HAF-CLAP: A HIERARCHICAL-AWARE MULTIMODAL CLAP SYSTEM FOR HETEROGENEOUS AUDIO CLASSIFICATION

Technical Report

Xiangyu Jing[1], *Yuandong Luo*[1], *Chaoyong Huang*[1], *Hongqing Liu*[1], *Liming Shi*[1], *Lu Gan*[2]

[1]School of Communications and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
hongqingliu@cqupt.edu.cn
[2]College of Engineering, Design and Physical Science, Brunel University, London, U.K.

ABSTRACT

This technical report describes the system submitted by Chongqing University of Posts and Telecommunications – Audio Lab (CQUPT-AUL) for DCASE 2026 Task 1: Heterogeneous Audio Classification. The proposed system, termed HAF-CLAP, is built upon LAION-CLAP and exploits both acoustic information and textual metadata under the Broad Sound Taxonomy (BST). To adapt pretrained audio-language representations to the target classification task, HAF-CLAP uses a hierarchical-aware multimodal classification framework with audio-text fusion. Several training and inference strategies are further applied to improve robustness and prediction stability. The final submission combines multiple complementary HAF-CLAP models. Experimental results on our internal validation split show that the submitted system achieves competitive performance under the hierarchical evaluation metric.

Index Terms— Heterogeneous audio classification, Broad Sound Taxonomy, HAF-CLAP, CLAP, multimodal learning

1. INTRODUCTION

Pretrained audio-language models provide a strong starting point for heterogeneous audio understanding, but their original contrastive objectives are not directly optimized for fixed-taxonomy supervised classification. In this submission, we develop HAF-CLAP, a Hierarchical Attention-Fusion CLAP system that uses LAION-CLAP as a multimodal backbone and adapts its audio and text representations to the BST classification task. The system is designed to bridge the gap between generic audio-text representation learning and hierarchy-aware classification under the DCASE 2026 Task 1 setting [1, 2].

DCASE 2026 Task 1 focuses on heterogeneous audio classification. The goal is to classify Freesound audio clips into second-level sound categories, where the Broad Sound Taxonomy (BST) consists of five top-level categories and twenty-three second-level categories. Compared with conventional acoustic scene classification tasks, this task covers a broader range of sound content, including musical sounds, instrument samples, speech-related sounds, sound effects, and soundscapes. The audio clips vary significantly in duration, recording conditions, and acoustic content, making robust representation learning important for this task. Another important characteristic of this task is the hierarchical structure of the labels. The official evaluation considers the relationship between predicted and ground-truth categories in the taxonomy. Therefore, errors across different top-level categories are more severe than errors within the same top-level category. In addition, the development data includes both a curated subset and a larger noisy crowd-sourced subset, and provides textual metadata such as titles, tags, and descriptions. These properties naturally encourage the use of multimodal information and hierarchy-aware modeling strategies.

Based on these observations, we build HAF-CLAP upon LAION-CLAP, which provides pretrained audio and text encoders learned from large-scale audio-language data [1]. Instead of treating CLAP embeddings as final classification outputs, we use them as multimodal feature representations and learn an additional classification module for the twenty-three BST second-level classes. The proposed framework combines acoustic and textual information and introduces hierarchy-aware feature modeling to better adapt the pretrained representations to the challenge task.

To further improve classification accuracy and prediction stability, we apply several training and inference strategies during system development. These strategies include hierarchy-aware supervision, feature-level regularization, robust model averaging, and the combination of complementary model variants. The final submitted system is obtained by averaging the prediction probabilities from multiple HAF-CLAP models. This design aims to improve robustness on heterogeneous and noisy audio data while keeping the system description concise.

2. DATASET AND TASK SETUP

DCASE 2026 Task 1 is based on BST, a two-level taxonomy for heterogeneous sound classification. The task provides two official development datasets derived from Freesound: BSD10k-v1.2 and BSD35k-CS. BSD10k-v1.2 is a curated subset with manually refined annotations, containing approximately 11,000 single-channel audio clips and about 35 hours of audio. BSD35k-CS is a larger crowd-sourced subset, containing approximately 35,000 audio clips and about 150 hours of audio. Compared with BSD10k-v1.2, BSD35k-CS provides more training data but also contains noisier user-provided labels.

Each audio clip is associated with a second-level BST label. The taxonomy contains five top-level categories and twenty-three second-level categories, covering music, instrument samples, speech, sound effects, and soundscapes. The official prediction target is the second-level category, while the corresponding top-level category is derived from the BST hierarchy and used by the hierarchical evaluation metric.

In addition to audio signals, the development datasets provide metadata fields such as Freesound sound ID, uploader information, license, title, tags, and description. These metadata fields provide useful semantic information for heterogeneous audio classification, especially for categories that are acoustically ambiguous. Therefore, this task naturally supports audio-based, metadata-based, and multimodal classification systems.

The official ranking metric is the macro-averaged hierarchical F-score(hF). In addition to the original hierarchical Precision, Recall, and F-score metrics (hP, hR, and hF) [3], the task uses a modified scoring rule with a parameter $\lambda = 0.75$. When the predicted second-level category is incorrect but its top-level parent matches the ground-truth top-level category, the prediction receives partial credit controlled by λ. Therefore, this metric encourages systems to perform both coarse-level recognition and fine-grained second-level discrimination.

For internal model selection, we use a fixed validation split throughout all experiments. All validation results reported in this report are computed on this split using the official-style hierarchical metrics. The same split is used for comparing model variants and selecting the final submitted systems.

3. SYSTEM DESCRIPTION

3.1. Overview

The overall framework of the proposed HAF-CLAP system is shown in Fig. 1. HAF-CLAP, short for Hierarchical Attention-Fusion CLAP, is a multimodal classification system built upon LAION-CLAP for heterogeneous audio classification. Different from the original CLAP formulation, which mainly learns a contrastive audio-text embedding space, our system uses CLAP as a pretrained feature extractor and learns a task-specific classification framework for the 23 second-level BST categories [1, 2].

As illustrated in Fig. 1, the system consists of three main stages: feature extraction, feature processing, and classification. The audio and text branches first extract high-level representations from the input audio and textual information. These representations are then transformed and refined by a task-specific feature processing module. Finally, the processed multimodal representation is used for second-level BST classification. During inference, prediction averaging is further applied to improve the stability of the final output.

3.2. Feature Extraction

LAION-CLAP provides pretrained audio and text encoders learned from large-scale audio-language data [1]. In HAF-CLAP, the audio encoder extracts acoustic representations from input waveforms, while the text encoder extracts semantic representations from textual inputs. During training, the text branch is constructed from the available metadata. During inference, the same multimodal pipeline is retained to keep the system consistent between training and evaluation.

The audio representation captures sound characteristics such as timbre, temporal patterns, and background acoustic information. The text representation provides complementary semantic cues from the available textual information. This multimodal setup is useful for heterogeneous audio classification, since some BST categories may have overlapping acoustic patterns while still being semantically different. At the same time, textual information may be incomplete or noisy, so the audio branch remains essential for robust prediction.

Instead of directly using CLAP similarity scores for zero-shot classification, we use the extracted CLAP embeddings as input features for supervised learning. This allows the system to retain the generalization ability of CLAP while adapting the representation to the fixed BST label space.

3.3. Feature Processing

The feature processing module is the main task-specific component of HAF-CLAP. Since the original CLAP embedding space is optimized for audio-text contrastive learning rather than supervised BST classification, the extracted audio and text embeddings are first mapped into a task-oriented latent space by modality-specific projection layers.

After projection, gated residual refinement is applied to both modalities. The residual connection helps preserve the semantic information inherited from CLAP, while the nonlinear transformation improves the discriminative ability of the features. The gating mechanism further allows the model to emphasize useful feature dimensions and suppress less relevant information for the target taxonomy.

The refined audio and text features are then combined by an attention-based fusion module. Instead of using a fixed concatenation or average of the two modalities, the fusion module adaptively balances the contribution of audio and text information for each sample. This design is useful because the reliability of acoustic and textual cues may vary across different audio clips. The output of this module is a fused multimodal representation for downstream classification.

During training, we also apply feature-level regularization in the CLAP embedding space. Inspired by mixup [4], the model is trained with interpolated feature representations and soft targets. This strategy regularizes the classification neck and smooths the decision boundary without changing the original CLAP pretraining objective.

3.4. Classification

The fused multimodal representation is fed into the final classifier to predict the 23 second-level BST categories. Since the BST labels have a two-level hierarchy, the top-level category is also considered during training. Inspired by label smoothing [5], we use a hierarchy-aware target distribution rather than ordinary one-hot labels. The smoothing is constrained within the same top-level group, which encourages the model to reduce severe cross-parent errors while still learning fine-grained second-level discrimination.

The model is optimized with the classification loss computed from the hierarchy-aware target distribution. During training, exponential moving average weights are maintained and used for inference to reduce checkpoint-level fluctuations. For single-model inference, test-time augmentation is applied to obtain multiple prediction views for each audio clip, and their posterior probabilities are averaged to produce a more stable prediction.

The final submitted system further combines several complementary HAF-CLAP variants. All variants follow the same general framework but differ in training configuration or initialization. The final prediction is obtained by averaging the posterior probability vectors from different models. We use probability-level averaging rather than validation-optimized ensemble weights, which keeps the inference procedure simple and reduces the risk of overfitting the ensemble coefficients to the internal validation split.

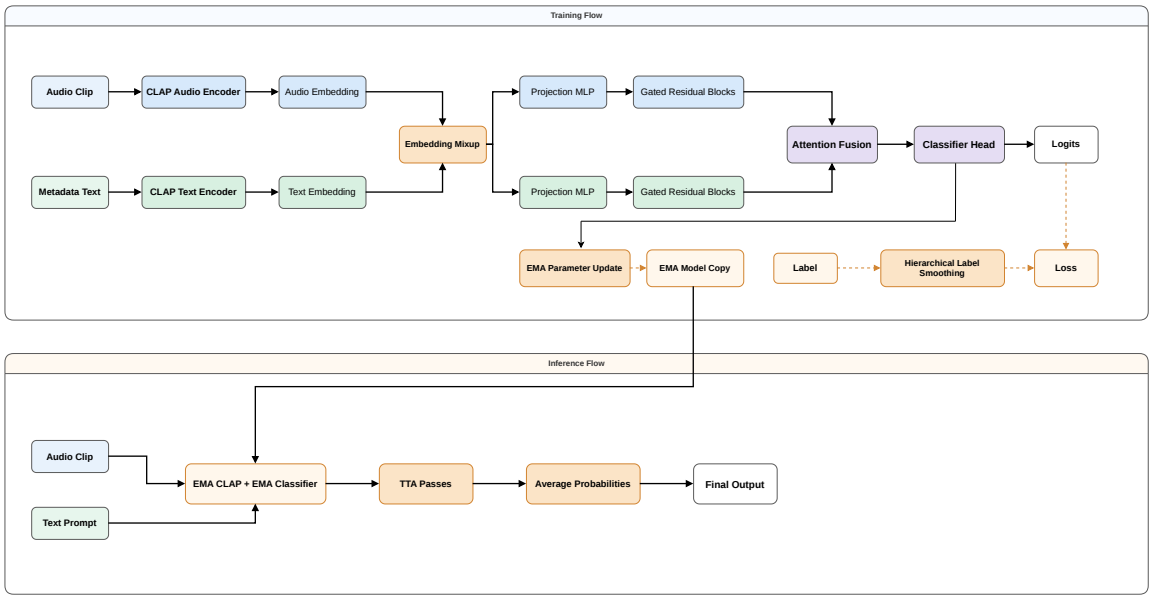

Figure 1: Overall framework of the proposed HAF-CLAP system. The upper part shows the training flow, where audio and textual information are encoded by CLAP and processed by the task-specific feature processing module. The lower part shows the inference flow, where EMA weights, test-time augmentation, and probability averaging are used to obtain stable predictions.

4. TRAINING AND RESULTS

All systems are trained on the official BSD10k-v1.2 and BSD35k-CS development sets. We use the LAION-CLAP checkpoint pretrained on LAION-Audio-630K as the main initialization [1]. The audio encoder is HTSAT-tiny [6] and the text encoder is RoBERTa [7]. For each audio clip, the waveform and metadata-derived text are encoded by CLAP, and the resulting audio and text embeddings are fed into the proposed attention-fusion classifier.

During fine-tuning, the CLAP backbone and the classifier are optimized jointly. We use Adam with a batch size of 64. The learning rate is set to 1×10^{-5} for CLAP and 5×10^{-5} for the classifier. A step learning-rate schedule and early stopping are adopted according to validation performance. Dropout is set to 0.3, and Gaussian feature noise with a standard deviation of 1×10^{-4} is applied to the embeddings.

Several training and inference strategies are used to improve robustness. Hierarchical label smoothing distributes the smoothing mass only among sibling classes that share the same top-level BST parent, with a smoothing factor of 0.1. This preserves the hierarchical structure of the target taxonomy and avoids assigning equal probability to unrelated classes. Embedding mixup is applied after CLAP feature extraction and before the classifier, with $\alpha = 0.2$ and a probability of 0.5. This encourages smoother decision boundaries in the audio-text embedding space.

EMA is used during training with a decay factor of 0.999. Instead of using the latest model parameters for evaluation, the EMA model maintains a smoothed version of the CLAP backbone and classifier parameters. At inference time, the EMA model is combined with TTA, where five stochastic forward passes are averaged to obtain the posterior probability of each class. This reduces the influence of unstable predictions caused by feature noise and random augmentation.

To further improve generalization, we use model ensemble averaging. Each ensemble member follows the same HAF-CLAP framework, but differs in random seed, training strategy, or initialization. Some models are trained from the original LAION-CLAP 630K checkpoint, while others use hierarchical label smoothing, embedding mixup, or different random seeds. In addition, two ensemble members use CLAP checkpoints obtained by continued pretraining on external audio-text datasets before DCASE fine-tuning.

For continued pretraining, we use FSD50K [8] and NSynth [9]. FSD50K provides diverse sound-event recordings with rich metadata, including labels, tags, and textual descriptions. These fields are converted into natural-language captions for audio-text contrastive training. NSynth contains musical notes from different instruments, pitches, and timbres. Its instrument-family and acoustic attributes are used to construct text descriptions. These datasets are selected because they cover sound effects, environmental sounds, and instrument-related content, which are relevant to the heterogeneous BSD taxonomy. The external datasets are used only for representation learning and do not contain any evaluation-set labels.

At inference time, each model produces a 23-dimensional posterior probability vector. For each model, TTA is first applied by averaging five stochastic forward passes. The final ensemble prediction is then obtained by equal-weight averaging, given by

$$\mathbf{p}_{\text{ens}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_m, \tag{1}$$

where M is the number of ensemble members and $\mathbf{p}_m$ is the posterior probability vector predicted by the m-th model. The final class is selected by the maximum value in $\mathbf{p}_{\text{ens}}$.

All experiments are evaluated on a fixed internal validation split using official-style hierarchical metrics. The main metric is macro-averaged hierarchical F-score (hF). We also report ordinary accuracy and top-level accuracy to analyze fine-level and coarse-level classification behavior. Table 1 summarizes the main validation results of different ensemble systems.

Table 1: Internal test results of different submitted systems.

System	M	Fusion	TTA	Acc.
task1-1	3	Equal	5	76.56
task1-2	4	Weighted	5	76.63
task1-3	2	Equal	5	78.92

In Table 1, M denotes the number of models used in each submitted system, Fusion denotes the probability-level fusion strategy, TTA denotes the number of test-time prediction views used by each model, and Acc. denotes the classification accuracy on our internal test set.

The submitted systems are constructed from different model combinations. The task1_1 system consists of three CLAP-based models and uses equal-weight probability averaging. The task1_2 system consists of four CLAP-based models and uses weighted probability fusion. The task1_3 system combines a PaSST model with a Tiny Text-Aux CLAP model, where both models use test-time augmentation and their prediction probabilities are fused with equal weights.

As shown in Table 1, task1_1 achieves 76.56

The experimental results show that model complementarity plays an important role in this task. Although the weighted fusion of multiple CLAP models provides a small improvement, introducing a structurally different PaSST model leads to a more significant gain. Therefore, the final submitted systems focus on combining complementary audio representations and stable test-time prediction averaging.

5. REFERENCES

[1] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," 2022.

[3] S. Kiritchenko, S. Matwin, and A. F. Famili, "Functional annotation of genes using hierarchical text categorization," in *Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 2005.

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[6] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," 2022.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019.

[8] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[9] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.