Audio-Text Representation Learning with Temporal and Semantic Alignment

Tatsuya Komatsu

Contents

Abstract

Audio-text representation learning has emerged as a foundational technology for multimodal AI, enabling applications from retrieval to language-based audio generation. While frameworks such as Contrastive Language-Audio Pretraining (CLAP) have established a shared latent space, the prevailing paradigm suffers from representational and practical bottlenecks. In terms of representation, current models handle audio and text sequences as global embeddings, compressing complex temporal sequences and rich semantics into single static vectors, which entangle multiple events and inherently discard fine-grained temporal and semantic structure. Practically, when such entangled global embeddings are the only representation available, the model can recover temporal and semantic regularities only indirectly from large and diverse collections of paired examples. Unlike in the image domain, however, obtaining rich descriptions for audio is intrinsically costly and non-scalable. As a result, models often fail to distinguish intricate temporal patterns, cannot fully disentangle mixed sound events, and struggle to generalize beyond the limited paired data.

This thesis identifies three challenges at the intersection of these representational and practical limitations: the loss of fine-grained temporal alignment between audio and text, the absence of explicit intra-modal relations, and a strong reliance on scarce paired data. First, conventional global pooling aggregates variable-length audio into a fixed vector, discarding essential temporal information and preventing precise align-

ment between specific sound events and textual phrases. Second, standard contrastive objectives focus solely on cross-modal matching, neglecting intra-modal relations (i.e., audio-audio and text-text similarities) within each modality. As a result, given two compositional scenes such as "violin over street noise" and "guitar over street noise," the model cannot articulate that the foreground instrument differs; it can only judge overall similarity. Third, high-quality audio-text pairs are extremely scarce compared to those in the computer vision domain. The data-hungry nature of contrastive learning means that even advanced architectures cannot reach their full potential without a way to exploit large collections of unlabeled audio.

To address these challenges, this thesis proposes a series of complementary methods that achieve fine-grained temporal and semantic alignment while improving data efficiency. To mitigate the loss of temporal granularity, we first introduce Aligned Contrastive Learning. By explicitly modeling frame-level audio features and token-level text embeddings through a frame- and token-wise similarity measure, the method achieves fine-grained temporal alignment. We empirically validate this approach on large-scale text-to-music retrieval, demonstrating that fine-grained alignment improves retrieval performance for temporally complex audio.

To address the lack of intra-modal relations, we then propose Audio Difference Learning. Moving beyond cross-modal discrimination, this generative framework trains the model to articulate the semantic difference between an input and a reference audio clip, thereby aligning audio-level differences with textual descriptions. We design a Diff Block, including a masking-based variant implemented via cross-attention, that isolates the residual information of one clip with respect to another. By training on synthetic mixtures and encouraging the difference representation to reconstruct the caption of the target source, the model learns to disentangle mixed acoustic events in the feature

space and acquire a compositional understanding of audio scenes. This framework improves captioning performance while enabling semantic alignment through explicit "audio arithmetic" in the latent space.

Finally, to alleviate the dependence on paired data, we introduce a Semi-Supervised Representation Learning framework that directly leverages large-scale unlabeled audio to update a CLAP-based backbone. The key innovation is an Audio-to-Text (a2t) Mapper that projects audio embeddings into the text token embedding space as continuous pseudo-token sequences, enabling implicit alignment between unlabeled audio and the text modality. This allows unlabeled audio to participate in the contrastive learning loop as if it were text, densifying the training distribution and regularizing the shared audio-text space without relying on synthetic captions. We explore both MLP- and Transformer-based mappers, as well as noise-injection strategies, and show that incorporating unlabeled audio in this way yields substantial gains in retrieval performance.

In summary, this thesis tackles the representational and practical limitations of audio-text representation learning by achieving temporal and semantic alignment while improving data efficiency. Fine-grained alignment enables temporally resolved representations, difference-based modeling provides compositionally structured semantic alignment, and the semi-supervised framework demonstrates that the data bottleneck can be mitigated by bridging the modality gap at the feature level. Together, these contributions establish a more expressive, interpretable, and scalable paradigm for audio-text representation learning, paving the way toward next-generation multimodal audio foundation models.

Chapter 1

Introduction

1.1 Background

Modeling real-world acoustic environments using natural language constitutes an essential step toward connecting human perceptual understanding with machine information processing. Audio-text modeling, which integrates audio signals and text within a shared representation space, has emerged as a core technology [1], [2]. It supports a wide range of applications, including audio retrieval [3], [4], audio captioning [5], zero-shot sound classification [6], [7], and sound generation [8].

Audio understanding has evolved significantly from traditional label-based classification to natural-language descriptions. Early approaches relied on closed-set taxonomies with fixed label vocabularies [9], [10], which inherently constrained the semantic expressiveness of learned audio representations. The shift toward natural language enables open-vocabulary understanding, whereby models can capture nuanced acoustic concepts that transcend categorical boundaries. This evolution mirrors the progression observed in computer vision, exemplified by the shift from image classification to image captioning and visual question answering.

The broader framework of multimodal representation learning, which encompasses audio-text approaches, has its origins in cross-modal contrastive representation learning methods developed in the 2010s, such as DeViSE [11] and Deep CCA [12]. These methods first systematized the idea of "mapping different modalities into a common latent space," primarily targeting correspondence learning between images and language. Subsequently, contrastive learning has generalized this concept to large-scale datasets, and with the advent of Contrastive Language-Image Pretraining (CLIP) [13], it has become the standard paradigm for multimodal representation learning. CLIP, trained on hundreds of millions of image-text pairs, has demonstrated strong generalization performance across diverse tasks, including zero-shot classification and caption generation.

Building on this success, methods such as Wav2Clip [14], AudioCLIP [15], and CLAP [1], [2] extended the contrastive framework to the audio-text domain, accelerating research on audio-text representation learning. However, unlike the vision-language domain, the audio-text domain faces unique challenges arising from the temporal continuity, transience, and superposition of acoustic events. Specifically, while images are spatially bounded and can be captured in a single snapshot, audio signals unfold over time with sound sources appearing, overlapping, and fading at arbitrary moments [16], [17]. Moreover, real-world sounds, ranging from human speech to environmental sounds, exhibit greater acoustic variability than the object-centric structures typically found in images [18]. These factors influence both data collection and representation design.

For instance, annotating audio data is considerably more costly than labeling images. While an image can be labeled at a glance, audio annotation requires sequential listening, often in real time or slower, making the process inherently more time-

consuming [19]. Furthermore, unlike images, where visual objects are spatially localized and can be independently annotated, sounds in an audio clip frequently overlap in time, requiring annotators to disentangle concurrent events [20], [21]. This fundamental difference in annotation effort results in a substantial gap between available image-text pairs (on the order of billions) and audio-text pairs (from tens of thousands to millions) [22], [23]. This practical bottleneck in annotation underlies data scarcity and amplifies the reliance of contrastive learning on large-scale data.

Furthermore, determining how to map the temporal structure of audio signals and text sequences into a fixed-size representation remains a central issue. Unlike images, which possess a fixed spatial resolution that naturally maps to a grid of patches [24], audio signals encompass diverse acoustic phenomena, including environmental sounds, human speech, and music, and vary drastically in duration, ranging from sub-second sound effects to multi-minute musical compositions. In practice, this variable-length nature is typically handled by first converting audio into fixed-length segments and then applying global pooling, or by applying temporal pooling to obtain a single global embedding [25], [26]. Such sequence-level aggregation, however, introduces representational trade-offs. It limits the granularity of cross-modal alignment and constrains how temporal relationships between sound events are preserved, thereby forming the backdrop for the limitations discussed in the following sections.

1.2 Audio-Text Contrastive Learning

1.2.1 Overview

To illustrate the setting of audio-text contrastive learning, consider the example shown in Fig. 1.1: an audio clip containing a short violin melody followed by audience

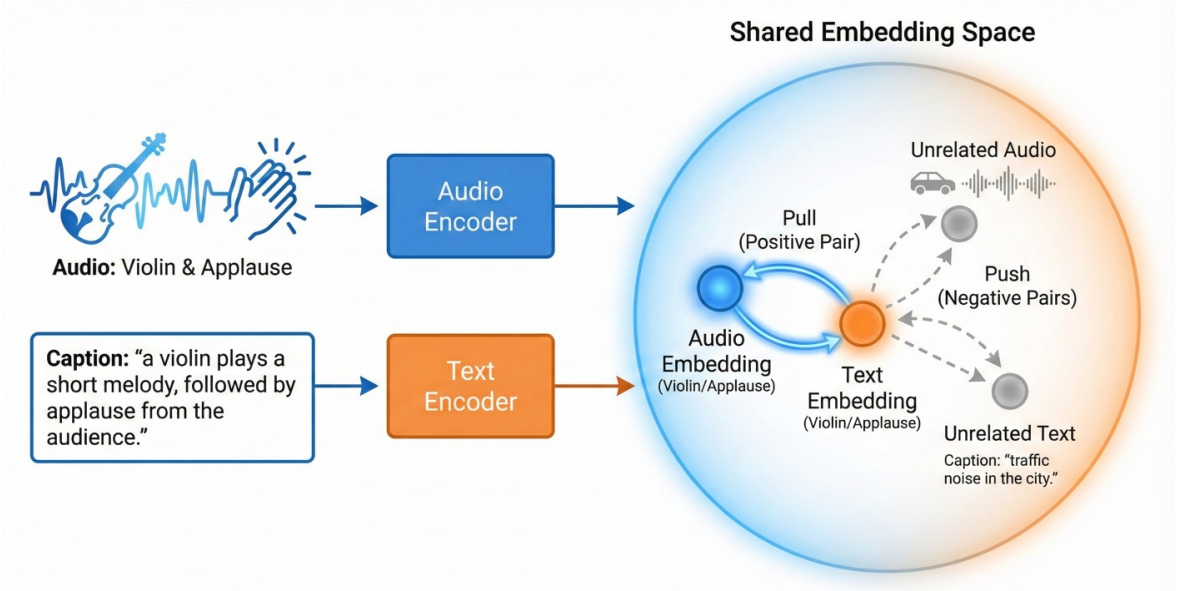

Figure 1.1: Concept of audio-text contrastive learning.

applause, together with a natural-language caption such as "a violin plays a short melody, followed by applause from the audience." The goal is to learn encoders that map both the audio clip and the caption into a shared embedding space, so that their embeddings are close to each other while being far from the embeddings of semantically unrelated audio clips and captions.

Audio-text contrastive learning formalizes this idea at scale. A training corpus is composed of paired audio clips and captions, and a minibatch of such pairs is processed jointly. For each audio clip in the batch, its own caption is treated as a positive example, while the captions of other clips in the batch are treated as negatives; the same construction is applied in the reverse, text-to-audio direction. The model is then optimized so that matching audio-text pairs achieve higher similarity scores than non-matching pairs [27].

We decompose the framework along two design axes: *representation design*, which specifies how audio and text are mapped from variable-length signals and token se-

quences into fixed-size embeddings, and *objective design*, which specifies how these embeddings are compared and optimized during training.

1.2.2 Representation Design

In audio-text contrastive models, an audio input is first converted into a sequence of frame- or segment-level features, and a text input is tokenized into a sequence of word or subword embeddings. Each modality-specific encoder (e.g., Transformer [28]) processes these sequences and produces modality-specific sequences of embeddings for audio and for text, respectively. To enable contrastive learning, these variable-length sequences are then compressed into a single global embedding for each modality, typically using mean pooling, attention pooling, or a designated aggregation token (such as the `[CLS]`token in Transformer models [28], [29]). Cross-modal similarity is then computed between these clip-level audio and sentence-level text embeddings, regardless of the original sequence lengths.

This *sequence-level aggregation* makes audio and text easily comparable in a shared space and supports scalable training. It also determines the granularity at which audio-text alignment is represented. In most contrastive models, similarity is defined for whole clips and whole sentences. The implications of this design choice for fine-grained audio-text alignment will be discussed in Section 1.3.1.

1.2.3 Objective Design

On top of these global embeddings, audio-text contrastive learning usually adopts an InfoNCE-type loss [27]. Given a minibatch of audio and text embeddings, the model constructs a similarity matrix between all audio-text pairs in the minibatch. For each

audio embedding, the paired text is treated as a positive example, while all other texts in the batch are treated as negatives; the same is done in the reverse direction. The loss then maximizes the similarity of each positive pair relative to all negatives using a softmax over similarity scores, in both audio-to-text and text-to-audio directions.

This optimization is purely relative. To make this more concrete, consider a batch containing two audio clips: (A) a solo violin playing a melody and (B) a dog barking, with corresponding captions "a solo violin plays a melody" and "a dog is barking". The loss only requires the embedding of (A) to be closer to its own caption "a solo violin plays a melody" than to "a dog is barking", and the embedding of (B) to be closer to "a dog is barking" than to "a solo violin plays a melody". The loss function does not specify where these four points should lie in the embedding space, as long as these relative inequalities are satisfied. More generally, the objective does not directly constrain absolute distances or the global geometry of the embedding space.

Moreover, the objective depends only on cross-modal similarities between audio and text [30]. There is no term that explicitly encourages acoustically similar sounds to be close to each other, or dissimilar sounds to be far apart, as long as each sound remains closer to its own caption than to the captions of other sounds. In other words, relations between sounds, specifically how two audio clips are similar or different, are not explicitly optimized; they can only be inferred indirectly through cross-referencing many audio-text pairs.

These characteristics of the standard contrastive objective underlie the limitations regarding intra-modal relations between sounds and the model's dependence on dataset diversity.

1.3 Limitations of Audio-Text Contrastive Learning

Despite its effectiveness, the standard formulation of audio-text contrastive learning faces three interlinked limitations: coarse-grained representations arising from sequence-level aggregation, unstructured intra-modal relations due to purely relative optimization, and a consequent heavy reliance on large-scale data. These issues are not merely practical hurdles; they arise from the design of current frameworks [31], [32] as analyzed in studies on the invariances learned by contrastive objectives [33]. In this section, we analyze how these design choices constrain the granularity, intra-modal relation, and data efficiency of the learned representations.

1.3.1 Representation: Loss of Granularity

The first limitation arises from the sequence-level aggregation discussed in Section 1.2. While compressing variable-length sequences into global vectors simplifies cross-modal comparison, it inevitably discards local temporal and structural information.

As illustrated in Figure 1.2, consider an audio clip containing "a violin melody followed by applause." Under standard global aggregation (e.g., mean pooling), the temporal alignment between events and their descriptions is lost: the model cannot preserve that the violin precedes the applause. A clip containing the reverse sequence ("applause followed by a violin melody") would yield a nearly identical global embedding, making the two scenarios indistinguishable in the representation space. Furthermore, this aggregation entangles distinct sound events into a "bag-of-events" representation [34], [35], where contributions from temporally and spectrally different events are collapsed

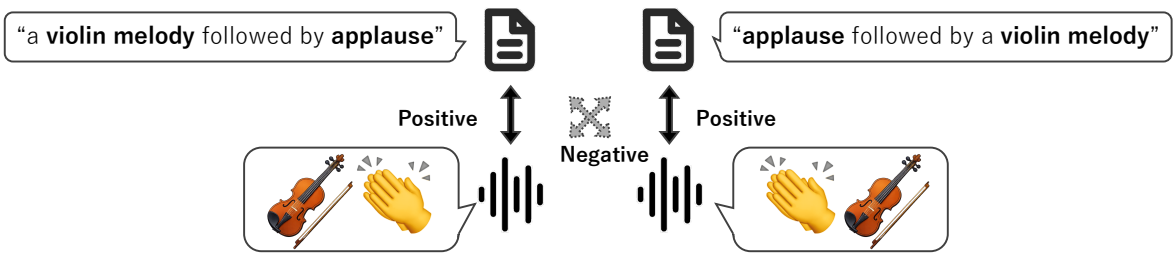

Figure 1.2: Illustration of the loss of fine-grained temporal alignment under global aggregation. Although the two audio clips share nearly identical content (violin and applause), they are treated as negatives because they are different samples. Meanwhile, global embeddings cannot distinguish the temporal order, making the positive audio-text pairs ambiguous.

into a single abstract vector. As a result, the mapping from event-level relation to the global embedding becomes underdetermined, such that many different configurations of sound events can be mapped to similar global vectors. This entanglement makes it difficult for the model to preserve fine-grained alignment between individual temporal frames and their corresponding textual phrases.

1.3.2 Objective: Unstructured Intra-Modal Relations

As illustrated in Figure 1.3, the second limitation stems from the relative nature of the contrastive objective. As noted in the previous section, the objective optimizes only the similarity between matching audio-text pairs, without explicitly constraining the relations within each modality, i.e., audio-audio and text-text relations [30].

To illustrate this, consider three audio clips: (A) a violin, (B) a guitar, and (C) a dog barking. Acoustically, the violin (A) and guitar (B) are similar, while the dog (C) is distinct. Ideally, the model should learn that A and B are close in the embedding

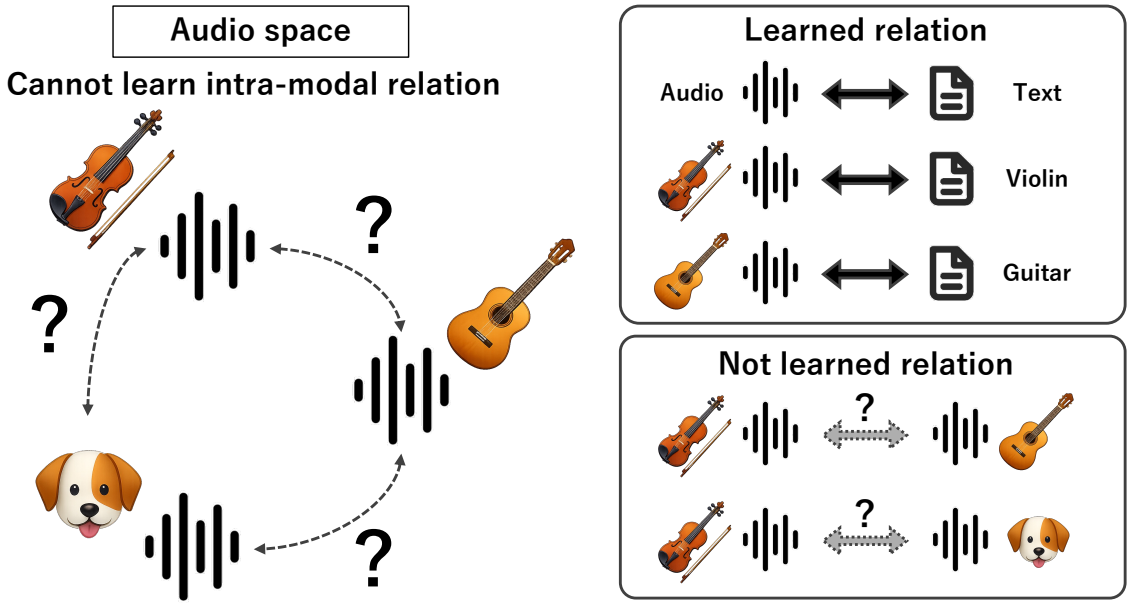

Figure 1.3: Illustration of unstructured intra-modal relations in contrastive learning. While audio-text correspondences are explicitly learned, relations between audio clips, such as whether violin is more similar to guitar than to dog barking, remain underdetermined.

space, while C is far away. However, since the standard objective does not measure audio-audio and text-text similarity directly, the model does not explicitly learn that A should be closer to B than to C. Instead, information about how close Sound A and Sound B should be must be inferred indirectly from their co-occurrence patterns with text.

Moreover, intra-modal relations encompass not just classifying two sounds as "similar clips" or "different clips," but articulating what specifically differs between them. In reality, audio clips typically contain multiple overlapping sound events, and captions mention multiple entities, actions, and attributes (e.g., "a dog barking over music in a crowded room"). A model with meaningful intra-modal understanding should identify

which components differ: given "violin over street noise" and "guitar over street noise," it should state that the foreground instrument differs, rather than merely judging overall similarity. However, as discussed in Section 1.3.1, the contributions of these components are entangled within a single audio embedding and a single sentence embedding. Without explicit alignment between audio components and their corresponding textual phrases, such fine-grained relational reasoning becomes difficult. Informally, each audio-text pair provides only a single constraint on many latent components, much like a single underdetermined equation in many unknowns.

To infer how Sound A and Sound B are related, the model must therefore observe many different combinations in which the corresponding words (such as "violin", "guitar", and "dog") appear in similar linguistic contexts. In effect, acoustic relations between A and B are reconstructed by cross-referencing massive numbers of audio-text pairs and statistically disentangling which parts of the captions correspond to which acoustic components. In the absence of such extensive cross-referencing, the audio embedding space tends to simply mirror the lexical structure of the text space [36].

1.3.3 Dependence on Data Scale

These representational and objective limitations directly amplify the dependence on large-scale data. As discussed in Section 1.1, paired audio-text data are inherently scarce and costly to annotate: annotators must listen to clips sequentially, overlapping sound events make labeling difficult, and the resulting audio-text corpora are orders of magnitude smaller than image-text datasets [19]. However, in the standard audio-text contrastive learning framework, the issue is not merely that data are scarce; the framework itself is data-inefficient.

Because sequence-level aggregation discards local temporal details and entangles

multiple sound events into a single pooled vector, the model cannot rely on the architecture alone to keep event-level information separate. Instead, it requires dense coverage of audio-text pairs to statistically reconstruct fine-grained relationships and to disentangle the event-level contributions inside the "bag-of-events" representations induced by global pooling. The model effectively needs to observe millions of examples to implicitly learn structures that could, in principle, be encoded explicitly in the architecture or the objective. This creates a negative loop: weaknesses in the model design demand data scales (e.g., millions to billions of pairs) that are attainable in vision-language domains but prohibitively difficult to obtain in the audio domain. Therefore, addressing the data-scarcity problem in audio-text learning requires not only collecting more data but also fundamentally rethinking representation and objective designs to be more data-efficient.

1.4 Research Objectives

The limitations identified above arise from two intertwined aspects of current audio-text contrastive learning: (i) design choices in representation and objective that compress multi-event audio and multi-phrase text into entangled global embeddings, leaving fine-grained cross-modal alignment and intra-modal relations underdetermined, and (ii) a strong dependence on large-scale paired data, which is difficult to satisfy in the audio domain. Together, these challenges motivate the following three research questions (RQs) that this thesis addresses.

RQ1: Fine-Grained Temporal Alignment. Current contrastive frameworks compare audio and text primarily through global embeddings, capturing cross-modal relationships only at a coarse, sequence-level resolution. As discussed in Section 1.3.1,

sequence-level aggregation entangles multiple sound events into a single "bag-of-events" representation and obscures which parts of the audio correspond to which textual phrases. **How can we represent and learn audio-text relationships at a finer granularity while maintaining the efficiency and scalability that make contrastive learning attractive?**

RQ2: Semantic Alignment for Intra-Modal Relations. Because the contrastive objective explicitly optimizes only cross-modal similarity, relations within each modality, such as audio-audio or text-text similarities, are not directly modeled. Consequently, these intra-modal relations are underdetermined and must be inferred indirectly through cross-referencing with the other modality, based on many entangled audio-text pairs. **How can we incorporate explicit intra-modal relational learning so that relations between audio clips are captured in a more semantically meaningful way within the learned representations?**

RQ3: Data Utilization Beyond Paired Data. The limitations above amplify the dependence of audio-text contrastive models on large-scale paired datasets, which are costly to collect and inherently scarce in the audio domain. Moreover, the vast majority of available audio is unlabeled and thus remains unused in standard contrastive frameworks, even though it contains rich event-level structures that could help disentangle acoustic factors. **How can we design a learning framework that effectively leverages both labeled and unlabeled audio, alleviating the reliance on massive paired datasets while preserving the quality of audio-text alignment?**

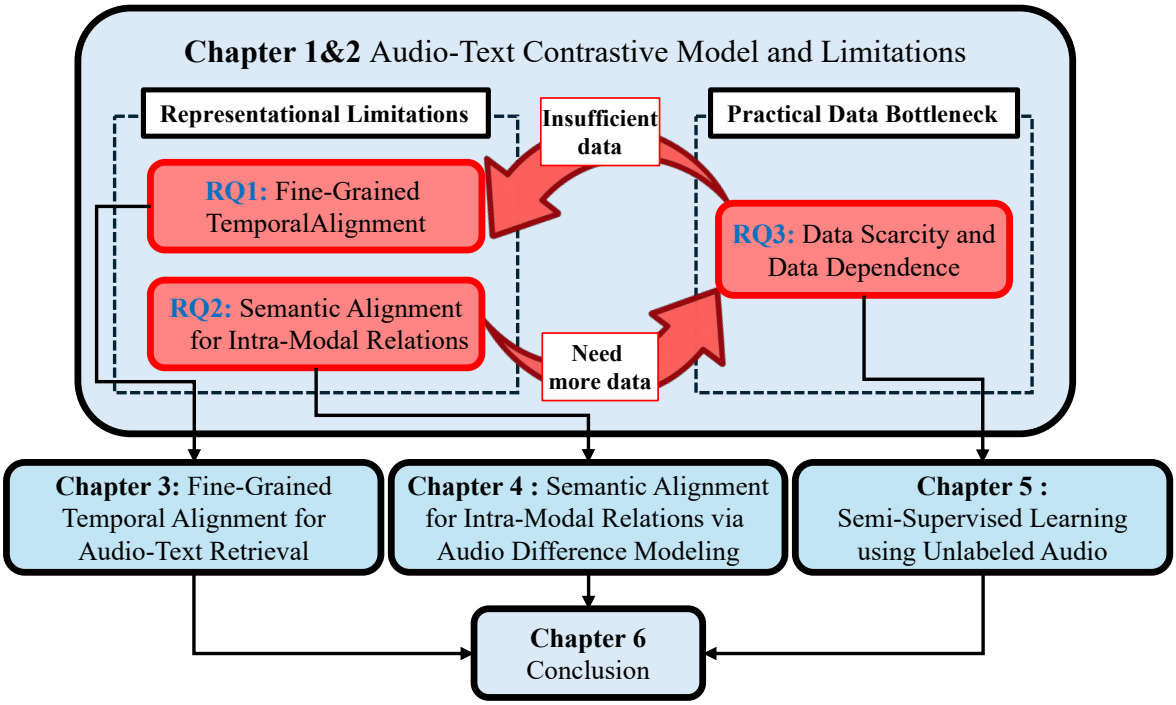

Figure 1.4: Overview of the thesis. Chapters 1 and 2 analyze contrastive audio-text models and identify three limitations: (1) loss of fine-grained alignment, (2) absence of intra-modal relations between sounds, and (3) data scarcity and dependence on paired audio-text data. Chapters 3-5 address these limitations respectively, and Chapter 6 concludes.

1.5 Thesis Overview

The overall structure of this thesis and the relationships between the three limitations and the proposed methods are summarized in Fig. 1.4.

Chapter 2 reviews the foundational literature and recent advancements in cross-modal representation learning. We begin by tracing the evolution of vision-language models, from early canonical correlation analysis to the breakthrough of CLIP. We then discuss how these paradigms have been adapted to the audio domain (e.g., CLAP) and

highlight the unique challenges posed by the temporal nature of audio and the scarcity of paired datasets. This chapter shows how existing global alignment approaches lead to three mutually reinforcing limitations, loss of fine-grained alignment, absence of intra-modal relations, and scarcity of paired audio-text data, and positions the methods proposed in the subsequent chapters.

Chapter 3 addresses the limitation of fine-grained temporal alignment (**RQ1**). We introduce *Aligned Contrastive Learning*, a framework that achieves fine-grained temporal alignment between frame-level audio features and token-level text embeddings. Moving beyond the conventional global pooling strategy, this method preserves local temporal details and reduces the entanglement of multiple events within a single global embedding. We evaluate this approach on a large-scale text-to-music retrieval task, demonstrating that fine-grained modeling significantly enhances retrieval performance for temporally dynamic audio content.

Chapter 4 tackles the lack of semantic alignment for intra-modal relations (**RQ2**). We propose *Audio Difference Learning*, a generative framework designed to disentangle mixed acoustic events in the feature space. By training the model to generate captions that describe the semantic difference between an input and a reference audio clip, we encourage the model to acquire a compositional understanding of sound, enabling semantic alignment through explicit audio arithmetic in the latent space. Experimental results on audio captioning benchmarks confirm that this approach enables the model to perform "semantic arithmetic" in the latent space, providing a way to model relations between sounds beyond purely cross-modal constraints.

Chapter 5 confronts the fundamental challenge of data scarcity (**RQ3**). We present a *Semi-Supervised Representation Learning* framework that effectively leverages large-scale unlabeled audio. By introducing an Audio-to-Text (A2T) Mapper, we project

unlabeled audio into the tokenized text embedding space as pseudo-token features, enabling implicit alignment between unlabeled audio and the text modality without additional captions. This chapter demonstrates that the use of unlabeled data significantly improves retrieval performance while alleviating dependence on massive paired datasets.

Finally, **Chapter 6** summarizes the main contributions of this thesis. We discuss the overarching insights gained from refining alignment, intra-modal relations, and data efficiency, and outline potential directions for future research towards unified audio-language foundation models.

Chapter 2

Related Work

2.1 Introduction

This chapter provides a comprehensive overview of the foundational frameworks and recent advancements relevant to audio-text representation learning. We begin by tracing the evolution of cross-modal learning in the vision-language domain, identifying the origins of the paradigms currently dominant in audio modeling. Next, we examine how these frameworks were adapted to the audio domain, specifically focusing on the architecture of Contrastive Language-Audio Pretraining (CLAP) and its derivatives. We then analyze the current data landscape, highlighting the critical disparity in scale between image and audio datasets. Finally, we synthesize the representational limitations of existing CLAP-style models, namely, the loss of temporal granularity, the lack of intra-modal relations, and data inefficiency, which serve as the primary motivation for the methodologies proposed in this thesis. Figure 2.1 illustrates the overall structure of this chapter.

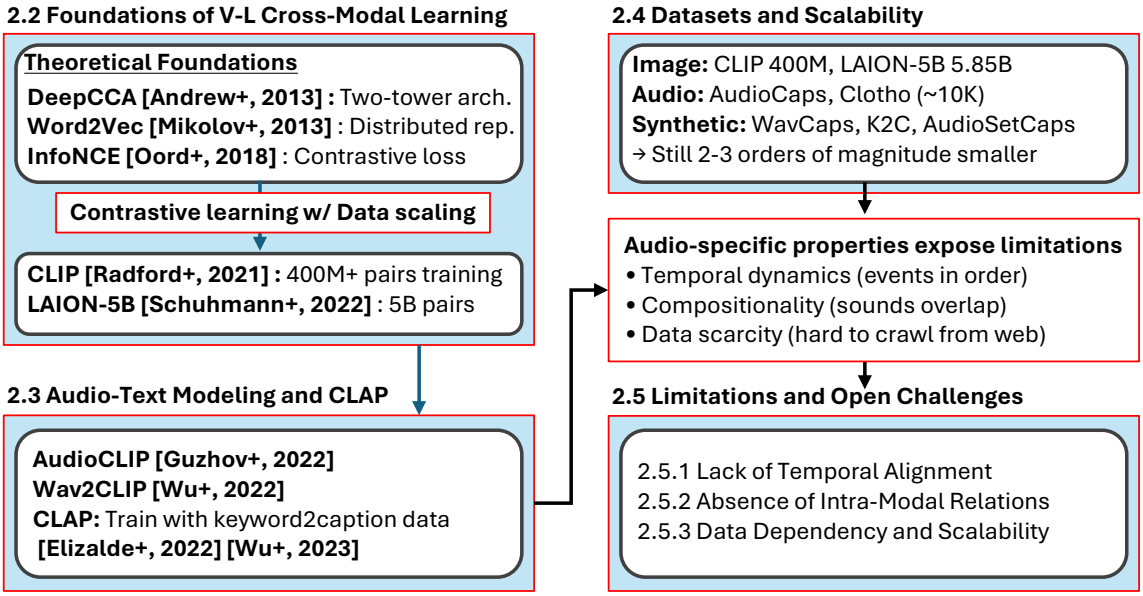

Figure 2.1: Overview of Chapter 2. Sections 2.2 and 2.3 trace the methodological lineage from vision-language foundations to CLAP. Section 2.4 summarizes the data landscape. Audio-specific properties expose the limitations discussed in Section 2.5.

2.2 Foundations of Vision-Language Cross-Modal Learning

Cross-modal representation learning, originally developed in the vision-language domain, serves as the theoretical foundation for audio-text modeling. This section outlines the progression of representative methodologies to clarify the architectural lineage inherited by current audio-text frameworks.

In the early 2010s, Deep CCA [12], an extension of Canonical Correlation Analysis (CCA) to deep learning, was proposed to align statistical structures between different modalities. Classical CCA [37] seeks projections that maximize the correlation between two sets of variables (e.g., image and text features) via linear transformations. However,

its expressiveness is limited when dealing with visual and linguistic data characterized by high-dimensional and non-linear structures. Deep CCA overcame this limitation by replacing linear transformations with deep neural networks, demonstrating a framework for projecting each modality into a shared embedding space with high correlation. Benton et al. further extended this to a multi-view setting with Deep Generalized CCA (DGCCA) [38], evolving it into a general representation learning framework capable of handling multiple modalities such as image, text, and audio simultaneously. These approaches can be regarded as the foundational precursors to the Two-Tower architecture, which employs modality-specific encoders to construct a shared embedding space, a design adopted later by CLIP and CLAP.

During the same period, vision-language representation learning based on mapping to semantic spaces also advanced. DeViSE (Deep Visual-Semantic Embedding) [11] addressed the limitation that traditional 1-of-K classification ignores semantic relationships between classes, introducing a method to map image features directly into a pretrained word embedding space. Distributed representations such as Word2Vec [39], based on the Skip-gram model, successfully embedded semantic relationships between words into a continuous space. By mapping the output of image classification models to these word vectors instead of one-hot labels, DeViSE enabled zero-shot classification for classes not observed during training. Developing this direction further, VSE++ (Visual-Semantic Embeddings with Hard Negatives) [40] introduced a margin-based loss focused on "hard negatives", samples the model finds most confusing, rather than randomly sampled negatives. This improved the ability to distinguish subtle semantic differences between images and text. The concept of hard negative mining anticipates the design of treating "all samples in a batch as negatives" in later large-scale contrastive learning, and shares the same fundamental principle as the contrastive loss

used in CLIP and CLAP.

Meanwhile, approaches to learning visual representations through generative tasks were also proposed. VirTex [41] and ICMLM [42] are methods that learn visual representations through tasks such as generating captions conditioned on images or solving Image-Conditioned Masked Language Modeling (MLM). These studies demonstrated the effectiveness of the Transformer architecture [28] and the rich semantic supervision provided by natural language descriptions. However, they also highlighted constraints such as high computational costs associated with language generation decoders and limited inference speeds. These bottlenecks suggest that for Web-scale multimodal data, a framework based on contrastive learning offers superior scalability compared to generative tasks.

Building on these research trends, CLIP (Contrastive Language-Image Pretraining) [13] was proposed. CLIP achieved scalable and generalizable representation learning through contrastive training on massive image-text pairs collected from the internet. Specifically, representations obtained from a Transformer-based text encoder and an image encoder based on ResNet [43] or Vision Transformer (ViT) [24] are projected into a common embedding space. The model is trained using an InfoNCE-type contrastive loss [27] to maximize the similarity of corresponding pairs while minimizing the similarity of other combinations. The success of CLIP is evidenced by its high zero-shot performance in downstream tasks such as ImageNet [44], establishing it as the de facto standard for multimodal representation learning. This scalability has been further validated by subsequent efforts like ALIGN [45] and the release of open datasets such as LAION-5B [22]. Notably, ALIGN demonstrated that scaling to 1.8 billion noisy image-text pairs yields strong generalization even without curated annotations. This finding underscores a critical principle: data quantity can compensate for label noise, a

strategy that audio-text modeling has yet to fully exploit due to the inherent scarcity of paired data.

Recent work has extended the two-modality paradigm to unified multi-modal spaces. ImageBind [46] binds six modalities (image, text, audio, depth, thermal, IMU) into a shared embedding space using images as the central anchor. LanguageBind [47] takes an alternative approach by using language as the binding anchor, freezing a pretrained LLM encoder and aligning other modalities to it. While these methods demonstrate impressive zero-shot transfer, they inherit limitations: ImageBind's image-centric alignment may underrepresent non-visual acoustic properties, whereas LanguageBind's frozen language encoder constrains the expressiveness of audio-specific semantics.

2.3 Audio-Text Modeling and CLAP Architecture

Following the success of CLIP, extensions to audio-text modeling have been actively explored. However, unlike the vision domain, the historical scarcity of large-scale audio-text paired data posed a fundamental constraint to applying the CLIP framework directly.

2.3.1 Indirect Approaches before CLAP

Prior to the establishment of direct audio-text contrastive learning, research focused on leveraging the rich supervision available in pretrained vision-language models. AudioCLIP [15] extended the pretrained CLIP framework into a tri-modal architecture handling image, text, and audio. By incorporating ESResNet [48] as an audio encoder and initializing the image and text heads with CLIP weights, it enabled cross-

modal retrieval primarily using tag-based supervision from AudioSet [9]. Similarly, Wav2CLIP [14] proposed a knowledge distillation approach, treating CLIP's multimodal embedding space as a high-quality semantic teacher. It trained an audio encoder to map audio signals into the shared space defined by CLIP's image embeddings derived from video frames.

These methods effectively circumvented data scarcity by indirectly borrowing knowledge from the vision-language domain. However, they inherently depend on visual correspondence or rigid tag supervision. Consequently, they are limited in their capacity to represent pure acoustic events detached from visual contexts or to capture the nuances of fine-grained free-form natural language.

2.3.2 The CLAP Framework

In contrast to visually-guided methods, Contrastive Language-Audio Pretraining (CLAP) aims for direct alignment between audio signals and natural language captions. Representative implementations, such as Microsoft's CLAP [1] and LAION's CLAP [2], adopt a dual-encoder (two-tower) architecture trained on large-scale audio-text pairs. This framework is characterized by two primary design components: representation design and objective design.

Representation Design

Formally, let an audio input be represented as a continuous, variable-length sequence $\mathbf{X}^{(\mathsf{A})} = (\mathbf{x}_1^{(\mathsf{A})}, \ldots, \mathbf{x}_{L_\mathsf{A}}^{(\mathsf{A})})$, and a text input as a discrete token sequence $\mathbf{X}^{(\mathsf{T})} = (\mathbf{x}_1^{(\mathsf{T})}, \ldots, \mathbf{x}_{L_\mathsf{T}}^{(\mathsf{T})})$. These sequences are processed by modality-specific encoders to pro-

duce embedding sequences:

$$\mathbf{Z}^{(\mathsf{A})} = (\mathbf{z}_1^{(\mathsf{A})}, \ldots, \mathbf{z}_{L_\mathsf{A}}^{(\mathsf{A})}) = \text{AudioEncoder}(\mathbf{X}^{(\mathsf{A})}), \tag{2.1}$$

$$\mathbf{Z}^{(\mathsf{T})} = (\mathbf{z}_1^{(\mathsf{T})}, \ldots, \mathbf{z}_{L_\mathsf{T}}^{(\mathsf{T})}) = \text{TextEncoder}(\mathbf{X}^{(\mathsf{T})}). \tag{2.2}$$

For the architecture selection, AudioEncoder is typically a CNN-based model such as PANNs [49] or a Transformer-based architecture such as AST [50] and HTS-AT [51], while TextEncoder employs models like BERT [29] or RoBERTa [52].

To enable contrastive learning, these variable-length embedding sequences must be mapped to a fixed-size common space. This is achieved through sequence-level aggregation, typically via mean pooling, attention pooling, or a designated `[CLS]` token, resulting in single global vectors $\bar{\mathbf{z}}^{(\mathsf{A})}$ and $\bar{\mathbf{z}}^{(\mathsf{T})}$. While computationally efficient, this aggregation sacrifices the fine-grained alignment between individual audio frames and text tokens.

Objective Design

The model is optimized to maximize the similarity of corresponding pairs while minimizing that of non-corresponding pairs. A widely used objective for this purpose is the symmetric InfoNCE loss [27]. Given a batch of N pairs, the loss is defined for both audio-to-text ($\mathsf{A}\to\mathsf{T}$) and text-to-audio ($\mathsf{T}\to\mathsf{A}$) directions as:

$$\mathcal{L}_{\text{InfoNCE}}^{\mathsf{A}\to\mathsf{T}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\bar{\mathbf{z}}_i^{(\mathsf{A})}, \bar{\mathbf{z}}_i^{(\mathsf{T})})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\bar{\mathbf{z}}_i^{(\mathsf{A})}, \bar{\mathbf{z}}_j^{(\mathsf{T})})/\tau)}, \tag{2.3}$$

$$\mathcal{L}_{\text{InfoNCE}}^{\mathsf{T}\to\mathsf{A}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\bar{\mathbf{z}}_i^{(\mathsf{T})}, \bar{\mathbf{z}}_i^{(\mathsf{A})})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\bar{\mathbf{z}}_i^{(\mathsf{T})}, \bar{\mathbf{z}}_j^{(\mathsf{A})})/\tau)}, \tag{2.4}$$

and the total loss is $\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2}(\mathcal{L}_{\text{InfoNCE}}^{\mathsf{A}\to\mathsf{T}} + \mathcal{L}_{\text{InfoNCE}}^{\mathsf{T}\to\mathsf{A}})$. Here, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a learnable temperature parameter that controls the sharpness of the similarity distribution.

Impact and Applications. With this formulation, CLAP has achieved zero-shot classification performance comparable to supervised learning in tasks such as ESC-50 [10] and UrbanSound8K [53]. Furthermore, it has established itself as a standard foundation model in the audio-text domain, serving as a conditioning representation for text-to-audio retrieval [2], audio captioning, and generative models like AudioLDM [8].

Theoretical Limitations of CLAP. While sequence-level aggregation facilitates efficient cross-modal comparison, it implicitly assumes that the semantic content is uniform across the sequence. It does not explicitly preserve alignment between temporal positions in audio and corresponding semantic content in text. Consequently, this design reduces the alignment granularity and leads to the loss of fine-grained frame-token alignment [34], a representational limitation we address in Chapter 3.

Regarding the objective, minimizing the InfoNCE loss implies that the gradient depends solely on cross-modal similarity scores (i.e., $\text{sim}(\mathbf{z}^{(\mathsf{A})}, \mathbf{z}^{(\mathsf{T})})$). As discussed in Section 1.3.2, this formulation encourages alignment between modalities but does not explicitly constrain the relationships within each modality (e.g., audio-audio and text-text similarity). From an information-theoretic perspective, minimizing this loss maximizes the lower bound of mutual information between audio and text representations [32]. While this effectively captures shared semantics, it treats modality-specific structures, such as the fine-grained acoustic relationships between different sound events, as irrelevant noise to be discarded. This deficiency in the standard objective serves as the theoretical basis for the Unstructured Intra-Modal Relations problem, which we address through the difference modeling framework in Chapter 4.

2.4 Datasets and Scalability in Audio-Text Learning

The performance and generalization capability of audio-text models depend heavily on the scale and quality of available paired datasets. However, in contrast to the image-text domain where Web-scale data is readily available, the audio domain suffers from a fundamental problem in data availability.

2.4.1 The Data Gap: Image vs. Audio

In computer vision, the success of foundation models is underpinned by massive datasets collected via Web crawling, such as CLIP's 400M pairs [13] and LAION-5B's [22] 5.85 billion pairs. Conversely, in the audio domain, high-quality human-captioned datasets remain scarce. Standard benchmarks like AudioCaps [54] and Clotho [5] contain only thousands to tens of thousands of clips. Even AudioSet [9], the largest general acoustic dataset, provides only tag information (e.g., "Dog", "Car") for approximately 2 million clips, lacking the rich natural language descriptions required to learn complex semantic relationships.

2.4.2 Synthetic Expansion and Its Limits

To bridge this gap, recent approaches have turned to data engineering and synthetic caption generation. Keyword-to-Caption (K2C) [2] generates pseudo-captions by feeding tag sequences into Large Language Models (LLMs), contributing to the construction of LAION-Audio-630K. WavCaps [23] compiled a weakly supervised dataset of approximately 400k clips by filtering and normalizing raw text descriptions from

the Web (e.g., filenames, metadata) using LLMs. Furthermore, projects like Auto-ACD [55] and AudioSetCaps [56] have attempted to construct caption datasets on the scale of millions by leveraging visual information from videos to infer audio contents.

Despite these efforts, the effective data scale remains 2 to 3 orders of magnitude smaller than that of the image domain. Moreover, the heavy reliance on synthetic captions introduces a dual constraint. First, synthetic data inevitably contains noise, hallucinations, and irrelevant tags associated with automated generation. Second, and more critically, there exists an intrinsic information asymmetry: textual descriptions generated from tags or vision often fail to capture non-linguistic acoustic properties, such as timbre, texture, and spatial characteristics. Consequently, current audio-text contrastive learning is forced to operate under conditions of both "limited scale" and "supervision with limited descriptive fidelity."

2.5 Limitations and Open Challenges

CLAP has enabled large-scale contrastive learning in the audio-text domain, successfully acquiring semantic representations for the audio modality [1], [2]. However, the framework typically relies on aggregating each modality into a single global vector. This architectural choice leaves three representational challenges unresolved, stemming from the unique temporal properties of audio and the current constraints of the data environment.

2.5.1 Lack of Fine-Grained Temporal Alignment

Most contrastive models, including CLAP, aggregate audio and text sequences into single global embedding vectors to optimize their similarity, i.e., global alignment. In

this process, high-temporal-granularity information, such as the temporal structure of acoustic events or the distinction between concurrent sounds, is inevitably averaged out.

Recent studies have attempted to improve fine-grained alignment. In the audio domain, T-CLAP [57] improves temporal alignment through data augmentation and loss calculation emphasizing temporal order, while MGA-CLAP [58] recovers local word-frame alignment during pooling. Parallel advancements in the video domain offer relevant insights; VideoCLIP [16] learns soft temporal alignment via loosely overlapping pairs, and TempCLR [17] achieves explicit sequence alignment via Dynamic Time Warping (DTW). More recently, CoLLAP [25] extended contrastive pretraining to long-form audio by augmenting musical temporal structures.

Despite these developments, most approaches remain task-specific modifications or incur high computational costs. A general framework that integrates "global semantics" and "local temporal structure" within a scalable representation learning pipeline has not yet been established. This gap motivates our work on Aligned Contrastive Learning in Chapter 3.

2.5.2 Absence of Intra-Modal Relations

While CLAP's contrastive loss effectively aligns audio with text at the global level, it does not establish alignment between audio components themselves, failing to preserve intra-modal relations. As discussed in Section 1.3.2, intra-modal relations encompass not just judging whether two sounds are similar or different, but articulating what specifically differs between them. Since the InfoNCE loss only discriminates between positive and negative pairs, the model cannot identify which components differ between two compositional scenes; it can only judge overall similarity.

In contrast, self-supervised learning (SSL) has excelled at modeling this intrinsic acoustic structure without textual supervision. Prominent models such as wav2vec 2.0 [59] and HuBERT [60] have demonstrated that predicting latent representations from masked inputs captures rich acoustic dependencies and clustering structures. Similarly, generative approaches like Audio-MAE [61] force the model to encode fine-grained spectro-temporal details to reconstruct masked spectrogram patches. However, a critical limitation of these SSL frameworks is that they operate purely within the audio domain; they inherently lack the mechanism to map these rich structures to semantic concepts defined in natural language. Although hybrid approaches like CAV-MAE [61] attempt to combine contrastive and masked autoencoding objectives, a fundamental trade-off persists: current models tend to be either "semantically aligned but structurally coarse" (CLAP) or "structurally rich but semantically unaligned" (SSL). Developing a method that integrates the benefits of both, semantic alignment and coherent intra-modal relations, is the focus of Chapter 4.

2.5.3 Data Dependency and Scalability

As discussed in Section 2.4, audio-text learning suffers from severe data scarcity compared to the image domain. Although datasets like WavCaps provide synthetic expansions, they rely on noisy, LLM-generated captions that cannot fully capture non-linguistic acoustic nuances. Given that contrastive learning fundamentally depends on data quantity and diversity [62], strategies relying solely on inflating paired data face diminishing returns.

Conversely, unlabeled raw audio exists in virtually inexhaustible quantities on the Web. As demonstrated by the audio-only SSL research mentioned above [59], [60], these unlabeled corpora contain rich acoustic information. Therefore, a critical challenge for

next-generation models is to break away from the reliance on expensive paired data. We need a new paradigm that effectively integrates the rich information held by unlabeled audio corpora directly into the cross-modal learning framework. This challenge is addressed by our semi-supervised framework in Chapter 5.

2.6 Summary

In this chapter, we reviewed the evolution of cross-modal learning and identified the representational limitations of current audio-text models. Specifically, we highlighted three critical gaps: (1) the loss of temporal granularity resulting from sequence-level aggregation, (2) the absence of coherent intra-modal relations due to the purely cross-modal contrastive objective, and (3) the fundamental performance bottleneck caused by data scarcity.

These unresolved challenges serve as the direct motivation for the research presented in the remainder of this thesis. **Chapter 3** addresses the first challenge by proposing a fine-grained alignment mechanism that transcends global pooling. **Chapter 4** tackles the second challenge by introducing a difference-based learning framework to explicitly capture intra-modal relations. Finally, **Chapter 5** confronts the data scarcity problem by proposing a semi-supervised framework that directly leverages unlabeled audio to refine the backbone representation.

Chapter 3

Fine-Grained Temporal Alignment for Audio-Text Retrieval

3.1 Introduction

As discussed in the previous chapters, audio-text representation learning has made significant strides, particularly with the advent of contrastive learning frameworks such as CLAP [1], [2]. These models typically project audio and text into a shared latent space to capture global semantic correspondences. However, as analyzed in Section 1.3.1, a fundamental limitation remains in the current modeling paradigm: the loss of granularity caused by sequence-level aggregation.

Conventional approaches compress variable-length audio sequences and text descriptions into single global vectors (e.g., via global average pooling or `[CLS]` token pooling) to compute similarity. While effective for capturing high-level concepts, this coarse-grained modeling inevitably discards the rich temporal dynamics and local substructures essential for detailed understanding. In particular, sequence-level aggregation entangles multiple sounds into a single "bag-of-events" representation, collapsing contributions from different time segments into one abstract embedding. For instance,

complex acoustic scenes often involve sequential events (e.g., "a violin melody followed by applause") or distinct foreground-background relationships. When such temporal details are aggregated into a single point in the embedding space, the mapping from frame-level patterns to the global representation becomes underdetermined. As a result, the model loses the ability to distinguish temporally distinct contents, limiting its precision in retrieval tasks that require fine-grained semantic understanding.

To address this limitation, this chapter introduces a framework termed Aligned Contrastive Learning. Moving beyond the global alignment paradigm discussed in Chapter 2, we propose a fine-grained modeling approach that explicitly achieves alignment between frame-level audio representations and token-level text embeddings. Instead of collapsing the time dimension into a single global vector, our method computes similarity matrices between local features of both modalities, allowing the model to identify which specific audio frames correspond to which textual words. By preserving local granularity during the training process, the model acquires a more expressive and precise representation that better respects the intrinsic temporal nature of audio signals.

The contribution of this chapter is to formulate this fine-grained alignment mechanism and demonstrate its superiority over standard global pooling approaches. In the following sections, we first formalize the limitations of conventional contrastive objectives in the context of temporal aggregation. We then detail the proposed Aligned Contrastive Learning framework, focusing on its specialized similarity measure that integrates frame-token interactions. Finally, we present experimental evaluations. To empirically validate this approach, we conduct experiments on a retrieval benchmark featuring an audio dataset with highly dynamic temporal structures. By targeting signals whose acoustic characteristics evolve significantly over time, we demonstrate

that preserving fine-grained details leads to substantial improvements in retrieval performance.

3.2 Related Work

3.2.1 Audio-Text Retrieval

Language-based audio retrieval has evolved from early linear ranking models [63] to modern contrastive joint embedding frameworks [64]. PAMIR [65] is widely regarded as a seminal work in this area: it formulated general audio retrieval from free-text queries as a learning-to-rank problem, learning a compatibility function between Bag-of-Words text representations and audio features, rather than relying on fixed tag vocabularies or purely metadata-based search. In parallel, semantic audio retrieval approaches such as Turnbull et al. [66] explored large-scale automatic tagging of music and sound effects, using supervised models to predict semantic tags and then retrieving items based on their predicted tag distributions. While these methods significantly advanced content-based audio retrieval, they still depended on a predefined tag vocabulary and could not fully support open-vocabulary natural language queries.

A key turning point came with the introduction of semantic embeddings and deep joint embedding architectures. Xie and Virtanen [7] leveraged distributed word representations to align audio features with a semantic embedding space, enabling zero-shot audio classification and retrieval for labels unseen during training. Elizalde et al. [67] further established a Siamese audio-text joint embedding architecture, in which audio and text are mapped into a shared latent space and ranked by cross-modal similarity, which is an architectural precursor to CLAP-style models. Subsequently, captioned datasets such as AudioCaps [54] and Clotho [5], provided standardized testbeds for

language-based audio retrieval. Building on these resources, CLAP models [1], [2] apply contrastive training to large collections of audio-text pairs, and have become practical foundation models for text-to-audio and audio-to-text retrieval. In this chapter, we take such CLAP-style contrastive retrieval as our starting point and focus on improving the granularity of alignment by moving from global embeddings to explicit frame-token interactions.

3.2.2 Fine-Grained Alignment in Vision-Language

The limitation of global pooling, often referred to as the "modality gap" or a granularity mismatch, has been extensively studied in vision-language research. Standard approaches like CLIP [13] rely on global alignment, which compresses spatial details into a single vector. To overcome this, fine-grained alignment methods leverage token-level interactions. FILIP [34] pioneered a "late interaction" mechanism that computes token-wise maximum similarity between image patches and text tokens, demonstrating that fine-grained correspondences can significantly improve zero-shot classification without additional object detectors. SCAV [26] operates on non-aggregated representation spaces, preserving full sequence information at the cost of increased memory. Similarly, ColBERT [68] in information retrieval demonstrated the efficacy of retaining token-level embeddings for precise matching. Recent work such as DenseAV [69] and CAV-MAE [70] extended this concept to audio-visual learning through self-supervised objectives, achieving fine-grained grounding of sound sources. Our work draws inspiration from these interaction-based mechanisms but adapts them to the audio-text domain, addressing the specific challenge of aligning variable-length audio frames with discrete textual tokens.

3.2.3 Token-Level Audio-Text Alignment

In the audio domain, shifting from global to local alignment is an emerging line of research [58], [71], [72]. Early attempts often relied on temporal pooling or attention mechanisms that still resulted in bottlenecked representations. More recent models have begun to explore explicit fine-grained objectives. GPA [72] introduces prototype alignment, which clusters audio frames into semantic prototypes before matching them with text. While effective, this approach introduces additional complexity through learnable prototypes, clustering, and prototype matching operations. MGA-CLAP [58] proposes a multi-granularity framework that simultaneously optimizes global, local, and instance-level objectives. However, balancing these multiple loss terms can be non-trivial. In contrast to these approaches, our Aligned Contrastive Learning framework focuses on a direct interaction mechanism. We employ a max-mean aggregation strategy that directly computes similarity between audio frames and text tokens. This design eliminates the need for auxiliary prototypes or complex multi-branch architectures, allowing the model to achieve fine-grained temporal alignment through direct interaction between feature sequences. Crucially, our approach introduces no additional learnable parameters and incurs no extra computational cost at inference time, as the aligned similarity is employed only during training.

3.3 Audio-Text Contrastive Learning Framework

We first formalize the standard audio-text contrastive learning framework. As illustrated in Figure 3.1, the baseline model consists of two separate encoders that map audio and text inputs into a shared latent space, optimized via the InfoNCE loss [27].

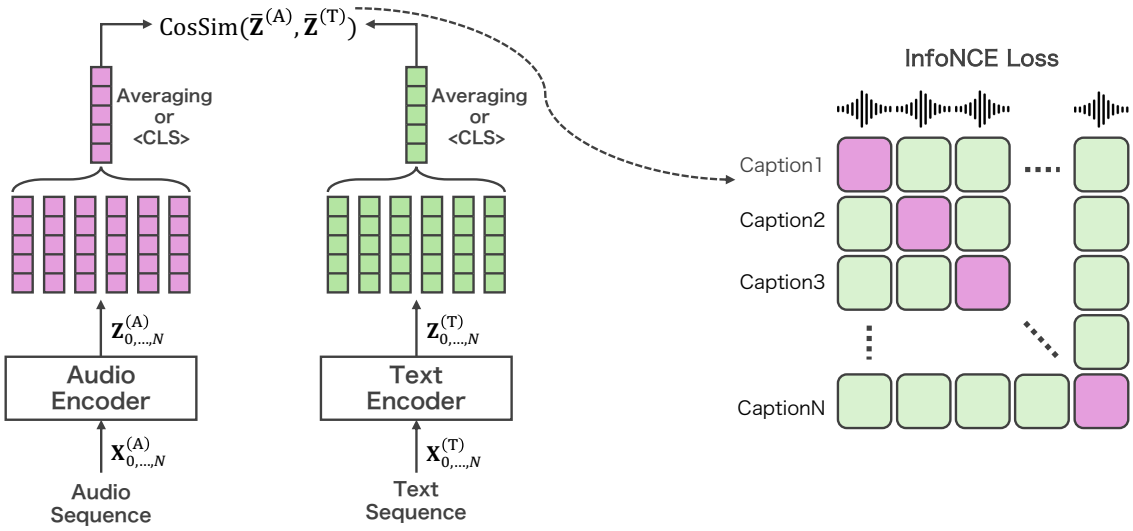

Figure 3.1: General framework of conventional audio-text contrastive learning with the InfoNCE loss.

3.3.1 Audio and Text Encoders

Let us denote an audio sequence by $\mathbf{X}^{(\mathsf{A})} \in \mathbb{R}^{L_\mathsf{A} \times D_\mathsf{A}}$, where L_A represents the number of temporal frames and D_A the audio feature dimension. Similarly, a text sequence is written as $\mathbf{X}^{(\mathsf{T})} \in \mathbb{R}^{L_\mathsf{T} \times D_\mathsf{T}}$, with L_T being the sequence length and D_T the text embedding dimension. The sequences are then passed through their modality-specific encoders to produce latent embedding sequences $\mathbf{Z}^{(\mathsf{A})}$ and $\mathbf{Z}^{(\mathsf{T})}$ as

$$\mathbf{Z}^{(\mathsf{A})} = \text{AudioEncoder}(\mathbf{X}^{(\mathsf{A})}), \tag{3.1}$$

$$\mathbf{Z}^{(\mathsf{T})} = \text{TextEncoder}(\mathbf{X}^{(\mathsf{T})}). \tag{3.2}$$

Regarding the encoder architecture, Transformer-based models have become the standard choice in recent audio-text representation learning, with audio encoders built on convolutional or Transformer backbones and text encoders initialized from pretrained language models such as BERT [1], [2], [29].

3.3.2 Contrastive Learning Objective

The model parameters are optimized using a cross-modal variant of the InfoNCE loss. We employ a similarity function $\mathrm{sim}(\cdot,\cdot)$ (typically cosine similarity) to compare global audio and text embeddings. Global embeddings are obtained by aggregating latent sequences into single vectors $\bar{\mathbf{z}}^{(\mathsf{A})}$ and $\bar{\mathbf{z}}^{(\mathsf{T})}$ (e.g., via a `[CLS]` token or mean pooling).

For the audio-to-text direction, the loss is defined as

$$\mathcal{L}_{\mathrm{InfoNCE}}^{\mathsf{A}\to\mathsf{T}} = -\log \frac{\exp\big(\mathrm{sim}(\bar{\mathbf{z}}^{(\mathsf{A})}, \bar{\mathbf{z}}^{(\mathsf{T})})/\tau\big)}{\sum_j \exp\big(\mathrm{sim}(\bar{\mathbf{z}}^{(\mathsf{A})}, \bar{\mathbf{z}}_j^{(\mathsf{T})})/\tau\big)}, \tag{3.3}$$

where $\tau > 0$ denotes the temperature parameter and j indexes other texts in the minibatch.

Similarly, the text-to-audio contrastive loss is

$$\mathcal{L}_{\mathrm{InfoNCE}}^{\mathsf{T}\to\mathsf{A}} = -\log \frac{\exp\big(\mathrm{sim}(\bar{\mathbf{z}}^{(\mathsf{T})}, \bar{\mathbf{z}}^{(\mathsf{A})})/\tau\big)}{\sum_j \exp\big(\mathrm{sim}(\bar{\mathbf{z}}^{(\mathsf{T})}, \bar{\mathbf{z}}_j^{(\mathsf{A})})/\tau\big)}. \tag{3.4}$$

The total InfoNCE loss is the sum of both retrieval directions:

$$\mathcal{L}_{\mathrm{InfoNCE}} = \mathcal{L}_{\mathrm{InfoNCE}}^{\mathsf{A}\to\mathsf{T}} + \mathcal{L}_{\mathrm{InfoNCE}}^{\mathsf{T}\to\mathsf{A}}. \tag{3.5}$$

Standard audio-text models compute sim between global embeddings derived from `[CLS]` tokens or pooled representations of audio and text Transformers [1], [2]. While effective for clip-level retrieval, this aggregation compresses the rich temporal and semantic structure of the sequences into single vectors and entangles multiple sound events and textual components into a single "bag-of-events" representation, as discussed in Section 1.3.1. As a result, frame-level and token-level alignments are not explicitly modeled. Our proposed method alleviates this limitation by explicitly modeling alignments at both temporal (audio) and token (text) resolutions.

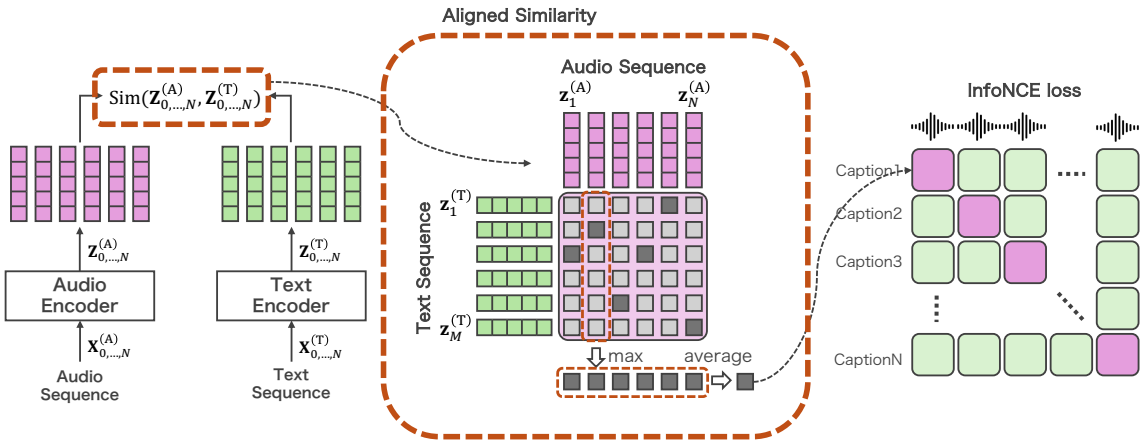

Figure 3.2: Proposed Aligned Contrastive Learning. For the similarity calculation between audio and text sequences, the proposed method first identifies, for each audio frame, the maximum cosine similarity over all text tokens, and then averages these maximum values over all audio frames.

3.4 Aligned Contrastive Learning

We propose a novel approach for audio-text representation learning, termed Aligned Contrastive Learning (Figure 3.2). This method introduces a fine-grained similarity measure designed to preserve the temporal granularity of audio and the semantic structure of text, addressing the limitations of sequence-level aggregation discussed in the previous sections.

3.4.1 Aligned Similarity Measure

The core of our approach is an aligned similarity score $\text{sim}_{\text{Align}}$ that aggregates frame-token interactions instead of globally pooled embeddings. Rather than collapsing each modality into a single vector, we explicitly operate on the sequence of audio frame embeddings and the sequence of text token embeddings.

Recall that the audio and text encoders produce embedding sequences $\mathbf{Z}^{(\mathsf{A})} = (\mathbf{z}_1^{\mathsf{A}}, \ldots, \mathbf{z}_{L_{\mathsf{A}}}^{\mathsf{A}})$ and $\mathbf{Z}^{(\mathsf{T})} = (\mathbf{z}_1^{\mathsf{T}}, \ldots, \mathbf{z}_{L_{\mathsf{T}}}^{\mathsf{T}})$. We first form a frame-token similarity matrix using a base similarity function $\mathrm{sim}(\cdot, \cdot)$ (typically cosine similarity),

$$s_{t\ell} = \mathrm{sim}\big(\mathbf{z}_t^{\mathsf{A}}, \mathbf{z}_\ell^{\mathsf{T}}\big), \quad 1 \le t \le L_{\mathsf{A}},\ 1 \le \ell \le L_{\mathsf{T}}.$$

The aligned similarity from audio to text is then defined as

$$\mathrm{sim}_{\mathsf{Align}}(\mathbf{Z}^{(\mathsf{A})}, \mathbf{Z}^{(\mathsf{T})}) = \frac{1}{L_{\mathsf{A}}} \sum_{t=1}^{L_{\mathsf{A}}} \max_{1 \le \ell \le L_{\mathsf{T}}} s_{t\ell}. \tag{3.6}$$

Interpretation. This formulation follows a "max-mean" scheme. For each temporal frame t in the audio sequence, the inner max operation selects the text token ℓ that yields the highest similarity $s_{t\ell}$. This realizes a latent alignment, assigning each audio frame to its most semantically relevant word (for example, aligning a frame containing a barking sound to the token "dog"). The outer average then aggregates these per-frame maxima over all audio frames. In this way, every temporal segment contributes to the final score through its best textual match, preserving the density of temporal information and reducing the entanglement of multiple events into a single global vector.

The aligned similarity $\mathrm{sim}_{\mathsf{Align}}$ is inherently audio-conditioned: it asks, for each audio frame, which word in the caption it is best aligned with. Intuitively, frames containing a dog bark will obtain high similarity to content words such as "dog" or "barking", whereas silent or purely background frames will not have a strong match to any token, contributing little to the average. Thus the definition "looks from frames to tokens" and naturally reflects how well the caption covers different parts of the audio over time. One could, in principle, define a text-conditioned variant that, for each token, selects the most similar frame; however, such a formulation tends to pick only a single representative frame for content words, discarding information about the duration and

frequency of events, and it also forces function words (e.g., "a", "in", "the") to be matched to some frame, introducing unnecessary noise. For these reasons, we adopt only the audio-conditioned similarity and use it directly as the scalar similarity for each audio-text pair within the InfoNCE loss, in place of the conventional global cosine similarity.

3.4.2 Optimization Objective

To incorporate fine-grained information into the learning process, we formulate an aligned contrastive loss, denoted as $\mathcal{L}_{\mathsf{Align}}$. This loss is obtained by substituting the standard global cosine similarity in the InfoNCE objective (Eq. (3.5)) with our proposed aligned similarity measure, $\text{sim}_{\mathsf{Align}}$ (Eq. (3.6)). Minimizing $\mathcal{L}_{\mathsf{Align}}$ explicitly encourages the encoders to maximize the alignment between local audio frames and their most relevant text tokens.

For the final training objective, we adopt a hybrid strategy that combines the conventional global loss $\mathcal{L}_{\text{global}}$ (based on `[CLS]` tokens or pooled representations) with the aligned loss $\mathcal{L}_{\mathsf{Align}}$. The total loss $\mathcal{L}_{\text{total}}$ is computed as the sum of these two objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\mathsf{Align}}. \tag{3.7}$$

This dual-objective framework allows the model to retain global semantic coherence via the traditional contrastive loss, while simultaneously refining local feature representations through fine-grained frame-token alignment.

3.4.3 Inference Efficiency

A key advantage of our framework is its efficiency during the inference phase. Calculating the fine-grained similarity $\text{sim}_{\mathsf{Align}}$ requires computing a frame-by-token inter-

action matrix, which is computationally expensive for large-scale retrieval. Therefore, we utilize the aligned similarity only during training. For retrieval inference, we rely solely on the cosine similarity between the global `[CLS]` embeddings of the audio and text, identical to the baseline CLAP-style model.

Crucially, although the inference mechanism remains unchanged, the global representations are enriched by the aligned training objective. The auxiliary alignment loss acts as a regularizer, encouraging the encoders to aggregate salient local features into the global `[CLS]` token more effectively to satisfy the fine-grained objective. Consequently, our method achieves the precision benefits of fine-grained modeling without incurring the computational overhead of late interaction during inference.

3.5 Experimental Settings

3.5.1 Evaluation Task: Text-to-Music Retrieval

Although the proposed Aligned Contrastive Learning framework is conceptually domain-agnostic and can be applied to various types of audio, in this chapter we focus on text-to-music retrieval as the primary evaluation task. Music signals exhibit highly complex temporal dynamics and compositional semantic attributes (e.g., genre, mood, instrumentation) that evolve over tens of seconds. Compared to short, event-based audio clips, music tracks feature intricate structural transitions and layered information, making them a particularly demanding testbed for evaluating fine-grained cross-modal alignment. We therefore regard text-to-music retrieval as a stress test of the proposed alignment mechanism, while leaving a systematic evaluation on other audio domains (e.g., environmental sounds) for future work.

Historically, cross-modal music retrieval has evolved from tag-based approaches [73],

[74] to caption-based methods [75], [76], [77]. Early deep learning work on music auto-tagging [78] established CNN-based architectures for predicting semantic tags from audio, and subsequent studies comparing metric learning and classification [79] highlighted the advantages of embedding-based representations for retrieval. Multimodal approaches that combine lyrics and audio features [80], [81] further demonstrated the value of aligning heterogeneous descriptors for music exploration. Building on these foundations, recent contrastive models such as MuLan [82] achieve strong performance by training on large-scale music-text pairs, and CLaMP [83] pioneers contrastive language-music pretraining using symbolic (MIDI/ABC) data. However, these state-of-the-art methods predominantly rely on global embeddings (e.g., average pooling or `[CLS]` tokens). Their reliance on sequence-level aggregation limits their ability to capture the temporal evolution characteristic of musical content, making them suitable baselines to demonstrate the benefit of our fine-grained alignment.

3.5.2 Dataset and Evaluation Metrics

We evaluate our model on the text-to-music retrieval task using the ECALS dataset [77]. ECALS is derived from the Million Song Dataset (MSD) [84], augmented with 500 Last.fm tags [74] and 1,402 AllMusic tag annotations [85]. The dataset consists of approximately 520,000 audio clips, each 30 seconds in duration. These clips are paired with textual descriptions, totaling approximately 140,000 unique captions generated from a vocabulary of 1,054 distinct tags.

We adopt two distinct evaluation protocols: tag-level and sentence-level retrieval.

- **Tag-level retrieval:** We measure performance using the macro-average of ROC-AUC and PR-AUC.

- **Sentence-level retrieval:** We assess performance using Recall at rank K (R@1, R@5, R@10), mean Average Precision at 10 (mAP@10), and median rank (MedR), following standard methodologies [77], [86].

These metrics quantify the model's ability to rank the ground-truth audio clip highly within the candidate pool given a text query.

3.5.3 Implementation Details

Audio encoder. The audio encoder processes 9.91-second audio clips sampled at 16 kHz, which are converted into 128-bin mel spectrograms. We utilize the Music Tagging Transformer [87] as the backbone. It first extracts local features via four convolutional layers, followed by four Transformer layers. The output sequence, including the `[CLS]` token, is finally projected into a 128-dimensional embedding space.

Text encoder. For the text encoder, we employ the pretrained `bert-base-uncased` model [29]. The token representations produced by the 12-layer Transformer are linearly projected into the same 128-dimensional space to align with the audio embeddings.

Training setup. All models are trained using the Adam optimizer with a learning rate of 5.0×10^{-5} and a batch size of 64. The temperature parameter τ is set to 0.2, and we train for 50 epochs. For comparison, we evaluate our method against baselines including Triplet-loss models [73] and standard InfoNCE models [77]. Unless otherwise noted, the "Baseline (Contrastive learning)" in our results refers to a conventional InfoNCE model trained from scratch under an identical experimental setup, ensuring a fair comparison.

Table 3.1: Comparison of tag-level and sentence-level retrieval results across various models. Our proposed Aligned Contrastive model demonstrates superior performance in most sentence-level retrieval metrics.

Model type	Used in	Tag-level retrieval		Sentence-level retrieval				
		50 tags	1054 tags	1000 captions				
		ROC/PR	ROC/PR	R@1	R@5	R@10	mAP@10	MedR↓
Reported in prior works								
Triplet with GloVe	[73], [74]	89.2 / 36.0	82.6 / 6.1	2.8	11.2	18.6	6.6	51.5
Triplet with BERT	[88]	87.7 / 35.0	78.8 / 5.4	6.7	23.6	36.6	14.1	16
InfoNCE	[76], [77], [82]	87.0 / 32.5	77.6 / 5.1	6.8	25.4	38.4	15.3	17
+ Stochastic sampling	[77]	89.8 / 38.0	84.8 / 7.7	**10.2**	29.8	42.8	18.7	13
In our environment								
Baseline (Contrastive learning)	[77]	89.7 / 38.4	84.2 / 7.6	9.5	28.7	42.9	18.0	12
Proposed (Aligned Contrastive)	Ours	**90.3** / **39.3**	**84.9** / **8.1**	9.6	**33.8**	**47.8**	**19.6**	**11**

3.5.4 Quantitative Results

Table 3.1 presents the performance comparison for text-to-music retrieval. Our proposed Aligned Contrastive model outperforms the baseline across most sentence-level retrieval metrics. In particular, we observe substantial gains in ranking accuracy, with R@5 improving from 28.7 to 33.8, R@10 from 42.9 to 47.8, and mAP@10 from 18.0 to 19.6.

These improvements indicate that explicitly learning frame-token alignment enables the model to better distinguish between semantically similar but distinct audio clips, thereby placing the ground-truth target closer to the top ranks. The tag-level metrics (ROC-AUC and PR-AUC) also show consistent improvements. Taken together, these results support our hypothesis that fine-grained modeling enriches the global embedding with detailed temporal information, which is essential for high-precision retrieval.

3.5.5 Qualitative Analysis: Visualization of Alignment

To qualitatively validate that the model acquires meaningful local correspondences, we visualize the frame-token alignment map in Figure 3.3. The heatmap displays the pairwise cosine similarity scores between the embeddings of specific text tokens (e.g., music tags) from the text encoder and the sequence of audio frames.

We observe distinct temporal activation patterns within the mixed-genre track. Specifically, the token "metal" exhibits high similarity with frames corresponding to heavy guitar riffs, whereas "electronic" localizes to segments featuring synthesized beats. This visualization demonstrates that our Aligned Contrastive Learning successfully decomposes global semantic concepts into their corresponding temporal segments. Crucially, it offers a level of interpretability regarding "which sound triggered which word", a fine-grained alignment capability that global-pooling models inherently lack.

3.6 Summary

This chapter addressed the critical limitation of temporal information loss in conventional audio-text modeling by introducing Aligned Contrastive Learning. Unlike traditional approaches that aggregate sequences into single global vectors, the proposed framework explicitly optimizes the alignment between audio frames and text tokens. Through experiments on the ECALS music dataset, we demonstrated that our method consistently outperforms global-only baselines in text-to-music retrieval. We chose music as the evaluation domain because its complex temporal dynamics and compositional structure provide a demanding testbed for fine-grained alignment. Qualitative visualizations further confirmed that the model acquires interpretable, fine-grained correspondences between sound events and textual descriptions. These findings

support our hypothesis that preserving local granularity and reducing the entanglement of multiple events into a single global embedding are decisive factors for improving the precision of cross-modal audio-text representations. Although the proposed framework is domain-agnostic, extending the evaluation to other audio domains such as environmental sounds remains future work.

While this chapter successfully enhanced the granularity of cross-modal alignment, a fundamental challenge remains regarding intra-modal relations. Even with fine-grained audio-text alignment, the standard contrastive objective does not explicitly enforce how acoustic relationships (e.g., differences between mixed sounds) are organized within the audio embedding space. The model may align audio to text, but it does not necessarily acquire a compositional understanding of the audio signal itself. Addressing this lack of explicit intra-modal relational modeling is the focus of the next chapter, where we introduce a method that targets relations between audio clips through descriptions of their differences.

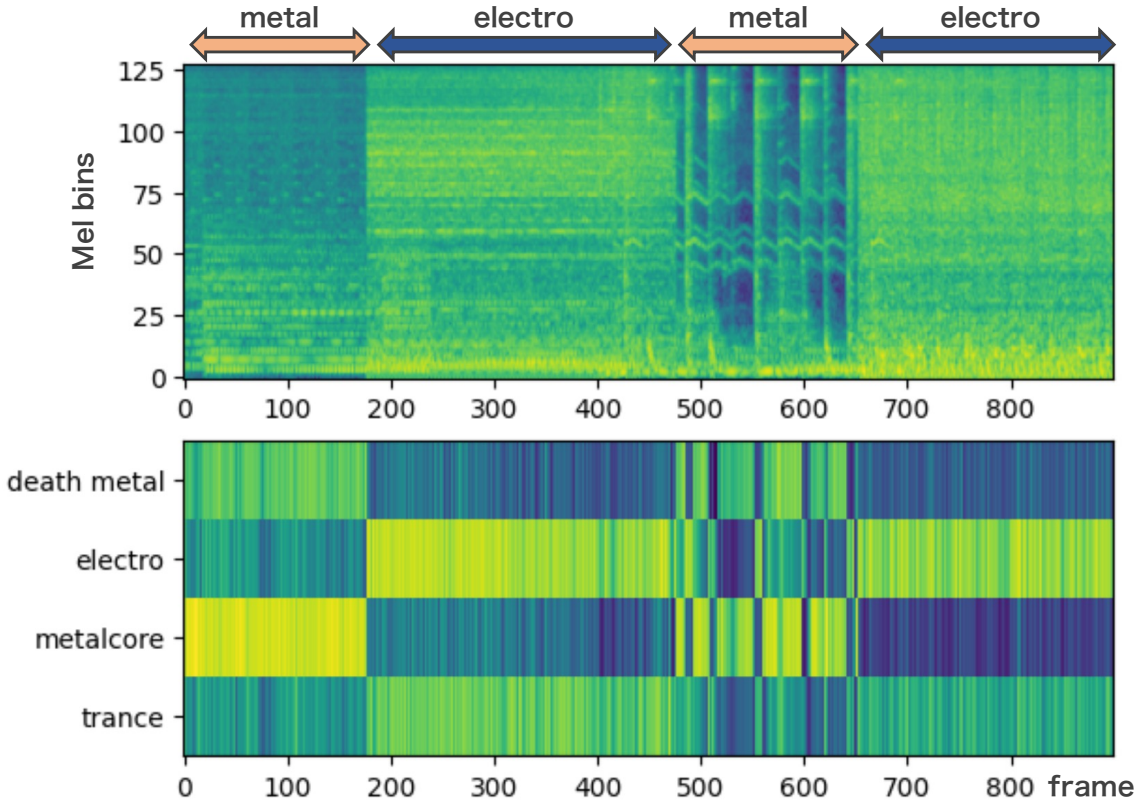

Figure 3.3: An example of the aligned similarity from our proposed method. The top figure displays the mel spectrogram of the input audio, while the bottom figure shows the similarity between the embeddings of music tags obtained from the text encoder and each audio frame. Brighter areas indicate higher similarity. The input audio is a mixture of metal and electronic music, and the visualization illustrates which parts of the music correspond to each tag (query).

Chapter 4

Semantic Alignment for Intra-Modal Relations via Audio Difference Modeling

4.1 Introduction

As discussed in the previous chapters, a fundamental limitation in current audio-text modeling lies in the lack of explicit intra-modal relations. Conventional encoders typically map an audio clip into a single global embedding, achieving audio-to-text alignment but failing to establish alignment between audio components themselves. While this strategy captures global semantics, it often fails to preserve the compositional nature of audio scenes within the latent space. For instance, a standard model may map "a dog barking" and "street noise" to distinct points, but it lacks a structural mechanism to represent their mixture ("a dog barking with street noise") as a compositional function of its parts. Overcoming this limitation requires training the model to perform semantic arithmetic within the audio modality itself, i.e., to understand how the addition or subtraction of acoustic events alters semantic meaning.

The contrastive objective employed in Chapter 3 improves cross-modal alignment but still does not explicitly model intra-modal relations. Because InfoNCE optimizes only a scalar similarity score between audio and text, it cannot capture how one sound differs from another in compositional terms. Moreover, semantic composition is often non-linear; the difference between "a mixed sound" and "background noise" cannot be represented by simple vector subtraction. To address this, we adopt a generative formulation: by training a model to describe the difference between two audio clips in natural language, we force the encoder to disentangle mixed acoustic events.

Motivated by this perspective, this chapter tackles intra-modal learning through the task of audio captioning. Audio captioning is the task of generating natural language descriptions for audio content [89], [90], [91]. While traditionally viewed as a standalone application for accessibility or retrieval, we reinterpret captioning here as a powerful probing mechanism for representation learning [41], [42]. If a model can accurately describe the difference between two sounds, this indicates that it has successfully disentangled the underlying acoustic events in its latent space. However, standard captioning datasets are relatively small and lack explicit instruction on acoustic relationships, making it difficult to learn such structure from existing data alone.

To address both the lack of intra-modal relations and the scarcity of captioning data, we propose a novel framework termed Audio Difference Learning. Unlike standard captioning, which maps a single input to text (Audio $\rightarrow$ Text), our approach introduces a reference audio and trains the model to describe the semantic difference between an input and a reference ($\text{Audio}_{\text{in}} - \text{Audio}_{\text{ref}} \rightarrow$ Text). This formulation explicitly imposes a compositional constraint on the representation space.

To realize this framework, we introduce two key technical contributions. First, we design a specialized Diff Block module that effectively captures the semantic discrepancy

between input and reference audio. We investigate architectures ranging from simple subtraction to attention-based masking mechanisms, enabling the model to handle complex, non-linear acoustic mixtures. Second, to address the lack of explicit difference annotations, we propose a new training strategy that constructs pseudo difference targets from existing caption data, effectively turning the problem into a self-supervised task. Through these mechanisms, we aim to achieve two goals: (1) enhancing intra-modal understanding by disentangling acoustic events in the feature space, and (2) providing a robust data augmentation method that improves captioning performance without additional human annotation cost.

4.2 Related Work

In this section, we review methodologies relevant to difference modeling. We first discuss how difference and change have been modeled in computer vision, which serves as a conceptual basis for our approach. We then examine recent attempts in the audio domain, specifically focusing on data augmentation and change captioning, to clarify the position of our proposed framework.

4.2.1 Difference Modeling in Computer Vision

The concept of describing differences or changes between inputs has been extensively explored in computer vision. Early work formulated change captioning, which aims to describe alterations between image pairs [92], [93], and demonstrated that attention mechanisms can effectively isolate changing regions. Subsequent studies advanced this line of research with more robust architectures and training strategies for viewpoint-agnostic change detection and captioning [94], [95], [96], [97].

Recent advancements have also leveraged pretrained cross-modal models. CLIP4IDC [98] adapts CLIP for image difference captioning, utilizing the robust embedding space of vision-language foundation models to describe changes. VisDiff [99] further generalizes this idea to describing differences between sets of images, capturing distributional rather than instance-level changes. These studies confirm that modeling "what changed" provides a powerful inductive bias for disentangling semantic concepts. Our work transfers this insight to the audio domain, aiming to acquire compositional representations by modeling acoustic differences rather than only absolute content.

4.2.2 Compositional and Difference Learning in Audio

Data augmentation via mixing. Attempts to model compositional semantics in audio often appear in the context of data augmentation. Inspired by MixGen in vision [100], recent audio captioning studies [101], [102], [103] synthesize training samples by mixing two audio clips and concatenating their captions. However, these methods typically rely on simple concatenation rules (e.g., joining captions with "and"), treating mixtures as a "bag of labels." This approach fails to establish alignment between individual mixed components and their corresponding captions, preventing explicit modeling of how one sound interacts with or differs from another.

Audio difference and commonality modeling. Directly addressing the relationships between audio clips is an emerging research direction beyond simple augmentation. Recent works [104], [105] have attempted to generate captions describing changes or differences between audio pairs. However, these approaches often rely on specialized datasets or focus solely on the change-captioning task itself, limiting their applicability as a general representation learning framework.

In contrast, our proposed Audio Difference Learning uses difference modeling not only as a downstream task, but as a pretext objective (or probe) to improve standard captioning performance and to acquire robust feature representations without requiring manual difference annotations. Another complementary perspective is offered by Audio Commonality Captioning [106], which trains models to describe shared semantics across clips. The duality between our difference modeling and such commonality modeling suggests that a complete model of audio composition should encompass both operations to fully disentangle the latent acoustic structure. Moreover, unlike image-based change captioning where spatial alignment is relatively stable, audio requires robustness to temporal continuity and misalignment; this motivates the masking-based difference mechanisms introduced in our framework.

4.3 Audio Captioning Framework

4.3.1 Task Definition

Automated Audio Captioning (AAC) is formulated as a cross-modal sequence generation task that maps general audio signals to free-form natural-language descriptions of acoustic events, scenes, and their temporal relations [89], [107]. Building on the foundational sequence-to-sequence framework [108], a common modeling paradigm is an encoder-decoder architecture in which log-mel spectrograms are encoded into a sequence of latent representations and decoded autoregressively into word sequences, often with an attention mechanism that provides soft alignment between time frames and generated tokens. Convolutional recurrent neural networks (CRNNs) extend this framework by combining convolutional layers for local time-frequency feature extraction with recurrent layers for temporal aggregation [91], and adversarial training has been

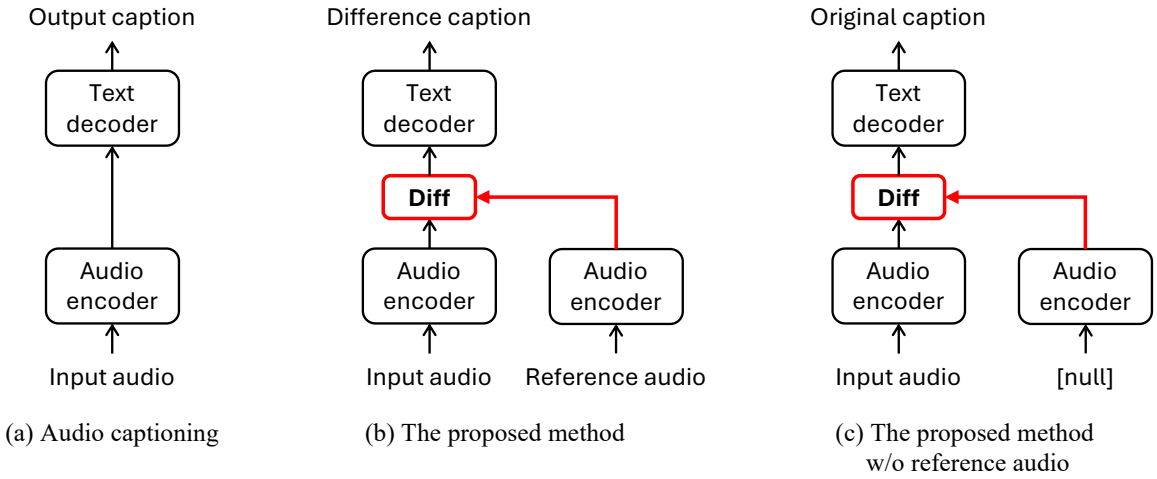

Figure 4.1: (a) Conventional audio captioning, which maps a single input to a caption. (b) The proposed Audio Difference Learning, which generates the difference between input and reference audio based on the difference of their encoded representations. (c) When the reference audio is absent (or null), the proposed framework reduces to the conventional audio captioning system.

explored to promote caption diversity and naturalness [109]. Pretrained audio tagging and audio-text models such as PANNs [49], AST [50], BEATs [110], HTS-AT [51], and CLAP [1], [2] are frequently adopted as encoders and fine-tuned for caption generation, yielding strong performance on benchmark datasets including Clotho and AudioCaps.

4.3.2 Encoder-Decoder Formulation

Audio captioning is typically formulated with an encoder-decoder architecture. Let the spectral feature of the input audio be denoted as $\mathbf{X}^{(\mathsf{A})} \in \mathbb{R}^{L_{\mathsf{A}} \times D_{\mathsf{A}}}$, and let the target text sequence be $\mathbf{y} \in \mathcal{V}^{L_{\mathsf{T}}}$, where $\mathcal{V}$ denotes the vocabulary. First, the input audio $\mathbf{X}^{(\mathsf{A})}$ is fed into the audio encoder, which transforms it into a sequence of latent features

$\mathbf{Z}^{(\mathsf{A})} \in \mathbb{R}^{L_\mathsf{A} \times D}$, where D is the feature dimension:

$$\mathbf{Z}^{(\mathsf{A})} = \mathrm{AudioEncoder}(\mathbf{X}^{(\mathsf{A})}). \tag{4.1}$$

The representation $\mathbf{Z}^{(\mathsf{A})}$ is intended to capture the semantic content of $\mathbf{X}^{(\mathsf{A})}$. Feeding $\mathbf{Z}^{(\mathsf{A})}$ into the decoder yields an estimate $\hat{\mathbf{y}}$ of the target caption:

$$\hat{\mathbf{y}} = \mathrm{TextDecoder}(\mathbf{Z}^{(\mathsf{A})}), \tag{4.2}$$

where $\hat{\mathbf{y}}$ represents the predicted probability distribution over the vocabulary at each time step. Cross-entropy between the predicted text sequence $\hat{\mathbf{y}}$ and the target text sequence $\mathbf{y}$ is commonly used as a loss function:

$$\mathcal{L} = \mathsf{CrossEntropy}(\mathbf{y}, \hat{\mathbf{y}}). \tag{4.3}$$

4.4 Proposed Method: Audio Difference Learning

4.4.1 Overview

We propose Audio Difference Learning to capture intra-modal relationships between an input $\mathbf{X}^{(\mathsf{A})}$ and a reference $\mathbf{X}^{(\mathsf{A})}_{\mathsf{Ref}}$. The core idea is to obtain a **difference representation** $\mathbf{Z}^{(\mathsf{A})}_{\mathsf{Diff}}$ in the feature space that corresponds, at a semantic level, to

$$\mathrm{Semantics}(\mathbf{X}^{(\mathsf{A})}) - \mathrm{Semantics}(\mathbf{X}^{(\mathsf{A})}_{\mathsf{Ref}}).$$

The overall structure is shown in Figure 4.1(b). We encode both the input $\mathbf{X}^{(\mathsf{A})}$ and the reference $\mathbf{X}^{(\mathsf{A})}_{\mathsf{Ref}}$ to obtain feature representations $\mathbf{Z}^{(\mathsf{A})}$ and $\mathbf{Z}^{(\mathsf{A})}_{\mathsf{Ref}}$, respectively. We then derive a difference representation $\mathbf{Z}^{(\mathsf{A})}_{\mathsf{Diff}}$ by applying a difference function $\mathrm{Diff}(\cdot, \cdot)$ to the two feature sequences:

$$\mathbf{Z}^{(\mathsf{A})}_{\mathsf{Diff}} = \mathrm{Diff}(\mathbf{Z}^{(\mathsf{A})}, \mathbf{Z}^{(\mathsf{A})}_{\mathsf{Ref}}), \tag{4.4}$$

where $\mathrm{Diff}(\cdot,\cdot)$ is designed to represent the semantic discrepancy between two encoded audio representations. The resulting difference representation $\mathbf{Z}_{\mathsf{Diff}}^{(\mathsf{A})}$ is then fed into the decoder to generate a caption $\hat{\mathbf{y}}_{\mathsf{Diff}}$:

$$\hat{\mathbf{y}}_{\mathsf{Diff}} = \mathrm{TextDecoder}(\mathbf{Z}_{\mathsf{Diff}}^{(\mathsf{A})}). \tag{4.5}$$

Note that if the reference audio $\mathbf{X}_{\mathsf{Ref}}^{(\mathsf{A})}$ is set to null (e.g., omitted or replaced with a neutral reference), the system reduces to the conventional audio captioning system in Figure 4.1(a,c).

4.4.2 Training Strategy: Latent Semantic Subtraction

A major hurdle in difference learning is the lack of datasets with explicit difference captions. We address this by constructing a synthetic training signal in which the difference is strictly defined.

Let $\mathbf{x}^{(\mathsf{A})}$ and $\mathbf{x}_{\mathsf{Ref}}^{(\mathsf{A})}$ denote waveform-level representations of the input and reference audio, respectively. We construct a new input $\mathbf{x}_{+}^{(\mathsf{A})}$ by superimposing the reference audio onto the original input in the time domain:

$$\mathbf{x}_{+}^{(\mathsf{A})} = \mathbf{x}^{(\mathsf{A})} + \mathbf{x}_{\mathsf{Ref}}^{(\mathsf{A})}. \tag{4.6}$$

After encoding, we obtain feature sequences $\mathbf{Z}_{+}^{(\mathsf{A})} = \mathrm{AudioEncoder}(\mathbf{X}_{+}^{(\mathsf{A})})$ and $\mathbf{Z}_{\mathsf{Ref}}^{(\mathsf{A})} = \mathrm{AudioEncoder}(\mathbf{X}_{\mathsf{Ref}}^{(\mathsf{A})})$. In the feature space, if the model correctly learns to subtract the reference information, the difference between the mixed input feature $\mathbf{Z}_{+}^{(\mathsf{A})}$ and the reference feature $\mathbf{Z}_{\mathsf{Ref}}^{(\mathsf{A})}$ should recover the semantics of the original $\mathbf{Z}^{(\mathsf{A})}$. We therefore compute

$$\mathbf{Z}_{\mathsf{Diff}}^{(\mathsf{A})} = \mathrm{Diff}(\mathbf{Z}_{+}^{(\mathsf{A})}, \mathbf{Z}_{\mathsf{Ref}}^{(\mathsf{A})}), \tag{4.7}$$

and decode it to obtain

$$\hat{\mathbf{y}}_{\mathsf{Diff}+} = \mathrm{TextDecoder}(\mathbf{Z}^{(\mathsf{A})}_{\mathsf{Diff}}). \tag{4.8}$$

We then optimize the model so that the decoded caption matches the original caption $\mathbf{y}$ of $\mathbf{X}^{(\mathsf{A})}$:

$$\mathcal{L}_{\mathsf{Diff}+} = \mathsf{CrossEntropy}(\mathbf{y}, \hat{\mathbf{y}}_{\mathsf{Diff}+}). \tag{4.9}$$

This strategy forces the $\mathrm{Diff}(\cdot)$ block and the encoder to learn a representation space in which acoustic mixing in the time domain corresponds to feature-space operations that can be approximately inverted (disentangled). In other words, the model is encouraged to organize its latent space so that additive mixtures of sounds can be decomposed into constituent semantic components through learned latent subtraction.

4.4.3 Design of Difference Calculation

Subtraction-based Approach

The simplest form of difference is element-wise subtraction (Figure 4.2):

$$\mathbf{Z}^{(\mathsf{A})}_{\mathsf{Diff}} = \mathbf{Z}^{(\mathsf{A})} - \mathbf{Z}^{(\mathsf{A})}_{\mathsf{Ref}}. \tag{4.10}$$

This implicitly assumes a linear relationship in the feature space and a one-to-one correspondence between time frames of the input and reference. While this approach is parameter-free and computationally cheap, it strictly requires temporal alignment between the mixed signal and the reference, which is a strong constraint for real-world intra-modal modeling where events may be shifted, stretched, or partially overlapping.

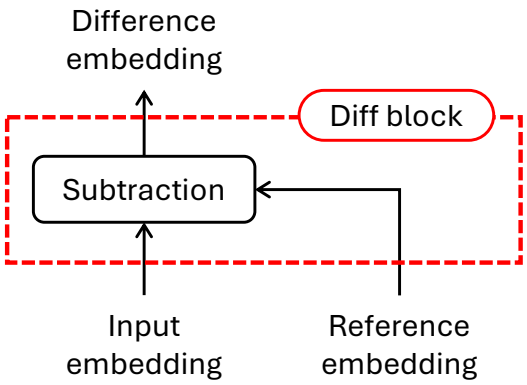

Figure 4.2: Diff Block with subtraction-based difference. The input embedding and reference embedding are subtracted element-wise to produce the difference embedding.

Masking-based Approach via Cross-Attention

To allow for more flexible intra-modal relationships (e.g., when the reference sound appears at a different time, with different duration, or with slight variation), we propose a masking-based approach (Figure 4.3). Instead of directly subtracting feature vectors, we first compute a similarity map between the input and reference using a cross-attention mechanism. Concretely, we obtain a similarity matrix $\mathbf{Z}_{\text{sim}}$ as

$$\mathbf{Z}_{\text{sim}} = \text{CrossAttention}(\mathbf{Z}^{(\text{A})}, \mathbf{Z}^{(\text{A})}_{\text{Ref}}), \tag{4.11}$$

where $\text{CrossAttention}(\cdot, \cdot)$ denotes a cross-attention block that measures how strongly each frame in $\mathbf{Z}^{(\text{A})}$ attends to frames in $\mathbf{Z}^{(\text{A})}_{\text{Ref}}$.

Based on this similarity, we generate a soft mask

$$\mathbf{M} = \mathbf{1} - \sigma(\mathbf{Z}_{\text{sim}}),$$

which assigns low weights to input frames that are well explained by the reference and high weights to frames that are dissimilar. We then apply this mask to the input

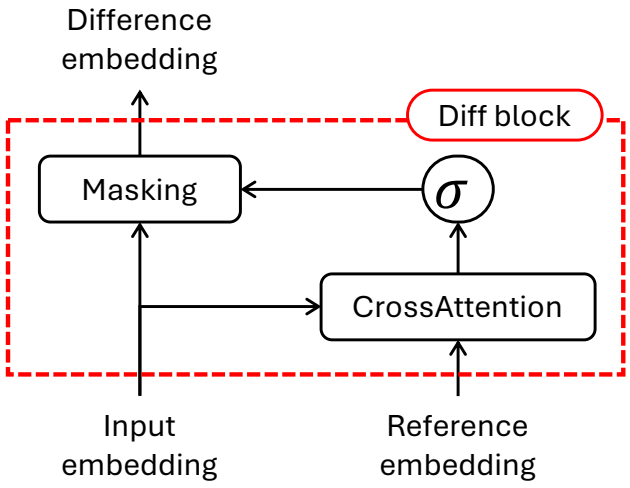

Figure 4.3: Diagram of the Diff block using the Masking method. Cross-attention calculates the similarity between the input and reference embeddings. The resulting weights are scaled by a sigmoid function and applied in the masking process to produce the difference embedding.

representation:

$$\mathbf{Z}^{(\mathrm{A})}_{\mathrm{masked}} = \mathbf{M} \odot \mathbf{Z}^{(\mathrm{A})}, \tag{4.12}$$

where $\odot$ denotes element-wise multiplication. This procedure effectively performs semantic subtraction by filtering out shared information, regardless of exact temporal alignment between input and reference. The resulting masked representation can then be used as $\mathbf{Z}^{(\mathrm{A})}_{\mathrm{Diff}}$ in Eq. (4.4). By relying on attention-based masking rather than strict frame-wise subtraction, this approach better matches the goal of robustly learning intra-modal relations in temporally continuous and loosely aligned audio signals.

4.5 Experimental Evaluations

4.5.1 Experimental Settings

We adopt the baseline audio captioning system from DCASE 2023 Task 6 as our backbone model. The input audio features are 64-dimensional mel spectrograms. The audio encoder is a pretrained 12-layer CNN, followed by a Transformer-based decoder (BART) [111]. In the masking-based approach, we insert the proposed cross-attention-based Diff Block between the encoder and decoder.

We compare the following training configurations: (i) the baseline model trained with the standard cross-entropy loss, (ii) Al-MixGen (PairMix) [101], a conventional mix-up augmentation method for audio captioning, and (iii) our proposed Audio Difference Learning with subtraction- or masking-based difference modeling.

For all models, we use the Adam optimizer with a learning rate of 1.0×10^{-4}, a batch size of 32, and train for 50 epochs.

4.5.2 Evaluation Metrics

Captioning performance is evaluated using standard metrics, including BLEU [112], METEOR [113], ROUGE [114], CIDEr [115], SPICE [116], and SPIDEr [117], along with task-specific extensions defined in the DCASE AAC challenges. BLEU measures the geometric mean of n-gram precision between a generated caption and one or more reference captions, with a brevity penalty to discourage overly short outputs. METEOR computes an F-score over unigram matches while incorporating stemming and synonym matching, and applies a fragmentation penalty to favor outputs with better word order. ROUGE is a recall-oriented family of measures that evaluates the overlap of n-grams or longest common subsequences, capturing how much of the reference

content is covered by the hypothesis. CIDEr is a consensus-based metric that uses TF-IDF-weighted n-gram similarity across multiple references to emphasize caption elements that are both frequent within a reference set and discriminative across images or audio clips. SPICE goes beyond surface n-gram overlap by parsing captions into scene-graph representations and comparing semantic tuples (objects, attributes, and relations), thereby assessing semantic fidelity. SPIDEr combines CIDEr and SPICE into a single score, balancing lexical similarity and semantic structure, and is widely adopted as a summary measure for caption quality.

4.5.3 Dataset Design for Difference Learning

To train and evaluate the model's ability to disentangle sources, we construct synthetic mixtures using the Clotho [5] and ESC-50 [10] datasets. Clotho provides the base audio captioning data, while ESC-50 provides diverse environmental sounds that we use as reference signals.

For training, we generate mixed inputs $\mathbf{X}_{+}^{(\mathsf{A})}$ by superimposing ESC-50 clips onto Clotho clips in the time domain. For the reference audio $\mathbf{X}_{\mathsf{Ref}}^{(\mathsf{A})}$, we prepare various conditions to test the robustness of intra-modal modeling, including: (i) cases where the reference is the same recording as in the mixture (Same/ Same), (ii) cases where the reference is temporally misaligned (Same/ Diff), and (iii) cases where the reference is a different recording of the same class (Diff/ Diff). We also investigate multi-reference settings (Multi/ Same, Multi/ Diff), where multiple ESC-50 clips are used as references for a single mixed input.

Table 4.1: Experimental results of the general audio captioning task setting: The proposed method employed the reference audio only during the training phase, and it was not used during the evaluation. These results highlight the impact of our proposed audio difference learning on the general audio captioning.

Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDEr	SPICE	SPIDEr
Baseline	0.585	0.379	0.251	0.161	0.179	0.386	0.399	0.120	0.259
AL-MixGen [101]	0.590	0.384	0.254	0.164	0.180	0.392	0.404	0.122	0.263
Proposed (Subtraction)	0.593	0.386	0.257	0.164	0.181	0.392	0.403	0.122	0.264
Proposed (Mask)	**0.606**	**0.395**	**0.266**	**0.173**	**0.187**	**0.405**	**0.431**	**0.129**	**0.280**

Table 4.2: Performance comparison of the proposed methods with various combinations of audio clips. Evaluations were conducted using subtraction- and masking-based Diff Blocks across different source and time alignment patterns. The results indicate that the masking-based approach consistently outperforms subtraction, maintaining high performance across all conditions.

Model (Source / Time)	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDEr	SPICE	SPIDEr
Subtraction (Same / Same)	0.590	0.383	0.253	0.162	0.180	0.390	0.403	0.121	0.262
Mask (Same / Same)	**0.606**	**0.395**	**0.266**	**0.173**	**0.187**	0.405	**0.431**	**0.129**	**0.280**
Mask (Same / Diff)	0.605	0.394	0.264	0.171	**0.187**	0.405	0.427	**0.129**	0.278
Mask (Diff / Diff)	0.605	0.392	0.261	0.168	0.186	**0.406**	0.419	0.126	0.273
Mask (Multi / Same)	**0.606**	0.394	0.264	0.171	**0.187**	0.404	0.419	0.127	0.273
Mask (Multi / Diff)	0.603	0.389	0.259	0.167	0.184	0.401	0.414	0.127	0.271

4.5.4 Results and Discussion

Table 4.1 presents the results on the Clotho test set. The proposed Audio Difference Learning outperforms both the baseline and the MixGen-based augmentation across all captioning metrics. In particular, the masking-based Diff Block achieves the highest

performance, yielding an improvement of approximately 8% in SPIDEr compared to the baseline. This indicates that the model successfully learns to disentangle mixed audio sources in the feature space, establishing semantic alignment between individual components and their descriptions, and leading to more accurate captions of the target content.

Robustness of intra-modal modeling. Table 4.2 further compares the subtraction- and masking-based Diff Blocks under different mixing and reference conditions. The masking-based block maintains high performance even when the reference audio is not temporally aligned (Same/ Diff) or is a different recording of the same class (Diff/ Diff), and remains robust in multi-reference settings (Multi/ Same, Multi/ Diff). In contrast, the subtraction-based block degrades markedly once strict alignment is violated. This confirms that the masking-based approach, which relies on cross-attention and soft masking, captures the abstract semantic difference between input and reference rather than merely performing frame-wise signal subtraction, thereby enabling robust learning of intra-modal relations.

4.5.5 Qualitative Analysis: Disentanglement Capability

Table 4.3 illustrates the model's inference capability on mixed-audio scenarios. When presented with a mixed audio $\mathbf{X}_{+}^{(\mathsf{A})}$, the standard baseline often fails to generate coherent captions for the target component, instead blending words from both sources. In contrast, our method, by leveraging the reference audio, successfully generates captions for the target component (e.g., $\mathbf{X}^{(\mathsf{A})}$) by explicitly subtracting the interference (e.g., the ESC-50 clip) via the difference representation.

Moreover, when we apply the difference function in the opposite direction, $\mathrm{Diff}(\mathbf{X}_{+}^{(\mathsf{A})}, \mathbf{X}^{(\mathsf{A})})$,

the proposed model can generate captions that focus on the added source (e.g., car horn, laughing). These qualitative results support the conclusion that our approach constructs a structured feature space in which audio components are effectively disentangled and can be selectively described via language.

4.6 Summary

In this chapter, we proposed Audio Difference Learning to address the limitation of understanding intra-modal relations in audio-text models. By training the model to describe the difference between an input and a reference audio, we established a representation space in which acoustic addition and subtraction can be explicitly modeled. The proposed masking-based difference block, implemented via cross-attention, demonstrated robustness to temporal misalignment and achieved state-of-the-art performance on standard audio captioning benchmarks. These results suggest that incorporating relative representations between audio inputs is a promising direction for moving beyond a monolithic view of audio features, leading to more interpretable and controllable audio-text modeling.

One limitation of the current approach is its design for pairwise difference computation. When audio clips contain three or more overlapping sources, the masking-based approach requires multiple sequential subtractions, which may accumulate errors and reduce interpretability. Extending difference learning to multi-source scenarios therefore remains an open challenge. More broadly, this chapter showed that language-supervised difference modeling can reveal and shape the latent structure of the audio space. In the next chapter, we turn to the complementary question of how to exploit large amounts of unlabeled audio to further improve audio-text representations, aiming to relax the dependence on paired data while preserving the structural benefits

established in Chapters 3 and 4.

Table 4.3: Examples of difference captioning results generated using difference representations between two audio clips. The proposed method can handle both mixed sounds and their differences.

(1) Input audio: *kids are playing as one child shrieks while birds are chirping*	
Baseline	birds are chirping and children are talking in the background
Proposed	birds are chirping and children are talking and playing in the background
(2) Caption for mixed audio: $\mathbf{X}_{+}^{(\mathsf{A})} = \mathbf{X}^{(\mathsf{A})} + [\texttt{Car Horn sound}]$	
Baseline	birds are chirping and children are talking in the background
Proposed	birds are chirping and children are talking in the background **as a car drives by**
(3) Caption with difference representation: $\mathrm{Diff}(\mathbf{X}_{+}^{(\mathsf{A})}, [\texttt{Car Horn sound}]) \Rightarrow (1)$	
Baseline	birds are chirping and children are talking to each other
Proposed	birds are chirping and children are talking in the background
(4) Caption with difference representation: $\mathrm{Diff}(\mathbf{X}_{+}^{(\mathsf{A})}, \mathbf{X}^{(\mathsf{A})}) \Rightarrow [\texttt{Car Horn sound}]$	
Baseline	a person is using a hard object to make a few seconds
Proposed	**an engine is whirring** and then it gets louder and louder
(1) Input audio: *a distorted drum or similar instrument is played*	
Baseline	a synthesizer is playing a musical instrument
Proposed	a synthesizer is playing a synthesizer with a musical instrument
(2) Caption for mixed audio: $\mathbf{X}_{+}^{(\mathsf{A})} = \mathbf{X}^{(\mathsf{A})} + [\texttt{Laughing sound}]$	
Baseline	a person is playing a synthesizer with a musical instrument in the background
Proposed	a person is playing a synthesizer **with a man talks in the background**
(3) Caption with difference representation: $\mathrm{Diff}(\mathbf{X}_{+}^{(\mathsf{A})}, [\texttt{Laughing sound}]) \Rightarrow (1)$	
Baseline	a synthesizer is playing a musical instrument
Proposed	a synthesizer is playing a musical instrument
(4) Caption with difference representation: $\mathrm{Diff}(\mathbf{X}_{+}^{(\mathsf{A})}, \mathbf{X}^{(\mathsf{A})}) \Rightarrow [\texttt{Laughing sound}]$	
Baseline	a person is playing a synthesizer while another person is speaking in the background
Proposed	a person is speaking and then **a child laughs**

Chapter 5

Semi-Supervised Representation Learning via Audio-to-Text Mapping

5.1 Introduction

The previous chapters focused on refining the architecture of audio-text models to capture fine-grained alignment (Chapter 3) and coherent intra-modal relations (Chapter 4). While these architectural improvements significantly enhance the expressiveness and interpretability of the learned representations, a fundamental bottleneck remains: the quality of the shared embedding space is heavily constrained by the scarcity of paired training data. Unlike the computer vision domain, which benefits from billion-scale datasets such as LAION-5B [22], audio-text datasets are scarce, typically containing only tens of thousands of pairs (e.g., Clotho [5], AudioCaps [54]). This substantial data gap leaves the shared latent space sparse and prone to overfitting, limiting the generalization capability of even the most sophisticated architectures.

Two mainstream strategies attempt to mitigate this scarcity: augmentation and syn-

thetic caption generation. Augmentation-based methods mix audio clips and compose their captions [101], [102], [103], sometimes inserting temporal connectors to encode ordering [57], [118], [119]. Generation-based methods employ large language models (LLMs) to synthesize captions from tags or metadata [23], [55], [120]. While effective to some extent, the former still relies on the availability of ground-truth captions for source clips, and the latter depends on textual side information, introducing biases or distributional drift inherent in LLM outputs [121], [122], [123] and often resulting in only modest improvements in downstream performance [124], [125].

Parallel to these paired-data strategies, leveraging unlabeled audio has gained attention in tasks such as weakly supervised captioning [126], [127], text-queried source separation [128], and audio-language modeling [129]. However, these approaches typically pass unlabeled audio through a frozen contrastive encoder and fine-tune task-specific heads. Consequently, they improve downstream task performance but leave the shared audio-text representation itself unchanged. How to exploit unlabeled audio to directly refine the contrastive encoder remains an open problem.

To tackle this challenge, we propose a semi-supervised framework that trains an audio-text contrastive model using large collections of unlabeled audio alongside available labeled clips. The core idea is to map the embeddings of unlabeled audio into the text token embedding space via an **Audio-to-Text (a2t) Mapper**, producing **continuous pseudo-token embeddings**. This process effectively treats unlabeled audio as if it were a valid textual description, enabling implicit alignment between unlabeled audio and the text modality within the contrastive objective. By integrating these pseudo-tokens into the training loop, we densify the training distribution and regularize the shared embedding space without relying on external caption generation.

Our approach is inspired by Pic2Word [130] in computer vision, which maps unla-

beled images to text tokens for zero-shot composed-image retrieval. A critical distinction, however, lies in the objective: whereas Pic2Word freezes the backbone to adapt it for a specific retrieval task, our framework jointly optimizes the audio encoder, text encoder, and the mapper. This transforms the mapping mechanism from a downstream adaptation tool into a representation learning component, enabling the backbone itself to learn richer features from unlabeled data. We explore two mapper designs: (i) a simple MLP-based mapper, the effectiveness of which we verified in our preliminary study [131], and (ii) a sequence-flexible Transformer-based mapper capable of attending to variable-length audio embeddings. To further enhance robustness, we introduce a noise-injection strategy in which Gaussian noise is added to audio embeddings before mapping, smoothing the pseudo-text representations and consistently improving retrieval accuracy.

The main contributions of this chapter are summarized as follows:

- We introduce, to the best of our knowledge, the first **semi-supervised framework that updates a CLAP-style backbone** using unlabeled audio, in contrast to prior approaches that keep the encoder frozen.
- We propose an **audio-to-text mapping** mechanism that projects audio features into the textual token space, enabling the utilization of unlabeled audio without the biases of synthetic caption generation.
- We demonstrate that incorporating unlabeled audio through our framework leads to substantial improvements in retrieval performance, validating that data density in the shared embedding space can be effectively increased without additional human annotation.

5.2 Related Work

Since general audio-text contrastive learning frameworks (e.g., CLAP [1], Wav2Clip [14]) were reviewed in Chapter 2, this section focuses specifically on methodologies addressing data scarcity. We categorize existing approaches into three groups: (1) augmentation and generation strategies that rely on paired supervision, (2) semi-supervised approaches utilizing unlabeled audio, and (3) cross-modal mapping techniques that serve as the theoretical basis for our proposed a2t mapper.

5.2.1 Strategies for Data Scarcity: Augmentation and Generation

To mitigate the shortage of human-annotated captions, two primary strategies have emerged: mixing-based augmentation and LLM-based generation.

Mixing-based Augmentation. Methods such as AL-MixGen [103] and PairMix [101] augment training data by mixing two audio clips and concatenating their captions (e.g., "sound A `and` sound B"). Subsequent work replaced the simple conjunction with temporal connectors such as "followed by" to encode ordering information [57], [118]. Although these methods effectively increase the diversity of training pairs, they still fundamentally require a ground-truth caption for every source clip. Our approach avoids this dependency by enabling the use of standalone unlabeled audio without requiring source captions.

Synthetic Caption Generation. Another prevalent strategy involves generating synthetic captions from tags or metadata using Large Language Models (LLMs). Datasets such as WavCaps [23], ClothoV2-GPT [120], and Auto-ACD [55] exemplify this ap-

proach in the audio domain. Related work in vision has explored text-level augmentation techniques such as paraphrasing and back-translation for image captioning [132], [133] and text-based augmentation for video captioning [134], underscoring the general appeal of expanding supervision through synthetic text. While cost-effective, these methods rely on the availability of textual side information (e.g., tags, titles) and can introduce bias or distributional drift inherent in LLM outputs [121], [122], [123]. Empirical studies have shown that such noise can lead to only modest improvements or even degradation in downstream tasks [124], [125]. In contrast, we take a different route: rather than generating discrete text strings, we directly align mapped audio vectors with the continuous token embedding space, bypassing the additional noise and bias brought by discrete text generation.

5.2.2 Learning from Unlabeled Audio

Recent studies have explored leveraging unlabeled audio for various audio-text tasks, though mostly in limited capacities.

Frozen Backbone Approaches. A common paradigm is to use a pretrained, frozen CLAP encoder to extract features from unlabeled audio, which are then used to train task-specific heads. This has been applied to weakly or unsupervised captioning [126], [127], text-queried source separation [128], and large audio-language models [129]. DR-Cap [135] further limits supervision by utilizing unlabeled audio clips solely at the inference stage. However, because these approaches keep the CLAP encoder frozen, the shared audio-text representation itself is not updated or improved. Our work differs fundamentally: we perform **semi-supervised contrastive pretraining** that jointly updates both the audio and text encoders, directly enhancing the quality of the back-

bone representation.

Hybrid Self-Supervised Learning. Some frameworks attempt to combine self-supervised objectives with contrastive learning. M2D-CLAP [136] integrates masked prediction (reconstruction) with CLAP training. While promising, these methods typically add a separate reconstruction head and objective, increasing architectural complexity. Our a2t mapper offers a streamlined alternative: by synthesizing pseudo-token representations, we allow unlabeled audio to participate directly in the contrastive loss without architectural modification to the backbone or auxiliary reconstruction tasks.

5.2.3 Cross-Modal Mapping and Inversion

Our concept of mapping audio to the text space draws inspiration from Textual Inversion techniques in computer vision. Pic2Word [130] proposed mapping unlabeled images to text tokens to enable zero-shot composed-image retrieval (e.g., "this image + 'dog"'). Similarly, in the text-to-image generation domain, techniques to invert images into pseudo-words have been widely adopted to personalize diffusion models [137], [138]. However, these methods typically freeze the backbone model and learn the mapping only for downstream adaptation or generation. We extend this mapping concept to representation learning: instead of adapting a frozen model, we use the mapping as a bridge to propagate gradients from unlabeled data back to the audio encoder, thereby refining the core representation space itself.

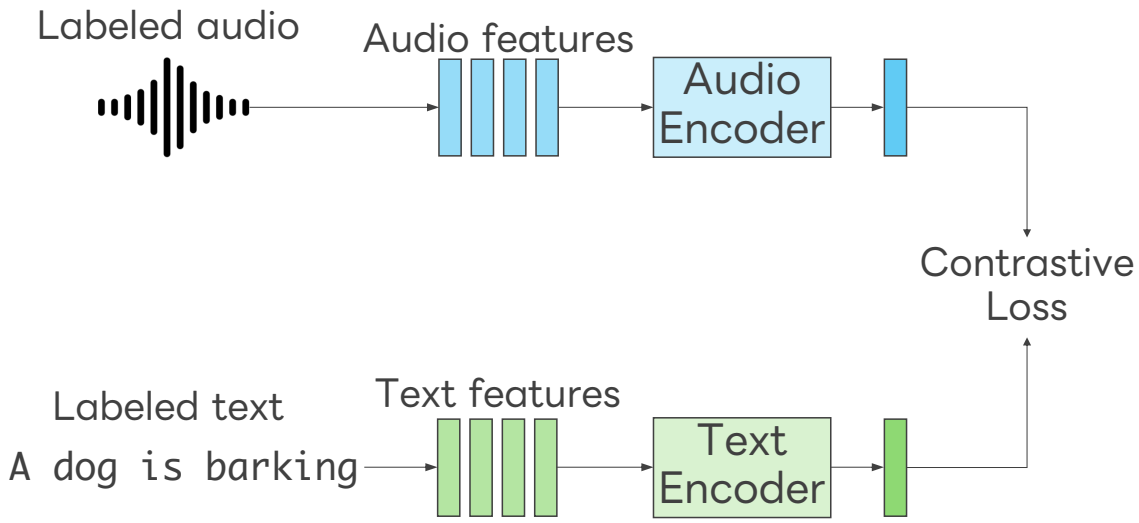

Figure 5.1: Standard audio-text contrastive learning pipeline. A paired clip is encoded into latent embeddings and optimized with the InfoNCE loss.

5.3 Audio-Text Contrastive Learning

We briefly review the contrastive framework introduced in Section 2.3; the next section extends it to the semi-supervised setting.

As shown in Figure 5.1, given paired samples $\left(\mathbf{X}^{(\mathsf{A})}, \mathbf{X}^{(\mathsf{T})}\right)$, modality-specific encoders map them to a shared latent space:

$$\mathbf{z}^{(\mathsf{A})} = \mathrm{AudioEncoder}\left(\mathbf{X}^{(\mathsf{A})}\right), \tag{5.1}$$

$$\mathbf{z}^{(\mathsf{T})} = \mathrm{TextEncoder}\left(\mathbf{X}^{(\mathsf{T})}\right), \tag{5.2}$$

with $\mathbf{z}^{(\mathsf{A})}, \mathbf{z}^{(\mathsf{T})} \in \mathbb{R}^D$. The model is trained with the symmetric InfoNCE loss. For the audio→text direction:

$$\mathcal{L}_{\mathrm{InfoNCE}}^{\mathsf{A}\to\mathsf{T}} = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{z}^{(\mathsf{A})}, \mathbf{z}^{(\mathsf{T})})/\tau\right)}{\sum_j \exp\left(\mathrm{sim}(\mathbf{z}^{(\mathsf{A})}, \mathbf{z}_j^{(\mathsf{T})})/\tau\right)}, \tag{5.3}$$

where sim denotes cosine similarity and τ the temperature. The text→audio counterpart is defined analogously:

$$\mathcal{L}_{\text{InfoNCE}}^{\mathsf{T}\to\mathsf{A}} = -\log \frac{\exp\big(\text{sim}(\mathbf{z}^{(\mathsf{T})}, \mathbf{z}^{(\mathsf{A})})/\tau\big)}{\sum_j \exp\big(\text{sim}(\mathbf{z}^{(\mathsf{T})}, \mathbf{z}_j^{(\mathsf{A})})/\tau\big)}. \tag{5.4}$$

The overall loss combines both directions:

$$\mathcal{L}_{\text{labeled}} = \mathcal{L}_{\text{InfoNCE}}^{\mathsf{A}\to\mathsf{T}} + \mathcal{L}_{\text{InfoNCE}}^{\mathsf{T}\to\mathsf{A}}. \tag{5.5}$$

5.4 Proposed Framework

We propose a semi-supervised contrastive method that leverages unlabeled audio by synthesizing a pseudo-token representation for it. The key idea is to mix a labeled clip with an unlabeled one and to generate a pseudo-token input through an audio-to-text mapper (a2t). This strategy turns every mixture into a valid audio-text pair, enabling InfoNCE training without additional annotations. Note that our framework does not generate natural-language captions; pseudo-token denotes latent vectors in the tokenized text embedding space.

5.4.1 System Overview

Figure 5.2 depicts the proposed semi-supervised contrastive pipeline. We assume mini-batches of paired samples $\big(x^{(\mathsf{A})}, x^{(\mathsf{T})}\big)$ and additional waveforms $x^{(\mathsf{A,Un})}$ (using only raw audio without class labels). First, we form the mixture

$$x^{(\mathsf{A,Mix})} \;=\; \tfrac{1}{2}\big(x^{(\mathsf{A})} + x^{(\mathsf{A,Un})}\big). \tag{5.6}$$

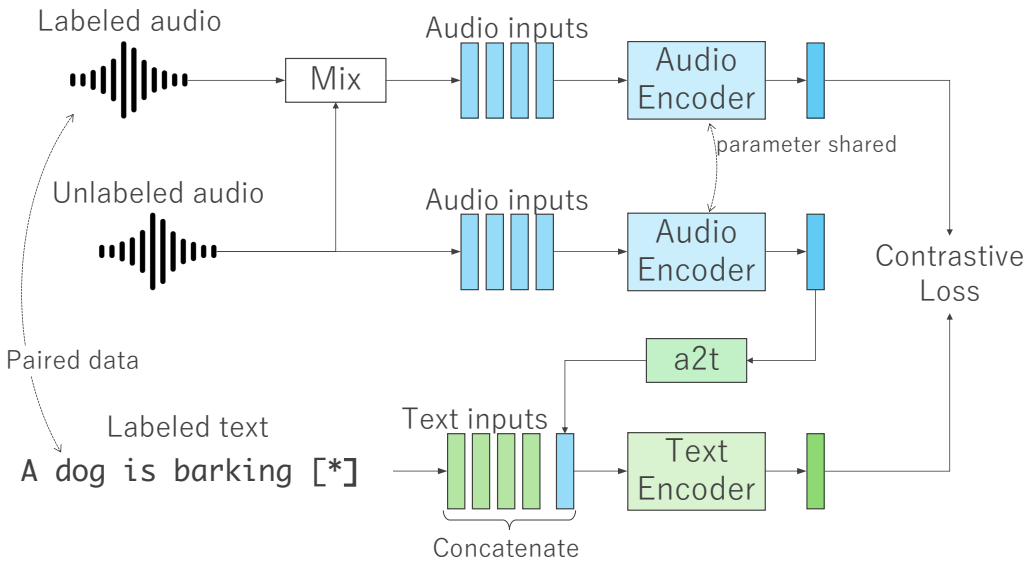

Figure 5.2: Proposed semi-supervised contrastive learning pipeline. A labeled audio clip $x^{(\mathsf{A})}$ and an unlabeled clip $x^{(\mathsf{A,Un})}$ are linearly mixed to obtain $x^{(\mathsf{A,Mix})}$. All audio inputs are processed by a shared audio encoder, producing embeddings $\mathbf{z}^{(\mathsf{A,Mix})}$ and $\mathbf{z}^{(\mathsf{A,Un})}$. The unlabeled embedding $\mathbf{z}^{(\mathsf{A,Un})}$ is transformed by an audio-to-text mapper (a2t) into a pseudo-token input, concatenated with the labeled text $x^{(\mathsf{T})}$, and encoded by the text encoder to yield $\mathbf{z}^{(\mathsf{T,Mix})}$. A contrastive loss is finally applied between the paired embeddings ($\mathbf{z}^{(\mathsf{A,Mix})}$, $\mathbf{z}^{(\mathsf{T,Mix})}$).

After converting to a time-frequency representation (e.g. a log-mel spectrogram), its embedding is obtained as

$$\mathbf{z}^{(\mathsf{A,Mix})} \;=\; \mathrm{AudioEncoder}\big(\mathbf{X}^{(\mathsf{A,Mix})}\big). \tag{5.7}$$

The unlabeled clip is embedded as $\mathbf{z}^{(\mathsf{A,Un})} = \mathrm{AudioEncoder}(\mathbf{X}^{(\mathsf{A,Un})})$ and passed to the a2t mapper f_{a2t}:

$$\hat{\mathbf{X}}^{(\mathsf{T,Un})} \;=\; f_{\mathrm{a2t}}\big(\mathbf{z}^{(\mathsf{A,Un})}\big), \tag{5.8}$$

where $\hat{\mathbf{X}}^{(\mathsf{T},\mathsf{Un})}$ lies in the text-token space. Section 5.4.2 details two variants (MLP-a2t, Transformer-a2t).

We concatenate the tokenized labeled text $\mathbf{X}^{(\mathsf{T})}$ with the pseudo input,

$$\mathbf{X}^{(\mathsf{T},\mathsf{Mix})} \;=\; \big[\mathbf{X}^{(\mathsf{T})};\, \hat{\mathbf{X}}^{(\mathsf{T},\mathsf{Un})}\big], \tag{5.9}$$

and encode it to obtain $\mathbf{z}^{(\mathsf{T},\mathsf{Mix})} = \text{TextEncoder}(\mathbf{X}^{(\mathsf{T},\mathsf{Mix})})$.

The mixed pair $\big(\mathbf{z}^{(\mathsf{A},\mathsf{Mix})}, \mathbf{z}^{(\mathsf{T},\mathsf{Mix})}\big)$ is contrasted against other pairs in the mini-batch using the InfoNCE loss:

$$\mathcal{L}_{\mathsf{Mix}} \;=\; \mathcal{L}_{\text{InfoNCE}}\big(\mathbf{z}^{(\mathsf{A},\mathsf{Mix})}, \mathbf{z}^{(\mathsf{T},\mathsf{Mix})}\big). \tag{5.10}$$

By augmenting labeled data with unlabeled clips in this way, the model sees a broader distribution of audio conditions. No human annotation is required for the additional examples, yet retrieval accuracy improves substantially.

5.4.2 Audio-to-Text Mapper (a2t)

The a2t module synthesizes a pseudo-token input from an audio embedding allowing unlabeled clips to participate in contrastive training. We investigate two variants (Figures 5.3 and 5.4): a simple MLP that emits a single vector and a Transformer-based decoder that produces a token sequence.

MLP-a2t (single vector output)

This variant employs the two-layer MLP mapper following our prior work [131] and originally inspired by Pic2Word [130], to convert each unlabeled audio embedding into a single-vector pseudo-token representation. It consists of two linear layers with ReLU

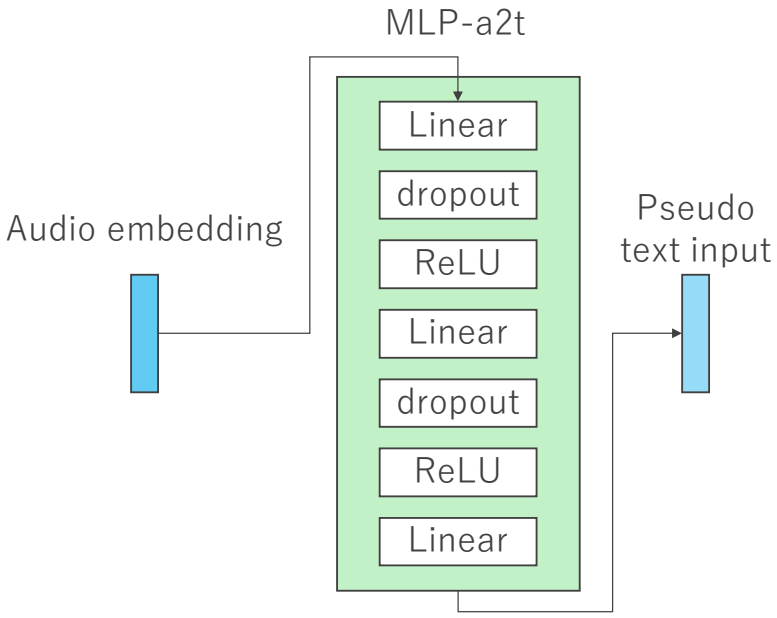

Figure 5.3: MLP-a2t. The global [CLS] audio embedding $\mathbf{z}^{(\mathsf{A})} \in \mathbb{R}^{D}$ is mapped to a single pseudo-token vector $\hat{\mathbf{X}}^{(\mathsf{T})} \in \mathbb{R}^{D_{\mathsf{T}}}$.

activation σ and dropout Drop:

$$\hat{\mathbf{X}}^{(\mathsf{T})} = W_2\,\sigma\big(\mathrm{Drop}\big(W_1\mathbf{z}^{(\mathsf{A})} + b_1\big)\big) + b_2, \tag{5.11}$$

where $W_1 \in \mathbb{R}^{H\times D}$, $W_2 \in \mathbb{R}^{D_{\mathsf{T}}\times H}$, H=1024 is the hidden dimension, and the dropout rate is set to 0.1. Because this mapper outputs a single vector, the resulting pseudo-token sequence has a fixed length of one.

Transformer-a2t (sequence output)

This variant employs a Transformer-based a2t mapper that converts frame-level audio embeddings into an L-token pseudo-token sequence.

Let $\mathbf{Z}^{(\mathsf{A,Un})} = [\,\mathbf{z}_1^{(\mathsf{A})}, \ldots, \mathbf{z}_{L_{\mathsf{A}}}^{(\mathsf{A})}\,] \in \mathbb{R}^{L_{\mathsf{A}}\times D_{\mathsf{A}}}$ denote the frame-level audio features output by the audio encoder before pooling. We introduce a trainable query matrix $\mathbf{Q} \in$

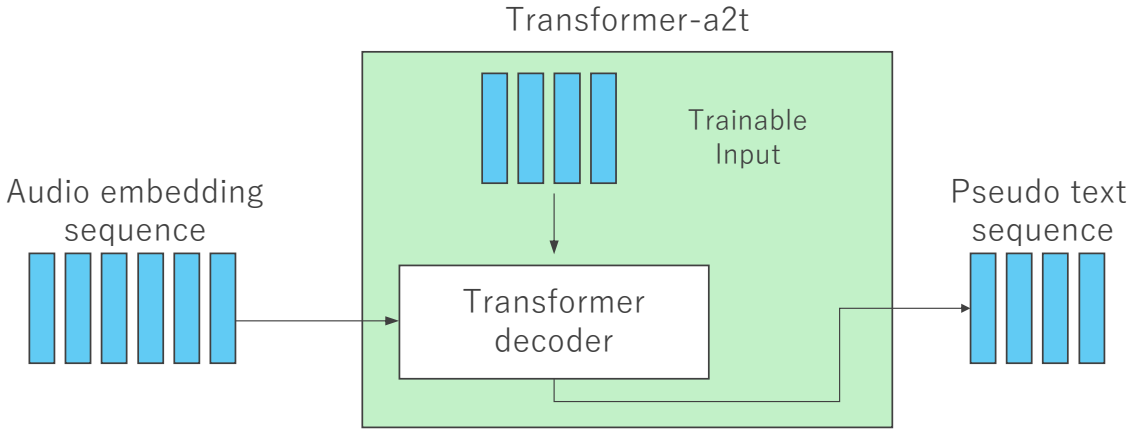

Figure 5.4: Transformer-a2t. A trainable query sequence $\mathbf{Q} \in \mathbb{R}^{L^{(\mathsf{Un})} \times D_{\mathsf{T}}}$ attends to frame-level audio embeddings $\mathbf{Z}^{(\mathsf{A},\mathsf{Un})} \in \mathbb{R}^{L_{\mathsf{A}} \times D_{\mathsf{A}}}$, producing a pseudo-token sequence $\hat{\mathbf{X}}^{(\mathsf{T})} \in \mathbb{R}^{L^{(\mathsf{Un})} \times D_{\mathsf{T}}}$.

$\mathbb{R}^{L^{(\mathsf{Un})} \times D_{\mathsf{T}}}$, where the sequence length $L^{(\mathsf{Un})}$ is a hyper-parameter. $\mathbf{Q}$ serves as a set of query vectors in the cross-attention of the Transformer-based a2t mapper and is updated end-to-end through backpropagation from the contrastive objective.

Unlike the MLP-based mapper, this learnable query sequence enables the model to generate a multi-token pseudo-token input. A stack of N Transformer decoder layers then performs cross-attention:

$$\hat{\mathbf{X}}^{(\mathsf{T})} = \mathrm{Decoder}_N\big(\mathbf{Q},\, \mathbf{Z}^{(\mathsf{A},\mathsf{Un})}\big) \tag{5.12}$$

We use N=2, eight attention heads, and a feed-forward network (FFN) width of $4D_{\mathsf{T}}$. Positional encodings are added to $\mathbf{Z}^{(\mathsf{A},\mathsf{Un})}$ to preserve temporal structure.

Although both mappers project an audio embedding into the text-token space, they operate at different levels of granularity. The MLP-a2t compresses an entire clip into a single vector ($L^{(\mathsf{Un})}$=1), which is adequate when the associated text is expected to describe a dominant foreground event (e.g., “a dog is barking”).

In contrast, Transformer-a2t produces a sequence of $L^{(\mathsf{Un})}$ learnable query tokens. Each token serves as a slot that attends to a distinct temporal or spectral pattern in the frame-level audio embeddings through cross-attention. This design enables the mapper to disentangle multiple acoustic events within the same clip and to encode them at different positions of the pseudo-token input. The empirical effect of these two granularities on retrieval performance is examined in our ablation experiments.

Noise-Perturbed Embedding

To regularize the mapper, we inject a small isotropic Gaussian perturbation into each unlabeled audio embedding before it is passed to the a2t module.

Given the audio embedding $\mathbf{z}^{(\mathsf{A},\mathsf{Un})} \in \mathbb{R}^{D_{\mathsf{A}}}$, we sample

$$\tilde{\mathbf{z}}^{(\mathsf{A},\mathsf{Un})} = \mathbf{z}^{(\mathsf{A},\mathsf{Un})} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\big(\mathbf{0},\, \sigma^2 \mathbf{I}\big), \tag{5.13}$$

and feed $\tilde{\mathbf{z}}^{(\mathsf{A},\mathsf{Un})}$ to Eq. (5.11) or Eq. (5.12). The noise scale σ is set to 1.0 in this study. The effect of the noise perturbation on the final performance is examined in the ablation study presented in Sec. 5.5.5.

5.4.3 Training Objective

The overall loss is the sum of three contrastive terms: (i) the labeled-pair loss in Eq. (5.5), (ii) the mixed-pair loss in Eq. (5.10), and (iii) a loss that aligns each unlabeled clip with its pseudo-token counterpart.

For every unlabeled waveform $\mathbf{x}^{(\mathsf{A},\mathsf{Un})}$ we obtain the audio embedding $\mathbf{z}^{(\mathsf{A},\mathsf{Un})} = \text{AudioEncoder}(\mathbf{x}^{(\mathsf{A},\mathsf{Un})})$, which the a2t mapper converts into a pseudo-token sequence $\hat{\mathbf{X}}^{(\mathsf{T},\mathsf{Un})}$. To inject a minimal natural-language prior, we prepend the fixed prompt ‘‘A

`sound of ''` and feed the result to the text encoder:

$$\mathbf{z}^{(\mathsf{T},\mathsf{Un})} = \text{TextEncoder}\big([\text{``}\texttt{A sound of }\text{''}; \hat{\mathbf{X}}^{(\mathsf{T},\mathsf{Un})}]\big), \tag{5.14}$$

where the prompt is tokenized in advance and concatenated in the same way as in (5.9). We then align the two views with an additional InfoNCE loss:

$$\mathcal{L}_{\text{unlabeled}} = \text{InfoNCE}\big(\mathbf{z}^{(\mathsf{A},\mathsf{Un})}, \mathbf{z}^{(\mathsf{T},\mathsf{Un})}\big). \tag{5.15}$$

The final objective combines all three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{labeled}} + \mathcal{L}_{\mathsf{Mix}} + \mathcal{L}_{\text{unlabeled}}. \tag{5.16}$$

All three losses are evaluated within the same mini-batch, and gradients are back-propagated through every module, so that the encoders and the a2t mapper are optimized jointly. The effect of Eq. (5.15) on the final performance is examined in the ablation study presented in Sec. 5.5.5.

5.5 Experimental Setup

We evaluate the proposed semi-supervised audio-text contrastive learning framework on audio-to-text retrieval tasks: Our experiments aim to assess (i) the effectiveness of incorporating unlabeled audio clips into the contrastive pipeline and (ii) the impact of the a2t architecture, including its sequence length $L^{(\mathsf{Un})}$ and the use of noise perturbation.

5.5.1 Datasets

We adopt two widely used captioning datasets, Clotho [5] and AudioCaps [54], as sources of labeled data, and AudioSet [9] as unlabeled audio. For each experiment,

we target either Clotho or AudioCaps. When the AudioCaps is the target, Clotho or AudioSet are treated as unlabeled clips; when Clotho is the target, {AudioCaps, AudioSet} are used as unlabeled clips.

5.5.2 Networks

All models fine-tune MS-CLAP (2023) [139], initializing its audio and text encoders with publicly available pretrained weights.

All network dimensions follow the MS-CLAP (2023) configuration. The shared embedding dimension of both audio and text encoders is d=512, and the text tokenizer (RoBERTa-base) uses an embedding dimension of D_{T}=768.

For the MLP-a2t mapper, both input and hidden layers operate in the same 512-dimensional space as CLAP, and the final linear layer outputs a 768-dimensional vector matching the text tokenizer.

The Transformer-a2t variant adopts the same hidden size as CLAP (D_{T}=768), with 8 attention heads and a feed-forward network width of $4D_{\mathsf{T}}$. Only the final output dimension is aligned to the text tokenizer size (768). $\mathbf{Q}$ in Eq. (5.12) is randomly initialized following the default initialization scheme of PyTorch linear layers. Transformer-a2t variants with query length $L^{(\mathsf{Un})} \in \{1, 2, 4\}$ are evaluated in Section 5.5.4.

We fine-tune for 30 epochs using Adam [140] with the scheduler-free optimizer [141]. The batch size is 256. All experiments are conducted on four NVIDIA A100 GPUs and complete within 1-2 hours each.

Evaluation metrics are Recall@1, Recall@5, Recall@10, and mean Average Precision at 10 (mAP@10) [142]. Higher scores indicate stronger alignment within the joint embedding space.

Table 5.1: **Retrieval: AudioCaps** AudioCaps supplies the labels; Clotho and/or AudioSet act as unlabeled audio. Other settings are the same as Table 5.2.

Method	labeled	unlabeled	R@1	R@5	R@10	mAP@10
Baseline methods						
CLAP [1] w/o finetune	-	-	15.79	46.21	62.94	28.67
CLAP [1] w/ finetune	AudioCaps	-	39.41	75.88	86.86	54.61
CLAP [1] w/ finetune	AudioCaps+Clotho		40.32	74.18	86.75	55.09
Proposed method						
MLP-a2t	AudioCaps	Clotho	36.01	75.76	87.43	52.57
MLP-a2t	AudioCaps	AudioSet	40.43	74.18	86.64	55.19
Transformer-a2t (length=2)	AudioCaps	Clotho	39.86	75.20	**88.45**	55.03
Transformer-a2t (length=2)	AudioCaps	AudioSet	**42.02**	**75.88**	88.00	**56.16**
Prior works with all labels						
AL-MixGen [103]	AudioCaps	Clotho	40.54	76.90	88.11	55.12
LAION-CLAP [2]			34.20	71.10	84.10	-
T-CLAP [57]			39.70	74.60	86.90	-

5.5.3 Overall retrieval performance

Table 5.1 presents the results when AudioCaps is used as the only labeled corpus, whereas Table 5.2 shows the symmetric setting in which Clotho serves as the target. In both scenarios, incorporating unlabeled clips enables our semi-supervised framework to consistently outperform the MS-CLAP baseline.

AudioCaps as the target. The best configuration, Transformer-a2t with $L^{(\mathsf{Un})}$=2 and unlabeled = AudioSet, achieves a 6.6% relative improvement in Recall@1 over the fine-tuned CLAP model, confirming that class-level AudioSet clips can be effec-

Table 5.2: **Retrieval: Clotho** Only Clotho provides (audio,text) pairs; AudioCaps and/or AudioSet are used without captions. Other settings are the same as Table 5.1.

Method	labeled	unlabeled	R@1	R@5	R@10	mAP@10
Baseline methods						
CLAP [1] w/o finetune	-	-	15.75	39.94	52.27	25.99
CLAP [1] w/ finetune	Clotho	-	19.71	45.55	59.62	30.80
CLAP [1] w/ finetune	Clotho+AudioCaps		19.52	45.75	59.64	31.00
Proposed method						
MLP-a2t	Clotho	AudioCaps	20.19	45.36	**59.71**	30.82
MLP-a2t	Clotho	AudioSet	**20.77**	45.36	58.66	31.43
Transformer-a2t (length=2)	Clotho	AudioCaps	20.19	45.93	59.52	31.22
Transformer-a2t (length=2)	Clotho	AudioSet	20.57	**46.60**	58.56	**31.79**
Prior works with all labels						
AL-MixGen [103]	Clotho	AudioCaps	21.34	46.99	59.33	32.52
LAION-CLAP [2]	-	-	15.30	38.40	51.20	-
T-CLAP [57]	-	-	17.30	39.90	53.00	-

tively exploited through a2t pairing. This configuration also surpasses a multi-corpus supervised baseline fine-tuned on both AudioCaps and Clotho.

Clotho as the target. For Clotho, the absolute gains are smaller yet still significant: Recall@1 improves by 4.4% relative to the fine-tuned CLAP baseline, again exceeding the multi-corpus supervised counterpart.

Comparison with larger models. Despite using far less labeled data, our semi-supervised model is competitive with AL-MixGen [103], which requires captions for every mixed clip, and even outperforms large open-domain models such as LAION-CLAP [2] and T-CLAP [57], both of which were pretrained on substantially larger

Table 5.3: **Impact of a2t design.** We compare an MLP-a2t (L=1) with Transformer-a2t at different query-sequence lengths L. All models are trained with target: Clotho and unlabeled: AudioSet. Higher values indicate better retrieval.

Method	Seq $L^{(\mathsf{Un})}$	AudioCaps retrieval				Clotho retrieval			
		R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
MLP-a2t	1	20.77	45.36	58.66	31.43	40.43	74.18	86.64	55.19
Transformer-a2t	1	**21.15**	46.03	**59.14**	31.77	41.56	73.39	87.09	55.33
Transformer-a2t	2	20.57	46.60	58.56	**31.79**	**42.02**	**75.88**	**88.00**	**56.16**
Transformer-a2t	4	20.19	**46.79**	57.89	31.34	41.22	74.75	86.52	55.71

datasets.

5.5.4 Effect of a2t architecture and sequence length

Table 5.3 examines how the pseudo-token length $L^{(\mathsf{Un})}$ affects performance. Replacing the MLP with a Transformer already improves results at $L^{(\mathsf{Un})}$=1, suggesting that cross-attention captures richer acoustic cues than simple projection. Extending the query sequence to $L^{(\mathsf{Un})} = 2$ yields the best overall scores on both datasets, whereas further increasing the length to $L^{(\mathsf{Un})} = 4$ provides no additional benefit and occasionally degrades performance, likely because excessively long pseudo texts deviate from the prompt template and add noise to the contrastive objective.

5.5.5 Ablation on noise and pseudo-token loss

Table 5.4 shows an ablation study on the noise perturbation and the unlabeled loss term in Eq. (5.15). Disabling the Gaussian perturbation reduces mAP@10 by roughly one point, indicating that the perturbation acts as a useful regularizer. Removing the

Table 5.4: **Contribution of regularisation components.** Starting from the full MLP-a2t model ($L^{(\mathsf{Un})}$=1, target: Clotho, unlabeled: AudioSet), we disable (i) noise perturbation of unlabeled embeddings and (ii) the unlabeled contrastive loss in Eq. (5.15). The drop in performance confirms the importance of each term.

Method	AudioCaps retrieval				Clotho retrieval			
	R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
MLP-a2t	20.77	45.36	58.66	31.43	40.43	74.18	86.64	55.19
w/o noise perturb.	20.38	42.68	56.46	30.27	39.98	74.97	87.88	54.86
w/o pseudo-loss	19.23	43.35	54.45	29.17	38.39	74.41	87.20	53.97

unlabeled loss in Eq. (5.15) has an even larger effect, eliminating most of the semi-supervised gain and reducing Recall@1 by about two points on both targets. These observations confirm that the unlabeled loss term is essential for binding each audio clip to its generated pseudo-token input and for preventing the mapper from drifting.

Note that prior work such as AL-MixGen [101] utilizes all available labels during training, whereas our method leverages unlabeled audio to supplement limited labeled data. This difference in data utilization should be considered when interpreting the performance comparison.

5.6 Conclusion

In this chapter, we addressed the issue of data scarcity in audio-text contrastive learning by proposing a Semi-Supervised Representation Learning framework. By introducing an Audio-to-Text (a2t) Mapper, we enabled unlabeled audio to effectively act as pseudo-tokens, allowing it to be integrated into the contrastive training loop. This

approach densifies the training data distribution and regularizes the shared embedding space without requiring expensive human annotation or error-prone LLM captioning. Experimental results demonstrated substantial improvements in retrieval performance, confirming that leveraging the vast scale of unlabeled audio is a viable pathway to broaden the generalization capability of audio-text models.

Chapter 6

Conclusion and Future Directions

6.1 Summary of Contributions

This thesis has addressed the fundamental challenges in audio-text representation learning, aiming to bridge the gap between human auditory perception and machine understanding. While the advent of contrastive frameworks like CLAP has revolutionized the field by establishing a shared latent space, conventional approaches have largely treated audio signals as global vectors trained on limited supervision. Consequently, they faced intrinsic limitations in capturing temporal granularity, intra-modal relations, and generalization to open-domain distributions.

The primary objective of this research was to transcend these limitations by refining the modeling architecture and learning paradigms. We tackled this challenge through three complementary approaches: enhancing alignment granularity (Chapter 3), explicating intra-modal relations (Chapter 4), and scaling via semi-supervised learning (Chapter 5). The specific contributions are summarized as follows:

1. **Fine-Grained Alignment for Temporal Complexity (Chapter 3):** We identified that the global pooling mechanism in conventional models inevitably discards

essential temporal dynamics. To resolve this, we proposed Aligned Contrastive Learning, a framework that explicitly achieves alignment between frame-level audio features and token-level text embeddings. Evaluating this method on text-to-music retrieval, a domain characterized by complex temporal evolution, demonstrated that preserving local granularity significantly improves the model's ability to distinguish semantically similar but structurally different audio scenes.

2. **Intra-Modal Relations via Difference Modeling (Chapter 4):** We addressed the limitation that standard cross-modal discriminative models fail to capture the compositional nature of audio. We proposed Audio Difference Learning, a paradigm that trains models to articulate the semantic difference between an input and a reference audio. By adopting a generative approach (Audio Captioning) with a masking-based cross-attention mechanism, we enabled the model to disentangle mixed acoustic events in the feature space. This work highlighted that learning relative representations enables semantic alignment and a compositional understanding of the audio modality.
3. **Implicit Alignment via Semi-Supervised Learning (Chapter 5):** Recognizing that architectural improvements are bounded by the scarcity of paired data, we introduced a semi-supervised framework that leverages vast amounts of unlabeled audio via an Audio-to-Text (a2t) Mapper. By projecting unlabeled audio into continuous pseudo-token embeddings, we enabled implicit alignment between unlabeled audio and the text modality within the contrastive training loop. This approach demonstrated that the "data scarcity wall" can be surmounted not merely by collecting more labels, but by bridging the modality gap at the feature level, densifying the training distribution effectively.

6.2 Discussion and Key Insights

Synthesizing the findings from the preceding chapters, we distill three key insights that transcend the specific methods and inform the future design of audio-language models. Each insight abstracts a core lesson from one chapter: fine-grained alignment (Chapter 3), difference learning (Chapter 4), and semi-supervised training with unlabeled audio (Chapter 5).

The success of Chapter 3 suggests that the prevailing one-vector-per-clip paradigm is insufficient for complex auditory scenes. Just as pixel-level alignment advanced computer vision, frame-level alignment is essential for deep audio understanding. Models must treat time not merely as a dimension to be collapsed, but as a structural component to be aligned with linguistic syntax. True understanding arises from fine-grained alignment of temporal events with semantic concepts.

Chapter 4 revealed a fundamental distinction between retrieval and generation. While retrieval models can exploit superficial correlations to match audio and text, generative tasks such as audio captioning, which requires producing a textual description from audio, force the model to understand the detailed representation of the content. Specifically, the ability to articulate differences implies that the model has learned a disentangled representation space where semantic arithmetic (e.g., $A - B = C$) is possible. This suggests that future representation learning should increasingly incorporate generative objectives to enforce understanding of differences between data samples.

Chapter 5 highlighted that architectural improvement alone cannot fully compensate for data sparsity. The audio-to-pseudo-text-token mapping aligns unlabeled audio with the text modality, effectively increasing the amount of training data without requiring additional annotations. For audio-text models to scale, semi-supervised or self-supervised signals must be integral components of the training objective.

6.3 Limitations

Despite the advancements presented, several limitations remain, pointing towards areas for further improvement.

Computational Complexity. The fine-grained alignment (Chapter 3) and cross-attention mechanisms (Chapter 4) introduce additional computational costs compared to simple global pooling. Scaling these methods to extremely long audio sequences or real-time applications remains a challenge, necessitating the exploration of efficient attention mechanisms or knowledge distillation techniques [143].

Reliance on Synthetic Mixtures. In Chapters 4 and 5, we relied on Mixup-based augmentation to create complex scenes or semi-supervised training signals. While effective, these linear mixtures do not fully capture the non-linear complexities of real-world acoustic environments, such as reverberation, source occlusion, and varying distances. Verifying these methods in "in-the-wild" recording conditions remains an open task.

Unified Approach. In this thesis, the three proposed methods, fine-grained alignment, difference learning, and semi-supervised training, were developed and evaluated in separate contexts to isolate their scientific contributions. A unified architecture that simultaneously incorporates all three elements has not yet been realized. Integrating these diverse objectives into a single foundation model without incurring training instability represents a non-trivial engineering challenge.

6.4 Future Directions

Based on the insights and limitations discussed in this thesis, we outline several promising directions for future research.

6.4.1 Towards Unified Audio-Language Foundation Models

A natural progression is to integrate the proposed methods into a unified audio-language foundation model. A model that combines fine-grained temporal understanding, disentangled structural representations, and large-scale training on unlabeled corpora would serve as a powerful backbone for diverse tasks, including captioning, retrieval, source separation, and audio generation. Recent audio-language models such as Pengi [144], LTU [145], SALMONN [146], Qwen2-Audio [147], and GAMA [148] have demonstrated the potential of large audio-language models for open-ended querying and reasoning. Incorporating fine-grained alignment, difference-based representations, and semi-supervised training into this emerging class of models is an important step toward truly unified audio-language foundation models.

6.4.2 Fine-Grained Text-to-Audio Generation

While this thesis focused primarily on understanding (audio-to-text), the learned representations also have strong potential for generation (text-to-audio). The fine-grained alignment introduced in Chapter 3 could enable controllable audio generation, where users specify the timing and evolution of sound events via natural language (e.g., “a dog barks at 3s, followed by a cuter dog barking at 5s”). State-of-the-art text-to-music systems such as MusicLM [149] and MusicGen [150] have achieved impressive generation quality, yet they still offer limited temporal and structural control. Bridging

fine-grained alignment with generative models is a promising direction for achieving temporally precise and semantically controllable audio generation.

6.4.3 Interactive Audio Editing

The difference modeling concept from Chapter 4 also opens the door to interactive audio editing. Instead of passive retrieval, users could interact with audio editing systems using differential prompts such as "remove the background noise" or "make the guitar sound more distorted". Developing such systems would represent a shift from static representation learning to dynamic audio manipulation, where models can apply and explain local edits in response to language instructions. Emerging approaches that use diffusion models for multimodal data augmentation [151] could further enhance the robustness and diversity of training data for interactive editing scenarios, especially when combined with difference- and commonality-based objectives.

6.5 Concluding Remarks

The field of audio-text learning is transitioning from simple tagging and retrieval towards a deep, structural understanding of acoustic scenes. This thesis has argued that achieving this depth requires moving beyond global, entangled, and fully supervised paradigms. By respecting temporal granularity, establishing fine-grained alignment, and embracing unlabeled data, we can build models that not only "hear" sound events but truly understand the rich information embedded within the audio signal.

Acknowledgements

This thesis would not have been possible without the guidance and support of many people who have contributed to my research throughout these years. I would like to take this opportunity to express my deepest gratitude to all of them.

First, I would like to express my sincere appreciation to Professor Kazuya Takeda of the Institutes of Innovation for Future Society, Nagoya University, Professor Tomoki Toda of the Information Technology Center, Nagoya University, and Professor Nobutaka Ono of Tokyo Metropolitan University, for kindly serving as examiners of this dissertation. I am deeply grateful that, despite their many commitments, they carefully read this thesis and provided me with many valuable comments and suggestions.

I am especially indebted to my supervisor, Professor Kazuya Takeda, for providing a free and supportive research environment in which I could fully devote myself to my work. Since my undergraduate years, he has consistently guided me and offered many insightful pieces of advice on how to think about various issues and how to frame problems from multiple perspectives. I vividly remember my first seminar presentation as an undergraduate, where I discussed the Levinson-Durbin algorithm with him. The joy of understanding not only the method itself but the ideas behind it remains with me to this day. The viewpoints and attitudes I learned from him have had a lasting influence not only on my research-related decisions but also on how I approach challenges in daily life, and they have become an essential part of my own way of thinking.

I am profoundly grateful for having been able to pursue my doctoral studies under his guidance and in such a stimulating environment.

I am also deeply grateful to Professor Tomoki Toda for his continuous guidance. He has always given me concrete and practical advice on all aspects of my research, ranging from the formulation of research topics to the structure of papers and the delivery of presentations. In regular seminars and individual meetings, he devoted substantial time to discussing ideas with me from their earliest stages, which helped me refine my thoughts and deepen the direction of my research through repeated trial and error. He also provided detailed comments and extensive support on my manuscripts and presentation materials, both in terms of content and expression, enabling me to shape my research outcomes into clearer and more convincing forms. It is thanks to his sustained guidance and warm encouragement that I have been able to continue my research to this day. I would like to express my heartfelt gratitude for his support over many years.

In particular, I wish to extend my sincere thanks to Professor Nobutaka Ono of Tokyo Metropolitan University, who carefully reviewed this dissertation. His concrete comments and suggestions were invaluable in refining the structure and arguments of this work. I have also had many opportunities to speak with him at conferences and to work alongside him in research communities including ASJ, DCASE, APSIPA, and AAAI. These experiences have greatly enriched my perspective as a researcher.

I am also deeply grateful to my colleagues, with whom I worked at LY Corporation, for their understanding and support regarding my decision to enter the doctoral program and for enabling me to balance my research with my professional duties. Even during busy periods at work, they provided opportunities for new challenges, which allowed me to continue both the research leading to this dissertation and my daily

responsibilities. I would like to express my sincere appreciation for providing such a rare environment in which I could pursue both my work and my development as a researcher. I am further indebted to Dr. Yusuke Fujita, Dr. Robin Scheibler, Dr. Taichi Nishimura, Mr. Hokuto Munakata, and Mr. Yuchi Ishikawa for their invaluable collaboration in joint research and co-authored papers. Their day-to-day discussions and advice have been a constant source of inspiration and new ideas, and their concrete support in the design and analysis of experiments greatly contributed to the development of the results reported in this dissertation. Beyond work, they have become close friends, and I treasure the time spent together.

I also thank Dr. Akihiko Sugiyama and Dr. Osamu Hoshuyama at NEC Corporation, who generously shared their time for countless discussions and advice during my early career. They taught me how to approach problems, identify their core, and communicate it clearly. These lessons have become the foundation of my entire career.

I would also like to express my deep gratitude to all the senior colleagues, peers, and junior members of the laboratories to which I have belonged. Through daily discussions and informal conversations, I gained a great deal of stimulation and learning. I am also grateful to Ms. Chika Ando, the laboratory secretary, who has supported me with administrative matters since my undergraduate years and always encouraged me with her cheerful presence. The time spent together in the laboratory community has made my life as a doctoral student immensely rich and fulfilling.

I am also grateful to all my co-authors and colleagues in the research community who have engaged in discussions with me over the years. These conversations have shaped my thinking and contributed to this work.

Finally, I would like to once again extend my profound gratitude to all those mentioned above for their generous support and encouragement.

References

[1] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," in *International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 1–5.

[2] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[3] H. Xie, S. Lipping, and T. Virtanen, "Language-based Audio Retrieval Task in DCASE 2022 Challenge," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2022, pp. 216–221.

[4] S. Lou, X. Xu, M. Wu, and K. Yu, "Audio-Text Retrieval in Context," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4793–4797.

[5] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An Audio Captioning Dataset," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 736–740.

[6] H. Xie and T. Virtanen, "Zero-Shot Audio Classification Based On Class Label Embeddings," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 264–267.

[7] H. Xie and T. Virtanen, "Zero-Shot Audio Classification Via Semantic Embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2021.

[8] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," in *Proc. International Conference on Machine Learning*, 2023, pp. 21 450–21 474.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 776–780.

[10] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM International Conference on Multimedia*, 2015, pp. 1015–1018.

[11] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Neural Information Processing Systems*, vol. 26, 2013, pp. 2121–2129.

[12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *Proc. International Conference on Machine Learning*, 2013, pp. 1247–1255.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning

Transferable Visual Models From Natural Language Supervision," in *Proc. International Conference on Machine Learning*, 2021, pp. 8748–8763.

[14] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning Robust Audio Representations from Clip," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4563–4567.

[15] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending Clip to Image, Text and Audio," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 976–980.

[16] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6787–6800.

[17] Y. Yang, J. Ma, S. Huang, L. Chen, X. Lin, G. Han, and S.-F. Chang, "TempCLR: Temporal alignment representation with contrastive learning," in *Proc. International Conference on Learning Representations*, 2023, 26 pages.

[18] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text," in *Proc. Neural Information Processing Systems*, vol. 34, 2021, pp. 24 206–24 221.

[19] I. Martín-Morató and A. Mesaros, "Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[20] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos, E. Quinton, G. Fazekas, and J. Nam, "The song describer dataset: A corpus of audio captions for music-and-language evaluation," *arXiv preprint arXiv:2311.10057*, 13 pages, 2023.

[21] S. Ghosh, A. Seth, S. Kumar, U. Tyagi, C. K. R. Evuru, R. S, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models," in *Proc. International Conference on Learning Representations*, 2023, 26 pages.

[22] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Neural Information Processing Systems*, vol. 35, 2022, pp. 25 278–25 294.

[23] 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," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 3339–3354, 2024.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. International Conference on Learning Representations*, 2020, 21 pages.

[25] J. Wu, W. Li, Z. Novack, A. Namburi, C. Chen, and J. McAuley, "CoLLAP: Contrastive Long-form Language-Audio Pretraining with Musical Temporal Struc-

ture Augmentation," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[26] I. Tsiamas, S. Pascual, C. Yeh, and J. Serrà, "Sequential Contrastive Audio-Visual Learning," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[27] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 13 pages, 2019.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conference of the North American Chapter of the ACL*, 2019, pp. 4171–4186.

[30] S. Goel, H. Bansal, S. Bhatia, R. Rossi, V. Vinay, and A. Grover, "CyCLIP: Cyclic Contrastive Language-Image Pretraining," in *Proc. Neural Information Processing Systems*, vol. 35, 2022, pp. 6704–6719.

[31] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On Mutual Information Maximization for Representation Learning," in *Proc. International Conference on Learning Representations*, 2020, 16 pages.

[32] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 6827–6839.

[33] S. Purushwalkam and A. Gupta, "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases," in *Proc. Neural Information Processing Systems*, vol. 33, 2020, pp. 3407–3418.

[34] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "FILIP: Fine-grained Interactive Language-Image Pre-Training," in *Proc. International Conference on Learning Representations*, 2021, 21 pages.

[35] A. S. Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio Retrieval With Natural Language Queries: A Benchmark Study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2023.

[36] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Proc. Neural Information Processing Systems*, vol. 35, 2022, pp. 17 612–17 625.

[37] *.HOTELLING*, "RELATIONS BETWEEN TWO SETS OF VARIATES*," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[38] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep Generalized Canonical Correlation Analysis," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 2019, pp. 1–6.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Proc. Neural Information Processing Systems*, vol. 26, 2013, pp. 3111–3119.

[40] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives," in *Proc. British Machine Vision Conference*, 2018, p. 12.

[41] K. Desai and J. Johnson, "VirTex: Learning Visual Representations from Textual Annotations," in *Proc. Computer Vision and Pattern Recognition*, 2021, pp. 11 157–11 168.

[42] M. B. Sariyildiz, J. Perez, and D. Larlus, "Learning Visual Representations with Caption Annotations," in *Proc. European Conference on Computer Vision*, 2020, pp. 153–170.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[45] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in *Proc. International Conference on Machine Learning*, 2021, pp. 4904–4916.

[46] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One Embedding Space To Bind Them All," in *Proc. Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.

[47] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, W. HongFa, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. W. Zhang, Z. Li, W. Liu, and L. Yuan, "LanguageBind: Ex-

tending Video-Language Pretraining to N-modality by Language-based Semantic Alignment," in *Proc. International Conference on Learning Representations*, 2023, 21 pages.

[48] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNet: Environmental Sound Classification Based on Visual Domain Models," in *Proc. International Conference on Pattern Recognition*, 2021, pp. 4933–4940.

[49] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[50] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[51] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 646–650.

[52] 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," *arXiv preprint arXiv:1907.11692*, 13 pages, 2019.

[53] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proc. ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[54] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating Captions for Audios in The Wild," in *Proc. Conference of the North American Chapter of the ACL*, 2019, pp. 119–132.

[55] L. Sun, X. Xu, M. Wu, and W. Xie, "Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning," in *Proc. ACM International Conference on Multimedia*, 2024, pp. 5025–5034.

[56] J. Bai, H. Liu, M. Wang, D. Shi, W. Wang, M. D. Plumbley, W.-S. Gan, and J. Chen, "AudioSetCaps: Enriched Audio Captioning Dataset Generation Using Large Audio Language Models," in *Proc. Neural Information Processing Systems*, 2024, 8 pages.

[57] Y. Yuan, Z. Chen, X. Liu, H. Liu, X. Xu, D. Jia, Y. Chen, M. D. Plumbley, and W. Wang, "T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, 2024, pp. 1–6.

[58] Y. Li, Z. Guo, X. Wang, and H. Liu, "Advancing Multi-grained Alignment for Contrastive Language-Audio Pre-training," in *Proc. ACM International Conference on Multimedia*, 2024, pp. 7356–7365.

[59] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 12 449–12 460.

[60] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked

Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, 2021.

[61] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass, "Contrastive Audio-Visual Masked Autoencoder," in *Proc. International Conference on Learning Representations*, 2022, 29 pages.

[62] T. Wang and P. Isola, "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere," in *Proc. International Conference on Machine Learning*, 2020, pp. 9929–9939.

[63] M. Slaney, "Semantic-audio retrieval," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. IV-4108-IV–4111.

[64] A.-M. Oncescu, A. S. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio Retrieval with Natural Language Queries," in *Proc. Interspeech*, 2021, pp. 2411–2415.

[65] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proc. ACM International Conference on Multimedia*, 2008, pp. 105–112.

[66] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic Annotation and Retrieval of Music and Sound Effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[67] B. Elizalde, S. Zarar, and B. Raj, "Cross Modal Audio Search and Retrieval with Joint Embeddings Based on Text and Audio," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 4095–4099.

[68] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 39–48.

[69] M. Hamilton, A. Zisserman, J. R. Hershey, and W. T. Freeman, "Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Sound and Language," in *Proc. Computer Vision and Pattern Recognition*, 2024, pp. 13 117–13 127.

[70] E. Araujo, A. Rouditchenko, Y. Gong, S. Bhati, S. Thomas, B. Kingsbury, L. Karlinsky, R. Feris, J. R. Glass, and H. Kuehne, "CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grained Alignment," in *Proc. Computer Vision and Pattern Recognition*, 2025, pp. 18 794–18 803.

[71] Y. Xin, Z. Zhu, X. Cheng, X. Yang, and Y. Zou, "Audio-text Retrieval with Transformer-based Hierarchical Alignment and Disentangled Cross-modal Representation," in *Proc. Interspeech*, 2024, pp. 1140–1144.

[72] Y. Xie, Z. Zhu, X. Zhuang, L. Liang, Z. Wang, and Y. Zou, "GPA: Global and Prototype Alignment for Audio-Text Retrieval," in *Proc. Interspeech*, 2024, pp. 5078–5082.

[73] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot Learning for Audio-based Music Classification and Tagging," in *Proc. International Society for Music Information Retrieval Conference*, 2019, pp. 67–74.

[74] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal Metric Learning for Tag-Based Music Retrieval," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 591–595.

[75] T. Chen, Y. Xie, S. Zhang, S. Huang, H. Zhou, and J. Li, "Learning Music Sequence Representation From Text Supervision," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4583–4587.

[76] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive Audio-Language Learning for Music," in *Proc. International Society for Music Information Retrieval Conference*, 2022, pp. 640–649.

[77] S. Doh, M. Won, K. Choi, and J. Nam, "Toward Universal Text-To-Music Retrieval," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[78] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic Tagging Using Deep Convolutional Neural Networks," in *Proc. International Society for Music Information Retrieval Conference*, 2016, pp. 805–811.

[79] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Metric learning vs classification for disentangled music representation learning," in *Proc. International Society for Music Information Retrieval Conference*, 2020, pp. 439–445.

[80] M. Müller, F. Kurth, D. Damm, C. Fremerey, and M. Clausen, "Lyrics-Based Audio Retrieval and Multimodal Navigation in Music Collections," in *Research and Advanced Technology for Digital Libraries*, 2007, pp. 112–123.

[81] K. Watanabe and M. Goto, "Query-by-Blending: A Music Exploration System Blending Latent Vector Representations of Lyric Word, Song Audio, and Artist," in *Proc. International Society for Music Information Retrieval Conference*, 2019, pp. 144–151.

[82] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A Joint Embedding of Music Audio and Natural Language," in *Proc. International Society for Music Information Retrieval Conference*, 2022, pp. 559–566.

[83] S. Wu, D. Yu, X. Tan, and M. Sun, "CLaMP: Contrastive Language-Music Pre-Training for Cross-Modal Symbolic Music Information Retrieval," in *Proc. International Society for Music Information Retrieval Conference*, 2023, pp. 157–165.

[84] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proc. International Society for Music Information Retrieval Conference*, 2011, pp. 591–596.

[85] A. Schindler and P. Knees, "Multi-Task Music Representation Learning from Multi-Label Embeddings," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019, pp. 1–6.

[86] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Learning Music Audio Representations Via Weak Language Supervision," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 456–460.

[87] M. Won, K. Choi, and X. Serra, "Semi-supervised Music Tagging Transformer," in *Proc. International Society for Music Information Retrieval Conference*, 2021, pp. 769–776.

[88] M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra, "Emotion Embedding Spaces for Matching Music to Stories," in *Proc. International Society for Music Information Retrieval Conference*, 2021, pp. 777–785.

[89] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, pp. 374–378.

[90] M. Wu, H. Dinkel, and K. Yu, "Audio Caption: Listen and Tell," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 830–834.

[91] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio Captioning Transformer," in *Proc. Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, 2021, pp. 211–215.

[92] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to Describe Differences Between Pairs of Similar Images," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4024–4034.

[93] D. H. Park, T. Darrell, and A. Rohrbach, "Robust Change Captioning," in *Proc. International Conference on Computer Vision*, 2019, pp. 4624–4633.

[94] M. Forbes, C. Kaeser-Chen, P. Sharma, and S. Belongie, "Neural Naturalist: Generating Fine-Grained Image Comparisons," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 708–717.

[95] H. Kim, J. Kim, H. Lee, H. Park, and G. Kim, "Viewpoint-Agnostic Change Captioning with Cycle Consistency," in *Proc. International Conference on Computer Vision*, 2021, pp. 2075–2084.

[96] X. Shi, X. Yang, J. Gu, S. Joty, and J. Cai, "Finding It at Another Side: A Viewpoint-Adapted Matching Encoder for Change Captioning," in *Proc. European Conference on Computer Vision*, 2020, pp. 574–590.

[97] L. Yao, W. Wang, and Q. Jin, "Image Difference Captioning with Pre-training and Contrastive Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3108–3116, 2022.

[98] Z. Guo, T.-J. Wang, and J. Laaksonen, "CLIP4IDC: CLIP for Image Difference Captioning," in *Proc. Conference of the Asia-Pacific Chapter of the ACL*, 2022, pp. 33–42.

[99] L. Dunlap, Y. Zhang, X. Wang, R. Zhong, T. Darrell, J. Steinhardt, J. E. Gonzalez, and S. Yeung-Levy, "Describing Differences in Image Sets with Natural Language," in *Proc. Computer Vision and Pattern Recognition*, 2024, pp. 24 199–24 208.

[100] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, and M. Li, "MixGen: A New Multi-Modal Data Augmentation," in *Proc. Winter Conference on Applications of Computer Vision*, 2023, pp. 379–389.

[101] E. Kim, J. Kim, Y. Oh, K. Kim, M. Park, J. Sim, J. Lee, and K. Lee, "Exploring train and test-time augmentations for audio-language learning," *arXiv preprint arXiv:2210.17143*, 5 pages, 2023.

[102] J.-H. Cho, Y.-A. Park, J. Kim, and J.-H. Chang, "Hyu submission for the dcase 2023 task 6a: Automated audio captioning model using al-mixgen and synonyms substitution," Tech. Rep., 2023, pp. 1–4.

[103] J. Kim, Y.-A. Park, J.-H. Cho, and J.-H. Chang, "Improving automated audio captioning fluency through data augmentation and ensemble selection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2023 Workshop*, 2023, pp. 86–90.

[104] S. T. Y. Kawaguchi, T. Nishida, K. Imoto, Y. Okamoto, K. Dohi, and T. Endo, "Audio-change captioning to explain machine-sound anomalies," in *Proc. Detection and Classification of Acoustic Scenes and Events 2023 Workshop*, 2023, pp. 201–205.

[105] D. Takeuchi, Y. Ohishi, D. Niizumi, N. Harada, and K. Kashino, "AUDIO DIFFERENCE CAPTIONING UTILIZING SIMILARITY-DISCREPANCY DISENTANGLEMENT," in *Proc. Detection and Classification of Acoustic Scenes and Events 2023 Workshop*, 2023, pp. 181–185.

[106] Y. Jia, X. Zhang, Y. Guo, Y. Chen, and S. Zhao, "From contrast to commonality: Audio commonality captioning for enhanced audio-text cross-modal understanding in multimodal llms," *arXiv preprint arXiv:2508.01659*, 5 pages, 2025.

[107] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, "Automated audio captioning: An overview of recent progress and new challenges," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 26, 2022.

[108] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Proc. Neural Information Processing Systems*, vol. 27, 2014, pp. 3104–3112.

[109] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Diverse Audio Captioning Via Adversarial Training," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8882–8886.

[110] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio Pre-Training with Acoustic Tokenizers," in *Proc. International Conference on Machine Learning*, 2023, pp. 5178–5193.

[111] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[112] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[113] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231.

[114] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[115] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-Based Image Description Evaluation," in *Proc. Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[116] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *Proc. European Conference on Computer Vision*, 2016, pp. 382–398.

[117] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient Optimization of SPIDEr," in *Proc. International Conference on Computer Vision*, 2017, pp. 873–881.

[118] H.-H. Wu, O. Nieto, J. P. Bello, and J. Salamon, "Audio-Text Models Do Not Yet Leverage Natural Language," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[119] Z. Xie, X. Xu, M. Wu, and K. Yu, "Enhance Temporal Relations in Audio Captioning with Sound Event Detection," in *Proc. Interspeech*, 2023, pp. 4179–4183.

[120] P. Primus, K. Koutini, and G. Widmer, "ADVANCING NATURAL-LANGUAGE BASED AUDIO RETRIEVAL WITH PASST AND LARGE AUDIO-CAPTION DATA SETS," in *Proc. Detection and Classification of Acoustic Scenes and Events 2023 Workshop*, 2023, pp. 151–155.

[121] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, 2024.

[122] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "The curse of recursion: Training on generated data makes models forget," *arXiv preprint arXiv:2305.17493*, 18 pages, 2024.

[123] S. Wyllie, I. Shumailov, and N. Papernot, "Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias," in *Proc. ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 2113–2147.

[124] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Training strategy of massive text-to-audio models and gpt-based query-augmentation," Tech. Rep., 2024.

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

[126] Y. Zhang, X. Xu, R. Du, H. Liu, Y. Dong, Z.-H. Tan, W. Wang, and Z. Ma, "Zero-Shot Audio Captioning Using Soft and Hard Prompts," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2045–2058, 2025.

[127] S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang, "Training Audio Captioning Models without Audio," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 371–375.

[128] K. Saijo, J. Ebbers, F. G. Germain, S. Khurana, G. Wiehern, and J. L. Roux, "Leveraging Audio-Only Data for Text-Queried Target Sound Extraction," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[129] W. Wang, R. Hou, H. Chang, S. Shan, and X. Chen, "MATS: An Audio Language Model under Text-only Supervision," in *Proc. International Conference on Machine Learning*, 2025, 20 pages.

[130] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2Word: Mapping Pictures to Words for Zero-Shot Composed Image Retrieval," in *Proc. Computer Vision and Pattern Recognition*, 2023, pp. 19 305–19 314.

[131] T. Komatsu, H. Munakata, and Y. Ishikawa, "Leveraging Unlabeled Audio for Audio-Text Contrastive Learning via Audio-Composed Text Features," in *Proc. Interspeech*, 2025, pp. 2600–2604.

[132] V. Atliha and D. Šešok, "Text Augmentation Using BERT for Image Captioning.," *Applied Sciences (2076-3417)*, vol. 10, no. 17, p. 5978, 2020.

[133] I. R. Turkerud and O. J. Mengshoel, "Image Captioning using Deep Learning: Text Augmentation by Paraphrasing via Backtranslation," in *Proc. IEEE Symposium Series on Computational Intelligence*, 2021, pp. 01–10.

[134] S. Li, B. Yang, and Y. Zou, "Utilizing Text-based Augmentation to Enhance Video Captioning," in *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2022, pp. 287–293.

[135] X. Li, W. Chen, Z. Ma, X. Xu, Y. Liang, Z. Zheng, Q. Kong, and X. Chen, "DRCap: Decoding CLAP Latents with Retrieval-Augmented Generation for Zero-shot Audio Captioning," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[136] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation," in *Proc. Interspeech*, 2024, pp. 57–61.

[137] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models," *ACM Trans. Graph.*, vol. 42, no. 4, 150:1–150:13, 2023.

[138] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in *Proc. Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[139] B. Elizalde, S. Deshmukh, and H. Wang, "Natural Language Supervision For General-Purpose Audio Representations," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 336–340.

[140] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. International Conference on Learning Representations*, 2015, 15 pages.

[141] A. Defazio, X. (Yang, H. Mehta, K. Mishchenko, A. Khaled, and A. Cutkosky, "The road less scheduled," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, vol. 37, 2024, pp. 9974–10 007.

[142] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. 2008.

[143] F. Paissan and E. Farella, "tinyCLAP: Distilling Constrastive Language-Audio Pretrained Models," in *Proc. Interspeech*, 2024, pp. 1685–1689.

[144] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An Audio Language Model for Audio Tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[145] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, "Listen, Think, and Understand," in *Proc. International Conference on Learning Representations*, 2023, 30 pages.

[146] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," in *Proc. International Conference on Learning Representations*, 2023, 23 pages.

[147] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 16 pages, 2024.

[148] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 6288–6313.

[149] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 15 pages, 2023.

[150] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Defossez, "Simple and Controllable Music Generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 704–47 720, 2023.

[151] C. Xiao, S. X. Xu, and K. Zhang, "Multimodal Data Augmentation for Image Captioning using Diffusion Models," in *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, 2023, pp. 23–33.

List of Publications

Journal Papers

[1] **T. Komatsu**, H. Munakata, Y. Ishikawa, K. Takeda, and T. Toda, "Semi-supervised text-audio contrastive learning method using pseudo-text input," APSIPA Transactions on Signal and Information Processing, 22 pages, 2026.

[2] **T. Komatsu**, K. Takeda, and T. Toda, "Audio Difference Learning Framework for Audio Captioning," APSIPA Transactions on Signal and Information Processing, vol. 14, no. 1, pp. 1–21, Nov. 2025, ISSN 2048-7703, DOI: 10.1561/116.20250021.

[3] M. Kato, A. Sugiyama, and **T. Komatsu**, "A stereo wind-noise suppressor with null beamforming and frequency-domain noise averaging," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 101, no. 10, pp. 1631–1637, 2018.

Peer-Reviewed Conference Papers

[1] T. Hasumi, **T. Komatsu**, and Y. Fujita, "Music tagging with classifier group chains," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, Apr. 2025, pp. 1–5.

[2] Y. Ishikawa, **T. Komatsu**, and Y. Aoki, "Pre-training with synthetic patterns for audio," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, Apr. 2025, pp. 1–5.

[3] Y. Ishikawa, S. Nakada, H. Munakata, K. Saito, **T. Komatsu**, and Y. Aoki, "Language-guided contrastive audio-visual masked autoencoder with automatically generated audio-visual-text triplets from videos," in Proc. INTERSPEECH, Rotterdam, Netherlands, Aug. 2025, pp. 2605–2609, DOI: 10.21437/Interspeech.2025-1054.

[4] **T. Komatsu**, H. Munakata, T. Hasumi, and Y. Fujita, "Aligned contrastive learning for text-to-music retrieval," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, Apr. 2025, pp. 1–5.

[5] **T. Komatsu**, H. Munakata, and Y. Ishikawa, "Leveraging unlabeled audio for audio-text contrastive learning via audio-composed text features," in Proc. INTERSPEECH, Rotterdam, Netherlands, Aug. 2025, pp. 2600–2604.

[6] H. Munakata, T. Nishimura, S. Nakada, and **T. Komatsu**, "Language-based audio moment retrieval," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, Apr. 2025, pp. 1–5.

[7] S. Nakada, T. Nishimura, H. Munakata, M. Kondo, and **T. Komatsu**, "Deteclap: Enhancing audio-visual representation learning with object information," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, Apr. 2025, pp. 1–5.

[8] Y. Fujita and **T. Komatsu**, "Audio fingerprinting with holographic reduced representations," in Proc. INTERSPEECH, Kos, Greece, Sep. 2024, pp. 62–66, DOI: 10.21437/Interspeech.2024-245.

[9] M. Hentschel, Y. Nishikawa, **T. Komatsu**, and Y. Fujita, "Keep decoding parallel with effective knowledge distillation from language models to end-to-end speech recognisers," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, Apr. 2024, pp. 10876–10880.

[10] **T. Komatsu**, Y. Fujita, K. Takeda, and T. Toda, "Audio difference learning for audio captioning," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, Apr. 2024, pp. 1456–1460.

[11] H. Munakata, T. Nishimura, S. Nakada, and **T. Komatsu**, "Pre-trained models, datasets, data augmentation for language-based audio retrieval," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Tokyo, Japan, Oct. 2024, pp. 86–90.

[12] T. Nishimura, S. Nakada, H. Munakata, and **T. Komatsu**, "Lighthouse: A user-friendly library for reproducible video moment retrieval and highlight detection," in Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), Miami, Florida, USA, Nov. 2024, pp. 53–60, DOI: 10.18653/v1/2024.emnlp-demo.6.

[13] R. Scheibler, Y. Fujita, Y. Shirahata, and **T. Komatsu**, "Universal score-based speech enhancement with high content preservation," in Proc. INTERSPEECH, Kos, Greece, Sep. 2024, pp. 1165–1169, DOI: 10.21437/Interspeech.2024-138.

[14] R. Shimizu et al., "Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, Apr. 2024, pp. 12672–12676.

[15] Y. Fujita, **T. Komatsu**, and Y. Kida, "Alternate intermediate conditioning with syllable-level and character-level targets for Japanese ASR," in Proc. IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, Jan. 2023, pp. 76–83.

[16] Y. Fujita, **T. Komatsu**, R. Scheibler, Y. Kida, and T. Ogawa, "Neural diarization with non-autoregressive intermediate attractors," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[17] A. Ito, **T. Komatsu**, and Y. Fujita, "Target vocabulary recognition based on multi-task learning with decomposed teacher sequences," in Proc. INTERSPEECH, Dublin, Ireland, Aug. 2023, pp. 1254–1258.

[18] **T. Komatsu** and Y. Fujita, "Interdecoder: Using attention decoders as intermediate regularization for CTC-based speech recognition," in Proc. IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, Jan. 2023, pp. 46–51.

[19] R. Scheibler, T. Hasumi, Y. Fujita, **T. Komatsu**, R. Yamamoto, and K. Tachibana, "Foley sound synthesis with a class-conditioned latent diffusion model," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Sep. 2023, pp. 156–160.

[20] A. Igarashi, K. Imoto, Y. Komatsu, S. Tsubaki, S. Hario, and **T. Komatsu**, "How information on acoustic scenes and sound events mutually benefits event detection

and scene classification tasks," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, Nov. 2022, pp. 7–11.

[21] **T. Komatsu**, "Non-autoregressive ASR with self-conditioned folded encoders," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, May 2022, pp. 7427–7431.

[22] **T. Komatsu**, Y. Fujita, J. Lee, L. Lee, S. Watanabe, and Y. Kida, "Better intermediates improve CTC inference," in Proc. INTERSPEECH, Incheon, South Korea, Sep. 2022, pp. 4965–4969, DOI: 10.21437/Interspeech.2022-11276.

[23] I. Kuroyanagi and **T. Komatsu**, "Self-supervised learning method using multiple sampling strategies for general-purpose audio representation," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, May 2022, pp. 3263–3267.

[24] Y. Nakagome, **T. Komatsu**, Y. Fujita, S. Ichimura, and Y. Kida, "Interaug: Augmenting noisy intermediate predictions for CTC-based ASR," in Proc. INTERSPEECH, Incheon, South Korea, Sep. 2022, pp. 5140–5144, DOI: 10.21437/Interspeech.2022-11284.

[25] R. Scheibler, **T. Komatsu**, Y. Fujita, and M. Hentschel, "On sorting and padding multiple targets for sound event localization and detection with permutation invariant and location-based training," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, Nov. 2022, pp. 1–6.

[26] R. Scheibler, **T. Komatsu**, Y. Fujita, and M. Hentschel, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel separation network," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Nancy, France, Nov. 2022, pp. 176–180.

[27] Y. Higuchi et al., "A comparative study on non-autoregressive modelings for speech-to-text generation," in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, Dec. 2021, pp. 47–54.

[28] **T. Komatsu**, T. Matsui, and J. Gao, "Multi-source domain adaptation with Sinkhorn barycenter," in Proc. European Signal Processing Conference (EUSIPCO), Dublin, Ireland, Aug. 2021, pp. 1371–1375.

[29] **T. Komatsu** and R. Scheibler, "Comparison of low complexity self-attention mechanisms for acoustic event detection," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, Dec. 2021, pp. 1139–1143.

[30] **T. Komatsu**, M. Togami, and T. Takahashi, "Sound event localization and detection using convolutional recurrent neural networks and gated linear units," in Proc. European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands (Virtual), Aug. 2021, pp. 41–45.

[31] **T. Komatsu**, S. Watanabe, K. Miyazaki, and T. Hayashi, "Acoustic event detection with classifier chains," in Proc. INTERSPEECH, Brno, Czech Republic (Virtual), Aug. 2021, pp. 601–605, DOI: 10.21437/Interspeech.2021-2218.

[32] J. Nozaki and **T. Komatsu**, "Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions," in Proc.

INTERSPEECH, Brno, Czech Republic (Virtual), Aug. 2021, pp. 3735–3739, DOI: 10.21437/Interspeech.2021-911.

[33] R. Scheibler, **T. Komatsu**, and M. Togami, "Multichannel separation and classification of sound events," in Proc. European Signal Processing Conference (EUSIPCO), Dublin, Ireland, Aug. 2021, pp. 1035–1039.

[34] S. Takeyama, **T. Komatsu**, K. Miyazaki, M. Togami, and S. Ono, "Robust acoustic scene classification to multiple devices using maximum classifier discrepancy and knowledge distillation," in Proc. European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands (Virtual), Aug. 2021, pp. 36–40.

[35] D. Xin, **T. Komatsu**, S. Takamichi, and H. Saruwatari, "Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual TTS," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada (Virtual), Jun. 2021, pp. 6608–6612.

[36] **T. Komatsu**, K. Imoto, and M. Togami, "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain (Virtual), May 2020, pp. 646–650.

[37] Y. Masuyama, M. Togami, and **T. Komatsu**, "Consistency-aware multi-channel speech enhancement using deep neural networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain (Virtual), May 2020, pp. 821–825.

[38] K. Miyazaki, **T. Komatsu**, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Tokyo, Japan, Nov. 2020, pp. 100–104.

[39] K. Miyazaki, **T. Komatsu**, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Weakly-supervised sound event detection with self-attention," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain (Virtual), May 2020, pp. 66–70.

[40] T. Takahashi, **T. Komatsu**, and K. Yamada, "Disentangling clustered representations of variational autoencoders for generating diverse samples," in Proc. IJCAI-PRICAI Workshop on Learning Data Representation for Clustering (LDRC), Kyoto, Japan, Jul. 2020, 8 pages.

[41] T. Takahashi, S. Takagi, H. Ono, and **T. Komatsu**, "Differentially private variational autoencoders with term-wise gradient aggregation," in Proc. Theory and Practice of Differential Privacy Workshop (TPDP), Virtual (with CCS), Nov. 2020, 8 pages.

[42] M. Togami, Y. Masuyama, **T. Komatsu**, and Y. Nakagome, "Unsupervised training for deep speech source separation with Kullback-Leibler divergence based probabilistic loss function," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain (Virtual), May 2020, pp. 56–60.

[43] M. Togami, Y. Masuyama, **T. Komatsu**, K. Yoshii, and T. Kawahara, "Computer-resource-aware deep speech separation with a run-time-specified number of BLSTM layers," in Proc. Asia-Pacific Signal and Information Processing As-

sociation Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand (Virtual), Dec. 2020, pp. 788–793.

[44] **T. Komatsu**, T. Hayashiy, R. Kondo, T. Todaz, and K. Takeday, "Scene-dependent anomalous acoustic-event detection based on conditional WaveNet and i-vector," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 2019, pp. 870–874.

[45] Y. Masuyama, M. Togami, and **T. Komatsu**, "Multichannel loss function for supervised speech source separation by mask-based beamforming," in Proc. INTERSPEECH, Graz, Austria, Sep. 2019, pp. 2708–2712, DOI: 10.21437/Interspeech.2019-1289.

[46] C. Narisetty, **T. Komatsu**, and R. Kondo, "Bayesian non-parametric multi-source modelling based determined blind source separation," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 2019, pp. 111–115.

[47] M. Togami and **T. Komatsu**, "Fast convergence algorithm for state-space model based speech dereverberation by multi-channel non-negative matrix factorization," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, Oct. 2019, pp. 244–248.

[48] M. Togami and **T. Komatsu**, "Variational Bayesian multi-channel speech dereverberation under noisy environments with probabilistic convolutive transfer function," in Proc. INTERSPEECH, Graz, Austria, Sep. 2019, pp. 106–110.

[49] T. Hayashi, **T. Komatsu**, R. Kondo, T. Toda, and K. Takeda, "Anomalous sound event detection based on WaveNet," in Proc. European Signal Processing Conference (EUSIPCO), Rome, Italy, Sep. 2018, pp. 2494–2498.

[50] T. Matsuyoshi, **T. Komatsu**, R. Kondo, T. Yamada, and S. Makino, "Weakly labeled learning using BLSTM-CTC for sound event detection," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Hawaii, USA, Nov. 2018, pp. 1918–1923.

[51] C. Narisetty, **T. Komatsu**, and R. Kondo, "Modelling of sound events with hidden imbalances based on clustering and separate sub-dictionary learning," in Proc. European Signal Processing Conference (EUSIPCO), Rome, Italy, Sep. 2018, pp. 847–851.

[52] **T. Komatsu** and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA, Mar. 2017, pp. 376–380.

[53] **T. Komatsu** et al., "An acoustic monitoring system and its field trials," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, Dec. 2017, pp. 1341–1346.

[54] **T. Komatsu**, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 2259–2263.

[55] **T. Komatsu**, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Sep. 2016, pp. 45–49.

[56] K. Ohtani, **T. Komatsu**, T. Nishino, and K. Takeda, "Adaptive dereverberation method based on complementary Wiener filter and modulation transfer function," in Proc. REVERB Workshop, Florence, Italy, May 2014, 4 pages.

[57] **T. Komatsu**, T. Nishino, G. W. Peters, T. Matsui, and K. Takeda, "Modeling head-related transfer functions via spatial-temporal Gaussian process," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013, pp. 301–305.

[58] **T. Komatsu**, G. Peters, T. Matsui, I. Nevat, and K. Takeda, "Modeling room impulse response via composites of spatial-temporal Gaussian processes," in Proc. Meetings on Acoustics, Acoustical Society of America, vol. 19, Montreal, Canada, May 2013, p. 040098.

[59] K. Kondo, Y. Takahashi, **T. Komatsu**, T. Nishino, and K. Takeda, "Computationally efficient single channel dereverberation based on complementary Wiener filter," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013, pp. 7452–7456.

[60] K. Ohtani, **T. Komatsu**, K. Kondo, T. Nishino, and K. Takeda, "Objective and subjective evaluation of complementary Wiener filter for speech dereverberation," in Proc. Meetings on Acoustics, Acoustical Society of America, vol. 19, Montreal, Canada, May 2013, p. 055038.

Preprints

[1] J. Sakuma, **T. Komatsu**, and R. Scheibler, "MLP-ASR: Sequence-length agnostic all-MLP architectures for speech recognition," arXiv preprint arXiv:2202.08456, 2022.

[2] J. Sakuma, **T. Komatsu**, and R. Scheibler, "MLP-based architecture with variable length input for automatic speech recognition," preprint, 2022.

[3] Y. Kida, **T. Komatsu**, and M. Togami, "Label-synchronous speech-to-text alignment for ASR using forward and backward transformers," arXiv preprint arXiv:2104.10328, 2021.

[4] Y. Okamoto et al., "Overview of tasks and investigation of subjective evaluation methods in environmental sound synthesis and conversion," arXiv preprint arXiv:1908.10055, 2019.

Technical Reports

[1] H. Munakata, T. Nishimura, S. Nakada, and **T. Komatsu**, "Training strategy of massive text-to-audio models and GPT-based query-augmentation," DCASE2024 Challenge, Tech. Rep., 2024.

[2] A. Ito, **T. Komatsu**, and Y. Fujita, "Vocabulary-set decomposition and multi-task learning for target vocabulary extraction in Japanese speech recognition," IEICE Technical Report, Tech. Rep. 389, 2023, pp. 159–164 (in Japanese).

[3] R. Scheibler, T. Hasumi, Y. Fujita, **T. Komatsu**, R. Yamamoto, and K. Tachibana, "Class-conditioned latent diffusion model for DCASE 2023 Foley sound synthesis challenge," Tech. Rep., 2023.

[4] R. Scheibler, **T. Komatsu**, Y. Fujita, and M. Hentschel, "3D CNN and Conformer with audio spectrogram transformer for sound event detection and localization," DCASE2022 Challenge, Tech. Rep., Jun. 2022.

[5] K. Miyazaki, **T. Komatsu**, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," DCASE2020 Challenge, Tech. Rep., Jun. 2020.

[6] M. Tani, **T. Komatsu**, C. Narisetty, and R. Kondo, "NEC acoustic situation awareness technology and its feasibility tests in Singapore," IEICE Technical Report, Tech. Rep. 106, 2017, pp. 29–32 (in Japanese).

Patents

[1] **T. Komatsu** and R. Kondo, "Anomaly detection apparatus, anomaly detection method, and program," US Patent 12,051,232, Jul. 30, 2024.

[2] **T. Komatsu** and R. Kondo, "Anomaly detection apparatus, anomaly detection method, and program," US Patent App. 18/756,474, Oct. 17, 2024.

[3] **T. Komatsu**, R. Kondo, and S. Mishima, "Learning device and pattern recognition device," US Patent 11,948,554, Apr. 2, 2024.

[4] R. Yamamoto, Y. Fujita, **T. Komatsu**, B. Park, and R. Scheibler, "Method for editing moving image, storage medium storing editing program, and editing device," US Patent App. 18/680,101, Dec. 5, 2024.

[5] **T. Komatsu** and R. Kondo, "Anomaly detection apparatus, anomaly detection method, and program," US Patent 11,715,284, Aug. 1, 2023.

[6] **T. Komatsu** and R. Kondo, "Pattern recognition robust to influence of a transfer path," US Patent 11,620,985, Apr. 4, 2023.

[7] **T. Komatsu** and R. Kondo, "Anomaly detecting device, anomaly detecting method, and recording medium," US Patent 11,397,792, Jul. 26, 2022.

[8] **T. Komatsu** and Y. Senda, "Device for estimating speed of moving sound source, speed monitoring system, method for estimating speed of moving sound source, and storage medium in which program for estimating speed of moving sound source is stored," US Patent 11,360,201, Jun. 14, 2022.

[9] **T. Komatsu** and R. Kondo, "Signal processing device, signal processing method, and storage medium for storing program," US Patent 11,200,882, Dec. 14, 2021.

[10] **T. Komatsu** and R. Kondo, "Signal processing device, signal processing method, and storage medium for storing program," US Patent App. 16/755,300, Jul. 22, 2021.

[11] **T. Komatsu**, R. Kondo, and T. Hayashi, "Anomaly detection apparatus, method, and program," US Patent App. 17/056,070, Aug. 19, 2021.

[12] R. Kondo and **T. Komatsu**, "Propagation path estimation apparatus, method, and program," US Patent App. 17/040,266, Jan. 14, 2021.

[13] C. Narisetty, R. Kondo, and **T. Komatsu**, "Information processing apparatus, method, and non-transitory storage medium," US Patent App. 16/969,868, Mar. 4, 2021.

[14] C. P. Narisetty, **T. Komatsu**, and R. Kondo, "A source separation device, a method for a source separation device, and a non-transitory computer readable medium," US Patent App. 17/286,095, Nov. 18, 2021.

[15] **T. Komatsu** and R. Kondo, "Signal processing device, signal processing method, and computer-readable recording medium," US Patent 10,817,719, Oct. 27, 2020.

[16] **T. Komatsu** and R. Kondo, "Signal processing device, signal processing method, and computer-readable recording medium," US Patent 10,679,646, Jun. 9, 2020.

[17] **T. Komatsu** and Y. Senda, "Signal detection device, signal detection method, and signal detection program," US Patent 10,650,842, May 12, 2020.

Academic and Professional Activities

Organizing Committee Member

- DCASE2020 Workshop, Technical Program Co-Chair
- DCASE2024 Workshop, Technical Program Co-Chair
- ICCV2025 Workshop, Foundation Data for Industrial Tech Transfer, Organizing Member
- AAAI2026 Workshop, Audio-Centric AI: Towards Real-World Multimodal Reasoning and Application Use Cases (Audio-AAAI), Organizing Member