

FlexSED: Towards Open-Vocabulary Sound Event Detection

Jiarui Hai, Helin Wang, Weizhe Guo, Mounya Elhilali

Department of Electrical and Computer Engineering
Johns Hopkins University, Maryland, USA

Abstract—Despite recent progress in large-scale sound event detection (SED) systems capable of handling hundreds of sound classes, existing multi-class classification frameworks remain fundamentally limited. They cannot process free-text sound queries, which enable more flexible and user-friendly interaction, and they lack zero-shot capabilities and offer poor few-shot adaptability. Although text-query-based separation methods have been explored, they primarily focus on source separation and are ill-suited for SED tasks that require precise temporal localization and efficient detection across large and diverse sound vocabularies. In this paper, we propose FlexSED, an open-vocabulary sound event detection system. FlexSED builds on a pretrained audio SSL model and the CLAP text encoder, introducing an encoder-decoder composition and an adaptive fusion strategy to enable effective continuous training from pretrained weights. To ensure robust supervision, it also employs large language models (LLMs) to assist in event query selection during training, addressing challenges related to missing labels. As a result, FlexSED achieves superior performance compared to vanilla SED models on AudioSet-Strong, while demonstrating strong zero-shot and few-shot capabilities. We release the code and pretrained models to support future research and applications based on FlexSED.

1. INTRODUCTION

Sound event detection (SED) focuses on identifying and temporally locating meaningful sounds in audio recordings. Given an input audio stream, SED systems aim to determine what happened and when, recognizing events such as speech, alarms, animal sounds, or environmental noises and marking their precise start and end times [1], [2]. This technology underpins a wide range of real-world applications, including smart home monitoring, health and safety surveillance, wildlife tracking, multimedia content analysis, and assistive technologies for the hearing impaired.

Advances in deep neural networks have enabled modern SED models to achieve strong performance, though mainly on small-scale corpora focused on domestic environments [2]–[4]. More recently, breakthroughs in audio spectrogram transformers, pretrained on large-scale unlabeled data via self-supervised learning (SSL), have significantly expanded SED capabilities, allowing models to address complex and diverse acoustic scenes involving over 400 sound classes [5]–[7]. However, these approaches still operate under a closed-vocabulary assumption, treating sound event detection as a multi-class classification task constrained to a predefined set of classes established during training. This limitation not only requires users to manually specify and map target classes for each use case, complicating real-world deployment, but also restricts the model’s ability to capture rich semantic and acoustic relationships among sound categories. In addition, traditional models do not support direct inference beyond the predefined label space without retraining and incorporating additional labeled data. Moreover, adapting to new events is challenging, as it requires modifying the classification head.

While these challenges continue to constrain the flexibility and scalability of current SED systems, related advances in computer vision offer a compelling contrast. Large-scale open-vocabulary detection (OVD) has advanced rapidly in computer vision [8], [9]. By enabling

automatic detection of arbitrary objects without predefined categories or extensive human annotation, OVD has driven image understanding toward more flexible, scalable, and adaptable recognition. These capabilities have unlocked a range of versatile applications, including evaluating text-to-image generation [10] and labeling data for tasks such as visual question answering [11]. However, such large-scale open-vocabulary functionality remains largely absent in current SED systems, limiting their generalization and hindering progress in broader audio tasks, including evaluating text-to-audio generation [12], [13], enhancing audio language models [14], [15], and serving as a reward signal for methods such as Direct Preference Optimization [16].

Although recent efforts have explored text-prompt-based separation models [17], [18], these approaches primarily aim to extract sound events rather than directly predict their temporal locations. Moreover, they often rely on computationally intensive methods, such as diffusion models, which are designed to improve separation quality when the target sound source is specified. This focus makes them less suitable for scenarios requiring detection across a large number of potential classes, as in SED. While text-query-based separation models have been applied to text-query-based SED [19], this approach was developed on a small-scale dataset, lacks demonstrated zero-shot and few-shot capabilities, does not offer true open-vocabulary support, and incurs substantial computational overhead due to its reliance on separation processes. As a result, it remains ill-suited for more challenging scenarios involving diverse and numerous sound classes.

Motivated by recent advances in large-scale OVD and the need to overcome the limitations of both conventional multi-class classification frameworks and separation-based text-query SED, we investigate the concept of **open-vocabulary sound event detection (OV-SED)** in this work. However, pursuing OV-SED introduces several key challenges. Unlike OVD in vision, where large-scale datasets are available, the largest temporally labeled dataset for SED, AudioSet-Strong [7], contains only about 100k audio clips, limiting the amount of training data. Moreover, negative queries play a crucial role in OV-SED training to improve classification accuracy. Yet, the presence of missing labels in AudioSet-Strong complicates the reliable construction of negatives. Poorly selected negatives risk introducing noise or incorrect supervision, which can mislead the model during training and ultimately degrade performance.

To address these challenges and advance OV-SED, we propose FlexSED, a framework specifically designed to tackle the core issues outlined above. Its key design components are as follows:

(1) Architecture design for prompt-aware audio modeling. To address data scarcity and enable efficient prompt-based audio modeling, we leverage two pretrained models: a frame-based audio SSL model [20] for fine-grained acoustic representation and the CLAP text encoder [21] for sound class prompt comprehension. We decompose the audio SSL model into an audio-only encoder and an audio-text fusion decoder specialized for prompt-audio integration, balancing performance and inference efficiency. Furthermore, we introduce a seamless adaptive fusion strategy to enable continued

This Work is supported by ONR N00014-23-1-2050 and N00014-23-1-2086.

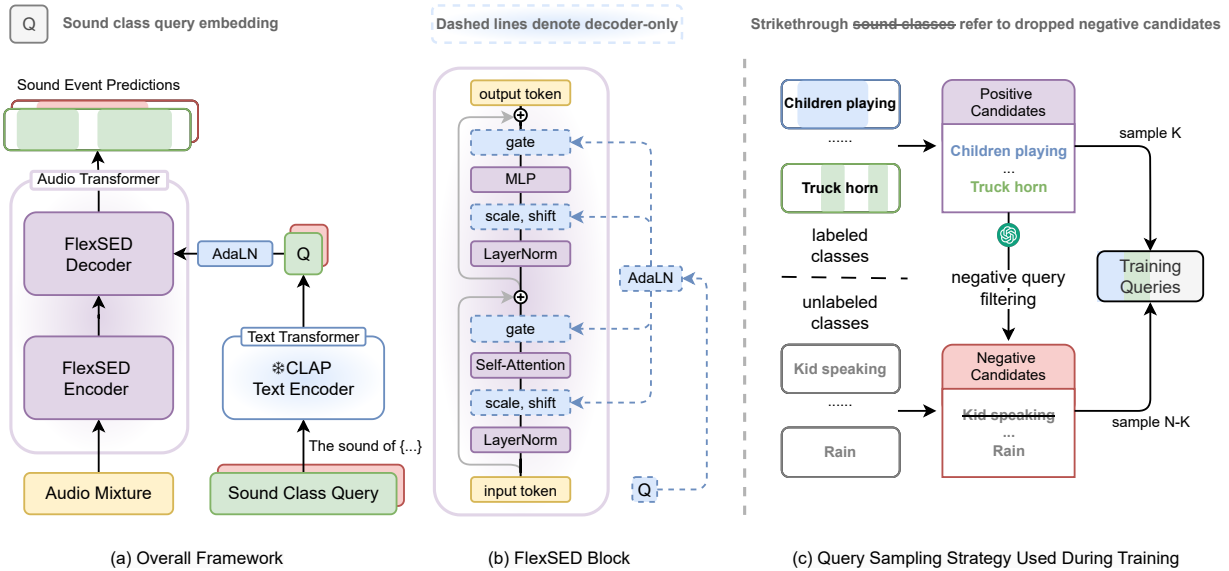


Fig. 1: The framework of FlexSED.

training from pretrained weights.

(2) **Mitigating missing label issues.** To mitigate the risk of incorrect negative queries arising from incomplete labels, we propose a negative candidate filtering strategy that leverages large language models (LLMs) [22] to infer semantic relationships among sound classes. This approach helps avoid conflicts between positive and negative queries, ensuring more reliable and consistent supervision.

(3) **Effectiveness and flexibility.** FlexSED surpasses conventional multi-class classification methods on the AudioSet-Strong evaluation set, while also demonstrating strong zero-shot capabilities and few-shot expandability.

(4) **Enabling practical applications.** We release¹ the code, model checkpoints, and user-friendly inference tools to support a wide range of OV-SED applications, including text-to-audio evaluation, audio question answering data annotation, and more.

2. PROPOSED METHOD

2.1. Overall Framework

As shown in Fig. 1 (a), OV-SED takes an audio clip along with a list of sound event candidate prompts as input. The audio signal is first processed by the encoder independently, without interacting with the candidate prompts. Meanwhile, the text prompts are processed by a text encoder. The decoder then fuses the audio-text embedding pairs in parallel, enabling interactions between candidate prompts, and produces the corresponding posterior probabilities.

2.2. FlexSED Model Architecture

To build a simple yet effective open-vocabulary sound event detector capable of identifying and localizing sound events from arbitrary text inputs, we propose a unified model that integrates audio and text representations. At its core, the model adopts a vanilla transformer-based SED architecture, modified to predict only the posterior corresponding to the input text prompt, rather than conventional multi-class outputs. This enables flexible, prompt-based sound event detection.

For text representation, we leverage CLAP [21], which produces rich and semantically aligned embeddings that bridge the gap between

textual descriptions and audio content. For audio processing, we adopt a frame-based audio spectrogram transformer [20] built on self-supervised learning, which provides strong audio representations, accelerates convergence on downstream audio and speech tasks, and operates at a latent sample rate of 25 Hz, making it well-suited for sound event detection.

However, the audio transformer does not natively support text fusion. To address this, inspired by AdaLN-Zero [23], commonly used in diffusion transformers for effective feature fusion, we introduce AdaLN into the pre-trained audio SSL blocks, as illustrated in Fig. 1 (b). Unlike AdaLN-Zero, which initializes the residual scaling gate to zero, we initialize this gate to one, creating a variant we call AdaLN-One. This design preserves the original SSL feature flow at initialization, while allowing the model to gradually learn to integrate text prompts during training.

Specially, given an input audio feature \mathbf{x} and a text prompt embedding \mathbf{p} , AdaLN fusion is defined as:

$$\mathbf{y} = \mathbf{x} + g(\mathbf{p}) \cdot \text{Layer}((1 + \gamma(\mathbf{p})) \cdot \text{LN}(\mathbf{x}) + \beta(\mathbf{p})), \quad (1)$$

where $\text{LN}(\cdot)$ is Layer Normalization, $\text{Layer}(\cdot)$ denotes the attention or feedforward layer, $\gamma(\mathbf{p})$, $\beta(\mathbf{p})$, and $g(\mathbf{p})$ are modulation parameters derived from the prompt condition, initialized to zero, zero, and one, respectively.

At initialization, it reduces to the standard residual block:

$$\mathbf{y} = \mathbf{x} + \text{Layer}(\text{LN}(\mathbf{x})). \quad (2)$$

This allows for seamless compatibility with the pre-trained model at the start of training. As training progresses, $g(\mathbf{p})$ is learned, enabling the model to dynamically adjust the influence of text prompt modulation based on the input prompt.

Additionally, while directly fusing prompts into each transformer block is feasible, this approach becomes inefficient when handling a large number of sound classes. To overcome this limitation, we propose an encoder-decoder architecture: the encoder focuses solely on audio representation, while the decoder integrates text prompts for fusion. This design enables encoder features to be cached during

¹Source Code: <https://github.com/JHU-LCAP/FlexSED>

inference, improving efficiency when evaluating numerous candidate sound classes.

In summary, we denote this method as FlexSED for its flexible design that leverages pre-trained models to enable efficient OV-SED.

2.3. Negative Query Filtering

When training FlexSED, we randomly compose positive and negative queries, as illustrated in Fig. 1 (c). Positive queries correspond to labeled sound events, while negative queries refer to unlabeled events. However, we observe that **AudioSet-Strong** suffers from potential missing labels. Treating such unlabeled yet present events as negative queries can confuse the model during training, undermine prompt-event alignment, and ultimately degrade performance.

Some missing labels arise from parent-child relationships defined in the AudioSet ontology, while others fall outside this structure. For example, a clip labeled as kid-playing may omit kid-speaking, despite both events likely being present. Similarly, classes like engine and truck often co-occur but are not explicitly linked in the ontology, making them difficult to filter automatically.

While manual identification of these missing labels is feasible, it becomes impractical at scale due to the more than 400 sound classes and the vast number of possible class pairs. To address this challenge, we leverage GPT-4² [22], which excels at understanding and reasoning over complex semantic relationships, including hierarchical structures, contextual similarities, and latent associations between concepts. Specifically, we use GPT-4 to identify various intra-class relations, such as subcategories, synonymous sound events, frequent co-occurrences, and safe negatives, classes that are unrelated or only rarely co-occurring. Guided by these insights, we sample negative queries primarily from safe negatives, thereby reducing the likelihood of conflicts with positive classes and enhancing training robustness.

3. EXPERIMENTAL SETUPS

3.1. Dataset

Strongly-labeled AudioSet is a subset of AudioSet, consisting of 100k frame-level labeled real audio clips for training, with 2k reserved for validation to assist in model selection. An additional 16k frame-level labeled clips are provided for evaluation. The dataset includes 456 distinct labels, with 407 audio classes shared between the training and evaluation sets. Training is performed on all classes, while evaluation focuses exclusively on the 407 shared classes.

3.2. Implementation Details

We conduct experiments on audio sampled at 16 kHz. The transformer backbone is based on Dasheng-base³, which operates on 64-dimensional log-Mel spectrograms extracted every 10 ms using a window size of 32 ms. For text encoding, we adopt the text encoder from Laion-CLAP⁴, employing simple prompts of the form “A sound of {class}” for sound event queries, where {class} refers to the sound class names defined in the AudioSet ontology.

The FlexSED transformer is initialized with the pre-trained Dasheng model, while the CLAP model remains frozen during training. FlexSED is optimized using AdamW with learning rates of 1×10^{-4} for the decoder and 1×10^{-5} for the encoder. A batch size of 16 audio samples and 20 prompts per audio sample is used. Among the 20 prompts, up to 10 are positive queries sampled from the labels—if more than 10 positives are available, 10 are randomly selected; if fewer, all are used. The remaining prompts are filled with negative

queries randomly sampled from negative candidates. The model is trained for 10 epochs, and the checkpoint with the best validation performance is selected for final evaluation. We apply simple Mel-spectrogram augmentations [24], [25], such as frame shift and filter augmentation, during training for the proposed FlexSED model⁵.

3.3. Evaluation Metrics

Following prior studies, we adopt the PSDS1 metric [26], [27], which captures the intersection between ground truth and detected events, emphasizing low reaction time and accurate sound event localization. Unlike earlier works that rely on coarser metrics with resolutions [7], we employ this finer-grained approach for greater temporal precision. Consistent with prior research [5], [6], we omit the variance penalty, as it was originally intended for datasets with fewer and more balanced classes. Furthermore, in line with previous work, we exclude PSDS2 from this study, as it is tuned to audio tagging rather than sound event detection and is impractical⁶ to compute on a server with 128 GB of memory when processing over 400 classes. We apply a median filter with a window length of 7 to smooth the model predictions.

We refer to PSDS1 computed across all classes as **PSDS1_A**. While this metric emphasizes accurate timing, it still penalizes false positives. In addition, we introduce a variant, **PSDS1_T**, which focuses solely on target sounds. In this setting, inference is restricted to the sound classes present in each audio clip, rather than spanning all 407 classes. This reflects a practical use case in which the target sounds are already known to the user, but precise temporal localization is still required. Such a scenario highlights the model’s potential for annotating weakly labeled data [28] or integrating with audio-tagging models [29].

4. RESULTS AND DISCUSSION

4.1. Comparison with Vanilla SED

We train two vanilla SED models for comparison. First, we reproduce ATST-Frame-SED on AudioSet-Strong following the procedure outlined in the original paper [5]. In parallel, we train Dasheng-SED using the same vanilla multi-class classification setup and training procedure. Notably, the Dasheng-based backbone demonstrates superior overall performance and exhibits reduced sensitivity to hyperparameters compared to ATST-Frame. Based on these observations, we adopt Dasheng as the backbone for FlexSED in our experiments. Additionally, we extend a target sound separation model for OV-SED by adding a classification layer on the pre-trained AudioSep [18].

As shown in Table 1, the proposed FlexSED significantly outperforms both the vanilla SED models and the target sound separation-based model. This improvement can be largely attributed to the integration of knowledge from the pre-trained CLAP and DaSheng models. CLAP enhances FlexSED’s ability to capture semantic relationships among sound classes compared to vanilla SED models, while DaSheng provides stronger temporal and acoustic modeling through its pre-trained audio transformer than AudioSep’s convolutional U-Net architecture. In addition to the performance boost, FlexSED also offers open-vocabulary capabilities for simpler usability. Furthermore, in Section 4.3, we demonstrate its effectiveness in zero-shot settings and its expandability in few-shot scenarios.

4.2. Ablation Studies

4.2.1. Prompt Fusion Method: As shown in Table 2, we compare several fusion strategies with AdaLN-One. Specifically, AdaLN-Zero initializes the gating parameter to zero; Token Fusion [30] directly

²<https://platform.openai.com/docs/models/gpt-4>

³<https://huggingface.co/mispeech/dasheng-base>

⁴<https://huggingface.co/laion/clap-htsat-fused>

⁵We leave the integration of mixup and mean-teacher into FlexSED for future work, as they are not directly compatible with the current framework.

⁶https://github.com/fgnt/sed_scores_eval/issues/5

Table 1: Performance comparison of sound event detection methods.

Method	OV-SED	PSDS1 _A ↑	PSDS1 _T ↑
ATST-Frame-SED	×	0.3846	0.5387
Dasheng-SED	×	0.3986	0.5349
AudioSep-SED	✓	0.3260	0.5332
FlexSED	✓	0.4484	0.5863
w/o Neg. Filtering		0.4366	0.5627

Table 2: Ablation study of model design choices.

Fusion	Enc. Blocks	Dec. Blocks	PSDS1 _A ↑	PSDS1 _T ↑
Token Fusion	First 6	Last 6	0.4378	0.5696
CrossAttn	First 6	Last 6	0.4430	0.5800
AdaLN-Zero	First 6	Last 6	0.4383	0.5717
AdaLN-One	First 4	Last 8	0.4374	0.5861
AdaLN-One	First 6	Last 6	0.4484	0.5863
AdaLN-One	First 8	Last 4	0.4393	0.5795
AdaLN-One	First 10	Last 2	0.4249	0.5826
AdaLN-One	First 12	Copied Last 2	0.4205	0.5728

inserts the prompt token into the decoder sequence; and Cross-Attention (CrossAttn) [31] introduces an additional cross-attention layer in the decoder blocks, using the audio embedding as the query and the prompt token as the key and value. Among these methods, AdaLN-One achieves the best performance. Cross-Attention performs slightly worse, likely due to the substantial architectural changes it introduces. AdaLN-Zero also lags behind AdaLN-One, highlighting the importance of seamless AdaLN parameter initialization for stable and effective continuous training. Token Fusion performs the worst, presumably because the pre-trained model is optimized to attend to audio embeddings and struggles to shift focus to the newly added prompt token.

4.2.2. Encoder-Decoder Design: We explore different configurations for splitting the original Transformer into encoder and decoder components.⁷ As shown in Table 2, dividing the encoder and decoder equally yields the best performance. Reducing the decoder size results in a performance drop, likely because fewer decoder layers limit the model’s ability to capture prompt-conditioned representations. On the other hand, increasing the decoder beyond half of the total blocks impairs performance, as too few encoder layers remain to learn prompt-independent acoustic features. Additionally, we experiment with adding two extra decoder blocks copied from the last two layers of the original model. This strategy performs worse than directly using the last two blocks as the decoder, suggesting that maintaining the original structure is important for effective continuous training.

4.2.3. Negative Query Filtering Strategy: As shown in Table 1, FlexSED without the proposed negative query filtering strategy exhibits degraded performance, underscoring the effectiveness of our method in removing potentially conflicting negative queries that could otherwise confuse training and impair model performance.

4.3. Zero-shot and Few-shot Performance

To evaluate the model’s zero-shot performance on unseen sound classes and its adaptability through fine-tuning with limited labeled data, we conduct zero-shot and few-shot experiments. Specifically, we randomly exclude K sound classes from training, ensuring these classes remain unseen for both positive and negative samples. We then assess the model’s performance on these excluded classes and compare it to its performance when trained on all classes. To ensure robustness, this procedure is repeated across three random folds of unseen classes,

⁷Due to computational constraints and inefficiency with large number of sound classes, we do not use all blocks for decoding.

Table 3: Zero-shot and few-shot performance with different numbers of excluded classes (K_{out}). Mean retained performance ratios (R-PSDS1) are computed relative to the full-model PSDS1_A^E and PSDS1_T^E on excluded classes across 3 folds.

K_{out}	Setting	R-PSDS1 _A ^E ↑	R-PSDS1 _T ^E ↑
20	zero-shot	64.99%	88.27%
	5-shot	80.23%	93.94%
	10-shot	84.58%	95.43%
	20-shot	87.02%	96.81%
40	zero-shot	55.04%	85.74%
	5-shot	75.74%	94.26%
	10-shot	77.41%	96.84%
	20-shot	82.83%	98.40%

and we report the mean retained PSDS ratios (R-PSDS1^E) relative to the full-model performance on the excluded classes.

For few-shot learning, we fine-tune the model on the excluded classes using a small number of labeled audio clips. The decoder is updated with a learning rate of 1×10^{-5} , and the encoder with 1×10^{-6} , over 25 epochs with a batch size of 8 audio clips.

As shown in Table 3, the results reveal several important trends regarding the model’s generalization and adaptability. In the zero-shot setting, the model retains a significant portion of its performance on unseen classes, particularly for PSDS1_T^E, where retention rates reach around 88% with $K_{\text{out}} = 20$ and 86% with $K_{\text{out}} = 40$. In contrast, PSDS1_A^E exhibits more pronounced degradation, dropping from around 65% to 55% as the number of excluded classes increases. This is expected, as PSDS1_A^E reflects both localization and classification performance, making it more sensitive to false positives and false negatives. Unseen classes can easily cause misclassification and missed detections, which jointly impact PSDS1_A^E more severely. On the other hand, PSDS1_T^E focuses primarily on timing accuracy and does not penalize false positives, making it more robust in zero-shot scenarios.

Once few-shot fine-tuning is applied, performance improves considerably. With as few as 5 shots, PSDS1_A^E increases to about 80% for $K_{\text{out}} = 20$ and 76% for $K_{\text{out}} = 40$. Additional labeled samples continue to boost performance, and with 20-shot fine-tuning, PSDS1_A^E reaches roughly 87% and 83%, while PSDS1_T^E surpasses 96% and 98%, nearly closing the gap to full training. Although larger K_{out} values naturally introduce greater challenges due to a broader unseen class space, fine-tuning effectively mitigates these difficulties.

Overall, these findings highlight the model’s ability to generalize reasonably well to unseen classes in zero-shot settings, especially for temporal localization, and its strong adaptability through few-shot learning to recover classification performance. This demonstrates the practicality and scalability of the proposed approach for sound event detection in dynamic and evolving environments.

5. CONCLUSIONS

In this work, we present FlexSED, a framework for open-vocabulary sound event detection. FlexSED effectively addresses key challenges in OV-SED by leveraging pretrained audio SSL and text models, introducing an adaptive fusion strategy for efficient prompt-audio integration, and mitigating label noise through LLM-guided pair filtering. These design choices enable FlexSED to outperform traditional classification approaches, while supporting zero-shot and few-shot capabilities for diverse real-world applications. For future work, we plan to explore strategies such as knowledge distillation [6] to further enhance model performance, scale training to larger and more diverse audio datasets, and investigate multimodal extensions that support richer query interactions beyond text prompts.

REFERENCES

- [1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.
- [2] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.
- [3] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.
- [4] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained atst model for sound event detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 911–915.
- [5] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1336–1351, 2024.
- [6] F. Schmid, T. Morocutti, F. Foscarin, J. Schlüter, P. Primus, and G. Widmer, "Effective pre-training of audio transformers for sound event detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [7] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.
- [8] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 393–14 402.
- [9] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 983–73 007, 2023.
- [10] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *Advances in Neural Information Processing Systems*, vol. 36, pp. 52 132–52 152, 2023.
- [11] K. Li, G. Vosselman, and M. Y. Yang, "Convincing rationales for visual question answering reasoning," *arXiv preprint arXiv:2402.03896*, 2024.
- [12] J. Hai, Y. Xu, H. Zhang, C. Li, H. Wang, M. Elhilali, and D. Yu, "Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer," *arXiv preprint arXiv:2409.10819*, 2024.
- [13] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint arXiv:2412.21037*, 2024.
- [14] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.
- [15] 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," *arXiv preprint arXiv:2406.11768*, 2024.
- [16] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, 2023.
- [17] H. Wang, J. Hai, Y.-J. Lu, K. Thakkar, M. Elhilali, and N. Dehak, "Soloaudio: Target sound extraction with language-oriented audio diffusion transformer," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [18] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [19] H. Yin, J. Bai, Y. Xiao, H. Wang, S. Zheng, Y. Chen, R. K. Das, C. Deng, and J. Chen, "Exploring text-queried sound event detection with audio source separation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [20] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," *arXiv preprint arXiv:2406.06992*, 2024.
- [21] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [23] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.
- [24] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugument: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.
- [25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, p. 2613, 2019.
- [26] Č. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.
- [27] J. Ebberts, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1021–1025.
- [28] J. F. Gemmeke, D. P. 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.
- [29] 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.
- [30] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 669–22 679.
- [31] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in neural information processing systems*, vol. 34, pp. 14 200–14 213, 2021.