ResiDual for Audio: Spectral Alignment in CLAP

August 15, 2025

Arianna Paolini

Abstract

In cross-modal transformer architectures, the outputs for various downstream alignment tasks are often obtained by training a linear transformation on top of the transformer's residual stream. However, recent findings suggest that the residual representation itself may already encode the information required to solve such tasks. This opens the possibility of using the residual stream as a general-purpose data embedding that, when properly modulated, can enhance the performance of a model without the need to fine-tune all its parameters. Building on this idea, the present work adapts the ResiDual spectral reweighting method (Basile et al., 2025) to the audio domain, by applying it to the CLAP model (Wu et al., 2024). A series of experiments were conducted to analyze the intrinsic structure of attention representations and to evaluate the effectiveness of injecting ResiDual modules in different layers of CLAP's architecture. The project code can be found at: `https://github.com/arianna011/Audio-ResiDual`.

1. Introduction

The ResiDual method introduced by (Basile et al., 2025) proposes a spectral reweighting of the Principal Components of transformer residual units, in particular of attention heads, whose observed low-dimensional structure likely reflect specialization on different input or task-specific features. By amplifying the effect of task-aligned components while suppressing task-irrelevant or noise ones, ResiDual can improve the model's performance in an interpretable and computationally efficient way. The method proved effective in the vision domain, when applied to models like CLIP, BLIP, ViT and DINOv2. The goal of this work is to investigate whether ResiDual could also be effectively employed in audio-text architectures, analysing the latent geometry of head manifolds generated by data in the audio domain and experimenting with direct application to the CLAP model (Wu et al., 2024).

Email: Arianna Paolini <paolini.1943164@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

2. Related Work

Previous work on parameter-efficient task adaptation of general purpose architectures involve the injection of trainable bottleneck adapters, both for NLP (Pfeiffer et al., 2020) and cross-modal vision-text tasks (Ebrahimi et al., 2024), into large pretrained transformer layers. However, the spectral approach to modality alignment based on residual decomposition proposed by (Basile et al., 2025) significantly differs from such methods, retaining the interpretability benefits of explicit geometric control over the residual representation space. Its effectiveness, nevertheless, has yet to be tested in the audio domain.

3. Method

In the ResiDual framework (Basile et al., 2025), the performance of a frozen zero-shot classifier is improved by modulating the transformer residual stream to maximize the alignment between the residual unit representations (attention heads, MLPs) and the target task subspace (e.g. text embeddings corresponding to class labels).

In this work, ResiDual spectral reweighting units are injected into the HTSAT model (Chen et al., 2022), CLAP's audio encoder. Its structure employs L Swin Transformer layers, each featuring a different number $depth_l$ (for $l \in [1, L]$) of Swin Transformer blocks (Liu et al., 2021) and a Patch-Merge layer which fuses adjacent latent tokens, to reduce sequence length and increase the embedding dimension, thus achieving a hierarchical structure. Transformer blocks leverage a Window Multi-Head Self-attention (Window-MSA) mechanism to optimize computational cost by only calculating attention on local windows, which are shifted and merged across layers, as better discussed in the Appendix (5.1).

In order to apply the ResiDual method to such complex structure, the following steps were taken:

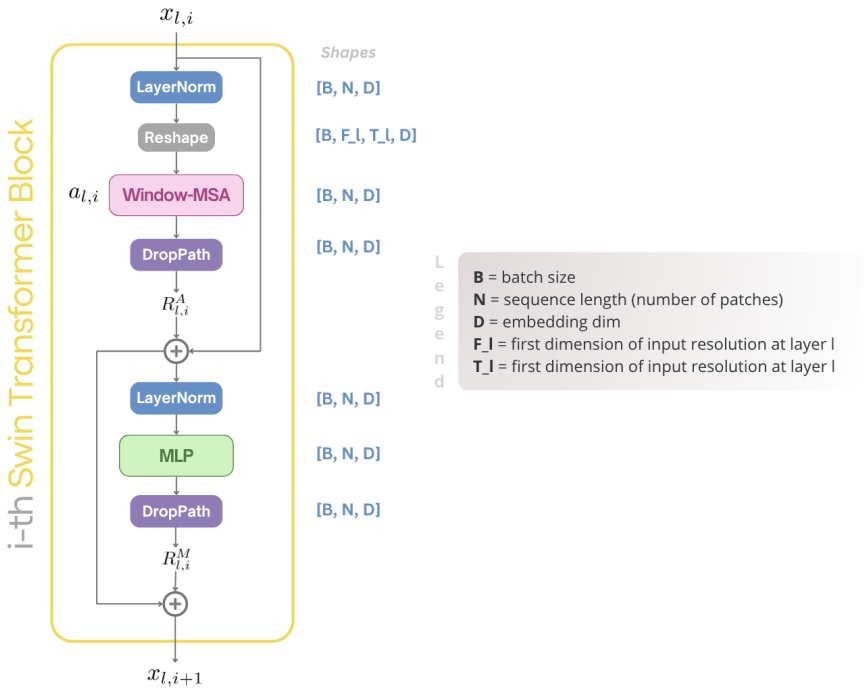

Figure 1. Representation of the i-th HTSAT Swin Transformer block at a generic layer l

1. given a list $layers$ of Swin transformer layers in which to inject ResiDual, attention representations $R^A_{l,i}$ (refer to Figure 1) are collected from each block $i \in depth_l$ for every $l \in Layers$ and aggregated into a single tensor R^A_l by concatenation on the sequence length dimension;

2. PCA decomposition is applied to R^A_l for every $l \in Layers$ to get a corresponding principal component basis ϕ_l and mean μ_l;

3. ResiDual anisotropic rescaling is applied to the attention representations R^A_l for every $l \in Layers$ independently, with λ being a learnable parameter intended to reweight the principal components of residual representations based on their task alignment:

$$RD_{\phi_l,\mu_l}(R^A_l, \lambda_l) = \phi_l^{-1}\, diag(\lambda_l)\, \phi_l\, (R^A_l - \mu_l)^T = Y^A_l$$

4. the rescaled residual representations Y^A_l from each layer $l \in Layers$ are used to reconstruct the transformer residual stream and allow the training of the ResiDual parameter λ on supervised audio classification.

The choice of considering only the attention representations R^A_l, rather than also including the MLP outputs R^M_l, is motivated by the observation that attention heads often produce low-dimensional, semantically specialized subspaces (Voita et al., 2019), thus being more suitable to interpretation and modulation with ResiDual. In contrast, MLP outputs tend to be higher-dimensional and more entangled, which makes them harder to target with ResiDual's spectral reweighting.

Moreover, ResiDual is applied at the granularity of transformer layers to enable coherent modulation of the entire temporal-spectral scale each layer specializes in, while avoiding redundant or conflicting per-block reweighting. This approach also reduces computational cost and mitigates the risk of overfitting.

4. Experimental Results

For all experiments, input data is drawn from the ESC50 dataset (Piczak, 2015), which contains 2000 five-second environmental audio recordings organized into 50 semantical classes, including animal sounds, natural soundscapes and human speech. The dataset was chosen for its relatively small size yet rich diversity of audio categories, enabling meaningful experimentation under limited computational resources.

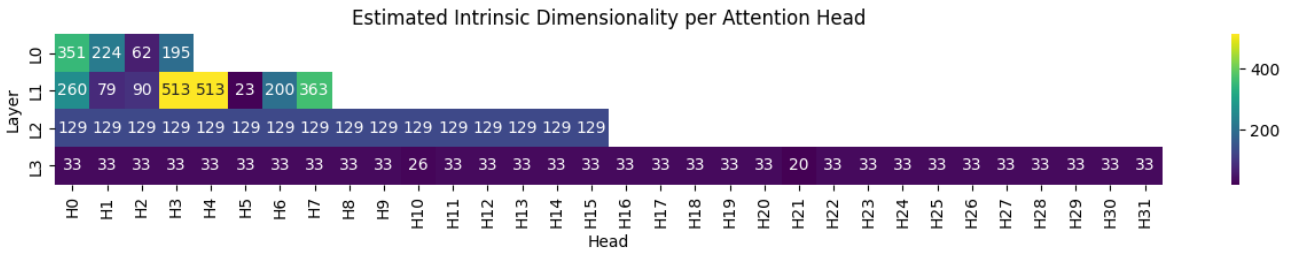

Figure 2. Linear estimation of the intrinsic dimensionality of each attention head in every layer of the HTSAT audio encoder. Values are obtained via PCA as the number of principal components needed to explain 99% of the variance in the data

Table 1. Performance comparison of CLAP pretrained checkpoint (Baseline), CLAP with injected ResiDual units (ResiDual) and CLAP with a linear classifier (Linear) on audio classification for the ESC50 dataset with five-fold cross-validation

	Baseline	ResiDual	Linear
Test accuracy	0.8790 ± 0.0245	0.8685 ± 0.0214	0.9775 ± 0.0073
Trained parameters	0	96	25,650

Before injecting ResiDual into the HTSAT audio encoder, the intrinsic structure of attention heads within the Window-MSA module was analyzed via PCA to assess the presence of head specialization. As shown in Figure 6, heads in the early layers exhibit diverse, high-dimensional structures, whereas in the later layers they collapse to a similar low dimensionality. This pattern suggests that attention in the deeper part of the network focuses on task-aligned features shared across heads, while in early stages heads assume heterogeneous roles, extracting diverse temporal-spectral features or more entangled information, likely reflecting a non-linear structure in some heads. More details can be found in the Appendix (5.2).

Based on this analysis, it could be hypothesized that ResiDual would be most effective when applied to early layers, where attention span richer and higher-dimensional representational spaces and modulation can act as a selective filter, amplifying useful head components so that refined features propagate through the rest of the network. This intuition was confirmed by a W&B sweep that searched for optimal hyperparameters, finding that the best classification accuracy is reached by applying ResiDual only on the first HTSAT layer (Appendix 5.3).

Thus, the ResiDual learnable factors were trained on the ESC50 dataset with 5-fold cross-validation, aiming to achieve supervised parameter-efficient adaptation to audio classification. Training details are provided in the Appendix (5.4). Performance was compared against the original pretrained CLAP checkpoint, which served as a zero-shot baseline, and against a linear classifier trained on top of the frozen HTSAT audio encoder. As reported in Table 1, ResiDual failed to improve classification accuracy in this setting, showing even a slight drop compared to the baseline, whereas the linear projection approach reached an almost perfect score.

This outcome could be explained by the fact that the pretrained cross-modal attention representations were already well aligned, making spectral reweighting more likely to introduce noise than to refine features, or by the extremely limited parameter budget of ResiDual, which may have been insufficient to capture meaningful task-specific information within the short training process. In contrast, the linear projection offered greater capacity to learn dataset-specific patterns, although with a higher risk of overfitting.

5. Conclusions

The ResiDual method, when applied to the CLAP model, was not enough to provide effective task adaptation, likely due to the complex hierarchical HTSAT architecture with shifted-window attention, which makes the encoder significantly different from the vision transformers on which ResiDual was originally evaluated.

Future work could extend ResiDual modulation to MLP representations, apply it to other audio-text architectures, and evaluate its performance in zero-shot and few-shot cross-modal audio tasks on broader and more diverse benchmarks. Despite the limitations observed in this study, the interpretability, parameter efficiency, and strong theoretical grounding of ResiDual still make it a promising direction for advancing lightweight adaptation methods in the audio domain.

References

Basile, L., Maiorca, V., Bortolussi, L., Rodolà, E., and Locatello, F. Residual transformer alignment with spectral decomposition, 2025. URL https://arxiv.org/abs/2411.00246.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection, 2022. URL https://arxiv.org/abs/2202.00874.

Ebrahimi, S., Arik, S. O., Nama, T., and Pfister, T. Crome: Cross-modal adapters for efficient multimodal llm, 2024. URL https://arxiv.org/abs/2408.06610.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. Adapterhub: A framework for adapting transformers, 2020. URL https://arxiv.org/abs/2007.07779.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pp. 1015–1018, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334594. doi: 10.1145/2733373.2806390. URL https://doi.org/10.1145/2733373.2806390.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019. URL https://arxiv.org/abs/1905.09418.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Nezhurina, M., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, 2024. URL https://arxiv.org/abs/2211.06687.

Appendix

5.1. CLAP's Architecture

CLAP's cross-modal architecture (Wu et al., 2024) includes a text encoding branch and an audio encoding one, which can be leveraged to perform audio classification in a zero-shot setting as illustrated in Figure 3. For the default pretrained checkpoint used in this work, the audio encoder f_{audio} corresponds to HTSAT (Chen et al., 2022), while f_{text} corresponds to RoBERTa (Liu et al., 2019).

In order to apply the ResiDual method to CLAP's audio encoder, a deep understanding of the HTSAT structure was needed.

HTSAT (Figure 4) leverages a hierarchical transformer structure to reduce the length of input tokens across subsequent transformer layers, thus optimizing GPU memory consumption and training time. Furthermore, a window attention mechanism replaces global attention to only focus on the relations between nearby tokens, corresponding to a certain range of continuous frequency bins and time frames. Such local attention scores are progressively merged together to build a coherent representation of the whole input audio waveform.

Each HTSAT layer l (for $l \in [1, L]$ with L corresponding to the total number of layers) comprehends $depth_l$ Swin transformer blocks (Liu et al., 2021), followed by a Patch-Merge block that reduces sequence size by merging adjacent patches and passing the result through a linear projection. Every Swin transformer block i in a layer l (for $i \in [1, depth_l]$) processes a batch of input tokens $x_{i,l}$ by applying layer normalization, window multi-head self attention (Window-MSA), dropout on entire sample paths (DropPath) and MLP, featuring skip connections as illustrated in Figure 5.

As already mentioned, the Window-MSA module (Figure 6) splits the 2D patch tokens into non-overlapping attention windows and computes a self-attention matrix for each window rather than globally. Since the Patch-Merge layer will merge adjacent windows going deeper in the network, the receptive field of window attention increases at each layer. The attention weights produced by the different attention heads in the module are analyzed via PCA in this work.

5.2. Analysis of attention heads

The attention weights produced by each head in the Window-MSA attention module were analyzed independently via PCA to study their intrinsic dimensionality and detect whether head specialization, which is considered as a theoretical prerequisite of Residual, actually occurs. In particular, attention matrices $a_{i,l}$ (Figure 1) were extracted for each layer $l \in [1, L]$ of the HTSAT encoder after feeding it with data from the ESC50 dataset (Piczak, 2015). For each layer, attention weights from individual blocks were aggregated as $a_l = \frac{1}{depth_l} \cdot \sum_{i=0}^{depth_l} a_{i,l}$.

The four plots in Figure 8 report the Participation Ratios

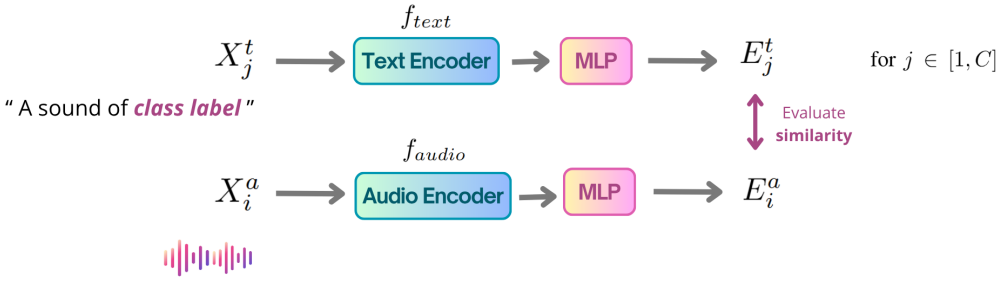

Figure 3. Representation of zero-shot audio classification for an audio sample X_i^a and a text prompt X_j^t corresponding to one of C classes

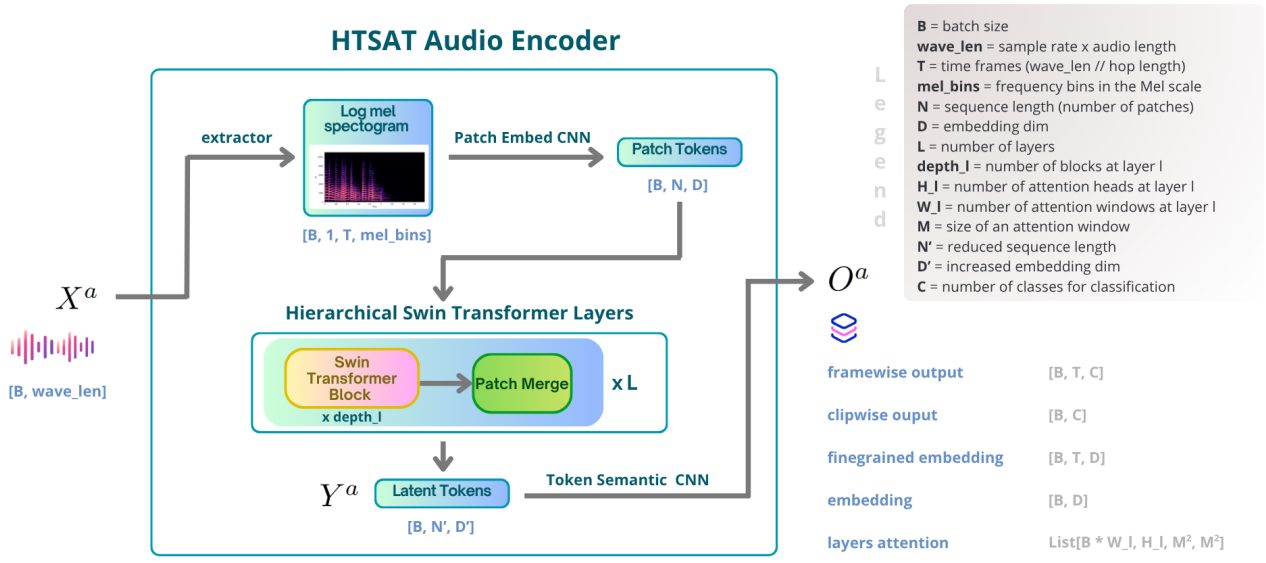

Figure 4. Representantion of data flow in the HTSAT architecture: a batch of audio waveforms X^a is converted into a set of log mel spectograms, which are passed to a Patch Embed CNN that cuts them into $P \times P$ patches and project them to D dimensional vectors to form Patch Tokens. Tokens are fed in order to the transformer layers leveraging window attention and progressively merged together into higher dimensional latent vectors, thus reducing total sequence size. The latent tokens Y^a from the last layer are fed to a Token Semantic CNN that converts them into activation maps which are semantically aware of class labels and produces the final output audio embeddings

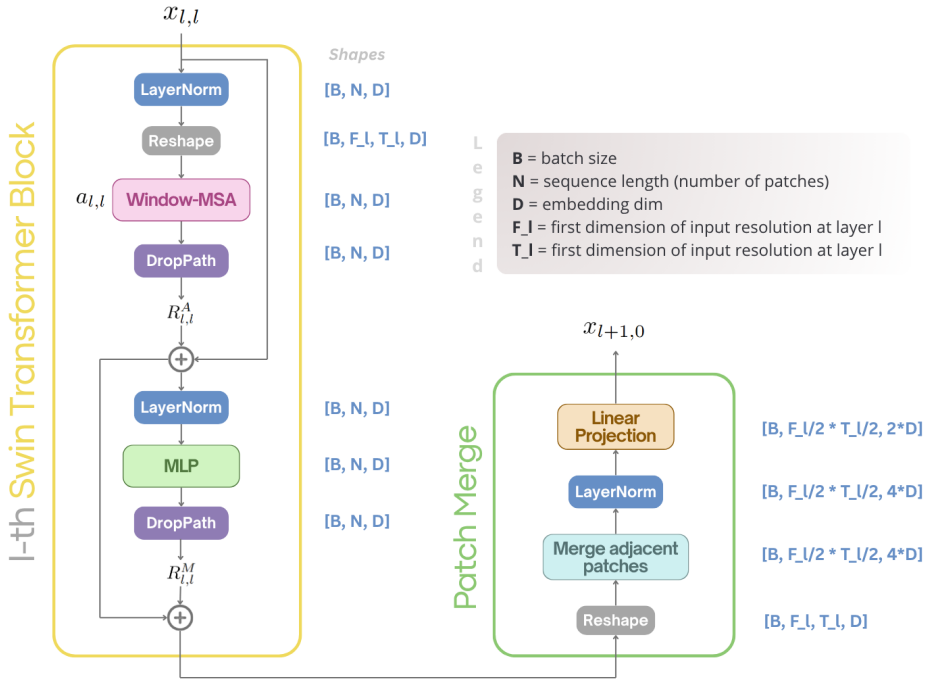

Figure 5. Representantion of data flow in the final block of a layer l in HTSAT: a batch of latent tokens $x_{l,depth_l}$ is passed to a layer normalization unit and reshaped to recover the 2D time-frequency audio structure. The result is fed to the Window Multihead Self Attention module, from which attention head values $a_{l,depth_l}$ can be extracted. The output of Window-MSA is passed through a DropPath unit to form attention representations $R^A_{l,depth_l}$, which are directly summed to the residual stream. The result then flows through another layer normalization module, MLP and DropPath to generate MLP representations $R^M_{l,depth_l}$, which are similarly summed to the residual stream. The output is fed to the Patch Merge block which merges adjacent latent tokens, normalizes them and linearly projects them to a lower dimension in order to produce the input for the next transformer layer $l+1$

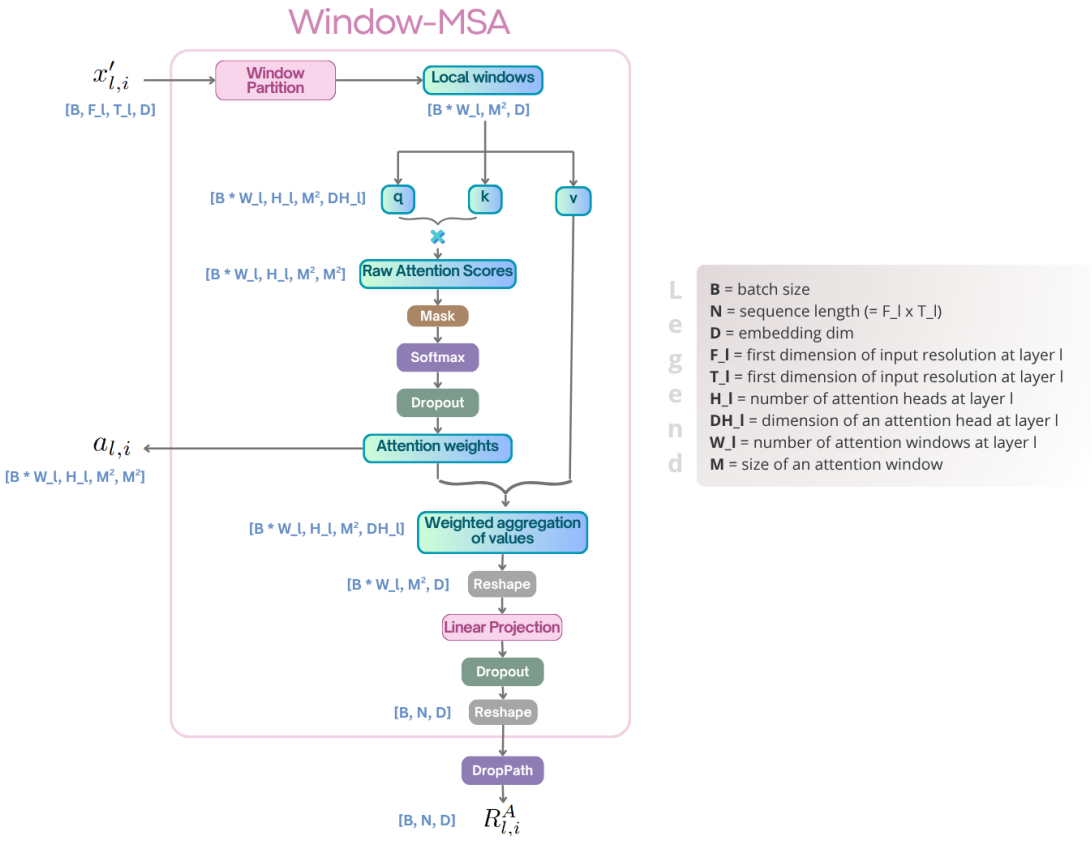

Figure 6. Representantion of data flow in a Window Multi-head Self-attention module in HTSAT: the partially processed input $x'_{l,i}$ of transformer block i at layer l is partitioned into W_l local windows of fixed size ($M \times M$). They are then split into queries, keys and values for H_l attention heads. Queries and keys are multiplied to produce attention weights $a_{l,i}$ for each attention head, which can be extracted to be later analyzed. The result of the weighted sum of the values is passed to a linear projection and a dropout unit, before going through DropPath and generating the final attention representations $R^A_{l,i}$

(PR) of individual attention heads in each layer of the HTSAT encoder. Higher PR values indicate richer and more distributed representations, while lower values suggest that attention is concentrated in a smaller set of dominant components. In layers 0 and 1, heads exhibit a wide spread of participation ratios, with some reaching high peaks while others remain low (especially in layer 1), revealing potential head specialization and heterogeneous representational roles. As depth increases, in layers 2 and 3, both the spread and the magnitude of PR values decrease, implying that the heads converge toward similar, lower-dimensional attention structures.

Figure 7 summarizes the average participation ratio across all heads for each layer, showing an overall decline in intrinsic dimensionality with depth, except for a pronounced peak in layer 1. This confirms the idea that while deeper layers focus on task-aligned features shared across heads, early layers retain greater head diversity and specialization, that could be selectively amplified by interventions such as ResiDual.

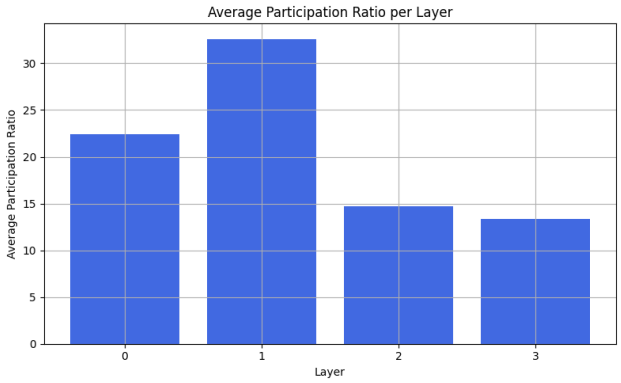

Figure 7. Average participation ratio of attention heads across all layers of the HTSAT audio encoder for the ESC50 dataset

5.3. Sweep results

A grid hyperparameter search was conducted via Weights&Bias to determine the best values for learning rate, number of epochs and target layers of the HTSAT audio encoder for ResiDual injection, using the ESC50 dataset with five-fold cross validation. The search identified a learning rate of 0.01 and 30 training epochs as optimal.

As shown in Figure 9, performance varied considerably across evaluation folds, but the highest median accuracy, with relatively low variance, was achieved when ResiDual was applied only to the first layer. This aligns with previous intuitions that early layers exhibit the richest and most diverse head subspaces, making them promising targets for modulation. In contrast, configurations involving deeper layers (e.g. inject_layers: [2,3] or [3]) yielded lower median accuracy and greater variance, suggesting less consistent benefits and potential instability when modulating low-dimensional, already task-aligned representations. Applying ResiDual to all layers ([0,1,2,3]) resulted in the largest variance and lowest median, indicating that widespread modulation may introduce noise or conflicting updates.

5.4. Training details

The ResiDual learnable parametes λ_l for each layer l in the target injection list $layers$ were trained using the Adam optimizer and cross-entropy loss on the ESC50 dataset with 5-fold cross validation. Training was performed with a learning rate of 0.01 for 30 epochs.

The other two versions of CLAP variants reported in Table 1 are:

- *Baseline*: the default pretrained CLAP checkpoint from `https://github.com/LAION-AI/CLAP`, featuring the HTSAT audio encoder and RoBERTA text encoder, evaluated in a zero-shot setting;
- *Linear*: the frozen original CLAP checkpoint with an added linear classifier on top of the HTSAT audio encoder, trained for 5 epochs and learning rate of 0.01 on the ESC50 dataset. A shorter training schedule was sufficient for the linear classifier, as its larger parameter count allowed it to fit the relatively small dataset faster than ResiDual.

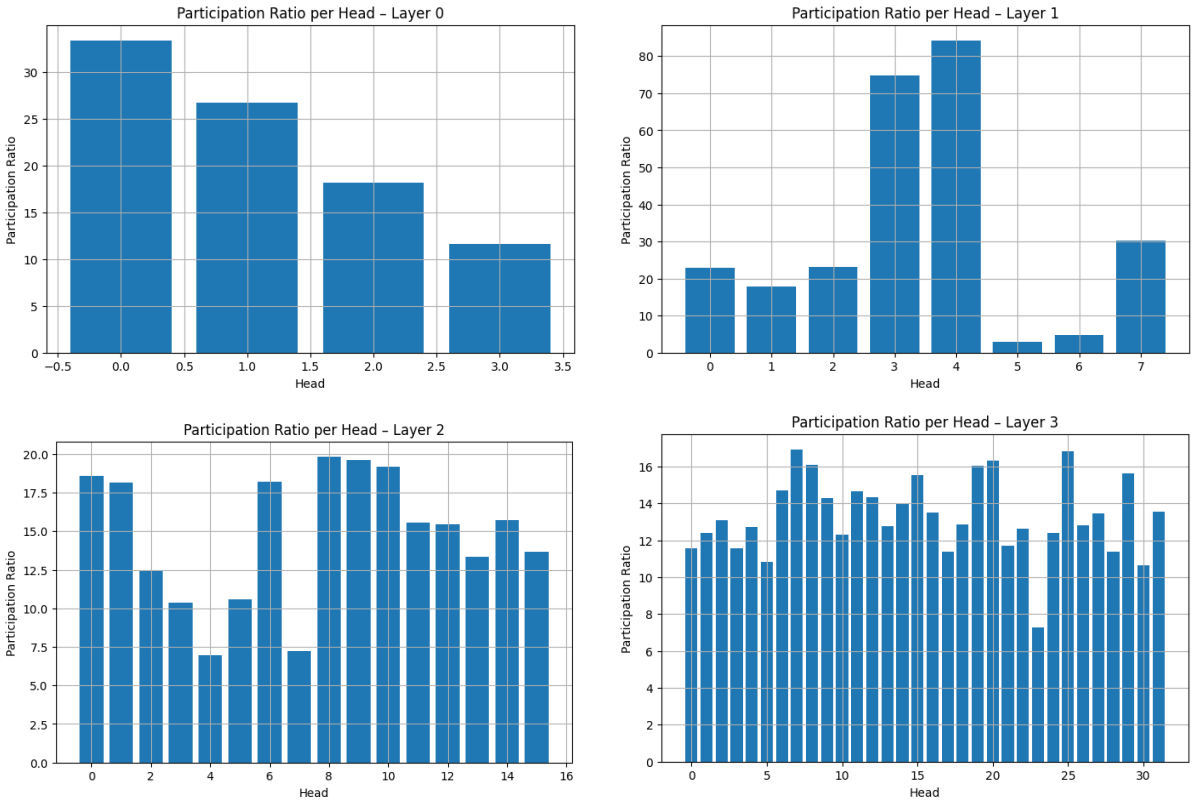

Figure 8. Participation ratios of individual attention heads in each layer of the HTSAT audio encoder for the ESC50 dataset

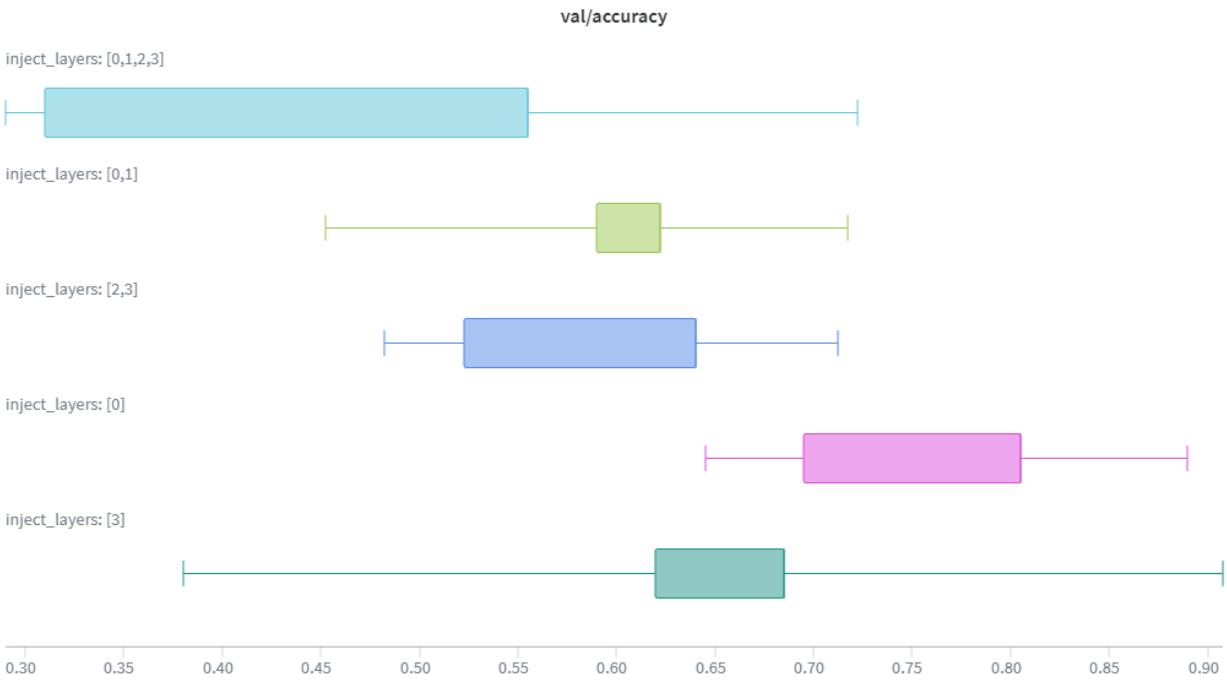

Figure 9. Validation accuracy achieved by ResiDual when injected at different layers of the HTSAT audio encoder