

# GLAP: General contrastive audio-text pretraining across domains and languages

Heinrich Dinkel<sup>1</sup>, Zhiyong Yan<sup>1</sup>, Tianzi Wang<sup>1</sup>, Yongqing Wang<sup>1</sup>, Xingwei Sun<sup>1</sup>, Yadong Niu<sup>1</sup>,  
Jizhong Liu<sup>1</sup>, Gang Li<sup>1</sup>, Junbo Zhang<sup>1</sup>, Jian Luan<sup>1</sup>

<sup>1</sup>MiLM Plus, Xiaomi Inc., China

dinkelheinrich@xiaomi.com, zhangjunbo5@xiaomi.com

## Abstract

Contrastive Language Audio Pretraining (CLAP) is a widely-used method to bridge the gap between audio and text domains. Current CLAP methods enable sound and music retrieval in English, ignoring multilingual spoken content. To address this, we introduce general language audio pretraining (GLAP), which expands CLAP with multilingual and multi-domain abilities. GLAP demonstrates its versatility by achieving competitive performance on standard audio-text retrieval benchmarks like Clotho and AudioCaps, while significantly surpassing existing methods in speech retrieval and classification tasks. Additionally, GLAP achieves strong results on widely used sound-event zero-shot benchmarks, while simultaneously outperforming previous methods on speech content benchmarks. Further keyword spotting evaluations across 50 languages emphasize GLAP’s advanced multilingual capabilities. Finally, multilingual sound and music understanding is evaluated across four languages.

**Index Terms:** contrastive language-audio pretraining, general pretraining, general audio encoders, large-language models

## 1. Introduction

In the field of computer vision, Contrastive Language-Image Pretraining (CLIP) [1] represents a significant breakthrough in extracting efficient representations that can be applied across various downstream tasks and domains. Similarly, Contrastive Language-Audio Pretraining (CLAP) [2, 3, 4] bridges text and audio domains, enabling zero-shot transfer learning i.e., testing the model on novel concepts that is has not seen during training. Notably, [5] trained on 4.6 Million pairs of audio and speech data, but has shown poor results for trivial speech classification tasks such as keyword spotting (see MSCLAP-2023 in Figure 1). While multilingual extensions [6] improved retrieval performance across eight languages, their approach still lacks basic speech understanding. CLAP embeddings primarily target sound and music, missing comprehensive speech representation (i.e., spoken language) - a critical aspect of audio processing. While there has been previous work focusing on speech-text embeddings using contrastive learning [3], a general approach that can be used between sound, music and speech domains is still missing. This work proposes general language audio pretraining (GLAP), an extension of previous CLAP works, aimed at aligning speech content with text, without compromising in sound and music performance. Our experiments demonstrate that GLAP achieves competitive performance in music and sound retrieval tasks while significantly improving speech understanding capabilities. GLAP also effectively generalizes its speech and sound understanding capabilities beyond English.

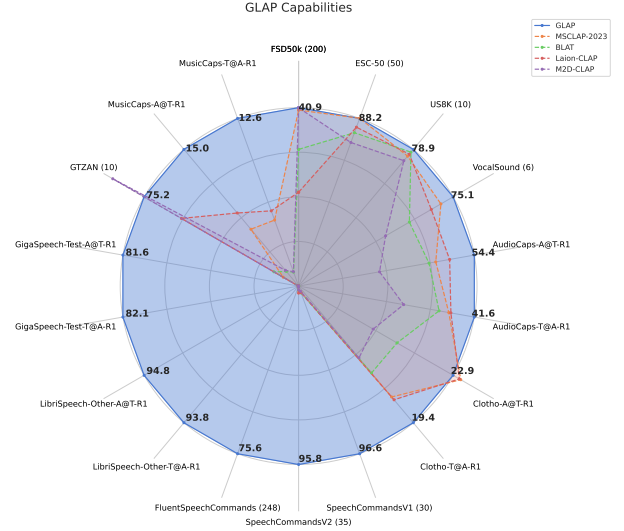


Figure 1: *GLAP’s retrieval and zero-shot performance. A@T and T@A represent retrieval tasks of Audio-to-Text and Text-to-Audio, respectively, others are zero-shot (number of labels in brackets). Missing baselines were evaluated by the authors.*

## 2. General language audio pretraining

To enable speech understanding in CLAP models, which are trained on sound and music data, one simple solution is to add speech data to the training dataset. However, as our analysis in Section 4.1 shows, this approach leads to compromised performance due to the absence of a unified audio encoder. The model either performs well on sound/music or on speech, but struggles to excel at both simultaneously. Thus, GLAP has two primary goals:

1. Deliver a unified encoder framework that maintains high performance for sound, music and speech retrieval tasks to enable alignment across these audio modalities.
2. Enable multilingual search capabilities for sound, music and speech content.

GLAP is trained with pairs of audio-text samples  $(a, t)$  in contrastive fashion. Features from these audio-text pairs are extracted through a pre-trained multi-lingual text encoder  $E_T$  and a pre-trained general audio encoder  $E_A$ :

$$e_a = \text{MLP}_A(E_A(a)), \quad e_t = \text{MLP}_T(E_T(t)),$$

A trainable multi-layer perceptron (MLP) is added to align the dimensions. Finally, the pair  $(e_a, e_t)$  is scored using cosine distance  $s = \frac{e_a \cdot e_t^T}{\|e_a\| \cdot \|e_t\|}$ . Unlike previous works, we use the

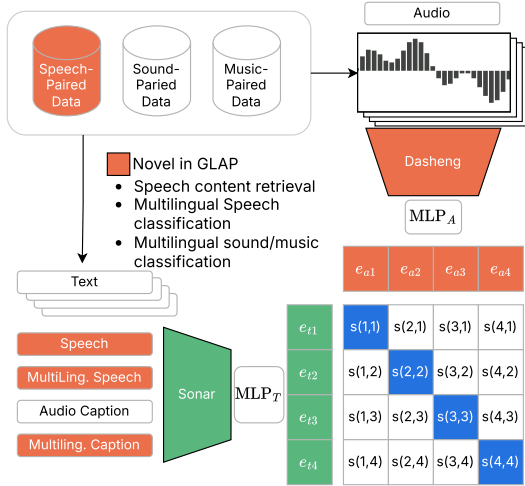


Figure 2: The GLAP framework. GLAP is trained via contrastive learning where positive pairs ( $\psi[i, j] = 1$ ) are shown in blue and negative pairs ( $\psi[i, j] = -1$ ) in white, with the added benefits of enabling multilingual speech-content retrieval, on-top of the standard sound/music capabilities.

sigmoid loss [7] as our main training objective  $\mathcal{L}$ , computed as:

$$\mathcal{L} = -\frac{1}{B} \sum_i \sum_j \log \sigma(s'(i, j) \cdot \psi[i, j]), \quad (1)$$

$$s'(i, j) = \frac{s(i, j) + \beta}{\tau}, \quad (2)$$

where  $\sigma$  is the sigmoid function,  $B$  is the batchsize,  $\beta, \tau$  are learnable parameters, “ $\cdot$ ” is the element-wise product, and

$$\psi[i, j] = \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{otherwise.} \end{cases}$$

The primary reason for choosing sigmoid loss over standard cross-entropy is its superior performance with large batch sizes and datasets, as we observed performance boosts of 1% to 5% across all retrieval tasks. An overview of the proposed framework can be seen in Figure 2.

### 3. Experimental Setup

The audio data is preprocessed by resampling all datasets to a single channel at 16 kHz. The trainable loss parameters  $\tau, \beta$  (Equation (2)) are initialized as 0.07 and  $-10$  respectively. For text embeddings, we use the text encoder from Sonar [8] as the default model, following [6]. We use a batch size of 128 per GPU across eight A800 GPUs, resulting in an effective batch size of  $B = 1024$ . Embeddings are gathered across all GPUs before calculating the loss. Each epoch is defined as processing 10,000 batches, with training running for a maximum of 20 epochs. Model training employs an 8-bit Adam optimizer with a cosine decay scheduler. The learning rate starts at 0, warms up to  $10^{-4}$  over the first two epochs, and decays to  $10^{-5}$  over the remaining training period. Training takes approximately 1.5 days, with the best models typically achieved within the first 15 epochs. The source code and checkpoints are publicly available<sup>1</sup>.

<sup>1</sup>[github.com/xiaomi-research/dasheng-glap](https://github.com/xiaomi-research/dasheng-glap)

Domain	Dataset	hours	#Pairs	#Lang
Speech	YODAS [9]	400 k	431 M	145
	GigaSpeech [10]	10 k	8 M	1
	LibriSpeech [11]	960	271 k	1
	AISHELL-1 [12]	180	131 k	1
Sound	Sound-VECaps <sub>A</sub> [13]	5200	1.6 M	1+7
	Auto-ACD [14]	5200	1.8 M	1+7
	AudiosetCaps [15]	5700	2.0 M	1+7
	WavCaps [16]	7544	400 k	1+7
	AudioCaps [17]	127	49 k	1+7
	Clothov2 [18]	35	4884	1+7
Music	MusicCaps [19]	7.3	2640	1+7
	Songdescriber [20]	12	360	1+7

Table 1: Training datasets with duration (hours), audio-text pairs (# Pairs), and languages (# Lang). Music and sound data, labeled in English, are auto-translated into seven other languages via Sonar.

#### 3.1. Datasets

**Training** In this work, we integrate a wide range of existing audio-text datasets, as outlined in Table 1. Our primary speech training dataset is YODAS [9], a 400k-hour YouTube corpus labeled mainly via automated speech-to-text pipelines. Due to its noisy labeling, we supplement it with cleaner English (GigaSpeech, LibriSpeech) and Chinese (AISHELL-1) datasets. While YODAS covers 145 languages, relying solely on multilingual speech data limits generalization in sound/music retrieval. To address this, we followed [6] and leveraged Sonar [8] to translate the original English captions of all sound and music datasets into other seven widely spoken languages: German, Chinese, Catalan, Spanish, Japanese, French, and Dutch. As shown in Table 1, the training data is skewed towards speech, leading to poor performance on non-speech tasks. To balance this, we categorize the data into four groups: sound + music, English speech (GigaSpeech + LibriSpeech + YODAS English), Chinese speech (AISHELL-1 + YODAS Chinese), and other languages in YODAS. During training, we sample equally from each group, ensuring a balanced training process.

**Evaluation** We evaluate retrieval performance across seven test sets, including sound datasets such as Auto-ACD (ACD)[14], AudioCaps (AC)[17], Clothov2 (Clotho) [18], music datasets such as MusicCaps (MC) [19] and the speech datasets LibriSpeech (LS) [11], GigaSpeech [10] and AISHELL-2 (AIS2) [23]. For zero-shot classification, we primarily follow [5] and evaluate on ESC-50, FSD50K, UrbanSound8K (US8K), CREMA-D, GTZAN, NSynth instruments, Beijing-Opera, VocalSound, as well as Speech Commands V1/V2 (SCV1/2) [24] and Fluent Speech Commands (FSC) [25]. All test datasets with the exception of AIS2 are labeled in English.

#### 3.2. Evaluation metrics

In audio-text retrieval tasks, performance is evaluated using recall at rank ( $R@k$ ), where  $R@k$  is 1 if the target item appears in the top  $k$  retrieved items, otherwise 0. We also use mean average precision at rank 10 (mAP10) for a more comprehensive comparison. For zero-shot inference, accuracy is used for single-class classification, and mean average precision (mAP) is used for multi-label classification, as is standard practice [5].

	LibriSpeechOther				MusicCaps				AISHELL2-Test			
	Text-to-Audio		Audio-to-Text		Text-to-Audio		Audio-to-Text		Text-to-Audio		Audio-to-Text	
	R@1	R@10	R@1	R@10	R@5	R@10	R@5	R@10	R@1	R@10	R@1	R@10
MSCLAP-2022 <sup>†</sup> [2]	0.1	0.6	0.0	0.4	4.3	7.2	5.2	7.4	0.0	0.2	0	0.18
MSCLAP-2023 <sup>†</sup> [5]	0.1	0.4	0.1	0.2	14.4	21.7	17.7	25.9	0.1	0.2	0.0	0.2
L-CLAP <sup>†</sup> [3]	0.1	0.8	0.1	0.5	17.2	25.5	<b>22.0</b>	31.1	0	0.2	0.0	0.2
L-CLAP <sup>†</sup> <sub>Speech-Music</sub> [3]	0.1	0.9	0.1	0.9	16.8	25.4	16.8	25.2	0	0.2	0.0	0.2
COLLAP-Roberta [21]	-	-	-	-	15.2	-	9.5	-	-	-	-	-
COLLAP-GPT [21]	-	-	-	-	17.4	-	10.3	-	-	-	-	-
BLAT <sup>†</sup> [4]	0.0	0.8	0	0.4	3.2	5.1	3.9	5.8	0.0	0.2	0.0	0.2
M2D-Clap <sup>†</sup> [22]	0.1	0.6	0.0	0.4	4.3	7.2	5.2	7.4	0.1	0.2	0.1	0.3
GLAP	<b>93.8</b>	<b>96.8</b>	<b>91.8</b>	<b>94.4</b>	<b>30.3</b>	<b>41.2</b>	15.0	<b>44.4</b>	<b>98.5</b>	<b>99.7</b>	<b>99.1</b>	<b>99.7</b>

Table 2: Retrieval results for music and speech datasets. <sup>†</sup> indicates evaluation from a public checkpoint. Best in bold; higher is better.

Task	Prompt
Speech	{label}
Music	The music in the style of {label}.
Sound	The sound of {label} can be heard.

Table 3: Prompts for zero-shot evaluation.

### 3.3. Prompting

GLAP supports zero-shot inference, allowing the model to generate outputs directly from text prompts without prior training on specific tasks. In zero-shot scenarios, crafting an effective text prompt is crucial for achieving optimal performance. This is particularly important for GLAP, as it lacks a dedicated token to distinguish between a spoken word (e.g., “cat”) and a sound event (e.g., “the sound of a cat”). Prompts used in this work are depicted in Table 3.

## 4. Results

### 4.1. Audioencoder investigation

Given the extensive range of previously explored CLAP methods, one might question why earlier approaches failed to achieve strong performance in general audio-text pretraining. In our view, a key limitation lies in the reliance on sound-event audio encoders for CLAP, which we believe represents a significant bottleneck for performance. In this section, we examine the role of audio encoders in general contrastive audio-text learning. For this purpose, we compare the proposed training framework using five different audio-encoders, being Dasheng [26], CED-Base [27], Beats [28], Whisper-Base [29] and WavLM [30]. Each encoder is selected for its approximately 90M parameters and support for variable-length inputs, a crucial requirement for processing speech recognition datasets—a capability lacking in many CLAP audio encoders [2, 5, 3]. Each encoder is initialized from a publicly available checkpoint and trained on the same dataset described in Section 3.1, using the previously described settings (Section 3), with Sonar serving as the default text encoder.

As it can be seen in our results in Table 4, the choice of audio-encoder is vital to achieve a well-balanced performance across tasks. Sound-event encoders, such as CED and Beats, perform well in sound and music retrieval tasks but struggle with speech-related tasks. Conversely, Whisper and WavLM excel in speech-related retrieval but underperform in sound event and music datasets. Dasheng, on the other hand, proves to

be the most versatile choice for general audio encoding, achieving competitive performance across sound, music, and speech domains. Based on these findings, all subsequent experiments utilize a *single* GLAP model with Dasheng as its audio encoder, without fine-tuning on target datasets.

Encoder	Sound		Music	Speech	
	AC	ACD	MC	LS-other	AIS2
CED-Base	58.6	62.0	25.1	87.8	70.6
Beats	55.1	64.3	23.9	91.8	44.0
Whisper-Base	46.5	52.9	15.8	98.9	99.4
WavLM	36.1	47.5	14.8	99.9	96.3
Dasheng	55.8	60.1	20.3	94.8	99.0

Table 4: Text-to-Audio retrieval performance across five datasets, categorized by domain. LS-other refers to the test-other subset of LibriSpeech, while AIS2 corresponds to the AISHELL-2 test set. All experiments were conducted using the proposed training dataset and configuration. Values indicate mAPI0, where higher scores represent better performance.

### 4.2. Sound retrieval results

The performance of GLAP on the widely used AudioCaps and Clotho datasets for English sound-event retrieval is presented in Table 5. GLAP demonstrates strong results on both benchmarks, surpassing other methods in Text-to-Audio retrieval (R@1) on AudioCaps while maintaining competitive performance on Clotho.

### 4.3. Music and speech retrieval results

For speech retrieval, we select the test-other dataset from LibriSpeech as a representative in-domain English speech benchmark, while the AISHELL-2 test set serves as an unseen Chi-

Method	AudioCaps				Clotho			
	Text-to-Audio R@1	Text-to-Audio R@10	Audio-to-Text R@1	Audio-to-Text R@10	Text-to-Audio R@1	Text-to-Audio R@10	Audio-to-Text R@1	Audio-to-Text R@10
BLAT [4]	33.3	82.4	40.4	85.7	12.3	46.1	13.9	48.2
LClap-Large [3]	34.2	84.1	43.1	90.1	15.3	51.2	20.8	60.0
MSCLAP-2022 [2]	33.5	80.2	47.8	90.7	16.2	51.4	23.6	60.3
MSCLAP2023 [5]	35.6	-	42.5	-	-	15.7	22.9	-
Wavcaps-CNN14 [16]	34.7	82.5	44.7	86.2	21.2	59.4	25.9	65.8
Wavcaps-HTSAT [16]	39.7	86.1	51.7	90.6	20.2	58.8	26.5	67.3
Auto-ACD [14]	39.5	85.4	53.7	91.7	15.3	52.1	17.7	52.6
T-CLAP [31]	39.7	86.9	49.8	91.9	17.3	53.6	21.8	57.4
MLCLAP [6]	40.7	87.8	50.1	92.8	18.8	59.0	21.1	62.5
Cacophony [32]	41.0	86.4	55.3	92.4	20.2	58.8	26.5	67.3
SoundVECaps [13]	41.2	85.3	53.3	93.0	-	-	-	-
GLAP	41.7	86.1	54.4	91.1	19.4	58.3	21.8	61.5

Table 5: Sound event retrieval results compared to baselines.

Method	Sound					Music			Speech		
	ESC50	FSD50K	US8K	CD	VS	GTZAN	NS	BO	SCV1	SCV2	FSC
BLAT [4]	80.6	31.3	77.3	17.6 <sup>†</sup>	53.9 <sup>†</sup>	10.0 <sup>†</sup>	9.03 <sup>†</sup>	31.4 <sup>†</sup>	3.9 <sup>†</sup>	2.2 <sup>†</sup>	0.4 <sup>†</sup>
MS-CLAP-2023 [5]	88.2	40.3	75.0	<b>29.7</b>	69.2	58.4	<b>47.9</b>	46.6	16.4 <sup>*</sup>	2.5 <sup>†</sup>	0.3 <sup>†</sup>
L-CLAP <sub>Speech-Music</sub> [3]	89.3	20.2 <sup>†</sup>	72.7 <sup>†</sup>	20.7 <sup>†</sup>	64.5 <sup>†</sup>	52.3 <sup>†</sup>	29.7 <sup>†</sup>	<b>57.2</b>	3.8 <sup>†</sup>	3.8 <sup>†</sup>	0.4 <sup>†</sup>
L-CLAP [3]	<b>91.0</b>	21.5 <sup>†</sup>	77.0	18.3 <sup>†</sup>	<b>79.3</b>	47.4 <sup>†</sup>	26.1 <sup>†</sup>	40.2 <sup>†</sup>	3.8 <sup>†</sup>	4.1 <sup>†</sup>	0.3 <sup>†</sup>
M2D-Clap [22]	75.5	40.8	72.4	17.7	42.3 <sup>†</sup>	<b>75.2</b>	23.4	47.0 <sup>†</sup>	3.0 <sup>†</sup>	2.1 <sup>†</sup>	0.4 <sup>†</sup>
GLAP	88.8	<b>40.9</b>	<b>78.9</b>	20.5	75.1	69.6	31.3	36.5	<b>96.6</b>	<b>95.8</b>	<b>75.6</b>

Table 6: Zero-shot evaluation performance. Results marked with <sup>†</sup> were obtained from a public checkpoint, while underlined entries used the corresponding training dataset and are therefore not truly zero-shot. Entries with \* used the 10-class variant instead of the 30-class version employed in this work. Best results are bolded and higher is better.

nese evaluation set. To assess music-related capabilities, we use the widely adopted MusicCaps dataset [19]. Since some baseline models operate at different sampling rates (16 kHz vs. 44/48 kHz), we resample the test datasets accordingly. As shown in Table 2, our approach significantly outperforms previous methods in both music and speech retrieval. GLAP demonstrates exceptional multilingual retrieval performance, achieving over 93% on the English LibriSpeech-other test set and 98% on the Chinese AISHELL-2 test set. While most compared CLAP models offer a competitive performance on the MC dataset.

#### 4.4. Zero shot evaluation

In this section, the zero-shot capabilities our approach are assessed using the prompts provided in Table 3. True zero-shot evaluation is difficult to determine, as some studies are trained on the respective datasets [3, 5] using its labels, meaning they are not strictly zero-shot. As shown in Table 6, our approach performs similarly to other baselines in sound-event and music classification. However, it excels in the keyword spotting task, significantly outperforming the baselines with accuracies of 95.8%, 96.6%, and 75.6% for the SCV1, V2, and FSC datasets, respectively. Notably, the FSC dataset requires understanding entire sentences, not just individual words, highlighting our model’s ability to align well with spoken content, as well as detecting single-word keywords (SCV1/2).

#### 4.5. Multilingual capabilities

GLAP’s multilingual spoken content capabilities are assessed through a zero-shot evaluation on the Multilingual Spoken Words (MSW) Corpus [33]. Results seen in Figure 3 show a strong zero-shot capability of GLAP on fifty languages. Oriya and Guarani achieve the best performance at 70% and 65.9% respectively. Notably, Chinese achieves 57.4% accuracy with 496 keywords, while Russian, containing 15,844 keywords, results in 39.6%. Additionally, we assess GLAP’s multilingual sound and music understanding by conducting zero-shot evaluations on the ESC-50, US8K and GTZAN datasets. For this, we use ChatGPT to translate the original (English) labels into the target language and adjust the prompts (Table 3) accordingly. As shown in Table 7, while performance drops compared to the English baseline, GLAP remains effective in multilingual sound-event classification. The model performs impressively in Russian (Ru) despite being trained only on Russian speech-text pairs from YODAS, not music/speech pairs (see Section 3.1). This suggests effective transfer of multilingual text-based knowledge to the audio domain.

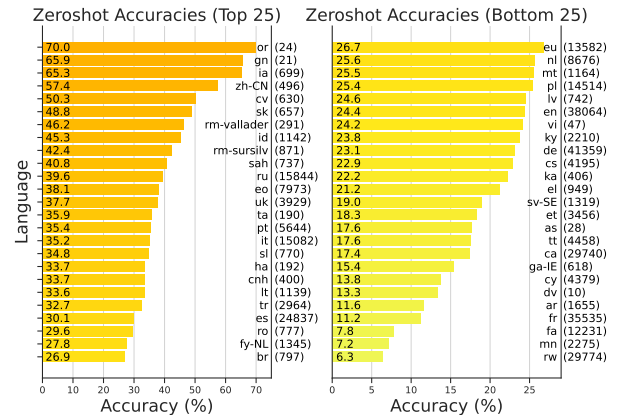


Figure 3: Multilingual zero-shot keyword spotting (KWS) performance across 50 languages. Only the test set for each language is used, and the accuracies are reported. The number of keywords (num) for each language is shown on the right.

Data	Language				
	En	De	zh-CN	Jp	Ru
US8K	78.9	74.8	66.1	72.2	49.0
ESC-50	88.8	64.3	71.4	74.3	62.1
GTZAN	69.6	68.3	62.5	63.2	65.3

Table 7: GLAP’s zero-shot evaluation for multilingual sound and music. Original labels (in gray) are translated into the target language using ChatGPT.

## 5. Conclusion

We introduce GLAP, a versatile language-audio pretraining framework that enables multilingual and multi-domain modeling of both audio and text. To the best of our knowledge, it is the first *single* system to integrate general audio and text embeddings into a unified contrastive framework. GLAP demonstrates competitive performance on well-established benchmarks like AudioCaps and Clotho, while surpassing previous methods in music and speech retrieval tasks. Zero-shot evaluations show strong results for English sound and music tasks, extending effectively to other languages. Inference on the Multilingual Spoken Words dataset highlights robust multilingual capabilities beyond English.

## 6. References

- [1] 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 *ICML*, 2021.
- [2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [3] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.
- [4] X. Xu, Z. Zhang, Z. Zhou, P. Zhang, Z. Xie, M. Wu, and K. Q. Zhu, "Blat: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2756–2764.
- [5] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 336–340.
- [6] Z. Yan, H. Dinkel, Y. Wang, J. Liu, J. Zhang, Y. Wang, and B. Wang, "Bridging language gaps in audio-text retrieval," in *Interspeech 2024*, 2024, pp. 1675–1679.
- [7] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.
- [8] P.-A. Duquenne, H. Schwenk, and B. Sagot, "Sentence-level multimodal and language-agnostic representations," *arXiv preprint arXiv:2308.11466*, 2023.
- [9] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, "Yodas: Youtube-oriented dataset for audio and speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [10] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.
- [11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [12] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.
- [13] Y. Yuan, D. Jia, X. Zhuang, Y. Chen, Z. Liu, Z. Chen, Y. Wang, Y. Wang, X. Liu, X. Kang *et al.*, "Sound-vecaps: Improving audio generation with visual enhanced captions," *arXiv preprint arXiv:2407.04416*, 2024.
- [14] L. Sun, X. Xu, M. Wu, and W. Xie, "Auto-acd: A large-scale dataset for audio-language representation learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5025–5034.
- [15] J. Bai, H. Liu, M. Wang, D. Shi, W. Wang, M. D. Plumbley, W.-S. Gan, and J. Chen, "Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models," *arXiv preprint arXiv:2411.18953*, 2024.
- [16] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.
- [17] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:174799768>
- [18] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.
- [19] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzett, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.
- [20] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos *et al.*, "The song descriptor dataset: a corpus of audio captions for music-and-language evaluation," *arXiv preprint arXiv:2311.10057*, 2023.
- [21] J. Wu, W. Li, Z. Novack, A. Namburi, C. Chen, and J. McAuley, "Collap: Contrastive long-form language-audio pre-training with musical temporal structure augmentation," *arXiv preprint arXiv:2410.02271*, 2024.
- [22] 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 *Interspeech 2024*, 2024, pp. 57–61.
- [23] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.
- [24] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *ArXiv e-prints*, Apr. 2018. [Online]. Available: <https://arxiv.org/abs/1804.03209>
- [25] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.
- [26] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Interspeech 2024*, 2024, pp. 547–551.
- [27] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 291–295.
- [28] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.
- [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.
- [30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [31] 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," *arXiv preprint arXiv:2404.17806*, 2024.
- [32] G. Zhu and Z. Duan, "Cacophony: An improved contrastive audio-text model," *arXiv preprint arXiv:2402.06986*, 2024.
- [33] M. Mazumder, S. Chitlangia, C. Banbury, Y. Kang, J. M. Ciro, K. Achorn, D. Galvez, M. Sabini, P. Mattson, D. Kanter *et al.*, "Multilingual spoken words corpus," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.