subsets of one, two, and three annotators over 30 trials. Agreement with the ten-annotator reference and fixed-model learnability generally improve as the number of annotators increases, providing internal evidence of convergence. Third, a noise-sensitivity experiment evaluates the robustness of majority aggregation to random annotation errors. Together, these analyses support the consensus labels as reliable, stable, and learnable operational supervision for perceived emotion; complementary construct-validity evidence from self-report and physiological information is discussed in Section B.1.3.

Theoretically, the finite-sample category proportions estimate the judgment distribution of a target observer population. Under annotator independence, population relevance, and a stable judgment process, additional observations reduce idiosyncratic error. Cultural consensus theory and latent-observer models provide related foundations for this interpretation (Romney et al., 1986; Dawid & Skene, 1979). Appendix B states the assumptions, convergence boundaries, levels of construct validity, and relationships among validation subsets. Figure 2 summarizes the nine-class labels and consensus-growth evidence; exact distributions and agreement results are reported in Tables 8–10.

3.3 Dataset Scale and Evaluation Protocol

The dataset contains 68,210 speech segments from 50 participants, with speaker-independent train/validation/test splits of 45,208/6,990/16,012. The nine-class counts are 47,911 neutral, 1,355 disgust, 7,714 joy, 694 fear, 756 sadness, 3,159 surprise, 2,962 anger, 1,970 frustration, and 1,689 anxiety. Collection tasks target eight emotions, while annotators may additionally assign neutral, yielding nine released labels. For comparison with common discrete-emotion settings, the seven-class task applies only two semantic merges: frustration into sadness and anxiety into fear. All other samples are retained, with the same speaker-independent splits as the nine-class task. Because neutral accounts for 70.24% of the data, Macro-F1 and UAR are primary metrics; Accuracy and Weighted-F1 are supplementary, and class-wise results are also reported.

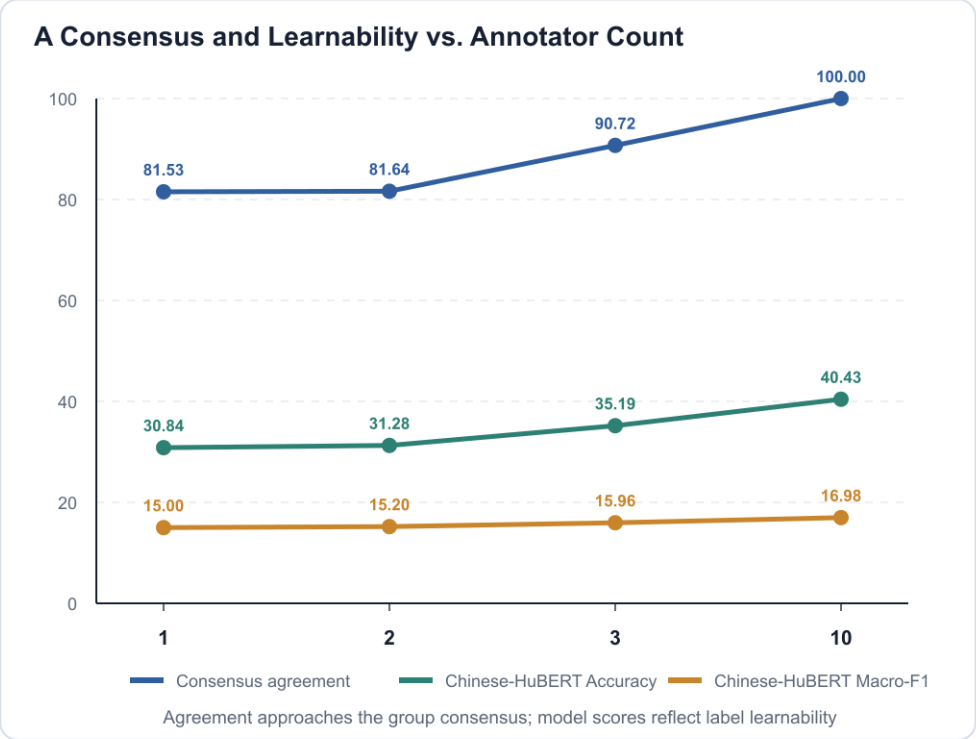


Figure 2: Nested consensus agreement and fixed-model learnability as the number of annotators increases. Results for one to three annotators are means over 30 random draws; the ten-annotator result is the full-consensus reference.

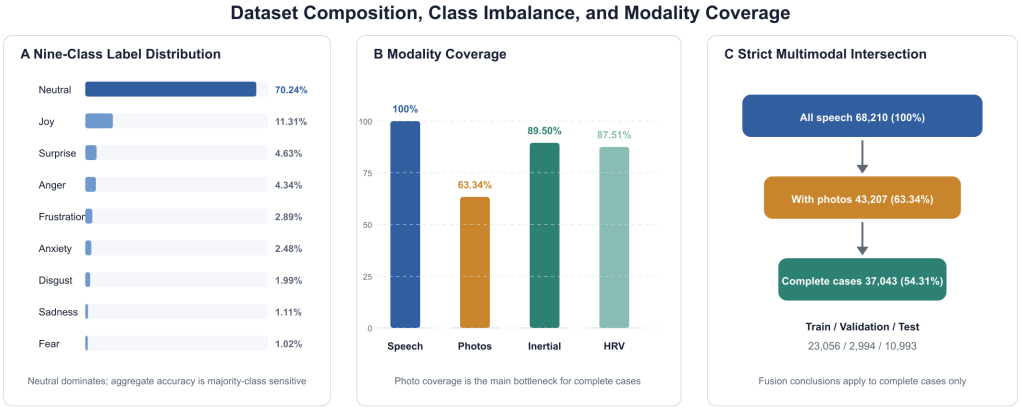


Figure 3: Nine-class label distribution, modality coverage, and the strict complete-modality intersection.

Coverage is 63.34% for photos, 89.50% for inertial signals, and 87.51% for HRV. The strict intersection in which speech, photos, HRV, and inertial signals are all available contains 37,043 segments (54.31%), split into 23,056/2,994/10,993 for training/validation/test. Figure 3 summarizes label distribution and modality coverage; Tables 8 and 14 provide complete counting rules.

The strict intersection is a complete-case subset, with photo coverage as the main bottleneck. Fusion results therefore characterize modality complementarity within samples where all four sources are

Table 2: Benchmark tasks and evaluation roles.

Track	Samples	Classes	Split	Primary purpose
Speech benchmark	68,210	7/9	Speaker-independent	Overall learnability and model comparison
Consensus growth	517	9	Fixed participant split	Reliability and label learnability
Curriculum diagnostic	68,210	7	Same held-out test set	Value of the progressive stage order
Multimodal benchmark	37,043	7	Strict-intersection split	Conditional modality complementarity

Table 3: Complete seven-class speech benchmark (%). FT: light fine-tuning; FR: frozen representation with a trained classifier.

Model	Adaptation	Accuracy	Macro-F1	Weighted-F1
Chinese-HuBERT-Large	FT	78.42	**52.59**	78.04
SenseVoice-Small	FR	70.31	43.77	71.93
Whisper-small	FT	76.39	43.00	74.84
XLS-R-300M	FT	76.69	38.52	74.47
HuBERT-Large	FT	74.50	38.08	73.33
WavLM-Base+	FT	72.30	36.09	71.33
PANNs CNN14	FR	72.61	32.50	70.31
data2vec-audio-base	FT	69.42	32.05	69.00
AST AudioSet	FR	70.06	31.70	69.29
Wav2Vec2-base-960h	FT	69.92	26.03	67.14
emotion2vec+ Large	FR	70.63	25.55	66.57
eGeMAPSv02 + SVM	Conventional	48.96	24.82	55.21
emotion2vec+ Base	FR	63.11	23.77	62.10

available. Appendix B defines this subset and reports unimodal, aggregate-fusion, and class-wise results; potential missingness bias relative to the full dataset is discussed in Section 5.

4 Experimental Results

4.1 Benchmark Tasks and Evaluation Protocols

Table 2 summarizes the four evaluation tracks. All seven-class speech experiments use the same speaker-independent split and label mapping; the test set is used only for final reporting. Fine-tunable pretrained encoders are trained for at most 15 epochs with per-GPU batch size 8, the first eight encoder layers frozen, and early stopping on validation Macro-F1 (patience 4). Models exposing fixed representations use a frozen encoder with a validation-selected classifier, while eGeMAPS uses a validation-selected SVM. Appendix B marks each adaptation regime and gives full implementation details. Macro-F1 and UAR are primary metrics because neutral accounts for 70.24% of the data; Accuracy and Weighted-F1 are supplementary.

4.2 Comprehensive Speech Benchmark

Table 3 reports all 13 speech baselines rather than only selected systems, including general self-supervised encoders (Baevski et al., 2020; Hsu et al., 2021), an emotion-specific representation (Ma et al., 2024b), and conventional acoustic features. Chinese-HuBERT-Large performs best, reaching 78.42% Accuracy, 52.59% Macro-F1, and 78.04% Weighted-F1. The gap between Accuracy and Macro-F1 is substantial across models, showing that aggregate correctness is strongly influenced by the neutral majority. The best model reaches 89.95% accuracy on neutral, 29.60% on fear, 41.69% on disgust, and 45.94% on sadness (Appendix Table 17), highlighting the value of reporting class-wise performance alongside aggregate Accuracy.

The ranking also shows that pretraining objective and language/domain match matter more than model size alone. Chinese-HuBERT-Large has the strongest balanced performance, whereas several

Table 4: Matched-model Macro-F1 (%) across datasets. IEMOCAP reports five-fold mean±standard deviation. Values are interpreted as benchmark behavior under each dataset's label space and protocol.

Model	MELD (7-class)	IEMOCAP (4-class)	Ours (7-class)
Chinese-HuBERT-Large	30.07	66.67±2.32	**52.59**
WavLM-Base+	27.56	65.58±5.49	36.09
HuBERT-Large	31.15	69.34±1.58	38.08
XLS-R-300M	26.26	55.30±8.45	38.52
data2vec-audio-base	26.36	65.95±2.03	32.05
Whisper-small	30.91	**70.90**±2.04	43.00

Table 5: Nine-class versus seven-class results under matched splits (%).

	Nine classes		Seven classes	
Model	Accuracy	Macro-F1	Accuracy	Macro-F1
Chinese-HuBERT-Large	77.97	46.26	78.42	52.59
WavLM-Base+	71.02	28.22	72.30	36.09
HuBERT-Large	73.34	31.70	74.50	38.08
XLS-R-300M	76.73	30.81	76.69	38.52

models with competitive Accuracy remain weak on minority classes. This pattern motivates reporting class-wise recall and confusion matrices in addition to aggregate scores.

4.3 Matched Comparison with Public Emotion Datasets

To separate dataset positioning from literature numbers obtained under incompatible implementations, we retrain the same baseline families on MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008). MELD uses its official seven-class train/dev/test split with 13,706 decodable utterances. IEMOCAP uses the standard four-class setting with five session-independent folds over 5,531 utterances and no speaker overlap. Table 4 reports Macro-F1, the metric common to these runs.

Three observations follow. First, absolute values are interpreted in the context of each dataset's language, label space, sample source, and evaluation protocol. Second, model rankings change across datasets: Whisper-small is strongest on IEMOCAP, HuBERT-Large leads this subset on MELD, and Chinese-HuBERT-Large is strongest on Ours. Third, the large Accuracy–Macro-F1 gap on Ours identifies class imbalance and fine-grained negative emotions as central benchmark challenges.

4.4 Nine-Class Labels and the Progressive Collection Structure

The released ontology contains nine labels. The seven-class task merges only frustration into sadness and anxiety into fear to support comparison with common SER settings. Table 5 compares identical model families under the two label spaces and the same speaker-independent split. The merge changes Accuracy by at most 1.28 points for these models but raises Macro-F1 by 6.33–7.87 points. This indicates that the seven-class protocol reduces difficulty concentrated in the two fine-grained negative categories, whereas the nine-class ontology preserves a more challenging diagnostic target.

We further evaluate the collection stages as a cumulative curriculum. Chinese-HuBERT-Large is trained for five epochs on individual elicitation, five epochs on individual plus dyadic data, and five epochs on all stages, initializing each stage from the previous best checkpoint while resetting the optimizer. Compared with random mixed training, the curriculum raises UAR by 2.44 points while changing Accuracy and Macro-F1 by -1.81 and -0.56 points (Table 6). The progressive structure therefore demonstrates practical value for class-balance analysis and learning-order diagnostics, with the UAR gain complementing aggregate Accuracy and Macro-F1.

Table 6: Final cumulative-curriculum comparison (%).

Training strategy	Accuracy	Macro-F1	UAR
Randomly mixed training	78.42	52.59	51.63
Three-stage cumulative curriculum	76.61	52.03	54.07
Change	-1.81	-0.56	$+2.44$

Table 7: Core fusion results on the strict multimodal intersection.

Setting	Accuracy	Macro-F1	ΔMacro-F1 vs. speech (pp)
Speech	74.25	50.89	Baseline
Speech + vision	75.08	50.99	+0.10 $[-0.41, +0.62]$
Speech + vision + sensors	75.08	50.77	-0.12 $[-1.10, +0.59]$

4.5 Multimodal Complementarity on the Strict Intersection

Multimodal experiments use the fixed 37,043-sample strict intersection and its 23,056/2,994/10,993 train/validation/test split. Speech features come from the best seven-class Chinese-HuBERT checkpoint. Frozen DINOv2-Small encodes at most 16 timestamped photos per segment followed by temporal aggregation, and sensor features comprise aligned HRV and inertial statistics. Hyperparameters are selected only on validation data. Confidence intervals use 2,000 paired participant-level bootstrap replicates.

Speech alone achieves 74.25% Accuracy and 50.89% Macro-F1. Speech+vision reaches 75.08% and 50.99%, corresponding to ΔMacro-F1 $= +0.10$ points with 95% CI $[-0.41, +0.62]$. The three-modality system reaches 75.08% and 50.77%, or -0.12 points with CI $[-1.10, +0.59]$. Both intervals include zero, placing the aggregate effect close to the speech baseline and directing interpretation toward class-conditional complementarity.

Class-wise analysis provides a more specific interpretation. Vision improves accuracy for disgust, joy, fear, sadness, surprise, and anger while reducing the dominant neutral category by 12.99 points. Figure 4 therefore reflects error redistribution and class-conditional correction across emotion categories. Sensor addition further helps some categories and changes others, consistent with lower signal-to-noise ratio, participant-specific baselines, motion artifacts, and physiological response delays. Exact weak-modality and class-wise results appear in Appendix Tables 18–20.

4.6 Main Findings

The benchmark yields four conclusions. First, the dataset is readily learnable by modern speech encoders, while the Accuracy–Macro-F1 gap and per-class results expose substantial long-tail difficulty. Second, matched experiments on MELD, IEMOCAP, and Ours show that model rankings depend on language, context, label space, and protocol; benchmark conclusions should therefore be dataset-specific. Third, the nine-class ontology retains fine-grained negative-emotion difficulty that is partly removed by the seven-class merge, while the cumulative stage order improves balanced recall and produces a distinct trade-off across evaluation metrics. Fourth, speech is the most stable modality, while visual and sensor signals contribute class-conditional complementary information.

5 Limitations and Conclusions

We present a Chinese multimodal emotion dataset with 68,210 speech segments from 50 participants and a progressive three-stage framework spanning individual elicitation, dyadic interaction, and daily-life collection under unified identities, labels, synchronization, and quality control. Independent multi-rater annotation, absolute-majority aggregation, and adaptive annotator expansion form an auditable consensus pipeline. Agreement statistics, annotator-count resampling, and fixed-model learnability experiments jointly validate inter-rater reliability, stability, and learnability of the resulting supervision. In the speaker-independent benchmark, Chinese-HuBERT-Large reaches 78.42% Accuracy and 52.59% Macro-F1 on the seven-class task. Cumulative curriculum learning increases UAR from

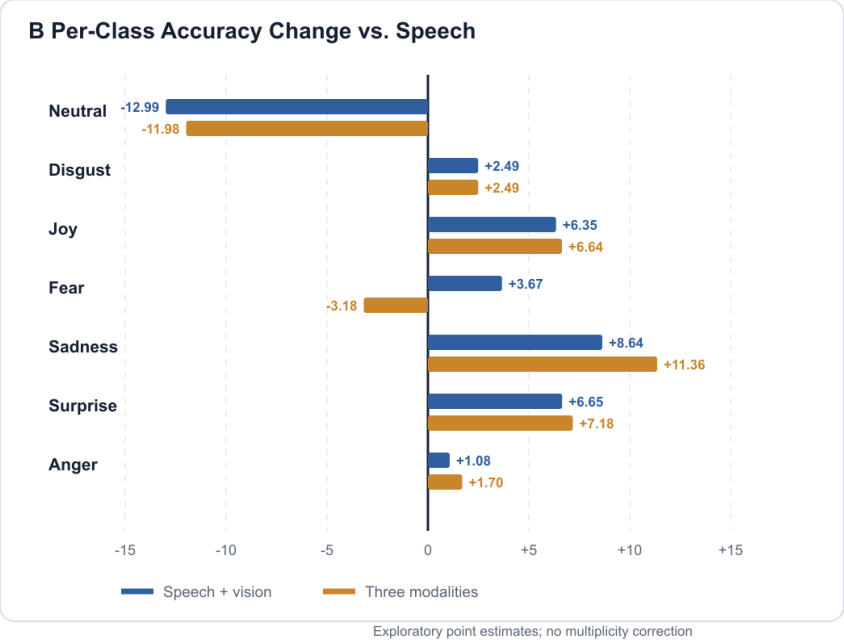


Figure 4: Class-wise accuracy changes relative to speech on the strict intersection. Positive values indicate classes improved after adding modalities; values summarize the observed class-conditional changes.

51.63% to 54.07%, with corresponding trade-offs in Accuracy and Macro-F1. Strict-intersection experiments further indicate that speech is the most stable modality, whereas visual and sensor information primarily provides class-conditional complementary evidence.

The current results also identify several directions for extension. The strict complete-modality intersection covers 54.31% of all segments, wearable signals are affected by signal quality and individual variability, the class distribution is highly imbalanced, and the relationship between perceived consensus and subjective experience can be strengthened through broader multi-source validation. Future work will analyze missingness bias, strengthen quality control and personalization for weak modalities, and evaluate external validity across cohorts and contexts. Under the current authorization, task definitions, aggregate statistics, model configurations, and evaluation code may be released; any external sharing of participant-level personal data requires renewed authorization and the applicable review procedures. Appendix C provides expanded limitations, ethics, privacy, and governance details.

AI Use Statement

Generative AI tools were used only for Chinese–English translation, wording refinement, and visualization of experimental results. All other content was completed by and remains the responsibility of the authors.

Ethics Statement

Written authorization was obtained from participants through a Personal Data Collection and Processing Authorization Form. The form describes the project purpose and the collection of voice, images, motion, and PPG/HRV information through phones and watches. Participants may refuse to provide the relevant data without adverse consequences and may withdraw their consent. The personal data are collected and processed in China, restricted to internal use within this project, stored in a dedicated isolated server area managed by the data-collecting institutions, and accessible only to project personnel. The authorization specifies retention through December 31, 2027, followed by deletion or anonymization; extended use or use by another project requires renewed consent. The

current authorization does not permit external disclosure or transfer of participant-level personal data. Appendix C details the retention, withdrawal, security, and governance arrangements.

Reproducibility Statement

The main paper specifies the speaker-independent splits, label mappings, evaluation metrics, modality coverage, strict-intersection cohort, and principal training protocol. Appendix A documents acquisition, synchronization, preprocessing, target-speaker filtering, and annotation quality control. Appendix B reports agreement definitions, consensus-growth sampling, benchmark configurations, complete aggregate and class-wise results, and uncertainty estimates. Release materials will provide label definitions, aggregate statistics, model configurations, and evaluation code; auditable manifests and participant-level data remain available only for internal reproduction and audit. Under the current authorization, participant-level personal data are not externally disclosed or transferred; any future sharing requires renewed authorization and the applicable procedures.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Proceedings of Interspeech*, pp. 1517–1520, 2005. doi: 10.21437/Interspeech.2005-446.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Chinese Linguistic Data Consortium. CASIA chinese emotional corpus (CLDC-SPC-2005-010). Dataset, Institute of Automation, Chinese Academy of Sciences, 2005.

Huang-Cheng Chou, Wei-Cheng Lin, Li-Chia Chang, Chi-Chun Li, Hsin-Pai Ma, and Chi-Chun Lee. NNIME: The NTHU-NTUA chinese interactive multimodal emotion corpus. In *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction*, pp. 292–298, 2017. doi: 10.1109/ACII.2017.8273615.

Alexander P. Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979. doi: 10.2307/2346806.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971. doi: 10.1037/h0031619.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Jean Kossaifi, Roland Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vikas Pandit, Alexandros Toisoul, Bjorn Schuller, Kim Star, Elnar Hajiyev, and Maja Pantic. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, 2021. doi: 10.1109/TPAMI.2019.2944808.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania, 2011.

Yan Li, Jianhua Tao, Lianhong Chao, Wei Bao, and Yazhu Liu. CHEAVD: A chinese natural emotional audio-visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8: 913–924, 2017. doi: 10.1007/s12652-016-0406-z.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mofan Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Bjorn W. Schuller, and Jianhua Tao. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9610–9614, 2023. doi: 10.1145/3581783.3612836.

Zheng Lian, Haiyang Sun, Licai Sun, Ziming Wen, Shun Zhang, Siyuan Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Ke Xu, Bin Liu, Erik Cambria, Guoying Zhao, Bjorn W. Schuller, and Jianhua Tao. MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48, 2024. doi: 10.1145/3689092.3689959.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lifeng Chen, Bin Liu, and Jianhua Tao. MERBench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5793–5810, 2026. doi: 10.1109/TPAMI.2026.3653457.

Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391.

Reza Lotfian and Carlos Busso. Curriculum learning for speech emotion recognition from crowdsourced labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):815–826, 2019. doi: 10.1109/TASLP.2019.2898816.

Ziyang Ma, Mingjie Chen, Han Zhang, Zhisheng Zheng, Wenxi Chen, Xie Li, Jiaxin Ye, Xie Chen, and Thomas Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proceedings of Interspeech 2024*, pp. 1580–1584, 2024a. doi: 10.21437/Interspeech.2024-788.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024b. doi: 10.18653/v1/2024.findings-acl.931.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050.

A. Kimball Romney, Susan C. Weller, and William H. Batchelder. Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88(2):313–338, 1986. doi: 10.1525/aa.1986.88.2.02a00020.

Shreya G. Upadhyay, Wei-Sheng Chien, Bo-Hao Su, Lucas Goncalves, Yu-Ting Wu, Ahmad N. Salman, Carlos Busso, and Chi-Chun Lee. An intelligent infrastructure toward large scale naturalistic affective speech corpora collection. In *Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction*, pp. 1–8, 2023. doi: 10.1109/ACII59096.2023.10388175.

Heng-Jui Wu, Huang-Cheng Chou, Kai-Wei Chang, Lucas Goncalves, Jun Du, Jyh-Shing Roger Jang, Chi-Chun Lee, and Hung-Yi Lee. Open-emotion: A reproducible EMO-SUPERB for speech emotion recognition systems. In *Proceedings of the 2024 IEEE Spoken Language Technology Workshop*, pp. 510–517, 2024. doi: 10.1109/SLT61566.2024.10832296.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208.

Jiaxin Zhao, Tianyi Zhang, Jianhui Hu, Yi Liu, Qin Jin, Xiaohang Wang, and Haizhou Li. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5699–5710, 2022. doi: 10.18653/v1/2022.acl-long.391.

A Detailed Dataset Construction and Processing

A.1 Participants, Scope, and Data Organization

The dataset was designed to connect controlled emotion elicitation with progressively more natural interaction while retaining a common participant cohort, label ontology, temporal reference, and quality-control procedure. The current audited release contains data from 50 participants. Each participant is represented by an anonymous identifier that remains consistent across collection stages and processing steps. These identifiers are also the unit used to construct the speaker-independent training, validation, and test partitions reported in the main paper.

The collection pipeline comprises participant preparation, multi-stage acquisition, multimodal synchronization, preprocessing, target-speaker filtering, independent annotation, and consensus-based quality control. Identity information is maintained separately from research records, and downstream manifests retain only the fields required for alignment and evaluation. The conclusions are scoped to the audited participant cohort and collection conditions. Any future release of demographic summaries will follow consent, de-identification, and minimum-disclosure requirements.

A.2 Emotion Ontology and Progressive Collection

The original annotation ontology contains nine perceived-emotion categories: neutral, disgust, joy, fear, sadness, surprise, anger, frustration, and anxiety. The nine-class benchmark preserves all categories. The seven-class benchmark merges frustration into sadness and anxiety into fear while retaining the remaining categories. This mapping is fixed before model training and is applied identically to the training, validation, and test partitions.

Data acquisition proceeds through three stages that differ in controllability, interaction, and ecological context.

Individual elicitation. Participants first completed standardized emotion-elicitation tasks in a relatively controlled recording environment. The protocol combined curated audiovisual prompts with autobiographical recall and scenario-based expression. The prompts were intended to establish a shared task context, whereas autobiographical recall allowed participants to connect the target category to personally meaningful experiences. Participants completed a self-check after each task; these self-reports are retained as auxiliary metadata and are not shown to external annotators or used in consensus voting. Recordings from this stage also provide clean reference speech for subsequent target-speaker filtering.

Dyadic interaction. The second stage introduced another participant and task-driven communication. Cooperation, competition, asymmetric-information exchange, and scenario-based discussion were used to elicit spontaneous turn taking, interruptions, negotiation, and context-dependent emotional change. The interaction partner and task context make this stage less constrained than individual elicitation while preserving a known acquisition protocol and participant identity.

Daily-life recording. The third stage extended acquisition to routine learning, work, social, and leisure contexts. Participants could pause recording or omit private situations. These recordings contain naturally occurring changes in topic, acoustic environment, interlocutors, and emotional intensity, and therefore require additional speech segmentation and target-speaker filtering. The three stages form a continuous design from high experimental control to greater contextual naturalism rather than three unrelated corpora.

A.3 Multimodal Acquisition and Temporal Alignment

Speech is the primary modality. The same acquisition sessions additionally provide timestamped photo sequences, heart-rate-variability (HRV) measurements, and inertial signals from wearable devices when available. Camera frames are sampled at fixed temporal intervals to form continuous photo sequences that can be aligned to speech segments without requiring frame-level manual annotation. HRV and inertial measurements retain device timestamps and session identifiers.

All devices are associated with a common session-level temporal reference. Preprocessing preserves the original start and end times of each speech segment, enabling lookup of the corresponding photo and sensor windows. Alignment tables record the availability of each modality rather than silently discarding incomplete cases. Consequently, the speech benchmark uses the complete audio cohort, whereas multimodal experiments use the explicitly audited strict intersection described in Appendix B. This separation prevents modality availability from being conflated with model performance.

A.4 Audio Standardization and Speech Segmentation

Raw audio is converted to a common waveform format and sampling configuration before segmentation. The preprocessing pipeline performs format validation, channel normalization, loudness normalization, and speech-activity detection. Candidate speech regions are then merged or split using duration and silence constraints so that each exported unit contains a coherent local utterance while preserving its source recording and timestamps. Corrupted files, empty regions, and segments that fail basic quality checks are recorded in the audit log rather than repaired without trace.

For controlled individual recordings, segmentation is applied directly after standardization. Dyadic sessions require additional handling of turn changes and overlapping speech. The exported manifest therefore records the source session, participant identifier, temporal boundaries, collection stage, and processing status for every retained segment. This metadata supports both reproducibility and later reconstruction of the local interaction context without exposing that context to annotators during the primary labeling task.

A.5 Long-Duration Audio and Target-Speaker Filtering

Daily-life recordings can contain long silent intervals, background speech, and multiple interlocutors. We therefore use a two-stage pipeline. First, speech-activity detection identifies candidate regions. Second, a target-speaker model compares each candidate with a reference representation derived from that participant's individual-stage recordings. The implementation uses a fixed speaker embedding and similarity threshold; rejected regions are retained in processing logs but are not presented as target-participant samples.

For efficient inspection and context management, retained target-speaker regions are first grouped into temporally ordered blocks of at most five minutes and are then segmented again into shorter utterance-level units. The second pass preserves the timestamps of the original recording. This procedure balances two requirements: annotators can work with compact speech units, while every unit remains traceable to a bounded source context and to the corresponding photo and sensor windows.

Target-speaker filtering reduces, but does not by itself eliminate, errors caused by background speakers, overlap, or acoustic mismatch. Quality checks therefore combine automated status flags with manual review of uncertain or low-quality segments. The final benchmark manifest includes only segments that pass the defined processing and annotation criteria.

A.6 Independent Annotation and Consensus Quality Control

Annotation is conducted through a web-based platform that records sample assignment, labeler identity, submission time, revision history, and final aggregation status. Each labeler receives an isolated speech segment together with the category definitions and annotation instructions. The interface does not reveal the participant's self-report, elicitation target, physiological measurements, visual data, other labelers' decisions, or the final consensus label. This separation is intended to measure perceived emotion from the presented behavior rather than agreement induced by shared metadata.

For each valid segment, labelers select a primary emotion category and an intensity level; an optional secondary category can record mixed expression. "Ambiguous" is available when a dominant category cannot be determined, and "Abnormal" is reserved for unusable material such as severe corruption or failed segmentation. These status labels are audited separately and are not treated as ordinary emotion classes in the reported benchmarks.

Production labels follow an absolute-majority rule with adaptive annotator expansion. A segment is initially assigned to three independent labelers. A label is accepted only when it receives more than half of the currently valid votes. If no absolute majority is present, the segment is assigned to additional independent labelers up to the predefined production limit. Segments that remain unresolved, or that are marked abnormal, are excluded from the discrete-classification benchmark rather than being forced into a category. Raw votes, assignment order, added-annotator decisions, and the final status are retained for audit.

This protocol distinguishes three objects that are often conflated: individual judgments, the aggregated perceived-emotion label, and auxiliary evidence such as self-report or physiology. Appendix B evaluates the aggregated labels through agreement coefficients, annotator-count resampling, fixed-model learnability tests, and noise-sensitivity analyses. Together with the complete vote history, these checks make the consensus process reproducible and allow later work to study alternative aggregation rules without modifying the original annotations.

B Complete Evaluation and Experimental Results

B.1 Evaluation Framework

We evaluate the dataset along four complementary dimensions. First, *inter-rater reliability* tests whether independent annotators make reproducible judgments on the same segment. Second, *model learnability* tests whether a fixed label system supports generalization to unseen participants. Third, *construct-validity evidence* relates observer consensus to the task context and available auxiliary metadata. Fourth, *ecological coverage* examines whether the shared protocol supports analysis across individual elicitation, dyadic interaction, and daily-life recording. These dimensions are evaluated jointly rather than being reduced to a single score.

B.1.1 Annotation Reliability

Because adaptive arbitration yields different numbers of valid ratings across samples, Krippendorff's alpha is the primary agreement coefficient (Krippendorff, 2011). For nominal labels,

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o denotes observed disagreement and D_e denotes the disagreement expected from the marginal label distribution. We additionally report Fleiss' kappa (Fleiss, 1971),

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e},$$

where $\bar{P}$ is mean observed agreement and P_e is chance agreement implied by the marginals. Fleiss' kappa is computed within equal-rater-count subsets and then summarized using the corresponding valid-sample counts.

B.1.2 Model Learnability

Learnability is evaluated with participant-independent splits. Accuracy measures the fraction of correct predictions, while unweighted average recall (UAR) gives equal weight to every class:

$$\text{Accuracy} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(\hat{y}_n = y_n), \qquad \text{UAR} = \frac{1}{K}\sum_{i=1}^{K} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}.$$

Macro-F1 is reported alongside UAR because the neutral class dominates the empirical distribution. Agreement and learnability address different questions: agreement characterizes reproducibility among observers, whereas model performance characterizes generalization of the resulting supervision.

Table 8: Nine-class label distribution over all 68,210 speech segments.

Emotion label	Number of samples	Percentage
Neutral	47,911	70.24%
Disgust	1,355	1.99%
Joy	7,714	11.31%
Fear	694	1.02%
Sadness	756	1.11%
Surprise	3,159	4.63%
Anger	2,962	4.34%
Frustration	1,970	2.89%
Anxiety	1,689	2.48%
Total	68,210	100.00%

Note: Category percentages are rounded independently to two decimal places; their displayed sum may therefore differ from 100.00% by 0.01 percentage points.

Table 9: Final agreement statistics on the auditable raw-annotation subset.

Agreement statistic	Number of valid samples	Value
Fleiss' kappa κ	65,456	0.4310
Krippendorff's Alpha α	65,456	0.4188

B.1.3 Construct-Validity Evidence

The consensus label represents perceived emotion: the stable categorical judgment formed by independent observers from the presented behavior. Task context, participant self-report, interaction structure, and physiological measurements are retained as separate evidence sources. This separation allows future analyses to test convergence among perception, context, subjective report, and physiological arousal without leaking those sources into the primary annotation decision. The progressive collection design additionally makes it possible to examine whether label and model behavior remain interpretable as contextual naturalism increases.

B.1.4 Dataset Statistics

Table 8 reports the nine-class distribution over the complete 68,210-segment audio cohort. The strong class imbalance motivates the joint use of Accuracy, Macro-F1, UAR, class-wise results, and class support throughout the benchmark.

The raw-annotation audit aligns 65,456 segments to at least two original ratings by session identifier and millisecond start time. The remaining 2,754 production segments are excluded from the agreement denominator because their raw vote records cannot be aligned at the same granularity. Agreement is therefore reported on the explicitly auditable subset rather than imputed for the unmatched records.

Both chance-corrected coefficients are positive and of moderate magnitude. The consensus-growth experiment below provides a second, independent view of stability by varying the number of ratings used to form the label.

B.2 Consensus-Label Validation

Emotion annotation is a subjective measurement problem: observers may differ because of experience, cultural background, and sensitivity to acoustic or behavioral cues. Multi-rater aggregation is therefore common in emotion datasets such as IEMOCAP, MSP-Podcast, and CMU-MOSEI (Busso et al., 2008; Lotfian & Busso, 2019; Zadeh et al., 2018). More generally, cultural-consensus theory formalizes how shared judgments can be recovered from multiple observers (Romney et al., 1986), while latent-label models such as Dawid–Skene treat annotator disagreement as estimable measurement variation (Dawid & Skene, 1979).

Table 10: Consensus-growth and Chinese-HuBERT results by number of annotators.

Annotators	Agreement with 10-rater consensus (%)	Accuracy (%)	UAR (%)	Macro-F1 (%)	ΔAccuracy (pp)
1	81.53 ± 1.73	30.84 ± 2.76	18.53 ± 1.69	15.00 ± 1.13	—
2	81.64 ± 1.53	31.28 ± 2.26	18.59 ± 1.56	15.20 ± 1.03	+0.43
3	90.72 ± 1.17	35.19 ± 2.24	18.89 ± 1.52	15.96 ± 1.14	+4.35
10	— (reference)	40.43	18.62	16.98	+9.58

Table 11: Data levels used for consensus validation and their independence audit.

Data level	Samples	Participants	Sessions	Retention rate	Relationship to other levels
Auditable raw-annotation set	65,456	50	—	—	At least two aligned ratings; count varies
Five-rater repeated subset	953	12	19	—	Dedicated repeated-annotation subset
Ten-rater complete subset	517	11	18	64.06% (517/807)	Fully nested within the five-rater collection

Our validation has three parts: annotator-count resampling tests convergence as more votes are used; a subset audit documents the sampling frame and dependence structure; and synthetic label perturbation tests robustness to random annotation errors.

B.2.1 Consensus-Growth Experiment

The experiment uses 517 complete cases from 11 participants and 18 sessions for which ten valid nine-class ratings are available. These cases were selected from 807 temporally aligned candidates. For each of 30 random repetitions, one-, two-, and three-rater labels are sampled from the same ten-rating pool. The ten-rater plurality is the fixed reference. Two-rater ties and three-rater all-different outcomes are resolved by a seeded random rule solely for this sensitivity analysis; the production annotation pipeline continues to require an absolute majority and adaptive expansion.

To test learnability, frozen Chinese-HuBERT-Large representations and the same linear-classifier protocol are used for all rater-count conditions. The participant-independent split and hyperparameter selection rule are held fixed. Table 10 reports agreement with the ten-rater reference, Accuracy, UAR, and Macro-F1. Values for one to three raters are mean±standard deviation over 30 resamples; the ten-rater condition is deterministic under the fixed labels and training configuration.

Relative to the one-rater condition, three and ten raters improve Accuracy by 4.35 and 9.58 percentage points and Macro-F1 by 0.96 and 1.98 points, respectively. Agreement with the ten-rater reference also rises from 81.53% with one rater to 90.72% with three raters. Because the smaller panels are sampled from the ten-rater pool, these agreement values quantify internal convergence; an independent external-validity design is reserved for future evaluation.

B.2.2 Validation-Subset Composition and Dependence Audit

Table 11 distinguishes the full auditable annotation set from the repeated-rating subsets. The ten-rater complete subset is nested within the repeated-annotation collection, so the analysis is interpreted as a hierarchical internal-convergence evaluation. Reporting this sampling structure defines the population to which the consensus-growth results apply.

The fixed complete-case subset supports a controlled comparison across rater counts because every condition uses the same samples. Its participant, session, and retention statistics are reported so that later work can repeat the analysis with a preregistered sampling frame or disjoint annotator panels.

B.2.3 Noise-Robustness Experiment

We additionally perturb 0%, 10%, 20%, and 30% of individual ratings by replacing selected labels with randomly chosen alternative categories. Labels are then re-aggregated using either five-rater consensus or a single annotator, and the same three fixed model families are evaluated. This experiment isolates sensitivity to random annotation error while holding the sample pool and modeling protocol constant.

Across all perturbation rates, five-rater aggregation yields higher mean Accuracy than the single-annotator condition. The gap grows from 4.04 points without synthetic corruption to 15.67 points

Table 12: Consensus robustness under different levels of synthetic label noise.

Noise rate	Setting	E2V-Large Accuracy (%)	E2V-Base Accuracy (%)	SV-Small Accuracy (%)	Mean Accuracy gain (pp)
0%	Five-rater consensus	86.03	80.67	75.31	+4.04
0%	Single annotator A1	81.91	76.63	71.35	
10%	Five-rater consensus	85.25	79.91	74.56	+9.05
10%	Single annotator A1	76.02	70.86	65.70	
20%	Five-rater consensus	82.86	77.56	72.26	+13.40
20%	Single annotator A1	69.18	64.16	59.14	
30%	Five-rater consensus	79.30	74.07	68.85	+15.67
30%	Single annotator A1	63.31	58.40	53.50	

Table 13: Matched-model comparison between BIIC-Podcast and Ours.

	emotion2vec-Base		emotion2vec-Large		SenseVoice-Small	
Indicators	BIIC-Podcast	Ours	BIIC-Podcast	Ours	BIIC-Podcast	Ours
Accuracy	0.447	0.529	0.470	0.561	0.470	0.626
UAR	0.190	0.251	0.206	0.274	0.180	0.189

at 30% corruption, providing direct evidence that aggregation suppresses independently injected random errors.

B.3 Public-Dataset and Content-Leakage Comparisons

B.3.1 BIIC-Podcast Comparison

BIIC-Podcast is a natural Chinese-language emotional-speech corpus derived from licensed online audio and covers monologues, conversations, interviews, and other heterogeneous contexts (Upadhyay et al., 2023). Table 13 reports a matched-model comparison using emotion2vec-Base, emotion2vec-Large (Ma et al., 2024b), and SenseVoice-Small. Accuracy and UAR are shown because both datasets are single-label multiclass tasks and are strongly imbalanced. The values characterize how the same model families behave under their respective label spaces and protocols.

Ours obtains higher Accuracy for all three matched models and higher UAR for the two emotion2vec variants, whereas the SenseVoice UAR values are close. These differences should be interpreted together with language variety, class definitions, source domains, sample scale, and evaluation splits.

B.3.2 Text-Content Leakage Audit

Elicitation tasks can introduce lexical cues correlated with target emotions. To quantify the relative contribution of acoustic and lexical information, SenseVoice ASR transcripts were generated for the 37,043 strict-intersection samples. All language, emotion, and event control tokens were removed before training a text-only classifier. Text alone reaches 63.54% Accuracy and 36.52% Macro-F1, below the matched audio result of 74.25% and 50.89%. Adding text to audio yields 74.69% Accuracy and 50.67% Macro-F1, a Macro-F1 change of -0.22 points. These results show that the audio representation captures substantial discriminative information beyond lexical content under this protocol.

B.4 Complete Benchmark Experiments

B.4.1 Data Splits and Evaluation Protocol

All experiments use the six audited original batches and their standardized speech segments. Derivative labels, intermediate correction products, and additional batches are excluded. The complete audio manifest contains 68,210 valid segments from 2,406 source recordings and 50 participants. Participant-disjoint training, validation, and test partitions contain 45,208, 6,990, and 16,012 segments, respectively.

Table 14: Modality-coverage audit over the six original collection batches.

Modality or subset	Number of samples	Coverage	Inclusion rule
Audio	68,210	100.00%	Full audio list
Continuous photos	43,207	63.34%	JPG/JPEG
Inertial signals	61,047	89.50%	Aligned to the speech time window
HRV	59,691	87.51%	Recent records do not exceed 120 s
Strict multimodal intersection	37,043	54.31%	Speech, photos, inertial signals, and HRV all available

Table 15: Complete seven-class speech-emotion benchmark.

Model	Adaptation Method	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
Chinese-HuBERT-Large	Light fine-tuning	78.42	52.59	78.04
SenseVoice-Small	Frozen representation + classifier	70.31	43.77	71.93
Whisper-small	Light fine-tuning	76.39	43.00	74.84
XLS-R-300M	Light fine-tuning	76.69	38.52	74.47
HuBERT-Large	Light fine-tuning	74.50	38.08	73.33
WavLM-Base+	Light fine-tuning	72.30	36.09	71.33
PANNs CNN14	Frozen representation + classifier	72.61	32.50	70.31
data2vec-audio-base	Light fine-tuning	69.42	32.05	69.00
AST AudioSet	Frozen representation + classifier	70.06	31.70	69.29
Wav2Vec2-base-960h	Light fine-tuning	69.92	26.03	67.14
emotion2vec+ Large	Frozen representation + classifier	70.63	25.55	66.57
eGeMAPSv02 + SVM	Traditional features + SVM	48.96	24.82	55.21
emotion2vec+ Base	Frozen representation + classifier	63.11	23.77	62.10

Because neutral accounts for 70.24% of the corpus, Macro-F1 and UAR are primary balanced metrics; Accuracy and Weighted-F1 are reported as complementary aggregate measures. Per-class accuracy and class support are included to expose failure modes hidden by the dominant category.

B.4.2 Modality Coverage and Strict Intersection

Coverage is audited by aligning each speech segment to timestamped photos, an inertial window, and the most recent HRV record within 120 seconds. Photos cover 43,207 of 68,210 segments (63.34%), inertial signals cover 61,047 (89.50%), and HRV covers 59,691 (87.51%). The strict intersection contains 37,043 segments (54.31%), split into 23,056/2,994/10,993 train/validation/test samples. Photo availability is the principal bottleneck.

Strict-intersection experiments are complete-case analyses and are therefore reported separately from the full-audio benchmark. The intersection may differ from the full cohort in stage, participant contribution, duration, and class composition; the modality audit and separate audio baselines make this selection boundary explicit.

B.4.3 Speech Baselines

The speech benchmark covers self-supervised speech encoders (Baevski et al., 2020; Hsu et al., 2021), general audio encoders, emotion-specific representations (Ma et al., 2024b), generative speech models, and conventional acoustic features. Chinese-HuBERT-Large is strongest across the reported aggregate metrics, reaching 78.42% Accuracy, 52.59% Macro-F1, and 78.04% Weighted-F1. The lower Macro-F1 of several high-Accuracy systems illustrates why aggregate accuracy alone is insufficient for this long-tailed benchmark.

For an external protocol comparison, the same seven model families are trained on MELD and IEMOCAP. MELD uses the official seven-class train/dev/test split with 13,706 decodable utterances (Poria et al., 2019). IEMOCAP uses the standard four-class, five-fold session-independent protocol over 5,531 utterances with disjoint train, validation, and test speakers (Busso et al., 2008). IEMOCAP values are five-fold mean$\pm$standard deviation. Differences in language, label space, source domain, and protocol prevent direct dataset-quality ranking.

All entries in Table 16 are percentages. The original mixed-precision Wav2Vec2 run on MELD produced NaNs; the table reports the independently retained FP32 rerun.

Table 16: Matched speech-baseline results on MELD and IEMOCAP (%).

Model	MELD			IEMOCAP		
	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
Chinese-HuBERT-Large	51.95	30.07	49.44	66.74±1.85	66.67±2.32	66.17±2.09
WavLM-Base+	48.47	27.56	46.62	65.47±4.43	65.58±5.49	64.77±4.91
HuBERT-Large	51.84	31.15	50.00	69.15±2.12	69.34±1.58	69.11±2.04
XLS-R-300M	53.64	26.26	47.16	57.46±8.08	55.30±8.45	54.44±9.43
data2vec-audio-base	47.20	26.36	45.05	65.43±1.87	65.95±2.03	65.01±1.82
Whisper-small	50.73	30.91	49.44	70.33±1.75	70.90±2.04	70.13±1.95
Wav2Vec2-base (FP32)	48.12	9.28	31.27	42.84±3.34	42.46±4.27	42.47±3.70

Table 17: Per-class accuracy of seven-class speech models.

Model	Neutral (%)	Disgust (%)	Joy (%)	Fear (%)	Sadness (%)	Surprise (%)	Anger (%)
Chinese-HuBERT-Large	89.95	41.69	59.54	29.60	45.94	46.95	47.72
SenseVoice-Small	78.97	21.94	62.36	30.46	39.58	44.37	52.10
Whisper-small	92.24	21.32	45.41	26.51	31.27	23.15	40.32
XLS-R-300M	92.82	1.57	49.56	13.77	21.38	28.46	49.32
HuBERT-Large	88.70	0.94	52.38	10.84	31.27	32.15	52.05
WavLM-Base+	88.81	17.87	38.45	9.81	31.27	31.67	29.04
PANNs CNN14	90.32	3.45	36.97	6.37	30.21	16.08	36.45
data2vec-audio-base	84.93	1.88	41.26	8.43	27.74	33.92	30.07
AST AudioSet	84.92	1.57	45.28	7.23	11.66	32.96	47.04
Wav2Vec2-base-960h	87.92	0.00	37.37	0.00	14.49	7.72	39.98
emotion2vec+ Large	90.04	0.00	38.98	1.03	9.19	10.13	23.50
eGeMAPSv02 + SVM	56.45	16.30	40.05	18.93	24.38	20.90	33.03
emotion2vec+ Base	79.58	0.31	33.42	2.07	10.42	25.56	20.77

B.4.4 Visual and Sensor Baselines

Visual and sensor baselines are evaluated to characterize the information available outside speech. The sensor results are substantially below the speech baseline. Plausible contributors include motion artifacts, participant-specific baselines, physiological response lag, window choice, cross-device alignment error, and mismatch between discrete emotion labels and physiological arousal. These hypotheses motivate signal-quality indices, participant-wise normalization, lag-window sensitivity analysis, and quality-aware fusion in future work.

B.4.5 Multimodal Fusion on the Strict Intersection

All fusion models are trained and evaluated on the same 37,043 complete cases. "Sensors" denotes the joint HRV and inertial representation; "three modalities" denotes speech, photo sequences, and the joint sensor representation. The matched speech baseline reaches 74.25% Accuracy and 50.89% Macro-F1. Speech+vision yields 75.08% and 50.99%, corresponding to a Macro-F1 change of +0.10 points with a 95% participant-level paired-bootstrap interval of $[-0.41, 0.62]$. Speech+sensors changes Macro-F1 by +0.03 points, and three-modality fusion by -0.12 points; all intervals include zero. Intervals use 2,000 participant-level paired resamples, and displayed deltas are computed from unrounded values.

B.4.6 Class-Wise Multimodal Changes

The near-zero aggregate change is accompanied by substantial class-specific shifts. Table 20 therefore presents a class-wise error-redistribution analysis. Relative to speech, speech+vision increases accuracy for all six non-neutral classes and decreases neutral accuracy by 12.99 points. These opposing shifts offset one another in aggregate metrics, showing how class-conditional complementarity can coexist with an aggregate effect close to the speech baseline.

Fusion hyperparameters are selected on the validation set and class-wise values are measured once on the independent test set. The observed differences provide directional evidence of class-conditional complementarity. Future confirmatory evaluation can preregister target classes, report paired error-transfer counts, and apply FDR control or a hierarchical model.

Table 18: Seven-class visual and sensor unimodal baselines.

Modality/model	Evaluation subset	Accuracy (%)	Macro-F1 (%)
DINOv2-Small + temporal aggregation	Photo-covered samples	71.23	18.47
HRV	Strict intersection	6.97	6.51
Inertial signals	Strict intersection	17.24	12.06
HRV + inertial signals	Strict intersection	16.55	12.36

Table 19: Seven-class multimodal results on the strict intersection.

Setting	Accuracy (%)	Macro-F1 (%)	ΔMacro-F1 vs. speech, 95% CI (pp)
Audio	74.25	50.89	Baseline
Visual	23.61	13.48	Baseline
Sensors	16.13	12.22	Baseline
Audio + Visual	75.08	50.99	+0.10 [-0.41, +0.62]
Audio + Sensor	74.19	50.91	+0.03 [-0.43, +0.46]
Visual + Sensor	28.64	14.93	Baseline
Audio + Visual + Sensor	75.08	50.77	-0.12 [-1.10, +0.59]

Adding sensors to speech+vision preserves gains for several categories and produces further class-specific changes, consistent with a modality whose utility depends on class, participant, and signal quality. The present evidence supports visual class-conditional complementarity and identifies sensor information as a more quality- and participant-dependent source of complementary evidence.

B.4.7 Results Discussion

Three factors organize the benchmark results. First, speech is the strongest and most stable modality. Second, multimodal fusion contributes through class-conditional error redistribution. Third, missingness and class imbalance affect interpretation: neutral constitutes 70.24% of all segments, and the complete multimodal cohort covers 54.31% of the audio data. For this reason, we report full-audio and strict-intersection baselines separately and prioritize Macro-F1, UAR, class-wise accuracy, confidence intervals, and class support alongside Accuracy.

B.4.8 Progressive-Stage Curriculum Results

Table 21 compares random mixing with cumulative training in the order individual $\rightarrow$ individual+dyadic $\rightarrow$ all stages. Both settings use Chinese-HuBERT-Large, the same seven-class ontology, and the same participant-independent test set. The curriculum increases UAR by 2.44 points while Accuracy and Macro-F1 change by -1.81 and -0.56 points. The stage order therefore changes the balance between average class recall and aggregate discrimination, demonstrating that the progressive collection structure supports a meaningful curriculum-learning and class-balance diagnostic.

C Extended Limitations and Data Governance

Construct validity. The current evidence validates the reliability, stability, and learnability of perceived-emotion consensus labels. Future work can strengthen construct validity through preregistered incident-level self-report, interaction-partner report, and independent behavioral measures, followed by explicit convergent- and discriminant-validity tests.

Modality missingness and selection. The strict speech–photo–HRV–inertial intersection covers 54.31% of the audio cohort, with photo availability as the primary bottleneck. Complete-case fusion results may therefore reflect both modality value and selection into the intersection. Future releases should report missingness by class, stage, participant, and segment duration, and should evaluate models that remain valid under missing modalities.

Table 20: Per-class accuracy and multimodal changes on the strict intersection.

Class	Speech	Speech + vision	ΔS+V (pp)	Three modalities	Δthree (pp)
Neutral	81.49	68.51	-12.99	69.52	-11.98
Disgust	43.15	45.64	+2.49	45.64	+2.49
Joy	67.76	74.11	+6.35	74.40	+6.64
Fear	43.28	46.94	+3.67	40.10	−3.18
Sadness	44.77	53.41	+8.64	56.14	+11.36
Surprise	52.93	59.57	+6.65	60.11	+7.18
Anger	52.47	53.55	+1.08	54.17	+1.70
Category average	55.12	57.39	+2.27	57.15	+2.03

Table 21: Random-mixing and progressive-curriculum results (%).

Training schedule	Accuracy	Macro-F1	UAR
Randomly mixed baseline	78.42	52.59	51.63
Progressive three-stage curriculum	76.61	52.03	54.07
Change (percentage points)	−1.81	−0.56	+2.44

Physiological and inertial measurement. Wearable signals are affected by motion artifacts, participant-specific baselines, window length, response lag, and device synchronization. Signal-quality indices, participant-wise normalization, lag sensitivity, and quality-aware gating are needed before attributing weak classification performance to the underlying modality.

Class imbalance. Neutral accounts for 70.24% of the corpus, and the rarest emotion classes have substantially larger uncertainty. Dataset users should retain participant-independent splits, report Macro-F1 or UAR together with Accuracy, include per-class support and uncertainty, and avoid changing the natural test distribution when applying class-balanced training.

Multimodal claims. Current fusion experiments identify class-conditional complementarity around an aggregate effect close to the speech baseline. Future work can quantify the generality of these shifts through preregistered target classes, paired error-transfer analysis, multiplicity control, and cross-stage evaluation.

C.1 Ethics, Privacy, and Data Governance

The Personal Data Collection and Processing Authorization Form identifies the institutions participating in this study as the data collectors. At signing, participants are informed of the processing purpose, collection methods and content, scope of use, security measures, retention period, withdrawal route, and contact channels. Participants may refuse to provide the relevant data without adverse consequences, and they may withdraw consent or raise questions about personal-data processing.

The authorized scope covers voice, images, motion, and PPG/HRV information collected through phones and watches. Personal data are restricted to internal use within this project, collected and processed in China, and are not transferred across borders or disclosed or transmitted to other projects, departments, affiliated companies, or commercial partners. The data are stored in a dedicated isolated server area managed by the data-collecting institutions, accessible only to project personnel, with technical and organizational safeguards proportionate to applicable data-protection requirements and processing risks.

The authorization specifies an expected retention period through December 31, 2027. Unless consent is withdrawn earlier, the project will delete or anonymize the personal data after that date. Any extension of use due to a change in the project schedule, or any use by another project, requires renewed participant consent and a newly signed authorization. Accordingly, external release under the current authorization is limited to task definitions, aggregate statistics, model configurations, evaluation code, and documentation that contains no participant-level personal data. Raw speech, images, physiological signals, and participant-linkable derived data are not externally disclosed or

transferred. Any future expansion of data sharing requires renewed authorization and the applicable procedures.