CACARA: Cross-Modal Alignment Leveraging a Text-Centric Approach for Cost-Effective Multimodal and Multilingual Learning

Anonymous ACL submission

Abstract

As deep learning models evolve, new applications and challenges are rapidly emerging. Tasks that once relied on a single modality – such as text, images, or audio – are now enriched by seamless interactions between multimodal data. These connections bridge information gaps: an image can visually materialize a text, while audio can add context to an image. Researchers have developed numerous multimodal models, but most rely on resource-intensive training across multiple modalities. Similarly, extending these models to new languages often follows the same resource-heavy training strategy. In this work, we propose a multimodal and multilingual architecture, CACARA, trained through emergent alignment learning, enabling the seamless integration of new modalities into an existing bimodal/multimodal model without requiring full retraining. Likewise, our approach extends the model's linguistic capabilities while preserving previously learned knowledge. Multimodal and multilingual properties emerge through alignment learning, leveraging prior training to enhance and synchronize multiple modalities and languages. Our strategy achieves up to a 14.24 percentage point (pp) improvement in R@1 audio-to-text retrieval, outperforming state-of-the-art multimodal models – all without the heavy computational cost of retraining across every modality and language.

1 Introduction

Deep learning has revolutionized multiple domains by enabling models to learn complex representations across diverse data types. Early breakthroughs in computer vision, driven by convolutional neural networks (Krizhevsky et al., 2012), were followed by advances in natural language processing, culminating in Transformer networks (Vaswani et al., 2017). Beyond images and text, deep learning has achieved state-of-the-art performance in audio (van den Oord et al., 2016), sensor data (Wang et al., 2019), and tabular data (Arik and Pfister, 2021), excelling in classification, retrieval, and generation tasks.

Real-world applications, however, often involve complex interactions between multiple data types. For instance, video understanding encompasses the joint processing of visual and auditory information (Goecke, 2005). These models integrate complementary information from different modalities, as exemplified by CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), which learns a joint representation space for images and text, enabling cross-modal retrieval and zero-shot classification. The benefits of multimodal learning extend beyond simple fusion, uncovering latent relationships and contextual cues that are not apparent in individual modalities (Baltrušaitis et al., 2018). However, training such models is challenging due to the need for synchronized data and the high cost of annotated datasets. One promising approach is implicit learning, where the model implicitly learns cross-modal relationships, even without strict temporal alignment, by leveraging the inherent correlations and statistical dependencies between modalities (Alayrac et al., 2020).

Multilingualism adds another layer of complexity. Supporting multiple languages not only expands accessibility but also enriches models with diverse linguistic structures (Conneau et al., 2020). Yet, low-resource languages remain underrepresented due to data scarcity and limited computational resources (Joshi et al., 2020). Current research disproportionately favors high-resource languages such as English, neglecting the needs of under-represented linguistic communities.

The intersection of multimodality and multilingualism presents both opportunities and challenges. A key concern is the computational cost of training and deploying large-scale models, restricting access to well-resourced institutions (Strubell et al., 2019, 2020). Therefore, there is a pressing need for

innovative training methodologies and model architectures that can effectively leverage multimodal and multilingual data while minimizing computational overhead.

In this work, we introduce a multimodal and multilingual model that addresses these challenges through two key strategies: emergent alignment learning and a modified Locked-image Text Tuning (LiT) protocol. These strategies reduce training costs while preserving high performance.

We develop a new modality integration approach that eliminates the need to retrain all encoders. By optimizing this alignment with English, we demonstrate that emergent alignment also benefits other languages. This aspect is not addressed in previous works. We fine-tune the audio encoder for English synchronization only to enable multilingual capabilities without incurring the high costs of multilingual audio pre-training. The text encoder, meanwhile, remains frozen throughout this process, capitalizing on its inherent cross-lingual capabilities. This enables multilingual audio-text alignment in languages beyond English with training costs comparable to a monolingual model.

Our findings show that multimodal models can learn language-agnostic concepts, improving R@1 text retrieval with audio by up to 14.24 percentage points (pp) and audio-to-text retrieval by 2.58 pp over existing multimodal approaches. Additionally, our method shows how to extend bilingual or multimodal models into a multilingual framework with minimal computational overhead while maintaining performance across modalities. It achieves an average classification accuracy of up to 66.5% across multiple languages without requiring retraining or explicit alignment. These results demonstrate a scalable approach for efficiently integrating multiple modalities and languages.

2 Related Work

Multimodal learning has expanded machine learning's scope, enabling models to process diverse data types. Foundational works like CLIP aligned images and text, inspiring extensions to other modalities such as audio, depth, and multilingual applications. CAPIVARA (Santos et al., 2023), a CLIP-based model, incorporates Portuguese in contrastive training to optimize performance in low-resource languages. Despite progress, challenges remain in efficient training and generalization, particularly in low-resource settings. This section reviews advances in multimodal and multilingual models.

ImageBind (Girdhar et al., 2023) extends CLIP's paradigm by introducing a unified embedding space for six modalities: images, text, audio, depth, thermal, and Inertial Measurement Unit (IMU) data. By leveraging contrastive learning and using images as an anchor modality, ImageBind showed that modalities can be effectively aligned through their natural pairing with images, eliminating the need for exhaustive paired data between all modality combinations. This approach achieves emergent cross-modal alignment without explicit supervision, demonstrating strong zero-shot transfer, enabling cross-modal retrieval and multimodal embedding arithmetic.

LanguageBind (Zhu et al., 2023) replaces images with language as the central modality for aligning different data types. Leveraging language's rich semantic structure, it aligns modalities within a shared embedding space using a frozen language encoder pre-trained on video-language data and contrastive learning for other modalities. Efficient training is achieved through Low-Rank Adaptation (LoRA) (Hu et al., 2022), demonstrating strong performance across video, audio, depth, and infrared modalities. LanguageBind outperforms ImageBind in infrared, depth, and audio classification tasks.

Vision-Audio-Language Omni-peRception (VALOR) (Liu et al., 2024) advances multimodal research by integrating vision, audio, and language within a tri-modal framework. It introduces two pretext tasks: Multimodal Grouping Alignment for fine-grained modality alignment and Multimodal Grouping Captioning for text generation based on different modality combinations. VALOR established robust alignment between modalities and support tasks such as retrieval, captioning, and question-answering.

Vision-Audio-Subtitle-Text omni-modality foundation model (VAST) (Chen et al., 2023b) expands multimodal learning by integrating vision, audio, subtitles, and text into a unified framework. By integrating subtitles and auxiliary modalities, VAST addressed the limitations of prior works, which often overlooked the role of additional information streams in video understanding, highlighting the importance of datasets and models that utilize multiple complementary sources of information.

Multilingual Multimodal Pre-training (MLMM) (Zhang et al., 2023) advances multilingual multimodal pre-training by addressing the predomi-

nance of English in existing models. It combines pre-training-based and generalization-based approaches. For pre-training, MLMM leverages large-scale multilingual image-text datasets with texts translated into multiple languages. It employs four key pre-training objectives: Image-Text Matching for coarse-grained alignment, Masked Language Modeling for fine-grained cross-modal understanding, Masked Region Feature Regression, and Masked Region Classification for vision-language alignment. The model demonstrates strong cross-lingual transfer ability, particularly when fine-tuning with languages from the same language family. MLMM also explores generalization-based approaches through multilingual knowledge distillation and multilingual acquisition as resource-efficient alternatives, achieving state-of-the-art performance across multilingual vision-language tasks while maintaining deployment flexibility.

As evidenced by the literature, the proliferation of both multimodal and multilingual models is undeniable. However, models that truly excel at simultaneously integrating multimodality and multilinguality remain relatively scarce. Moreover, the few existing models that attempt this integration are often plagued by substantial training costs, long training times, and significant computational demands. CACARA directly addresses these limitations by introducing a novel, efficient approach based on an implicit learning strategy. This strategy not only facilitates seamless transfer learning across diverse modalities but also enables robust cross-lingual performance. This approach allows CACARA to be trained exclusively on English data while still achieving strong performance on tasks in all supported languages, thus drastically reducing the resource requirements typically associated with multilingual multimodal models.

3 Methodology

This section presents the CACARA model's overall framework, covering its architectural design, training and evaluation datasets, and selected hyperparameters. We then detail the training workflow, highlighting the emergent alignment strategy responsible for its multimodal and multilingual capabilities.

3.1 CACARA Model

The CACARA model integrates multimodal and multilingual learning through three encoders: image, text, and audio (Figure 1). The image encoder is based on a Vision Transformer (ViT) (Dosovitskiy et al., 2021), while the text encoder utilizes XLM-RoBERTa (base version) (Conneau et al., 2020). The audio encoder, built upon BEATs (Chen et al., 2023a), is incorporated via emergent alignment learning. The image and text encoders are initialized with pre-trained OpenCLIP (Ilharco et al., 2021) encoders.

Training is performed using contrastive learning with the InfoNCE (van den Oord et al., 2018) loss function to align the audio and text encoders. We keep the image and text models frozen to preserve the high-quality image-text alignment from OpenCLIP pre-training, optimizing only the audio encoder's alignment within the shared feature space. A key aspect of CACARA's training procedure is that, unlike LiT, which adapts a text encoder for downstream tasks while keeping the image encoder fixed, CACARA maintains the pre-trained image-text representation and aligns the new audio encoder to this existing joint space.

3.2 Multimodal and Multilingual Emergent Aligment

CACARA training leverages the emergent learning capacity of pre-trained models. Since the image and text encoders are pre-aligned in a shared feature space due to their pre-training, training the newly incorporated audio model against the text encoder implicitly aligns it with the shared space of all three. This emergent alignment enables the new audio model to describe untrained image modality features without training.

The text modality is the primary anchor for emergent learning within this architecture. As depicted in Figure 1, solid lines indicate direct training pairings (audio-text data presented to the model), while dotted lines indicate emergent learning, the emergent relationship between audio and image.

Due to the text encoder's role as anchor and the freezing of pre-trained text and image encoders, it remains fixed, synchronizing text with additional modalities. This eliminates the need for explicit multilingual training, as XLM-RoBERTa's inherent multilingual capabilities extend to additional modalities through emergent alignment. Integrating a new modality synchronizes it with the multilingual features of the textual model.

This strategy enables non-multilingual models, such as BEATs, to acquire linguistic capabilities from the text model. As illustrated in Figure 1,

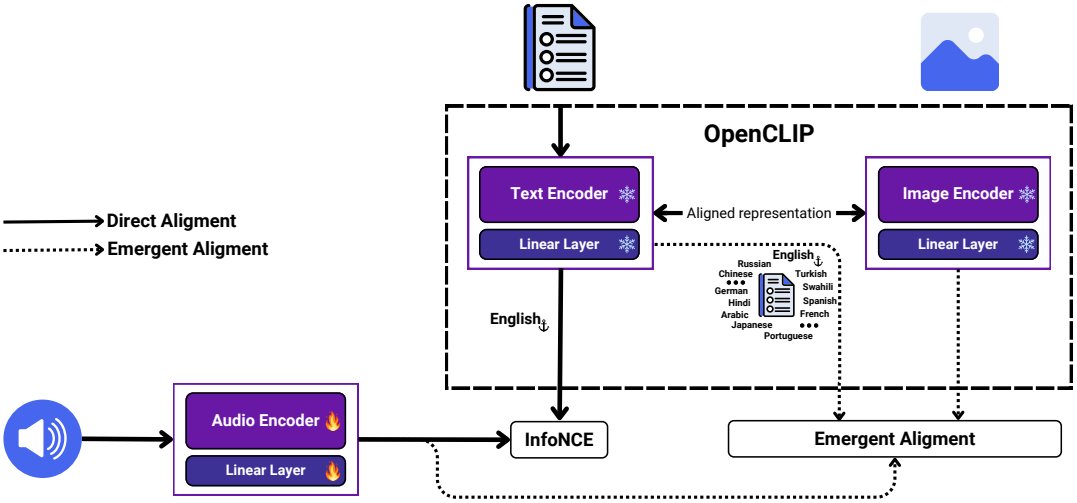

Figure 1: Flow of adding a new modality (Audio) to the text-image bimodal model. Linguistic expansion and alignment of these languages for the new modality. During audio model training, alignment is performed only with the anchor encoder (textual), and at this stage, no image information is processed by the model. In addition to the highlighted 12 languages selected for translation and evaluation, the model supports 88 more languages.

multiple languages align with other modalities via this emergent synchronization, significantly reducing training costs and time. Unlike conventional approaches that train models on each language separately, CACARA requires training only in English, leveraging its higher data availability and model quality. This contrasts with existing literature, which often incurs multiplicative computational costs by training on multilingual datasets individually.

Although CACARA's text encoder supports 100 languages, we selected 12 languages for evaluation: English, Portuguese, Spanish, French, German, Chinese, Japanese, Russian, Turkish, Hindi, Arabic, and Swahili. Due to the lack of multilingual test datasets, test data was translated into these languages using Google Translator from English.

3.3 Optimization Pipeline

The final CACARA model was developed via a four-stage optimization pipeline (Figure 2). The first stage involved a comparative evaluation of different audio encoders. We selected four encoders: BEATs (Chen et al., 2023a), HTS-AT (Chen et al., 2022), AudioMAE (Huang et al., 2022), and MAE-AST (Baade et al., 2022). These models were chosen based on their state-of-the-art performance in sound event detection and audio tagging, and their relatively recent introduction to the field, as established in the existing literature. Appendix A.1 provides a detailed analysis of the models built with these encoder combinations.

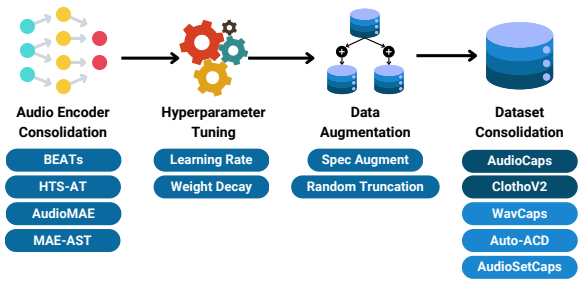

Figure 2: The CACARA model's training and optimization steps are divided into four main stages: consolidation of the newly added encoder, hyperparameter tuning, data augmentation, and consolidation of the training datasets.

The second stage focused on identifying optimal hyperparameters. While many parameters were investigated, the learning rate and weight decay substantially influenced training and model performance. We conducted a systematic search to determine values that ensured a fair comparison across all encoders. Appendix A.2 describes the best combination of hyperparameters.

The third stage incorporated data augmentation to improve model robustness. We used two techniques: SpecAugment (Park et al., 2019a), which masks blocks of frequency and time channels, and Random Truncation (Elizalde et al., 2023a), which divides training data into smaller segments. Appendices A.2 and A.3 detail these techniques.

The fourth stage involved a deliberate selection of datasets, focusing on quantity, diversity, multi-

modal decoupling, and label quality. We chose five datasets, and more information about these datasets is presented in Appendix A.4: AudioCaps (Kim et al., 2019a), ClothoV2 (Drossos et al., 2020), WavCaps (Mei et al., 2024a), Auto-ACD (Sun et al., 2024), and AudioSetCaps (Bai et al., 2024).

AudioCaps and ClothoV2, both featuring human-annotated captions, served as the gold standard and were included in all dataset configurations. The remaining datasets, featuring machine-generated annotations, were analyzed in various combinations to assess their impact on model performance. Section 4.1 presents the specific configurations and corresponding results.

4 Experiments and Results

Our analysis of CACARA focuses on three primary characteristics: (1) its multimodal capabilities, (2) its multilingual performance, and (3) the efficiency and scaling of its underlying resources. The first two characteristics and their results are discussed in Sections 4.1 and 4.2. In addition, to compare and understand the model's capabilities regarding resources and performance, we have performed extended tests that show how this model can behave in scenarios with more computational capacity, as detailed in Section 4.3. Intermediate training results and ablation studies on model components are presented in Appendices A.1, A.2, and A.3, as well as extended results with other sets and combinations of data. Qualitative visualizations are provided in Appendix A.5.

For the results presented in this section, we used the training datasets AudioSetCaps (ASC), Auto-ACD (AA), WavCaps (WC), AudioCaps (AC), and ClothoV2 (C). For the CACARA model, different combinations of datasets have been evaluated, with various results depending on the task and due to a distribution similar to the training data. We also applied data filtering based on CLIP similarity, where the filtering percentage x% is specified as f 0.x.

4.1 Multimodal Evaluation

We compare CACARA with established bimodal and multimodal models. While most prior work has focused on bimodal architectures, these models inherently lack the flexibility to handle inter-domain scenarios and emergent learning, where modalities independently acquire new conceptual representations. In contrast, CACARA is designed to leverage multimodality, expanding its applicability and enhancing adaptability across diverse tasks.

We selected three representative bimodal models for comparison: CLAP (Microsoft) (Elizalde et al., 2023b), CLAP (LAION) (Wu et al., 2023), and WavCaps Model (Mei et al., 2024b). These models, all focusing on audio-text modalities, provide a relevant benchmark for evaluating the audio-centric capabilities introduced in CACARA. This comparison aims to leverage these bimodal models' high degree of alignment and reported performance as a reference point for achievable results within a constrained modality space. For this reason, bimodal models cannot be directly compared to multimodal models. Thus, the tables highlight the best results specifically for multimodal models.

For multimodal comparison, we selected ImageBind, VAST, and LanguageBind models, which represent the state-of-the-art in multimodal learning to the best of our knowledge. A direct comparison with MLMM, the sole identified work in our review to integrate both multimodality and multilinguality, was not feasible due to the lack of publicly accessible code and implementation details, hindering reproducibility. Thus, our comparison remains direct and comprehensive, focusing on models that effectively align multiple modalities.

We evaluated the model on two core tasks: retrieval and classification. We used two distinct datasets for each task: AudioCaps and ClothoV2 for retrieval, and ESC-50 (Piczak) and UrbanSounds8K (Diment et al., 2017) for classification. Performance was measured using standard information retrieval metrics (R-precision at ranks 1, 5, and 10, and mean average precision) for retrieval and mean classification accuracy for classification.

Table 1 shows the retrieval results. For AudioCaps, CACARA, trained on all datasets with a 0.2 filtering threshold, achieved the highest R@1 (33.98%) among multimodal models, surpassing the best existing multimodal models by 14.23 percentage points (pp) while keeping competitive performance against bimodal models. For ClothoV2, the best-performing model was CACARA trained with WavCaps, with R@1 of 17.26%.

Table 2 shows the classification results for ESC-50 and UrbanSounds8K datasets. The performance differences across models in this task are relatively small. For ESC-50, LanguageBind achieved the highest mean accuracy (94.75%), followed closely by CACARA trained with WavCaps (94.37%) and the best-performing bimodal model, WavCaps (94.25%), with a maximum difference of only

	Model	Audio to Text				Text to Audio			
		R@1	R@5	R@10	R@Avg	R@1	R@5	R@10	R@Avg
		AudioCaps							
BM	CLAP (Microsoft)	15.75	44.7	61.62	40.69	6.32	24.9	38.11	23.11
BM	CLAP (LAION)	34.58	70.8	83.69	63.02	9.31	35.52	51.68	32.17
BM	WavCaps	38.70	73.32	86.05	66.02	10.57	38.38	53.21	34.05
MM	ImageBind	8.59	27.58	40.49	25.55	2.25	9.90	16.69	9.61
MM	VAST	19.75	46.30	57.98	41.34	4.99	19.89	29.25	18.04
MM	LanguageBind	13.05	45.96	63.76	40.92	4.61	18.94	30.80	18.12
CACARA	CACARA$_{AA/WC/AC/C}$	31.03 ± 0.51	64.57 ± 0.74	79.34 ± 0.53	58.31 ± 0.43	6.76 ± 0.45	25.52 ± 1.65	38.61 ± 1.30	23.63 ± 1.13
CACARA	CACARA$_{WC/AC/C}$	30.08 ± 0.31	64.95 ± 0.42	78.71 ± 0.75	57.91 ± 0.41	**7.57** ±0.16	28.02 ± 0.23	42.43 ± 0.14	26.00 ± 0.09
CACARA	CACARA$_{WC/AC/C/f\ 0.1}$	30.96 ± 0.05	64.78 ± 0.91	78.54 ± 0.35	58.10 ± 0.41	7.37 ± 0.20	27.88 ± 0.26	41.72 ± 0.61	25.66 ± 0.35
CACARA	CACARA$_{WC/AC/C/f\ 0.2}$	31.03 ± 0.60	65.32 ± 0.68	78.80 ± 1.10	58.38 ± 0.77	7.41 ± 0.09	**28.40** ± 0.58	**42.68** ± 0.27	**26.17** ± 0.31
CACARA	CACARA$_{ASC/AA/WC/AC/C}$	33.27 ± 0.33	67.46 ± 0.33	81.51 ± 0.30	60.75 ± 0.29	6.91 ± 0.15	26.18 ± 0.18	39.55 ± 0.15	24.21 ± 0.06
CACARA	CACARA$_{ASC/AA/WC/AC/C/f\ 0.1}$	33.64 ± 0.24	**68.40** ± 0.24	**81.84** ± 0.17	61.29 ± 0.23	7.34 ± 0.10	27.69 ± 0.32	41.27 ± 0.29	25.43 ± 0.05
CACARA	CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$	**33.98** ± 0.64	68.30 ± 0.64	81.81 ± 0.21	**61.36** ± 0.26	7.30 ± 0.15	27.87 ± 0.32	41.21 ± 0.29	25.46 ± 0.16
		ClothoV2							
BM	CLAP (Microsoft)	15.46	38.74	51.52	35.24	4.61	16.90	26.43	15.98
BM	CLAP (LAION)	14.64	37.28	49.68	33.87	3.77	15.27	23.79	14.28
BM	WavCaps	18.78	45.15	57.72	40.55	4.38	18.76	28.61	17.25
MM	ImageBind	5.11	16.17	24.96	15.41	1.51	5.57	9.15	5.41
MM	VAST	11.02	26.64	35.48	24.38	2.30	8.73	13.99	8.34
MM	LanguageBind	16.11	**41.07**	53.05	36.74	3.75	**16.38**	24.65	**14.93**
CACARA	CACARA$_{AA/WC/AC/C}$	14.42 ± 0.88	37.21 ± 1.74	49.95 ± 1.60	33.86 ± 1.40	2.90 ± 0.32	12.17 ± 1.49	19.40 ± 1.75	11.49 ± 1.18
CACARA	CACARA$_{WC/AC/C}$	14.39 ± 0.19	35.87 ± 0.21	48.08 ± 0.32	32.78 ± 0.14	3.88 ± 0.13	14.69 ± 0.41	22.49 ± 0.12	13.69 ± 0.18
CACARA	CACARA$_{WC/AC/C/f\ 0.1}$	15.38 ± 0.49	37.40 ± 1.62	49.63 ± 2.13	34.13 ± 1.40	3.92 ± 0.35	14.99 ± 0.41	23.10 ± 0.54	14.00 ± 0.42
CACARA	CACARA$_{WC/AC/C/f\ 0.2}$	**17.26** ± 2.33	40.91 ± 5.84	**53.85** ± 6.74	**37.34** ± 4.96	**4.03** ± 0.36	**15.60** ± 1.25	**24.78** ± 2.36	14.80 ± 1.32
CACARA	CACARA$_{ASC/AA/WC/AC/C}$	11.40 ± 2.17	29.82 ± 6.17	41.12 ± 7.79	27.45 ± 5.37	2.22 ± 0.61	8.71 ± 2.41	14.42 ± 3.66	8.45 ± 2.22
CACARA	CACARA$_{ASC/AA/WC/AC/C/f\ 0.1}$	13.04 ± 3.12	33.53 ± 6.95	46.00 ± 8.70	30.86 ± 6.25	2.56 ± 0.87	9.84 ± 2.98	16.15 ± 4.43	9.52 ± 2.76
CACARA	CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$	13.21 ± 3.16	33.63 ± 6.91	45.67 ± 8.51	30.84 ± 6.19	2.59 ± 0.74	10.07 ± 2.87	16.62 ± 4.68	9.76 ± 2.76

Table 1: Multimodal results, divided into two datasets: AudioCaps and ClothoV2, and two retrieval tasks: Audio-to-Text and Text-to-Audio. The evaluated models are categorized as BM (Bimodal Models), MM (Multimodal Models), and CACARA (CACARA Multimodal Models). **Bolded** results indicate the best performance among multimodal models, while underlined results represent the second-best values for the same set.

	Model	Accuracy	
		ESC-50	UrbanSounds8K
BM	CLAP (Microsoft)	93.85	82.74
BM	CLAP (LAION)	83.10	80.91
BM	WavCaps	94.25	82.28
MM	ImageBind	64.15	48.20
MM	VAST	76.80	68.00
MM	LanguageBind	**94.75**	79.24
CACARA	CACARA$_{AA/WC/AC/C}$	89.95 ± 2.14	77.04 ± 2.98
CACARA	CACARA$_{WC/AC/C}$	94.15 ± 0.10	77.40 ± 1.07
CACARA	CACARA$_{WC/AC/C/f\ 0.1}$	94.37 ± 0.49	**79.51** ± 1.06
CACARA	CACARA$_{WC/AC/C/f\ 0.2}$	94.00 ± 1.39	77.83 ± 0.11
CACARA	CACARA$_{ASC/AA/WC/AC/C}$	82.45 ± 0.70	70.04 ± 1.48
CACARA	CACARA$_{ASC/AA/WC/AC/C/f\ 0.1}$	89.97 ± 0.38	75.40 ± 1.19
CACARA	CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$	91.35 ± 0.69	74.99 ± 0.55

Table 2: Multimodal results, evaluated on ESC-50 and UrbanSounds8K datasets for the classification task. The evaluated models are categorized as BM (Bimodal Models), MM (Multimodal Models), and CACARA (CACARA Multimodal Models. **Bolded** results indicate the best performance among multimodal models, while underlined results represent the second-best values for the same set.

0.38 pp. For UrbanSounds8K, CACARA trained on WavCaps (79.51%) was the best multimodal model, while LanguageBind achieved 79.24%, a difference of only 0.27 pp. However, the bimodal model CLAP (Microsoft) achieved a result of 82.74%, demonstrating an advantage in this task and dataset.

4.2 Multilingual Evaluation

Beyond its multimodal capabilities, CACARA is inherently multilingual, supporting approximately 100 languages without explicit training for aligning new languages to audio data. We selected twelve languages to evaluate this emergent alignment and analyzed performance in audio-text retrieval and classification tasks. Due to many results in multiple languages, we selected two versions of CACARA for each task to provide a focused comparison.

For the retrieval task, Table 3 presents results for the models CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$ in the Audiocaps dataset and CACARA$_{WC/AC/C/f\ 0.2}$ in the ClothoV2 dataset. The audio encoder was trained only in English, which leads to better performance in this language. Therefore, we use English results as the upper bound. Although it has never been trained for other languages, the results can be quite satisfactory, depending on the

	CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$ in Audiocaps							
	Audio to Text				Text to Audio			
Language	R@1	R@5	R@10	R@Avg	R@1	R@5	R@10	R@Avg
English	33.98 ± 0.64	68.30 ± 0.43	81.81 ± 0.21	61.36 ± 0.27	7.30 ± 0.15	27.87 ± 0.32	41.21 ± 0.29	25.46 ± 0.16
Portuguese	21.62 ± 0.70	51.27 ± 1.31	66.34 ± 1.23	46.41 ± 1.00	5.74 ± 0.18	21.46 ± 0.18	33.47 ± 0.35	20.22 ± 0.11
Spanish	24.06 ± 0.48	53.94 ± 0.87	68.05 ± 0.35	48.68 ± 0.51	5.85 ± 0.16	22.58 ± 0.29	34.31 ± 0.25	20.91 ± 0.08
French	21.87 ± 0.70	51.20 ± 1.51	66.31 ± 1.29	46.46 ± 1.17	5.94 ± 0.05	22.97 ± 0.27	35.19 ± 0.47	21.37 ± 0.16
Russian	20.73 ± 0.47	48.96 ± 0.60	63.03 ± 0.74	44.24 ± 0.58	5.55 ± 0.31	19.48 ± 0.47	30.70 ± 0.39	18.58 ± 0.39
Arabic	15.25 ± 0.73	39.70 ± 1.05	53.90 ± 1.15	36.28 ± 0.97	4.71 ± 0.15	18.45 ± 0.23	28.27 ± 0.31	17.15 ± 0.10
Hindi	14.08 ± 0.41	37.47 ± 0.30	51.51 ± 0.80	34.35 ± 0.42	3.71 ± 0.08	14.97 ± 0.09	23.56 ± 0.07	14.08 ± 0.02
German	23.95 ± 0.41	54.36 ± 0.06	68.26 ± 0.14	48.86 ± 0.12	6.00 ± 0.09	23.09 ± 0.37	35.04 ± 0.45	21.38 ± 0.25
Chinese(zh)	18.24 ± 0.28	46.59 ± 1.11	62.04 ± 0.42	42.29 ± 0.47	5.45 ± 0.12	20.52 ± 0.20	31.70 ± 0.23	19.23 ± 0.09
Swahili	1.11 ± 0.23	3.96 ± 0.11	6.48 ± 0.24	3.85 ± 0.19	0.65 ± 0.07	2.29 ± 0.06	3.67 ± 0.08	2.20 ± 0.02
Japanese	21.36 ± 0.49	51.46 ± 0.24	65.64 ± 0.38	46.15 ± 0.22	5.59 ± 0.24	21.87 ± 0.48	33.82 ± 0.20	20.42 ± 0.30
Turkish	17.04 ± 0.40	42.65 ± 0.63	56.80 ± 0.45	38.84 ± 0.47	4.42 ± 0.23	17.22 ± 0.32	27.15 ± 0.51	16.27 ± 0.33
	CACARA$_{WC/AC/C/f\ 0.2}$ in ClothoV2							
	R@1	R@5	R@10	R@Avg	R@1	R@5	R@10	R@Avg
English	17.26 ± 2.33	40.91 ± 5.84	53.85 ± 6.74	37.34 ± 4.96	4.03 ± 0.36	15.60 ± 1.25	24.78 ± 2.36	14.80 ± 1.32
Portuguese	10.83 ± 0.60	28.98 ± 1.56	39.72 ± 1.93	26.51 ± 1.35	3.09 ± 0.16	11.61 ± 0.09	18.57 ± 0.12	11.09 ± 0.05
Spanish	11.37 ± 0.39	29.88 ± 1.73	40.54 ± 2.09	27.26 ± 1.34	3.18 ± 0.12	12.09 ± 0.17	19.38 ± 0.05	11.55 ± 0.05
French	10.83 ± 0.66	28.75 ± 1.43	39.55 ± 1.35	26.38 ± 1.10	3.00 ± 0.14	11.50 ± 0.23	18.05 ± 0.12	10.85 ± 0.12
Russian	8.94 ± 0.51	24.45 ± 1.88	34.62 ± 2.55	22.67 ± 1.64	2.78 ± 0.14	10.60 ± 0.33	16.82 ± 0.51	10.06 ± 0.24
Arabic	6.84 ± 0.61	20.34 ± 1.22	29.13 ± 1.31	18.77 ± 1.02	2.34 ± 0.08	8.85 ± 0.04	14.50 ± 0.08	8.56 ± 0.05
Hindi	6.02 ± 0.43	17.30 ± 1.24	25.69 ± 1.60	16.34 ± 1.05	2.04 ± 0.16	7.80 ± 0.30	12.33 ± 0.54	7.39 ± 0.31
German	11.55 ± 0.72	29.99 ± 1.50	40.96 ± 1.49	27.49 ± 1.21	3.21 ± 0.15	12.45 ± 0.01	19.55 ± 0.27	11.74 ± 0.13
Chinese(zh)	9.55 ± 0.38	25.94 ± 1.73	36.59 ± 1.80	24.03 ± 1.29	2.79 ± 0.17	11.02 ± 0.15	17.64 ± 0.43	10.48 ± 0.19
Swahili	1.03 ± 0.03	3.35 ± 0.16	5.39 ± 0.27	3.26 ± 0.14	0.52 ± 0.05	1.70 ± 0.13	2.65 ± 0.20	1.62 ± 0.09
Japanese	10.75 ± 0.73	29.19 ± 1.89	40.24 ± 2.10	26.72 ± 1.55	2.97 ± 0.19	11.50 ± 0.37	18.32 ± 0.34	10.93 ± 0.22
Turkish	7.93 ± 0.74	22.53 ± 1.23	32.12 ± 2.69	20.86 ± 1.54	2.51 ± 0.07	9.61 ± 0.20	15.51 ± 0.37	9.21 ± 0.17

Table 3: Recall value for the Audio-to-Text and Text-to-Audio retrieval tasks on two different CACARA models (CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$ for the AudioCaps dataset and CACARA$_{WC/AC/C/f\ 0.2}$ for the ClothoV2 dataset) across the twelve evaluated languages.

language used.

Some languages – Spanish, German, Portuguese, French, Russian, and Japanese – perform well, achieving R@1 above 20 for the CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$ model on AudioCaps. On average, the other languages achieve R@1 above 13, with performance variations primarily influenced by the quantity and quality of textual data used during the pre-training of the text model (Geigle et al., 2024). This suggests that improving low-resource language performance does not require retraining the entire model. Only Swahili was below average because it is a language with very few available resources. We observed the same behavior from the CACARA$_{WC/AC/C/f\ 0.2}$ model in ClothoV2.

For the classification task, Table 4 shows results for two models, CACARA$_{WC/AC/C/f\ 0.1}$ and CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$, evaluated on ESC-50 and UrbanSounds8K. As expected, languages with more resources showed strong classification performance, similar to retrieval results. In addition, Mandarin showed good results compared to the previous task, while Portuguese showed a drop in results. However, the overall classification average of the different languages is 66.5% for CACARA$_{WC/AC/C}$ using ESC-50.

	CACARA$_{WC/AC/C/f\ 0.1}$		CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$	
Language	ESC-50	UrbanSounds8K	ESC-50	UrbanSounds8K
English	94.37 ± 0.49	79.51 ± 1.06	91.35 ± 0.69	74.99 ± 0.55
Portuguese	79.63 ± 0.81	66.49 ± 0.79	80.92 ± 0.98	71.38 ± 1.35
Spanish	86.25 ± 0.40	72.02 ± 1.78	85.60 ± 0.78	71.02 ± 0.67
French	83.60 ± 0.69	69.56 ± 0.65	82.45 ± 0.49	67.72 ± 1.12
Russian	81.15 ± 0.33	71.77 ± 1.00	78.30 ± 1.08	67.16 ± 1.32
Arabic	65.28 ± 0.95	63.13 ± 0.64	63.65 ± 1.09	63.77 ± 0.60
Hindi	64.30 ± 0.97	58.19 ± 0.84	60.12 ± 0.83	58.11 ± 1.15
German	82.12 ± 0.87	74.99 ± 0.12	75.92 ± 1.08	72.38 ± 0.10
Chinese(zh)	83.47 ± 1.68	70.14 ± 0.12	81.23 ± 1.07	67.22 ± 1.15
Swahili	20.55 ± 0.41	42.99 ± 1.08	20.48 ± 2.17	38.35 ± 1.97
Japanese	84.08 ± 1.63	68.79 ± 0.02	81.78 ± 0.75	73.03 ± 0.74
Turkish	74.12 ± 1.83	64.94 ± 0.68	69.85 ± 1.97	63.91 ± 0.82

Table 4: Classification retults on two different CACARA models (CACARA$_{WC/AC/C/f\ 0.1}$ and CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$) in the datasets ESC-50 and UrbanSounds8K across the twelve evaluated languages.

	Optimized Basic Model												Expanded Resources Model											
	Audio to Text						Text to Audio						Audio to Text						Text to Audio					
	Audiocaps																							
Language	R@1		R@5		R@10		R@Avg		R@1		R@5		R@10		R@Avg		R@1		R@5		R@10		R@Avg	
English	33.98 ± 0.64		68.30 ± 0.43		81.81 ± 0.21		61.36 ± 0.27		7.30 ± 0.15		27.87 ± 0.32		41.21 ± 0.29		25.46 ± 0.16		31.45 ± 0.14		66.09 ± 0.31		79.61 ± 0.45		59.05 ± 0.28		7.64 ± 0.09		28.78 ± 0.21		42.95 ± 0.14		26.46 ± 0.12	
Portuguese	21.62 ± 0.70		51.27 ± 1.31		66.34 ± 1.23		46.41 ± 1.00		5.74 ± 0.18		21.46 ± 0.18		33.47 ± 0.35		20.22 ± 0.11		18.54 ± 0.98		46.57 ± 0.75		60.96 ± 0.53		42.02 ± 0.74		5.99 ± 0.10		22.07 ± 0.43		33.75 ± 0.06		20.60 ± 0.19	
Spanish	24.06 ± 0.48		53.94 ± 0.87		68.05 ± 0.35		48.68 ± 0.51		5.85 ± 0.16		22.58 ± 0.29		34.31 ± 0.25		20.91 ± 0.08		21.03 ± 0.38		50.85 ± 0.26		65.32 ± 0.80		45.73 ± 0.45		6.04 ± 0.10		22.46 ± 0.21		34.62 ± 0.05		21.04 ± 0.06	
French	21.87 ± 0.70		51.20 ± 1.51		66.31 ± 1.29		46.46 ± 1.17		5.94 ± 0.05		22.97 ± 0.27		35.19 ± 0.47		21.37 ± 0.16		20.33 ± 0.42		48.61 ± 0.36		62.67 ± 0.38		43.87 ± 0.27		6.25 ± 0.08		23.44 ± 0.22		35.66 ± 0.23		21.78 ± 0.06	
Russian	20.73 ± 0.47		48.96 ± 0.60		63.03 ± 0.74		44.24 ± 0.58		5.55 ± 0.31		19.48 ± 0.47		30.70 ± 0.39		18.58 ± 0.39		17.58 ± 0.35		44.19 ± 1.45		58.16 ± 0.89		39.97 ± 0.89		5.50 ± 0.32		19.88 ± 0.34		31.28 ± 0.42		18.89 ± 0.30	
Arabic	15.25 ± 0.73		39.70 ± 1.05		53.90 ± 1.15		36.28 ± 0.97		4.71 ± 0.15		18.45 ± 0.23		28.27 ± 0.31		17.15 ± 0.10		11.18 ± 0.47		31.96 ± 0.86		45.17 ± 0.51		29.44 ± 0.58		4.71 ± 0.17		18.37 ± 0.19		28.10 ± 0.18		17.06 ± 0.07	
Hindi	14.08 ± 0.41		37.47 ± 0.30		51.51 ± 0.80		34.35 ± 0.42		3.71 ± 0.08		14.97 ± 0.09		23.56 ± 0.07		14.08 ± 0.02		11.70 ± 0.30		33.33 ± 0.71		46.76 ± 0.76		30.60 ± 0.38		3.83 ± 0.11		15.10 ± 0.44		23.28 ± 0.23		14.07 ± 0.24	
German	23.95 ± 0.41		54.36 ± 0.06		68.26 ± 0.14		48.86 ± 0.12		6.00 ± 0.09		23.09 ± 0.37		35.04 ± 0.45		21.38 ± 0.25		21.33 ± 0.13		50.33 ± 0.90		64.41 ± 0.90		45.36 ± 0.62		5.85 ± 0.08		23.06 ± 0.38		35.64 ± 0.46		21.52 ± 0.30	
Chinese(zh)	18.24 ± 0.28		46.59 ± 1.11		62.04 ± 0.42		42.29 ± 0.47		5.45 ± 0.12		20.52 ± 0.20		31.70 ± 0.23		19.23 ± 0.09		15.27 ± 0.74		41.17 ± 1.37		56.36 ± 0.43		37.60 ± 0.80		5.57 ± 0.31		20.79 ± 0.16		32.16 ± 0.52		19.51 ± 0.05	
Swahili	1.11 ± 0.23		3.96 ± 0.11		6.48 ± 0.24		3.85 ± 0.19		0.65 ± 0.07		2.29 ± 0.06		3.67 ± 0.08		2.20 ± 0.02		0.61 ± 0.18		2.89 ± 0.32		5.10 ± 0.18		2.87 ± 0.21		0.59 ± 0.03		2.21 ± 0.06		3.49 ± 0.10		2.10 ± 0.06	
Japanese	21.36 ± 0.49		51.46 ± 0.24		65.64 ± 0.38		46.15 ± 0.22		5.59 ± 0.24		21.87 ± 0.48		33.82 ± 0.20		20.42 ± 0.30		18.35 ± 0.52		45.76 ± 0.96		59.78 ± 0.80		41.30 ± 0.68		5.76 ± 0.18		22.74 ± 0.30		34.45 ± 0.05		20.98 ± 0.10	
Turkish	17.04 ± 0.40		42.65 ± 0.63		56.80 ± 0.45		38.84 ± 0.47		4.42 ± 0.23		17.22 ± 0.32		27.15 ± 0.51		16.27 ± 0.33		12.76 ± 0.15		35.07 ± 0.38		47.96 ± 0.51		31.93 ± 0.34		4.52 ± 0.16		17.64 ± 0.22		27.81 ± 0.29		16.66 ± 0.06	
	ClothoV2																							
English	13.21 ± 3.16		33.63 ± 6.91		45.67 ± 8.51		30.84 ± 6.19		2.59 ± 0.74		10.07 ± 2.87		16.62 ± 4.68		9.76 ± 2.76		15.67 ± 0.50		38.46 ± 0.44		51.25 ± 0.06		35.13 ± 0.31		3.28 ± 0.17		13.57 ± 0.14		21.51 ± 0.05		12.79 ± 0.08	
Portuguese	11.16 ± 0.08		30.27 ± 0.56		41.59 ± 0.48		27.68 ± 0.33		2.26 ± 0.17		9.47 ± 0.43		15.82 ± 0.50		9.18 ± 0.31		11.17 ± 0.20		30.45 ± 0.28		41.94 ± 0.72		27.85 ± 0.40		2.64 ± 0.10		10.92 ± 0.06		17.91 ± 0.10		10.49 ± 0.06	
Spanish	11.67 ± 0.24		30.95 ± 0.33		43.00 ± 0.63		28.54 ± 0.21		2.30 ± 0.11		9.72 ± 0.37		15.67 ± 0.16		9.23 ± 0.19		11.67 ± 0.40		31.33 ± 0.25		42.58 ± 0.39		28.53 ± 0.23		2.63 ± 0.18		10.87 ± 0.18		17.66 ± 0.46		10.39 ± 0.15	
French	10.30 ± 0.14		28.72 ± 0.06		40.27 ± 0.13		26.43 ± 0.06		2.28 ± 0.18		9.63 ± 0.39		15.67 ± 0.15		9.20 ± 0.16		10.45 ± 0.74		28.07 ± 1.20		39.78 ± 0.82		26.10 ± 0.80		2.56 ± 0.21		10.68 ± 0.32		17.36 ± 0.12		10.20 ± 0.19	
Russian	9.53 ± 0.21		26.88 ± 0.35		38.57 ± 0.23		25.00 ± 0.15		2.11 ± 0.10		8.29 ± 0.24		13.70 ± 0.76		8.03 ± 0.35		10.09 ± 0.52		27.71 ± 0.76		39.13 ± 0.77		25.64 ± 0.66		2.54 ± 0.09		9.79 ± 0.31		15.85 ± 0.29		9.40 ± 0.20	
Arabic	7.90 ± 0.19		23.31 ± 0.52		34.68 ± 0.03		21.96 ± 0.20		1.71 ± 0.26		7.06 ± 0.29		12.02 ± 0.34		6.93 ± 0.29		7.99 ± 0.10		23.64 ± 0.12		34.04 ± 0.28		21.89 ± 0.03		2.04 ± 0.02		8.22 ± 0.14		13.47 ± 0.21		7.91 ± 0.12	
Hindi	6.55 ± 0.27		19.04 ± 0.26		28.74 ± 0.44		18.11 ± 0.29		1.56 ± 0.12		6.07 ± 0.20		10.26 ± 0.32		5.97 ± 0.10		6.81 ± 0.66		19.51 ± 0.89		28.32 ± 0.85		18.22 ± 0.79		1.85 ± 0.09		6.98 ± 0.32		11.46 ± 0.07		6.76 ± 0.11	
German	10.95 ± 0.25		29.53 ± 0.63		40.91 ± 0.83		27.13 ± 0.57		2.62 ± 0.09		10.11 ± 0.22		16.54 ± 0.37		9.76 ± 0.21		11.22 ± 0.39		30.31 ± 0.74		41.44 ± 1.40		27.66 ± 0.81		2.72 ± 0.07		11.12 ± 0.31		18.30 ± 0.48		10.72 ± 0.27	
Chinese(zh)	9.90 ± 0.35		27.41 ± 0.56		39.20 ± 0.55		25.51 ± 0.31		2.26 ± 0.05		8.76 ± 0.24		14.25 ± 0.20		8.43 ± 0.13		9.98 ± 0.15		27.28 ± 0.14		38.59 ± 0.80		25.28 ± 0.36		2.48 ± 0.10		9.71 ± 0.17		15.53 ± 0.21		9.24 ± 0.14	
Swahili	0.78 ± 0.06		3.44 ± 0.26		5.61 ± 0.12		3.28 ± 0.11		0.42 ± 0.05		1.54 ± 0.07		2.43 ± 0.05		1.46 ± 0.02		0.75 ± 0.10		2.98 ± 0.07		5.20 ± 0.38		2.98 ± 0.10		0.55 ± 0.01		1.65 ± 0.12		2.54 ± 0.23		1.58 ± 0.12	
Japanese	10.49 ± 0.35		28.92 ± 0.41		40.76 ± 0.76		26.72 ± 0.44		2.06 ± 0.14		8.86 ± 0.30		14.51 ± 0.43		8.48 ± 0.25		10.48 ± 0.23		28.87 ± 0.26		40.19 ± 0.44		26.51 ± 0.31		2.35 ± 0.20		9.80 ± 0.22		16.03 ± 0.16		9.39 ± 0.17	
Turkish	8.60 ± 0.48		24.53 ± 0.82		35.75 ± 0.55		22.96 ± 0.60		1.74 ± 0.09		7.04 ± 0.09		12.13 ± 0.00		6.97 ± 0.06		8.45 ± 0.06		24.02 ± 0.50		34.61 ± 0.75		22.36 ± 0.41		2.14 ± 0.07		8.06 ± 0.24		13.56 ± 0.19		7.92 ± 0.13	

Table 5: Retrieval task results for expanded resources experiments, divided into two datasets: AudioCaps and ClothoV2, and two retrieval tasks: Audio-to-Text and Text-to-Audio across the twelve evaluated languages.

4.3 Expanded Resources

To achieve a single robust model applicable across tasks, we trained an expanded-resource version of CACARA: CACARA$_{ASC/AA/WC/AC/C/f\ 0.2}$. Unlike previous models trained with 3 epochs, a batch size of 64, and different datasets, this version was trained with 10 epochs, a batch size of 110, and using all datasets together, with a data filtering of 0.2, for a more complete training.

A direct comparison of this expanded model against the best-performing optimized models is in Table 5, evaluating audio-to-text and text-to-audio retrieval across 12 languages. This task in the AudioCaps dataset did not improve when given more resources and training time and continues to show values lower than those obtained with the optimized model. This is due to the proximity of the trained sets to the distribution of the test set. However, it still improved over the same model with fewer resources. For the same task, in the ClothoV2 dataset, a general improvement is observed for the audio-to-text retrieval task, but text-to-audio retrieval remained below the optimized model's results.

For the classification task, comparing both datasets through Table 6, results are generally improved when using more resources with more varied data. This shows that this task benefits from more robust training. Despite benefiting from a more robust structure, the presented model continues to demonstrate excellent results with a modest computational structure but with the capacity to obtain better results with training with greater computational power and data availability. In this structure, we continued to train the model only in English; the other languages obtained an improvement due to emergent learning.

	OB Model	ER Model	OB Model	ER Model
	ESC-50		UrbanSounds8K	
Language	Accuracy			
English	91.35 ± 0.69	93.50 ± 0.18	74.99 ± 0.55	78.33 ± 2.75
Portuguese	80.92 ± 0.98	81.70 ± 0.88	71.38 ± 1.35	68.15 ± 1.63
Spanish	85.60 ± 0.78	87.65 ± 0.43	71.02 ± 0.67	72.72 ± 1.50
French	82.45 ± 0.49	83.85 ± 0.17	67.72 ± 1.12	70.47 ± 1.82
Russian	78.30 ± 1.08	80.62 ± 1.36	67.16 ± 1.32	69.32 ± 0.22
Arabic	63.65 ± 1.09	63.67 ± 0.77	63.77 ± 0.60	62.40 ± 1.48
Hindi	60.12 ± 0.83	62.57 ± 0.88	58.11 ± 1.15	56.55 ± 0.78
German	75.92 ± 1.08	78.37 ± 0.64	72.38 ± 0.10	74.44 ± 1.24
Chinese(zh)	81.23 ± 1.07	82.80 ± 0.93	67.22 ± 1.15	68.50 ± 0.59
Swahili	20.48 ± 2.17	21.80 ± 0.22	38.35 ± 1.97	36.28 ± 3.61
Japanese	81.78 ± 0.75	85.83 ± 1.46	73.03 ± 0.74	72.18 ± 1.88
Turkish	69.85 ± 1.97	72.92 ± 0.81	63.91 ± 0.82	64.52 ± 0.82

Table 6: Classification task results, evaluated on ESC-50 and UrbanSounds8K datasets for expanded resources experiments cross the twelve evaluated languages.

5 Conclusions

Despite numerous efforts in the literature to develop multimodal and multilingual models, many approaches fail to leverage prior knowledge and optimize training efficiency effectively. In this work, we proposed CACARA, an architecture and model based on emergent alignment learning capable of integrating a new modality into an existing bimodal/multimodal architecture without requiring full retraining. Additionally, our approach enables the multilingual expansion of the newly added modality to all supported languages. By leveraging emergent alignment, our method simplifies training. It significantly reduces computational costs, eliminating the need for retraining all components while enhancing conceptual complementarity across modalities. Our results demonstrate a high degree of alignment between the integrated modalities, achieving superior R@1 performance compared to most state-of-the-art multimodal models in the literature.

Limitations

While our model demonstrates the capacity for seamless integration of additional modalities without degradation of performance in existing aligned modalities, the present study's scope was limited to the incorporation of the audio modality. Extending the framework to incorporate further modalities (e.g., video, thermal, wearable sensor data, and depth data) would provide a more comprehensive assessment of the model's scalability and ability to generalize across diverse representational spaces.

Furthermore, our experimental design was constrained to base-level encoder models. A more rigorous evaluation of the model's effectiveness would necessitate an investigation across a spectrum of encoder sizes. This would entail not only increased computational resources (in terms of data quantity, batch size, and training duration) but also a systematic exploration of the relationship between model parameterization and performance gains. A larger parameter count would allow one to assess the scalability.

Finally, while our evaluation encompassed 12 languages, demonstrating a degree of multilingual capability, the generalizability of these findings could be further strengthened by expanding the linguistic scope. A more comprehensive evaluation should include languages exhibiting greater typological diversity and, crucially, languages with varying levels of available digital resources. This would allow for a more nuanced understanding of the model's performance in low-resource language settings, which are often underrepresented in current research.

Ethics Statement

This work focuses on enhancing multimodal and multilingual models by leveraging emergent alignment through implicit learning to reduce computational overhead and enhance accessibility. We fully comply with the terms of use and licensing agreements associated with all datasets used for training, evaluation, or testing our models. This work does not involve human subjects; however, we recognize the ethical and societal responsibilities of deploying such models, including the potential for misuse (e.g., generating harmful or misleading text, audio, or images). Despite efforts to improve multilingual capabilities of a multilingual model, our models may still exhibit biases or under representation of specific languages, cultures, topics, or applications, particularly those with limited data resources that can be inherited from a already pre-trained language model. While designed for beneficial applications and scientific advancement, these models could be repurposed for unintended uses.

References

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-Supervised Multimodal Versatile Networks. *Advances in Neural Information Processing Systems*, 33:25–37.

Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687.

Alan Baade, Puyuan Peng, and David Harwath. 2022. MAE-AST: Masked Autoencoding Audio Spectrogram Transformer. In *Interspeech 2022*, pages 2438–2442.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2024. AudioSetCaps: Enriched Audio Captioning Dataset Generation Using Large Audio Language Models. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal Nachine learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. VGGSound: A Large-Scale Audio-Visual Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 721–725.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 646–650. IEEE.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023a. BEATs: Audio Pre-Training with Acoustic Tokenizers. In *40th International Conference on Machine Learning*, pages 5178–5193.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Aleksandr Diment, Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2017. TUT Rare Sound Events, Development Dataset.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An Audio Captioning Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 736–740. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023a. Clap learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2023b. Natural Language Supervision for General-Purpose Audio Representations. *Preprint*, arXiv:2309.05767.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. Babel-imagenet: Massively multilingual evaluation of vision-and-language representations. *Preprint*, arXiv:2306.08658.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space to Bind them All. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Roland Goecke. 2005. Current Trends in Joint Audio-Video Signal Processing: A Review. In *Eighth International Symposium on Signal Processing and its Applications*, pages 70–73. IEEE.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked Autoencoders that Listen. *Advances in Neural Information Processing Systems*, 35:28708–28720.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019a. AudioCaps: Generating Captions for Audios in The Wild. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019b. AudioCaps: Generating captions for Audios in the Wild. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024. VALOR: Vision-Audio-Language Omni-Perception

Pretraining Model and Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024a. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024b. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–15.

Irene Martin Morato and Annamaria Mesaros. 2021. MACS - Multi-Annotator Captioned Soundscapes.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019a. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019b. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*, page 2613.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Karol J Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *23rd ACM international conference on Multimedia*, pages 1015–1018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *22nd ACM International Conference on Multimedia*, pages 1041–1044.

Gabriel O. dos Santos, Diego A. B. Moreira, Alef I. Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. 2023. CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages. In *Workshop on Multi-lingual Representation Learning (MRL), Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 184–207.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13693–13696.

Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. 2024. Auto-ACD: A Large-Scale Dataset for Audio-Language Representation Learning. In *32nd ACM International Conference on Multimedia*, pages 5025–5034.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep Learning for Sensor-Based Activity Recognition: A Survey. *Pattern Recognition Letters*, 119:3–11. Deep Learning for Pattern Recognition.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Liang Zhang, Ludan Ruan, Anwen Hu, and Qin Jin. 2023. Multimodal Pretraining from Monolingual to Multilingual. *Machine Intelligence Research*, 20(2):220–232.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *The Twelfth International Conference on Learning Representations*, pages 1–22.

A Appendix

A.1 Encoders

As the first step in the CACARA model optimization pipeline, in the first phase, we focused on selecting the audio encoder to be incorporated into the multimodal model, for which four different encoders were selected, including BEATs, HTS-AT, AudioMAE, and MAE-AST:

- **BEATs** (Chen et al., 2023a) is a transformer-based model that employs an iterative framework that integrates acoustic tokenization with self-supervised learning (SSL) to support the development of robust audio representations. The process alternates between two stages: (1) an acoustic tokenizer quantizes continuous audio features into discrete labels, which are used to train an SSL model through masked prediction, and (2) the SSL model supervises the tokenizer's optimization via knowledge distillation. This iterative process enables the components to improve together with each cycle.

- **Hierarchical Token-Semantic Audio Transformer (HTS-AT)** (Chen et al., 2022) is a transformer-based model that employs a hierarchical design to process mel-spectrograms. The architecture reduces sequence length progressively through patch-merge operations across transformer groups, processing audio data along temporal and frequency dimensions to capture relevant patterns. A localized window attention mechanism is introduced, limiting attention calculations to small, defined regions rather than the entire input. This approach reduces computational complexity while preserving essential contextual relationships. The model also includes a token-semantic module, implemented as a CNN layer after the final transformer block. This module combines frequency information and maps features to event classes, generating event presence maps for use in both classification and temporal event localization tasks.

- **AudioMAE** (Huang et al., 2022) adapts the Masked Autoencoder framework to mel-spectrograms modeling through an asymmetric encoder-decoder architecture. The encoder processes approximately 20% of unmasked spectrogram patches, while the decoder reconstructs the complete spectrogram. The architecture implements local window attention mechanisms in the decoder to capture temporal and frequency correlations in audio signals.

- **MAE-AST** (Baade et al., 2022) combines the Masked Autoencoder architecture with the Audio Spectrogram Transformer for audio processing. The system utilizes a deep encoder for unmasked tokens (approximately 25% of input) and a shallow decoder for reconstruction using encoded features and mask tokens. This configuration reduces computational requirements compared to architectures that process all tokens through each layer. Experimental results indicate strong performance in generative pre-training but potential limitations in tasks requiring detailed audio analysis such as speaker identification, suggesting specific trade-offs in the architecture's feature representation capabilities.

The results obtained for each of these models are listed in Table 7, for the retrieval task, and for classification in Table 8. From this set of results, we selected the BEATs model because it performed better after initial integration with the other two encoders, image and text.

	Audio to Text				Text to Audio			
	AudioCaps							
Model	R@1	R@5	R@10	R@Avg	R@1	R@5	R@10	R@Avg
AudioMAE	27.29	59.08	74.74	53.70	5.47	22.41	34.11	20.66
HTSAT	10.21	30.55	43.22	27.99	1.69	7.63	14.31	7.87
MAE_AST	26.84	61.08	76.04	54.65	5.49	22.34	34.02	20.61
BEATs$_{wc/f\ 0.2}$	30.44	64.99	78.22	57.89	7.40	28.08	42.56	26.01
BEATs$_{ASC/AA/W}$	34.69	68.75	81.57	61.67	7.24	27.85	41.17	25.42
	ClothoV2							
AudioMAE	10.55	27.85	39.46	25.95	2.16	8.63	14.30	8.36
HTSAT	4.19	14.30	22.12	13.54	0.61	2.64	4.90	2.72
MAE_AST	9.91	26.81	37.89	24.87	2.22	9.21	15.06	8.83
BEATs$_{wc/f\ 0.2}$	19.94	47.42	61.33	42.90	4.42	17.04	27.51	16.32
BEATs$_{ASC/AA/W}$	9.45	25.70	36.13	23.76	1.56	6.59	10.94	6.36

Table 7: Ablation results between the different encoders tested in the optimization models phase, for the Audio to Text and Text to Audio retrieval task. On the AudioCaps and ClothoV2 datasets.

A.2 Hyperparameters and Computing Resources

The hyperparameters used for fine-tuning the general BEAT models are shown in Table 9. The basic model has a batch size of 64 and a number of epochs of 3.

To train the base models, a 48GB Quadro RTX 8000 GPU was used. On average, training took 90

	ESC-50	UrbanSounds8K
Model	Accuracy	
AudioMAE	67.8	65.34
HTSAT	34.2	48.16
MAE_AST	67.25	66.92
BEATs$_{wc/f\ 0.2}$	93.25	77.87
BEATs$_{ASC/AA/W}$	90.9	75.62

Table 8: Comparison between the different encoders tested in the optimization models phase, for classification task. On the ESC-50 and UrbanSounds8K datasets.

hours to complete. The models with expanded computational resources were trained on an NVIDIA A100 GPU with 80GB, with an average training time of 255 hours.

A.3 Augmentation

Two strategies were used for data augmentation: Random Truncation and SpecAugument. We carried out a preliminary stage to evaluate both augmentations. The different combinations of these augmentations are shown in Tables 10 and 11 for the Retrieval and Classification tasks.

- **Random Truncation (RT)** Truncates or pads the audio input to a fixed duration (in this work we use 10 seconds). For audio clips shorter than the target length, padding is applied in two stages: random padding with silence at the beginning, followed by additional padding to reach the target duration. For longer clips, a random segment of the required length is extracted from the audio. This method introduces variability by exposing the model to different temporal sections of the same audio during training, reducing overfitting while ensuring consistent input dimensions across the dataset.

- **SpecAugment (SpecAug)** (Park et al., 2019b) operates directly on the log mel spectrogram of input audio rather than the raw waveform. Initially developed for speech recognition tasks, this method has been successfully adopted for sound event detection and audio classification, as demonstrated in several recent studies (Kong et al., 2020; Gong et al., 2021; Chen et al., 2023a). The method comprises three main operations: (1) time warping, which deforms the time-series along the time direction, (2) frequency masking, where f consecutive mel frequency channels $[f_0, f_0 + f)$ are masked, with f chosen from a uniform distribution from 0 to the frequency mask parameter F, and f_0 selected from $[0, \nu - f)$ where ν is the number of frequency channels, and (3) time masking, where t consecutive time steps $[t_0, t_0 + t)$ are masked, with t chosen uniformly from 0 to the time mask parameter T, and t_0 selected from $[0, \tau - t)$. An upper bound prevents time masks from exceeding p times the number of time steps. Since the spectrograms are normalized to zero mean, setting masked values to zero is equivalent to setting them to the mean value. For the experiments that used SpecAugment we used $F = 48$ for the frequency masking parameter and $T = 96$ for the time masking parameter. These parameters control the maximum width of the frequency and time masks respectively.

A.4 Datasets

This section details the datasets used for training, validation, and testing our model. Our training strategy leverages a combination of large-scale, automatically annotated datasets and smaller, high-quality, human-annotated datasets. This approach allows us to benefit from the breadth of data provided by the larger datasets while also incorporating the precision and accuracy afforded by human labeling. Specifically, we utilize AudioSetCaps, WavCaps, Auto-ACD, AudioCaps, and Clotho v2 as our training data.

The rationale behind this selection stems from a need to balance data quantity and quality. Datasets like WavCaps, Auto-ACD, and AudioSetCaps offer substantial amounts of data, crucial for training robust and generalizable models. While these datasets are automatically annotated and thus potentially contain some noise, their sheer size compensates for this limitation. Prior work has demonstrated the effectiveness of training on these datasets individually, achieving promising results. Our approach builds upon this by combining them, hypothesizing that the combined data will lead to even better performance.

Complementing these large-scale datasets, we incorporate AudioCaps and Clotho v2. These datasets are meticulously annotated by human labelers, providing a "gold standard" of data quality. While smaller in size compared to the automatically

Hyperparameter	Expanded Resources	Basic Model
Batch size	110	64
Maximum text token length	77	
Maximum audio length	10 seconds	
Optimizer	Adam	
Weight decay	1e-6	
Adam ϵ	1e-8	
Adam β	[0.9, 0.98]	
Learning rate schedule	CosineWarmupLR	
Maximum learning rate	5e-5	
Minimum learning rate	1e-5	
# Epochs	10	2

Table 9: Hyperparameters for the Expanded Resources and Basic Model configurations.

	Audio to Text				Text to Audio			
	AudioCaps							
Model	R@1	R@5	R@10	R@Avg	R@1	R@5	R@10	R@Avg
BEATs + No-Augument	31.09	66.07	79.71	66.07	6.25	24.86	38.11	24.86
BEATs + SpecAug	31.05	65.60	79.55	65.60	6.59	25.53	38.45	25.53
BEATs + RT	30.33	65.15	80.09	65.15	6.21	25.51	38.13	25.51
BEATs + SpecAug + RT	32.01	66.28	79.80	66.28	6.32	25.40	39.46	25.40
	ClothoV2							
BEATs + No-Augument	10.12	28.98	41.63	28.98	2.51	9.68	15.90	9.68
BEATs + SpecAug	11.25	30.18	41.93	30.18	2.76	10.32	17.03	10.32
BEATs + RT	10.41	29.00	41.17	29.00	2.49	10.30	16.59	10.30
BEATs + SpecAug + RT	11.12	29.89	41.97	29.89	2.51	9.97	15.87	9.97

Table 10: Ablation results for retrieval task between the different datasets tested in the training dataset selection phase, for the Audio to Text and Text to Audio. On the AudioCaps and ClothoV2 datasets.

	ESC-50	UrbanSounds8K
Augumentation	Accuracy	
BEATs + No-Augument	76.15	62.31
BEATs + SpecAug	79.05	62.76
BEATs + RT	80.95	60.49
BEATs + SpecAug + RT	82.90	63.86

Table 11: Ablation results for classification task between the different datasets tested in the training dataset selection phase. On the ESC-50 and ClothoV2 datasets.

generated datasets, their high accuracy is essential for refining the model's understanding of complex audio-sound relationships and ensuring accurate caption generation. By training on a combination of these high-quality and large-scale datasets, we aim to create a model that is both comprehensive in its understanding of audio and accurate in its descriptions.

- **ESC-50** (Piczak, 2015) (Environmental Sound Classification) comprises 2,000 audio clips, each with a duration of 5 seconds, distributed across 50 distinct classes. These classes are grouped into five broader categories: animal sounds, natural soundscapes and water sounds, human non-speech sounds, domestic sounds, and urban noises. Each of these five categories contains 10 specific sound classes (for a total of 50), with each class represented by 40 audio clips. To facilitate consistent evaluation, the dataset provides predefined splits for 5-fold cross-validation.

- **UrbanSound8K** (Salamon et al., 2014) focuses specifically on urban environmental sounds, containing 8,732 labeled sound excerpts under 4 seconds from 10 distinct urban sound sources. The sounds include air conditioners, car horns, playing children, dog barking, drilling, engine idling, gunshots, jackhammering, sirens, and street music. The dataset is organized into 10 folds for cross-validation, making it a standard benchmark for urban sound classification.

- **VGG-Sound** (Chen et al., 2020) is an audio-visual dataset containing over 200k video clips of 10 seconds each, spanning 309 distinct sound classes. These classes include musical instruments, human sounds, animal vocalizations, environmental noises, and mechanical sounds, with each class containing 200 to 1,000 clips. The clips were collected from diverse, unconstrained environments to reflect real-world acoustic conditions. The dataset was curated using a multi-stage verification process involving visual classification, audio validation, and noise filtering, ensuring high-quality and consistent data.

- **AudioSet** (Gemmeke et al., 2017) comprises of over 2 million human-labeled 10-second YouTube video excerpts. It's organized in a hierarchical ontology of 527 sound classes. While extremely comprehensive, it has an unbalanced distribution, with some classes having significantly more samples than others. The dataset provides both balanced and unbalanced training sets, along with a consistent

evaluation set.

- **AudioCaps** (Kim et al., 2019b) builds on AudioSet, containing 46K audio-caption pairs with varying caption density across splits. The training set includes 38,118 clips with single captions, while validation and test sets have 500 and 979 clips respectively, each with five captions. The dataset's curation process deliberately excluded music categories, visually dependent sounds, and expert knowledge categories. During caption collection, annotators received AudioSet labels as word hints, with video hints available as a last resort. The dataset emphasizes describing auditory content over visual elements.

  Clotho (Drossos et al., 2020) represents a focused effort on audio captioning with 4,981 audio samples of 15 to 30 seconds in duration, and 24,905 captions total. Drawing from the Freesound platform, the audio samples cover diverse environmental and acoustic content. During data collection, annotators wrote captions based solely on audio signals, without access to visual cues or word tags. The dataset underwent post-processing to remove named entities, speech transcription, and words appearing only once, while retaining natural language descriptions of sound events, acoustic scenes, and spatial-temporal relationships.

- **MACS** (Morato and Mesaros, 2021) (Multi-annotator Captioned Soundscapes) contains approximately 4,000 audio samples with multiple human annotations per clip. Each audio is restricted to a 10-second duration. What distinguishes MACS is its use of professional annotators and a structured annotation process that ensures high-quality, consistent captions focused purely on auditory content.

- **WavCaps** (Mei et al., 2024a) represents the largest scale effort with approximately 400,000 audio-caption pairs sourced from FreeSound, BBC Sound Effects, SoundBible, and AudioSet. What sets it apart is its innovative three-stage processing pipeline. First, it filters out clips shorter than one second and removes repetitive descriptions. Then, it employs ChatGPT to transform raw descriptions into proper captions. Finally, it removes named entities and extremely brief captions. While it's considered weakly labeled due to its automated processing, WavCaps maintains caption quality through this structured approach, making it valuable for large-scale audio-language training.

- **Auto-ACD** (Sun et al., 2024) is a large-scale audio-language dataset containing 1.5M audio-caption pairs. Each audio clip is paired with a detailed caption that averages 18 words in length, drawing from a vocabulary of approximately 23K words. The captions encompass comprehensive descriptions of acoustic events, environmental context, and scene settings. The dataset uses audio clips from YouTube videos and provides rich descriptive text that goes beyond simple sound labels to include detailed acoustic and environmental information.

- **AudioSetCaps** (Bai et al., 2024) comprises 1.9M audio-caption pairs built upon AudioSet recordings. The dataset provides extensive coverage of audio content through detailed captions that describe not only the primary sound events but also their characteristics and environmental context. The captions are enriched with fine-grained audio information including spoken language details, speech emotions, musical instruments, and music genres. The dataset maintains high caption quality through a refinement process that ensures accuracy and relevance to the audio content.

A.5 Qualitative Analysis

For a qualitative evaluation of our model's cross-modal retrieval capabilities, we conducted experiments on the AudioCaps test set. We present representative examples of successful and unsuccessful retrieval outcomes for both text-to-audio and audio-to-text tasks in Figures 3, 4, 5, and 6. Furthermore, we also investigate the emergent capabilities of our model in audio-to-image and image-to-audio retrieval tasks, showing its ability to implicitly align the audio and image modalities, that were not explicitly aligned during training. Specifically, we leveraged the test subset of the VGG-Sound dataset, extracting audio and the central video frame following the methodology outlined in (Guzhov et al., 2022). These results are visualized in Figures 7 and 8.

Figure 3 illustrates successful results of the text-to-audio retrieval task. In each example shown,

the ground truth audio clip corresponding to the input text query is present within the top-3 ranked retrieval results. These ground truth matches are highlighted in red. For each retrieved audio clip, we display a representative frame extracted from the corresponding YouTube video segment and the original caption associated with the audio clip. This visualization allows for a direct comparison between the textual query, the retrieved audio content (as represented by the video frame and caption), and the ground truth.

Figure 4 showcases examples of unsuccessful text-to-audio retrieval. In these instances, the ground truth audio clip corresponding to the input text query is absent from the top-3 ranked retrieval results, indicating a mismatch between the text and the retrieved audio.

Figure 5 demonstrates successful audio-to-text retrieval. Given an audio clip as a query, the model retrieves the top-ranked textual descriptions. The retrieved text is displayed, with the ground truth description highlighted in red. To provide context for the audio query, we include a representative frame extracted from the corresponding YouTube video segment. This frame, along with the original text caption associated with the audio, helps to clarify the content of the audio query.

Figure 6 presents examples of unsuccessful audio-to-text retrieval. Here, the ground truth textual description corresponding to the audio query is not found within the top-3 ranked retrieval results, indicating a failure to accurately capture the audio's content in the retrieved text.

Figure 7 illustrates successful audio-to-text retrieval. Given an audio clip as a query, the model retrieves the top-ranked images. The retrieved image is displayed, with the ground truth description highlighted in red.

Figure 8 presents successful image-to-audio retrieval. Given an image as a query, the model retrieves the top-ranked audios. The retrieved audio is represented by a representative frame from the corresponding video and the associated caption. The ground truth audio is highlighted in red.

Figure 3: Examples of successful text-to-audio retrieval. The red highlighted image and text indicates the ground truth audio clip retrieved within the top 3 results for the given text query. The image and caption provide context for the retrieved audio.

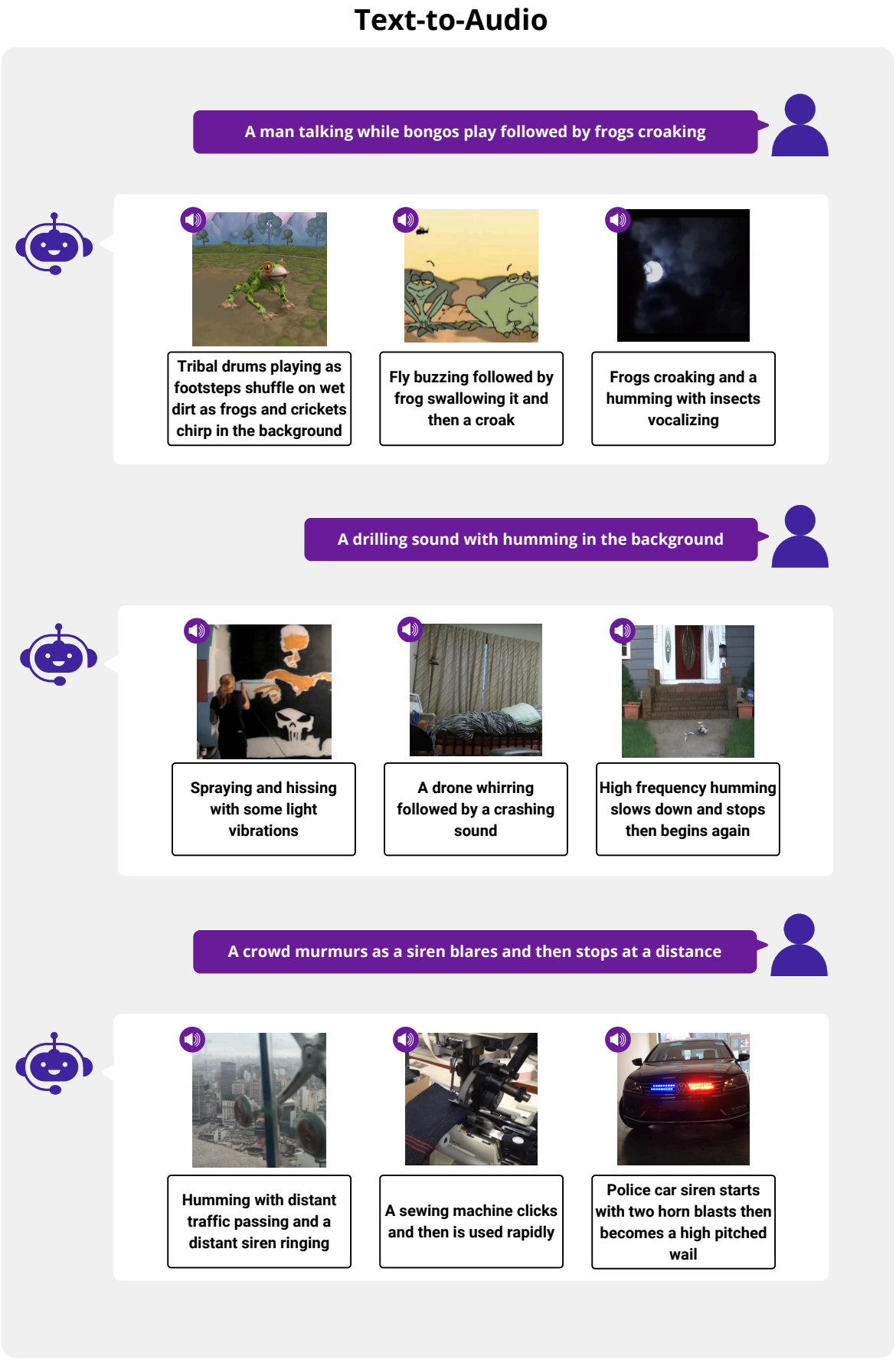

Figure 4: Examples of unsuccessful text-to-audio retrieval. The ground truth audio was not found within the top 3 results for the given text query.

Figure 5: Examples of successful audio-to-text retrieval. The red highlighted text indicates the ground truth text description retrieved for the given audio query. The image and caption provide context for the audio.

Figure 6: Examples of unsuccessful audio-to-text retrieval. The ground truth text was not found within the top 3 results for the given audio query.

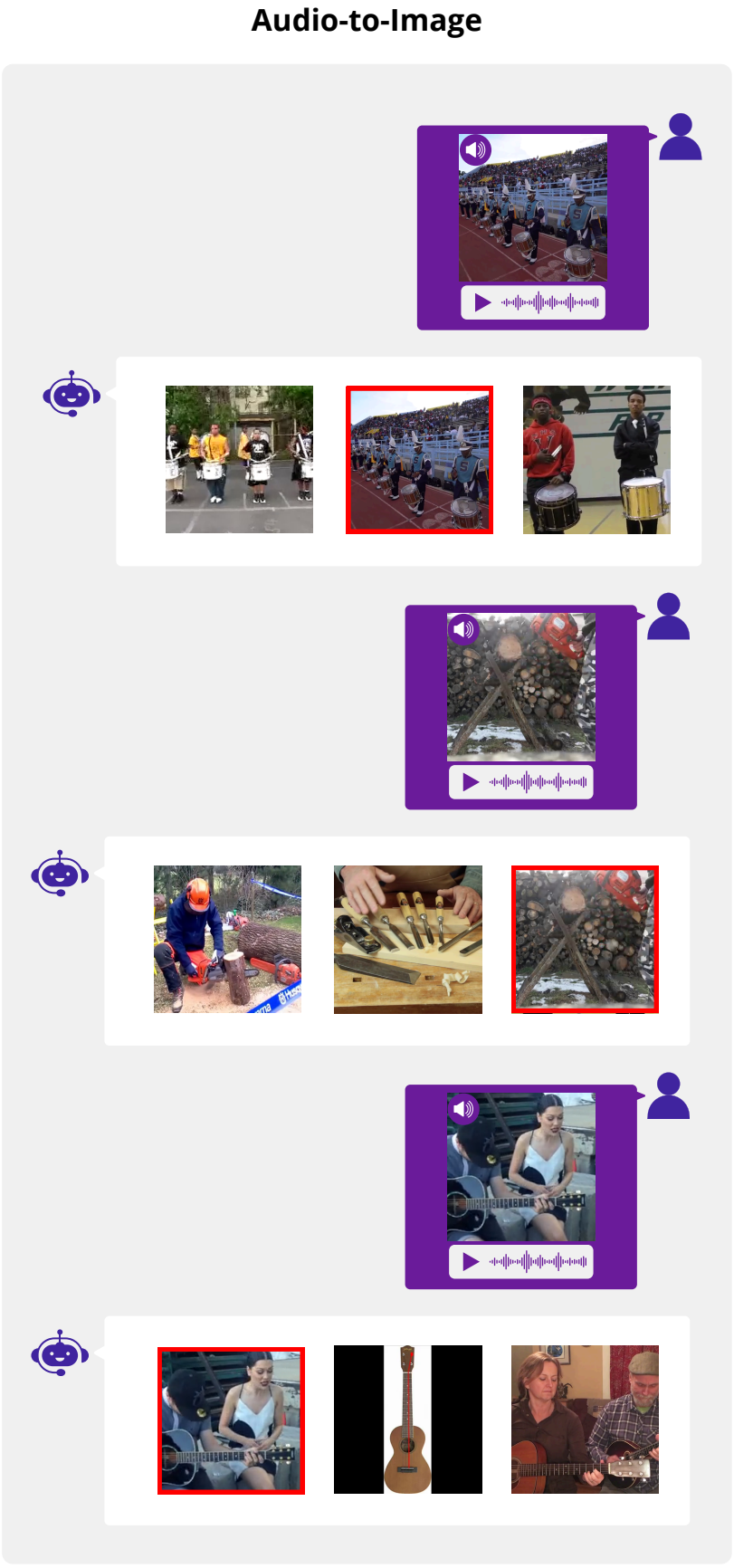

Figure 7: Examples of successful audio-to-image retrieval. The red highlighted image indicates the ground truth image retrieved for the given audio query.

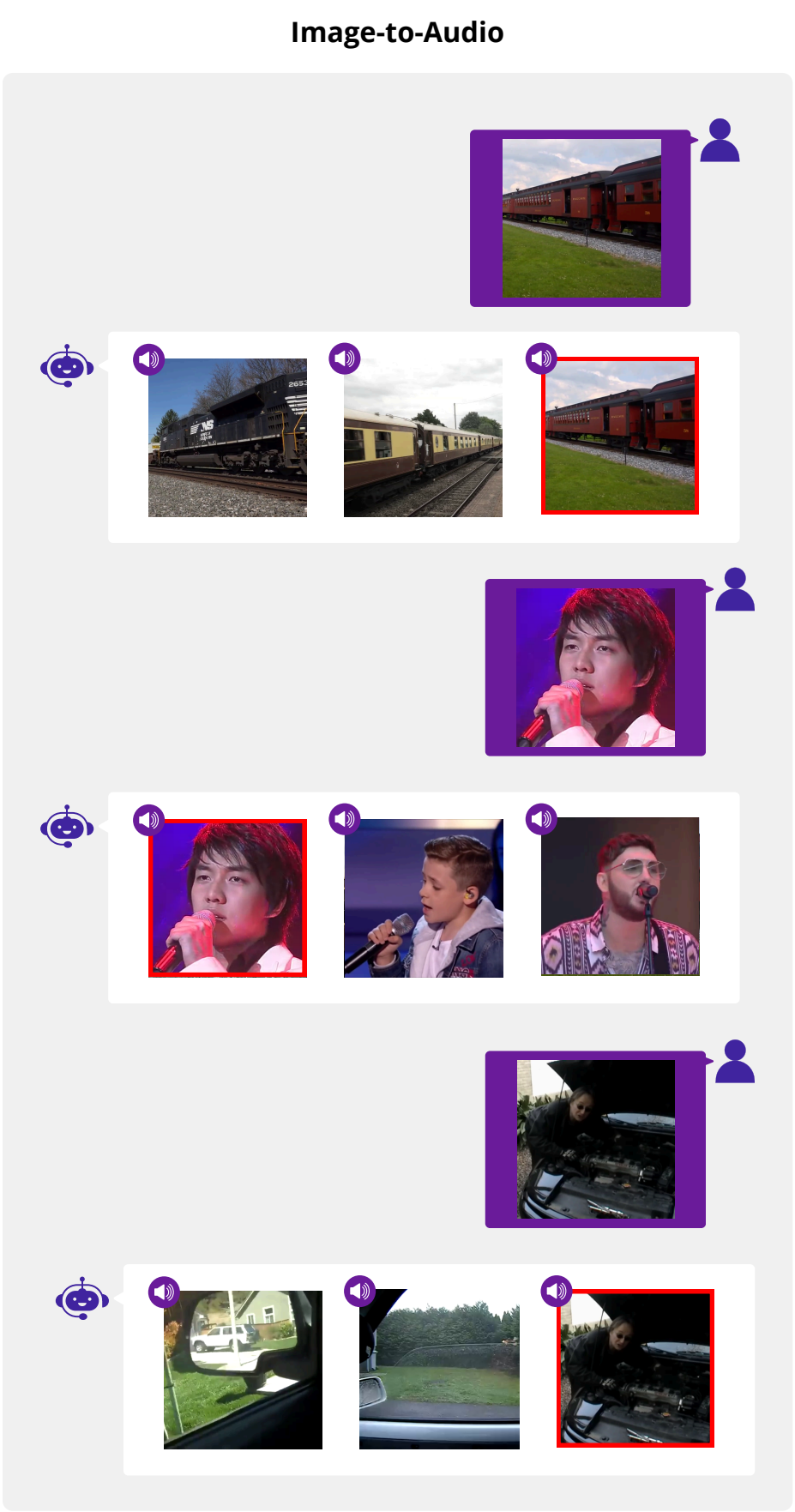

Figure 8: Examples of successful image-to-audio retrieval. The red highlighted text indicates the ground truth audio retrieved for the given image query.