Cross-modal Attention Architectures for Language-Based Audio Retrieval

Óscar Calvet[1], *Doroteo T. Toledano*[1],

[1] AUDIAS, Universidad Autónoma de Madrid, Madrid, Spain

***Abstract*—We present our submission to Task 6 of the DCASE 2025 Challenge on language-based audio retrieval, where our team ranked second overall, with our best single system achieving the fifth-highest score among all individual submissions. Our approach investigates multiple cross-modal architectures, including both standard dual-encoders and attention-based models that leverage fine-grained interactions between audio and text embeddings. All models are trained with contrastive learning on a combination of large-scale captioned audio datasets, using PaSST and RoBERTa as backbone encoders. While each system achieves competitive results on its own, we observe consistent improvements when combining them in an ensemble, suggesting that the architectures capture complementary audio-text relationships. We support this finding with initial representational analyses, which point to differences in how these models structure the shared embedding space. Our results highlight the benefits of architectural diversity in modeling semantic similarity across modalities.**

***Index Terms*—Language-based Audio Retrieval, Audio transformer, Cross-modal attention**

1. INTRODUCTION

Language-based audio retrieval (LBAR) systems aim to retrieve relevant audio recordings from a large corpus based on free-form natural language queries. Unlike traditional sound event detection or tagging systems that rely on predefined taxonomies, LBAR enables flexible and intuitive access to audio content by aligning auditory signals with descriptive semantics. This task remains technically challenging due to the need to bridge heterogeneous modalities—raw waveforms and text—within a unified embedding space where semantic similarity is meaningfully preserved.

At the core of this task lies a challenging multimodal alignment problem: systems must learn to associate raw audio signals with textual descriptions, despite the inherent differences in structure and modality. The dominant approach involves contrastive learning within a dual-encoder framework, where separate encoders project audio and text inputs into a shared embedding space. Relevance is then computed based on the cosine similarity between these embeddings. While this method has proven effective, its reliance on global representations limits its ability to capture fine-grained semantic alignments, such as correspondences between specific acoustic events and words or phrases.

Recent work [1], [2] suggests that richer modeling of cross-modal interactions, particularly via attention mechanisms, can address this limitation by incorporating Token-level correspondences into the retrieval process. Motivated by these insights, we explore a diverse set of architectures for LBAR, including both standard dual-encoders and attention-based models that explicitly model interactions between audio and text embeddings.

Our results demonstrate that while each individual model performs competitively, their combination in an ensemble yields consistent improvements across metrics. We further analyze the representations learned by these systems and find evidence that they capture complementary aspects of audio-text semantics. Our findings underscore the value of architectural diversity in modeling cross-modal semantic similarity and contribute insights towards the development of more robust and expressive LBAR systems.

2. CROSS MODAL ATTENTION IN TEXT TO AUDIO RETRIEVAL

We present a cross-modal attention-based method for text-to-audio retrieval that aims to effectively align textual and audio representations following already existing proposals [1].

Given textual embeddings $T \in \mathbb{R}^{N \times d_t}$ and audio embeddings $A \in \mathbb{R}^{M \times d_a}$ where N denotes the number of embeddings of a sentence, d_t the dimension of the text embeddings, M the number of embeddings of an audio and d_a the dimension of the audio embeddings, we first project them into a common space of dimension d. Specifically, textual embeddings T are projected through a linear transformation to obtain the query vector (Q):

$$Q = W_q T \quad Q \in \mathbb{R}^{N \times d} \tag{1}$$

where $W_q \in \mathbb{R}^{d \times d_t}$ are learned parameters.

Audio embeddings are projected independently using two separate linear transformations to generate key (K) and value (V) vectors:

$$K = W_k A, \quad K \in \mathbb{R}^{M \times d} \tag{2}$$

$$V = W_v A, \quad V \in \mathbb{R}^{M \times d} \tag{3}$$

where $W_k, W_v \in \mathbb{R}^{d \times d_a}$ are also trainable parameters.

Next, we apply multi-head attention, defined as:

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, \dots, \text{head}_h) W^O \tag{4}$$

where h indicates the number of heads, $W^O \in \mathbb{R}^{d \times d}$ are trainable parameters and each attention head is implemented following the standard definition [3]:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

where W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d \times d_h}$ are all trainable parameters and d_h denotes the head size.

The output of the attention module is passed through another linear layer:

$$H = W_o \cdot \text{MultiHead}(Q, K, V), \quad H \in \mathbb{R}^{N \times d} \tag{6}$$

with parameters $W_o \in \mathbb{R}^{d \times d}$ also being trainable.

Finally, we compute the similarity between these refined cross-modal representations H and the original textual embeddings T using the average cosine distance between them as follows:

$$\text{sim}(T, H) = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i \cdot H_i}{|T_i||H_i|} \tag{7}$$

To train our models we rely on the normalized temperature cross-entropy loss [4], which transforms the similarities into conditional probabilities using a temperature-scaled softmax.

3. EXPERIMENTAL SETUP

To evaluate the effectiveness of our retrieval systems and to ensure fair comparison with state-of-the-art (SOTA) methods, we adopt a comprehensive experimental protocol that follows the standards established in recent benchmark efforts. Our setup covers training data, preprocessing, model implementation, and evaluation strategy, including both the conventional Clotho split and the extended evaluation with improved caption-to-audio correspondences.

3.1. Datasets

All models are trained using a combination of four large-scale captioned audio datasets: ClothoV2, AudioCaps, WavCaps, and TACOS. This unified training set provides both human-annotated and synthetic captioning data, covering a broad range of sound types and textual styles, thereby supporting robust and generalizable retrieval capabilities.

ClothoV2 [5] serves as our primary benchmark for evaluation. It consists of audio clips ranging from 10 to 30 seconds, each paired with five natural language captions. The dataset is divided into training, validation, and test sets, comprising 3840, 1045, and 1045 audio files, respectively. We train only on the training subset, monitor performance on the validation split, and report results on the test split. In addition to the standard evaluation setup, we also utilize the improved caption correspondences recently introduced, which provide human-verified many-to-many relevance annotations between queries and audio files. This enhancement allows for a more fine-grained and realistic assessment of retrieval performance.

AudioCaps [6] contributes over 50,000 audio clips, each paired with a single human-written caption. The audio is derived from AudioSet and spans a wide variety of acoustic scenes and events. We aggregate the training, validation, and test splits of AudioCaps into a single dataset, using it as part of the pretraining corpus to increase model generalization.

WavCaps [7] extends the scale of training by providing weakly labeled captions for over 400,000 audio clips collected from multiple online repositories. The captions are synthetically generated using a large language model (GPT-3.5) and capture high-level sound descriptions. Despite their automatic nature, the diversity and volume of WavCaps provide valuable learning information when combined with human-annotated corpora.

TACOS (Temporally-Aligned Audio CaptiOnS) [8] introduces frame-level supervision by aligning captions to specific regions of audio clips. While the dataset includes detailed time-segment annotations, our current experiments use only the weak labels—i.e., global clip-level caption associations—to remain consistent with the global retrieval task. TACOS contains more than 12,000 real-world audio files with approximately 48,000 region-level captions.

To ensure fair benchmarking and avoid evaluation leakage, we follow dataset-specific guidelines for removing overlap with Clotho test sets during training. This step is particularly relevant for synthetic datasets like WavCaps, which aggregate clips from sources such as AudioSet and Freesound.

3.2. Pretrained Embedding Models

Audio is processed using the PaSST [9] transformer encoder, a pretrained audio model based on the vision transformer architecture. PaSST is designed to efficiently handle long audio sequences through patch-wise input and temporal patch dropout. Each audio clip is transformed into a sequence of token embeddings, with one embedding approximately every 10 seconds. During training, audio inputs longer than 30 seconds are truncated, while shorter ones are zero-padded to a fixed maximum duration to ensure consistent batch sizes. The tokens produced by PaSST have length $d_a = 768$.

Text inputs are encoded using RoBERTa-large [10], a transformer-based language model with 24 layers and 355 million parameters. RoBERTa is well-established for Sentence-level and Token-level representation learning and has shown strong results in multimodal retrieval. Captions are normalized to lowercase, stripped of punctuation, and tokenized using the RoBERTa tokenizer. Token sequences are padded or truncated to a fixed length of 32 tokens, which we found sufficient to capture most captions without loss of information. The text tokens produced by the model have length $d_t = 1024$.

3.3. Model Training and Optimization

We trained three different configurations consisting of the standard dual-encoder architecture (hereafter referred to as Dual-encoder) which is exactly the same as the one used for the challenge benchmark [11], a cross-attention system which only used the Sentence-level embedding token from RoBERTa (Sentence-level attention model from now on), and finally another cross-attention system which utilized all the text embeddings (Token-level attention model from now on). We used 8 attention heads and a joint embedding space of dimension 1024.

To properly isolate the impact of architectural design while enabling a fair comparison with state-of-the-art methods, we conducted experiments both using only the Clotho dataset and the full combined training set comprising Clotho, AudioCaps, WavCaps, and TACOS.

We trained both encoders, the embeddings projection layers and the cross-modal attention heads employing the Adam optimizer. When training on Clotho alone, we used a fixed batch size of 64 audio-text pairs; for the combined datasets experiments, batch sizes of 128 were used for the PaSST–RoBERTa Large and Sentence-level embedding systems and 96 for the full text embeddings model due to memory limitations. After one warm-up epoch, a cosine annealing schedule was used to decrease the learning rate from 2×10^{-5} to 10^{-7} over ten epochs. All models are trained end-to-end, including the encoders and cross-modal attention layers where applicable.

After pretraining, fine-tuning with knowledge distillation was performed following the approach of [12]. For models trained on the full dataset, distillation was applied using ClothoV2, AudioCaps, and TACOS weak, with the same training setup. In contrast, models trained solely on Clotho were fine-tuned exclusively on that dataset. In all cases, predictions from the three models were averaged to estimate caption–audio correspondences, using a temperature parameter $\tau = 0.05$ and a loss balancing factor $\lambda = 1$.

4. EMBEDDING SPACE ANALYSIS

To better understand how each architecture encodes and aligns audio and textual information, we analyze the structure of the shared embedding space across the three model variants. For each model, we extract the final representations of audio and text from the ClothoV2 test set and visualize their projections using dimensionality reduction techniques.

In all models, PaSST produces a sequence of patch-based embeddings for the input audio, while RoBERTa yields either a single CLS token (for Sentence-level models) or a sequence of token embeddings (for the Token-level attention model). For the dual-encoder and Sentence-level attention models, a single embedding per modality is directly available or easily extracted, and cosine similarity between these vectors is used for retrieval. However, in the Token-level model, where multiple token or frame-level embeddings exist, we compute the mean of the audio and text embeddings respectively to obtain a single global representation suitable for analysis and comparison.

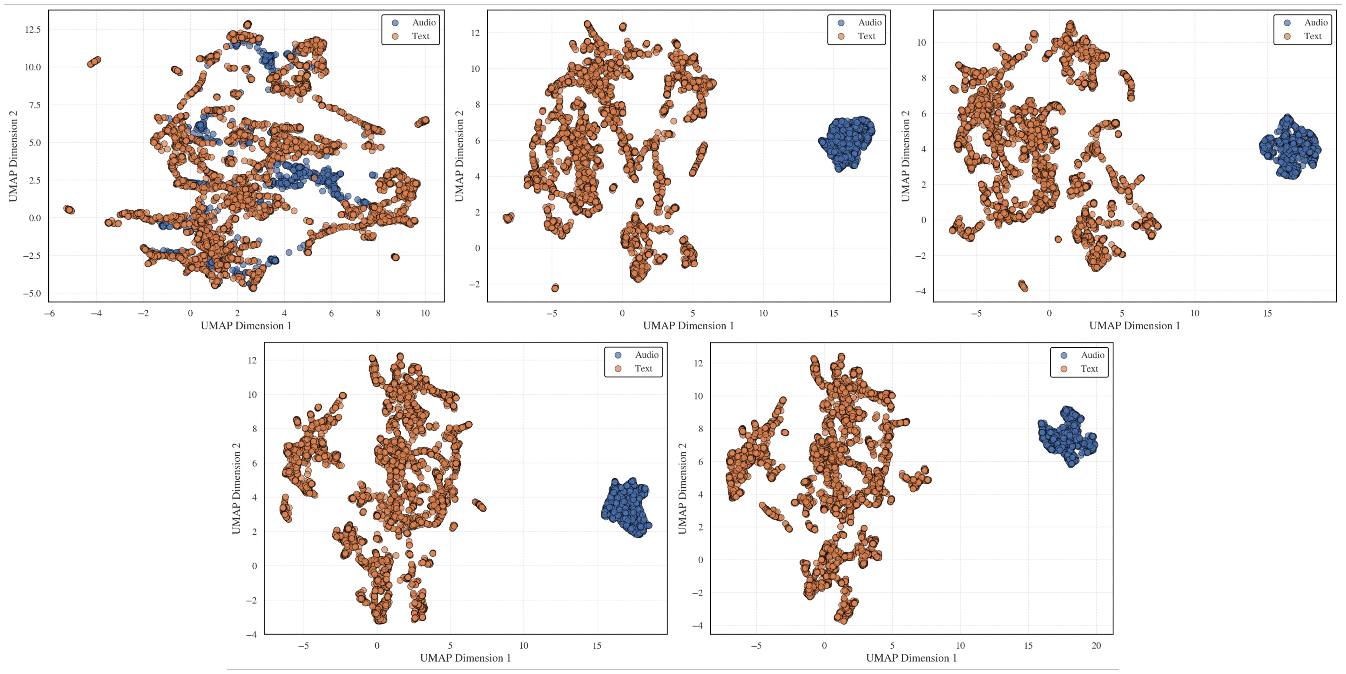

Fig. 1: UMAP projections of the shared embedding space for each model architecture. Top-left: dual-encoder baseline showing highly intermixed audio (blue) and text (orange) embeddings. Top-center and top-right: Sentence-level attention model's key and value projections, respectively, showing strong modality separation. Bottom-left and bottom-center: Token-level attention model's key and value projections, likewise exhibiting distinct clustering by modality. For attention-based models, audio embeddings were projected separately as keys and values; when needed, audio and text representations were averaged to produce global embeddings for visualization.

We analyze the structure of the learned embedding spaces by projecting the global audio and text embeddings from each model into two dimensions using Uniform Manifold Approximation and Projection (UMAP) [13]. These visualizations expose notable differences in how each architecture organizes multimodal data. In the case of the dual-encoder model, audio and text embeddings form relatively well-aligned clusters, with no sharp separation between modalities, indicating that the shared space effectively captures coarse semantic alignment. However, the modality boundaries remain somewhat diffuse, and visual inspection alone is insufficient to evaluate alignment quality. To quantify the degree of semantic correspondence, we perform a statistical comparison of cosine distances between matching audio–caption pairs and randomly sampled mismatched pairs. A two-sample t-test reveals a highly significant difference ($p < 10^{-16}$), confirming that matched pairs are substantially closer in the embedding space, and that the model has learned a discriminative structure that reflects semantic similarity.

In contrast, both attention-based models exhibit clearly segregated clusters for audio and text. We include visualizations of the key (K) and value (V) projections not because we expect them to align directly with the textual embeddings (Q), but to illustrate how the attention mechanism organizes modality-specific intermediate representations. By construction, K and V are audio-derived projections that mediate the interaction with textual queries Q, and their structure is informative about the degree of specialization of the audio pathway prior to attention. This behavior differs from the dual-encoder setup, where training directly enforces cross-modal embeddings to be close in cosine distance. In that case, similarity is imposed at the level of the raw embeddings, which pushes audio and text to occupy a common latent space and leaves little room for modality-specific clustering. By contrast, in attention-based models alignment is not enforced on K and V themselves, but only after the attention operation, through the output H compared to Q. As a result, K and V remain organized in modality-specific subspaces, and alignment with text emerges only after this intermediate transformation. The observed clustering should therefore not be interpreted as evidence against cross-modal learning, but rather as an indication that attention preserves distinct pathways internally while deferring semantic alignment to later stages. More broadly, this suggests that cross-modal attention layers might increase representational capacity by allowing each modality to maintain its own structure while still producing compatible semantic representations when required for training.

These differences suggest that each architecture encodes complementary aspects of cross-modal similarity—global fusion versus local alignment—which likely contributes to the performance gains observed in our ensemble system. Nevertheless, this analysis is an initial exploration and further experiments will be needed to corroborate the observed behaviors.

5. RESULTS & DISCUSSIONS

5.1. Comparison to state-of-the-art systems

To contextualize the performance of our proposed models, we compare them against the current state-of-the-art systems presented at the DCASE 2024 Challenge [14]. The winning system from that challenge achieved a mAP@10 of 41.91, R@1 of 29.33, R@5 of 59.311, and R@10 of 71.923 on the ClothoV2 test set using an ensemble of three distilled models each using a different audio encoder: PaSST [9], AST [15], and MobileNetV3 [16] architectures, and RoBERTa as the text encoder.

In comparison, our best ensemble system, combining Dual Encoder, Sentence-level attention, and Token-level attention models trained with knowledge distillation, achieves 40.423 mAP@10, R@1 of 27.732, R@5 of 58.201, and R@10 of 71.732 when trained on the full dataset. While these results are below the 2024 SOTA across all metrics,

Table 1: Retrieval performance of different model architectures, including a baseline dual-encoder and two attention-based systems (Sentence-level and Token-level), trained with and without knowledge distillation. Results are reported for models trained solely on the Clotho dataset and on a combined dataset (Clotho, AudioCaps, WavCaps, and TACOS). Metrics include mAP@16 and mAP@10 using improved caption correspondences, as well as mAP@10, R@1, R@5, and R@10 for original caption–audio pairs. Distilled models and their ensemble demonstrate the performance gains from architectural diversity and teacher-student training.

	Clotho Only						**All Datasets**					
	Improved Captions		*Original Captions*				*Improved Captions*		*Original Captions*			
	mAP@16	mAP@10	mAP@10	R@1	R@5	R@10	mAP@16	mAP@10	mAP@10	R@1	R@5	R@10
Dual Encoder	34.943	32.532	30.413	18.756	**45.876**	60.268	42.324	39.664	37.389	24.976	54.086	68.172
Sentence-level Attention	35.151	32.730	29.431	18.086	44.708	58.947	42.434	39.827	35.769	23.655	52.766	66.258
Token-level Attention	35.195	32.931	29.582	18.201	45.378	59.273	41.375	38.908	35.770	23.828	52.057	65.742
Knowledge Distillation												
Dual Encoder^{1}	34.354	32.119	29.918	18.220	45.378	59.273	44.203	41.662	38.293	25.263	56.000	69.282
Sentence-level^{2}	33.042	31.009	28.816	17.952	43.254	57.627	43.926	41.332	37.495	24.784	54.947	68.650
Token-level^{3}	33.000	31.034	28.722	17.589	43.502	57.914	42.457	40.061	37.901	25.818	54.43	67.617
Ensemble^{1,2,3}	**35.304**	**33.161**	**30.758**	**19.254**	45.856	**60.785**	**46.864**	**44.176**	**40.423**	**27.732**	**58.201**	**71.732**

Table 2: Retrieval performance (mAP@16 and mAP@10) for the two attention architectures presented on this paper trained with both the text and audio encoders frozen. The models are only trained on the Clotho dataset and evaluated on the improved captions.

Model Architecture	**mAP@16**	**mAP@10**
Sentence-level Attention	23.732	21.901
Token-level Attention	26.130	24.184

Table 3: Retrieval performance (mAP@16 and mAP@10) for pairwise model ensembles using improved caption correspondences. All models are trained on the combined dataset.

Model Ensemble	**mAP@16**	**mAP@10**
Dual Encoder + Sentence-level Attention	46.521	43.894
Dual Encoder + Token-level Attention	46.005	43.323
Sentence-level + Token-level Attention	44.872	42.270

particularly in terms of early precision (R@1), they demonstrate that our approach remains competitive, especially considering that it relies on a single audio encoder (PaSST) across all models and does not leverage multiple specialized backbones.

Although our system does not surpass the 2024 SOTA in raw performance, it achieves strong results without relying on encoder heterogeneity or handcrafted ensemble tuning.

5.2. Importance of Encoder Fine-Tuning

A central design consideration in multimodal retrieval is whether fine-tuning large pretrained encoders is necessary, or whether task performance can be achieved by training only the cross-modal interaction layers. This question is particularly relevant for our attention-based architectures, which preserve distinct embedding spaces for audio and text. One might hypothesize that the attention mechanism alone could serve as a sufficient alignment module.

To test this hypothesis, we conducted experiments where the audio and text encoders (PaSST and RoBERTa) were frozen, and only the projection and attention layers were trained. The results, summarized in Table 2, show a consistent and significant drop in retrieval performance across all architectures when fine-tuning is disabled.

These findings suggest that while attention mechanisms do enable cross-modal interaction, they are not sufficient on their own to fully bridge the modality gap, especially when the encoders remain fixed to their pretraining objectives. Fine-tuning allows the encoders to adapt their representations to the retrieval task and dataset characteristics, resulting in more semantically aligned embeddings and better overall performance.

5.3. Pairwise Ensemble Analysis

To further investigate the individual contributions of each model architecture within the ensemble, we conducted an ablation study by evaluating all possible pairwise combinations of the three systems. Specifically, we assessed the performance of the following two-model ensembles: Dual Encoder + Sentence-level Attention, Dual Encoder + Token-level Attention, and Sentence-level Attention + Token-level Attention. The results are presented in Table 3.

This analysis revealed that the combination of the Dual Encoder and Sentence-level Attention models consistently achieved the highest retrieval performance among the pairwise configurations. This suggests that these two architectures capture highly complementary cross-modal features, likely owing to their distinct alignment mechanisms. These findings are consistent with our embedding space visualizations in 4, which indicate that the Dual Encoder produces a more intermixed embedding space, whereas the Sentence-level Attention model preserves stronger modality-specific structures.

The ensemble composed solely of the Sentence-level and Token-level Attention models yielded smaller gains compared to individual performance. This outcome suggests a degree of representational redundancy, potentially due to the shared attention-based architecture and similar alignment strategies. Overall, these pairwise ensemble results support our central hypothesis: architectural heterogeneity, particularly combining models with fundamentally different cross-modal interaction paradigms, plays a critical role in enhancing retrieval accuracy through ensembling.

6. CONCLUSION

This work explored a set of cross-modal architectures for language-based audio retrieval, including a standard dual-encoder model and two attention-based variants. All models were built using a common set of backbone encoders (PaSST and RoBERTa) and trained with contrastive learning on a diverse collection of captioned audio datasets.

While none of the individual models outperform the current state-of-the-art, they each demonstrated competitive performance. Importantly, combining them in an ensemble led to consistent performance improvements, suggesting that the models capture complementary aspects of the retrieval task. Pairwise ensemble analysis supported this interpretation, with the most notable gains observed when combining the Dual Encoder with the Sentence-level Attention model.

Our results also emphasized the necessity of fine-tuning the pretrained encoders, as freezing them during training significantly reduced retrieval accuracy across all model types.

REFERENCES

[1] Y. Xin, D. Yang, and Y. Zou, “Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss,” 2023. [Online]. Available: https://arxiv.org/abs/2303.05681

[2] Y. Xin and Y. Zou, “Improving audio-text retrieval via hierarchical cross-modal interaction and auxiliary captions,” 2025. [Online]. Available: https://arxiv.org/abs/2307.15344

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[5] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: an audio captioning dataset,” in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2020, pp. 736–740.

[6] C. D. Kim, B. Kim, H. Lee, and G. Kim, “Audiocaps: Generating captions for audios in the wild,” in *NAACL-HLT*, 2019.

[7] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, “Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research,” *arXiv preprint arXiv:2303.17395*, 2023.

[8] P. Primus, F. Schmid, and G. Widmer, “Tacos: Temporally-aligned audio captions for language-audio pretraining,” 2025. [Online]. Available: https://arxiv.org/abs/2505.07609

[9] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, “Efficient training of audio transformers with patchout,” in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. ISCA, 2022, pp. 2753–2757. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-227

[10] 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. [Online]. Available: https://arxiv.org/abs/1907.11692

[11] DCASE Challenge Task 6, https://dcase.community/challenge2025/task-language-based-audio-retrieval/.

[12] P. Primus, F. Schmid, and G. Widmer, “Estimated audio-caption correspondences improve language-based audio retrieval,” 2024. [Online]. Available: https://arxiv.org/abs/2408.11641

[13] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[14] P. Primus and G. Widmer, “A knowledge distillation approach to improving language-based audio retrieval models,” DCASE2024 Challenge, Tech. Rep., June 2024.

[15] X. Li, N. Shao, and X. Li, “Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks,” 2023. [Online]. Available: https://arxiv.org/abs/2306.04186

[16] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, “Searching for mobilenetv3,” 2019. [Online]. Available: https://arxiv.org/abs/1905.02244