Feature-Centric Late-Fusion Approach for Heterogenous Audio Classification

Technical Report

Nipun Sharma[*]

Independent Researcher
Madhya Pradesh, India
nipunsharma30100@gmail.com

Aditya Sharma[†]

Independent Researcher
Bangalore, India
adi1129311@gmail.com

Abstract

This technical report describes the submitted system for DCASE 2026 Task 1, Heterogeneous Audio Classification, which is defined over the Broad Sound Taxonomy with 5 top-level and 23 second-level categories. The proposed approach uses a late-fusion architecture built on frozen foundation-model embeddings, combining CLAP audio embeddings, PANNs audio embeddings, and CLAP text embeddings derived from metadata. Each modality is projected into a shared hidden space and fused with a lightweight Transformer encoder. The resulting sequence is then aggregated via mean-pooling to create a unified representation. Training uses a hierarchical objective that combines fine-level cross-entropy with an auxiliary coarse-level loss, together with label smoothing and class weighting. Evaluation is performed using the hierarchical F-score adopted for the task, with additional reporting of top-level and second-level accuracy. For the final submission, logits from seven selected checkpoints were averaged in a checkpoint ensemble to improve robustness and reduce fold-specific variance. The final system utilizes this streamlined three modalities configuration.

Index Terms--- Heterogeneous audio classification, DCASE 2026, Broad Sound Taxonomy(BST), multimodal fusion, CLAP, PANNs.

1. INTRODUCTION

The Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge provides a common benchmark for developing and comparing methods for audio understanding under realistic task constraints. In DCASE 2026, Task 1 focuses on Heterogeneous Audio Classification, a problem defined over the Broad Sound Taxonomy (BST), which contains 5 top-level categories and 23 second-level sound classes. Compared with earlier DCASE Task 1 editions centered on acoustic scene classification, the 2026 task expands the scope toward a broader and more heterogeneous taxonomy of sound content. This setting introduces two main difficulties: first, the target label space spans semantically diverse sound types; second, the evaluation protocol rewards not only exact second-level predictions but also partial hierarchical correctness at the top taxonomy level. As a result, effective systems must capture complementary acoustic and semantic cues while remaining aligned with the hierarchical structure of the label space.

Recent foundation models provide strong pretrained representations for audio and language, making them attractive building blocks for heterogeneous sound classification. Instead of training a large end-to-end model from raw waveforms, this work adopts a feature-centric strategy based on frozen pretrained embeddings, the submitted system combines CLAP audio, PANNs audio, and CLAP text embeddings in a late-fusion architecture that is designed to capture complementary acoustic and semantic information while keeping trainable parameters compact.

The final submission uses a three-modality late-fusion system utilizing frozen CLAP audio, PANNs audio, and CLAP text features. This specific configuration was deliberately selected to maximize parameter efficiency and maintain strict modality alignment without the architectural bloat of additional foundation models. Final predictions for final_submission.csv were obtained by averaging logits from seven selected PyTorch Lightning checkpoints generated during a stratified 5-fold cross-validation procedure. This temporal ensembling approach significantly reduced variance across training folds and ensured high stability at inference time.

The remainder of this report is organized as follows. Section 2 describes the task setup and data processing pipeline, Section 3 presents the proposed late-fusion architecture and training objective, Section 4 summarizes the experimental configuration, and Section 5 discusses the submitted system and its inference strategy.

2. TASK SETUP AND DATA PROCESSING

DCASE 2026 Task 1 addresses heterogeneous audio classification using the Broad Sound Taxonomy (BST), which organizes sounds into a two-level hierarchy consisting of 5 top-level categories and 23 second-level classes. This hierarchical structure is important for both training and evaluation, since the official task emphasizes not only exact fine-grained prediction but also partial correctness at the broader semantic level.

The submitted system is built around precomputed embeddings extracted from the audio files and associated

metadata rather than training a large end-to-end waveform model. During development, the data pipeline supports multiple embedding sources, including CLAP audio, PANNs audio, and CLAP text, all of which are loaded from disk and concatenated into a single feature vector for each sound instance.

The dataset loader reads one or more metadata CSV files and constructs file paths for all required embeddings using the sound_id field. To enforce strict adherence to the fine-grained 23-class problem formulation, a taxonomic filtering step is applied before training. Concretely, samples whose class_idx corresponds to broad top-level parent categories (ending in 00) or ambiguous 'other' categories (ending with 99) are deliberately discarded before model development.

For model development, the pipeline supports either stratified K-fold cross-validation or a stratified holdout split. In the default configuration, stratified 5-fold cross-validation is used so that each split preserves the class distribution of the full dataset. This design is especially important for heterogeneous sound classification, where class imbalance can be substantial across broad semantic categories. To compensate further for imbalance, inverse-frequency class weights are computed from the training partition of each fold and supplied to the fine-grained classification.

The final submission pipeline uses the same embedding-based preprocessing strategy at evaluation time. For the blind evaluation set, CLAP audio, PANNs audio, and CLAP text embeddings are extracted and stored separately before being consumed by the classifier ensemble. This ensures consistency between the development and evaluation pipelines while keeping feature extraction modular and reproducible.

Training-time augmentation is also applied in the embedding space rather than directly on the waveform. The augmentation wrapper can add Gaussian noise and randomly mask embedding dimensions, providing a simple but effective regularization for the fused representation. Together, these preprocessing choices produce a clean and flexible multimodal input pipeline that supports both single-system training and ensemble-based evaluation.

3. PROPOSED METHOD

The submitted system is based on a late-fusion architecture that combines multiple frozen foundation-model embeddings into a single classifier. Rather than fine-tuning large pretrained backbones jointly, each modality is first converted into a fixed-length representation, and only the fusion network and classification layers are trained. This design keeps the trainable model compact while allowing complementary acoustic and semantic information to be integrated for heterogeneous sound classification.

3.1. Multimodal embedding representation

The full development pipeline supports three embedding sources: CLAP audio embeddings of dimension 512, PANNs audio embeddings of dimension 2048, and CLAP text embeddings of dimension 512. These embeddings are extracted independently using frozen pretrained models and stored offline, after which they are loaded and concatenated into a single vector during training and inference.

CLAP audio embeddings provide semantically informed acoustic representations, while PANNs embeddings contribute complementary audio-tagging features from a pretrained Cnn14-based model. CLAP text embeddings are extracted from metadata text and serve as an additional semantic modality aligned with the target taxonomy.

3.2. Late-fusion classifier

Let $z^{(c)} \in \mathbb{R}^{512}$, $z^{(p)} \in \mathbb{R}^{2048}$, and $z^{(t)} \in \mathbb{R}^{512}$ denote the CLAP audio, PANNs audio, and CLAP text embeddings, respectively. Each modality is first projected into a shared hidden space of dimension 512 by a modality-specific linear layer. The resulting projected embeddings are treated as a short sequence of modality tokens and passed through a lightweight Transformer encoder with one encoder layer, 8 attention heads, and a feed-forward dimension equal to four times the hidden size.

After Transformer fusion, the output tokens are averaged across the modality dimension to obtain a pooled representation. This pooled vector is then processed by a classification head composed of a fully connected layer, layer normalization, GELU activation, dropout, and two residual multilayer perceptron blocks, followed by a final linear layer that outputs logits for the 23 second-level BST classes. The use of mean pooling over modalities provides a simple and stable aggregation strategy that works effectively with a small number of heterogeneous embedding tokens.

3.3. Hierarchical supervision

Because the target label space follows a two-level taxonomy, the model is trained with hierarchical supervision. Each of the 23 second-level classes is mapped to one of 5 coarse parent categories, and the network is optimized with a joint loss consisting of a fine-level cross-entropy term and an auxiliary coarse-level cross-entropy term. Fine-level logits are converted to coarse logits by summing the logits of all child classes that belong to the same top-level category.

The total training loss is defined as

$$L = L_{\text{fine}} + \alpha L_{\text{coarse}}$$

where L_{fine} is weighted cross-entropy with label smoothing over the 23 second-level classes, L_{coarse} is label-smoothed cross-entropy over the 5 top-level categories, and α is the coarse-loss weight. In implementation, inverse-frequency class weights are used only for the fine-level term, while the coarse-level loss is left unweighted because the number of parent categories is small and more balanced. This hierarchical objective encourages the model not only to predict the exact class correctly, but also to preserve broader semantic consistency when fine-level discrimination is difficult.

3.4. Regularization

To mitigate overfitting, regularization mechanisms are actively employed at both input and loss levels. At the embedding level, the dataloader applies random dimension masking with masking probability 0.1, encouraging the classifier to rely on distributed multimodal cues rather than memorization of specific dimensions. Furthermore, label smoothing (0.1) is applied consistently to both fine- and

coarse-level cross-entropy loss terms to prevent overconfidence and improve generalization.

3.5. Final ensemble inference

The final final_submission.csv is generated using the 3-modality late-fusion model and a 7-checkpoint ensemble. For each evaluation sample, logits are computed independently by seven selected PyTorch Lightning checkpoints and then averaged before decoding the final class prediction. The selected checkpoints exhibited validation hF scores ranging from 0.733 to 0.838.

This checkpoint-level logit averaging reduces sensitivity to fold-specific variation and improves prediction robustness over a single model. The resulting ensemble serves as the final submitted system for DCASE 2026 Task 1.

4. EXPERIMENTAL SETUP

4.1. Data split and validation protocol

The experiments follow the DCASE 2026 Task 1 development setting for heterogeneous audio classification, while the blind evaluation set is reserved for final submission generation. During model development, stratified cross-validation is used so that each fold preserves the class distribution over the 23 second-level BST categories. In the main training pipeline, the datamodule supports both a stratified holdout split and stratified K-fold cross-validation; the final system uses the cross-validation regime together with checkpoint ensembling for robust model selection.

The final submission system is based on checkpoints obtained from a stratified 5-fold training procedure, followed by selection of seven high-performing checkpoints for logit averaging at inference time. This strategy allows the model to exploit multiple training trajectories while reducing sensitivity to the variance of any individual fold. The evaluation embeddings for the blind set are extracted separately and processed using the same modality-specific feature pipeline employed during development.

4.2. Training configuration

The embedding-fusion classifier is trained in PyTorch Lightning. The main hyperparameters exposed by the training script include a maximum of 120 epochs, batch size 256, initial learning rate 1×10^{-3}, dropout 0.3, label smoothing 0.1, and a coarse-loss weight of 0.3. Optimization is performed using AdamW with weight decay 1×10^{-4}, while learning-rate adaptation is handled by ReduceLROnPlateau monitored on the validation hierarchical F-score val_hF. A linear warmup schedule is applied during the first 10 training epochs, and gradient clipping with value 1.0 is used to stabilize optimization.

4.3. Preliminary Design

Preliminary experiments also considered AST and WavLM embeddings in addition to the final three-modality configuration. However, these additional branches increased model complexity and did not improve validation performance, and were therefore excluded from the submitted system. Alternative aggregation with a learnable CLS token was also explored, but mean pooling over modality tokens provided more stable validation behavior. The training framework further supports Gaussian-noise augmentation and mixup, but the final system relied on random dimension masking and label smoothing as the primary regularization mechanisms.

4.4. Evaluation metrics and model selection

Model selection is based primarily on the hierarchical F-score val_hF, which is also the metric used for checkpoint monitoring during training. The implementation follows a hierarchical precision/recall/F-score formulation in which partial credit is assigned when the predicted class matches the ground truth at the top BST level but not at the second level, using $\lambda = 0.75$. In addition to hierarchical F-score, the framework computes top-level accuracy and second-level accuracy for further analysis of taxonomy-aware behavior.

5. CONCLUSION

This technical report presented the submitted system for DCASE 2026 Task 1, Heterogeneous Audio Classification, formulated over the Broad Sound Taxonomy with 5 top-level and 23 second-level categories. The final submission uses a 3-modality late-fusion architecture based on frozen CLAP audio, PANNs audio, and CLAP text embeddings, followed by a lightweight Transformer fusion module and a hierarchy-aware classification head.

The proposed system was designed to align closely with the hierarchical nature of the task. In particular, the combination of fine-level and coarse-level supervision, together with hierarchical evaluation using macro-averaged hF, provides a natural framework for handling semantically related errors within the BST label space. Development choices further showed that careful modality selection was more effective than simply increasing the number of input branches. The final prediction pipeline also employs a 7-checkpoint ensemble, where logits from multiple selected models are averaged before generating the submission output. This strategy improves robustness by reducing variance across folds and training trajectories, which is particularly useful in challenge settings where stable hierarchical ranking behavior matters more than isolated single-model peaks.

Overall, the submitted approach demonstrates that compact multimodal fusion of frozen foundation-model embeddings is an effective strategy for heterogeneous audio classification. Future work may focus on more explicit taxonomy-aware fusion, stronger per-class calibration, and systematic ablation analysis across modality combinations and ensemble strategies.