Interspeech 2024
1-5 September 2024, Kos, Greece

DiffATR: Diffusion-based Generative Modeling for Audio-Text Retrieval

*Yifei Xin, Xuxin Cheng, Zhihong Zhu, Xusheng Yang, Yuexian Zou**

School of ECE, Peking University, Shenzhen, China
xinyifei@stu.pku.edu.cn

Abstract

Existing audio-text retrieval (ATR) methods are essentially discriminative models that aim to maximize the conditional likelihood, represented as $p(candidates|query)$. Nevertheless, this methodology fails to consider the intrinsic data distribution $p(query)$, leading to difficulties in discerning out-of-distribution data. In this work, we attempt to tackle this constraint through a generative perspective and model the relationship between audio and text as their joint probability $p(candidates, query)$. To this end, we present a diffusion-based ATR framework (DiffATR), which models ATR as an iterative procedure that progressively generates joint distribution from noise. Throughout its training phase, DiffATR is optimized from both generative and discriminative viewpoints: the generator is refined through a generation loss, while the feature extractor benefits from a contrastive loss, thus combining the merits of both methodologies. Experiments on the AudioCaps and Clotho datasets with superior performances, verify the effectiveness of our approach. Notably, without any alterations, our DiffATR consistently exhibits strong performance in out-of-domain retrieval settings.

Index Terms: audio-text retrieval, diffusion model, joint probability distribution, out-of-domain retrieval

1. Introduction

Given a query in the form of an audio clip or a caption, audio-text retrieval (ATR) seeks to identify a matching item from a collection of candidates in another modality. Existing ATR techniques [1–3] predominantly interpret cross-modal interactions via discriminative modeling based on contrastive learning [4–6]. However, from a probabilistic viewpoint, discriminative models only capture the conditional probability distribution, denoted as $p(candidates|query)$. This leads to a limitation that discriminative models fail to model the underlying data distribution [7], and their latent space includes fewer intrinsic data characteristics $p(query)$, which challenges their ability to generalize on unseen data [8]. In contrast, generative models capture the joint probability distribution between a query and candidates as $p(candidates, query)$, ensuring a more semantic-based latent mapping based on the data's inherent characteristics. As a result, generative models show superior generalizability and transferability compared to the discriminative counterparts.

Recent advancements in generative modeling have been notable, with breakthroughs in image synthesis [9], natural language processing [10], and audio/speech generation [11–13]. Among the leading generative techniques is the diffusion model [14, 15], which is a type of likelihood-based approach that refines signals by iteratively eliminating noise through a learned denoising model. The diffusion model's unique ability to refine from coarse to fine-grained stages makes it adept at capturing the relationship between audio and text. With this in mind, we employ the diffusion model to represent the audio-text correlation via their joint probability.

To this end, we introduce DiffATR, a diffusion-based ATR framework, to tackle the shortcomings of discriminative approaches from a generative perspective. We attempt to model the ATR task as an iterative procedure that progressively crafts a joint distribution from noise. Given a query and a collection of candidates, we adopt the diffusion model to generate the joint probability distribution $p(candidates, query)$. To enhance the ATR performance, we optimize our method from both generative and discriminative angles. During the training stage, the generator is refined using a common generation loss, i.e., Kullback-Leibler (KL) divergence [16], while the feature encoder is optimized by the contrastive loss, i.e., NT-Xent loss [17, 18], ensuring a blend of discriminative accuracy and generative flexibility.

DiffATR presents two key strengths. Firstly, its generative nature makes it inherently generalizable and adaptable to out-of-domain samples without additional adjustments. Secondly, the iterative refinement property of the diffusion model allows DiffATR to progressively enhance the retrieval results from coarse to fine. Experimental results on two benchmark ATR datasets, including AudioCaps and Clotho, demonstrate the advantages of DiffATR. To further evaluate the generalization of our method to unseen data, we introduce an out-of-domain ATR task, where the goal is to retrieve relevant audio/text data from a dataset that belongs to a domain different from the one the model is originally trained on. Our method not only represents a novel effort to promote generative approaches for in-domain ATR, but also shows the potential of generative solutions in out-of-domain ATR settings.

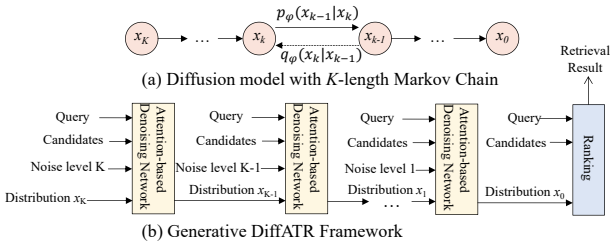
The primary contributions of this paper include:

- To the best of our knowledge, we are the first to boost ATR task from a generative viewpoint. We introduce a novel diffusion-based ATR framework (DiffATR) that models ATR as an iterative procedure to progressively generate joint distribution from noise.
- DiffATR effectively improves the retrieval performance on multiple ATR benchmarks and can generalize to various baseline models.
- More encouragingly, DiffATR consistently excels in out-of-domain ATR without extra adjustments.

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001).

* Yuexian Zou is the corresponding author.

10.21437/Interspeech.2024-405



(a) Diffusion model with K-length Markov Chain

(b) Generative DiffATR Framework

Figure 1: *(a) The diffusion model for joint probability generation. (b) Overview of our DiffATR framework.*

2. Proposed Method

2.1. Existing Discriminative Modeling

For ATR methods, similarity learning [19–21] is often employed. Typically, both text and audio embeddings are mapped to a shared latent space, where their similarity is calculated as the dot product of their corresponding representations. In text-to-audio retrieval, such approaches [4, 22] utilize logits to calculate the posterior probability:

$$p\left(a|t;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)=\frac{\exp\left(f_{\boldsymbol{\theta_t}}(t)^\top f_{\boldsymbol{\theta_a}}(a)/\tau\right)}{\sum_{a'\in\boldsymbol{A}}\exp\left(f_{\boldsymbol{\theta_t}}(t)^\top f_{\boldsymbol{\theta_a}}\left(a'\right)/\tau\right)}, \quad (1)$$

where τ is the temperature hyper-parameter, $\boldsymbol{A}$ represents the corpus of audio clips. $\boldsymbol{\theta_t}$ and $\boldsymbol{\theta_a}$ denote the text and audio encoder parameters, respectively. Then, existing approaches rank all audio candidates based on the posterior probability $p\left(a|t;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)$. Similarly, in audio-to-text retrieval, they rank all text candidates based on $p\left(t|a;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)$. The parameters $\{\boldsymbol{\theta_t},\boldsymbol{\theta_a}\}$ of text and audio encoders are optimized by minimizing the contrastive loss:

$$\begin{aligned}\boldsymbol{\theta_t}^*,\boldsymbol{\theta_a}^*=\underset{\boldsymbol{\theta_t},\boldsymbol{\theta_a}}{\arg\min}-\tfrac{1}{2}\mathbb{E}_{(t,a)\in\mathcal{D}}\left[\log p\left(a|t;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)\right.\\ \left.+\log p\left(t|a;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)\right]\end{aligned} \quad (2)$$

where $\mathcal{D}$ is a collection of text-audio pairs (t,a). Such learning method is equivalent to maximizing conditional likelihood, i.e., $\prod_{(t,a)\in\mathcal{D}}p(a|t)+\prod_{(t,a)\in\mathcal{D}}p(t|a)$, which is called discriminative training [23, 24].

Since current approaches focus solely on the conditional probability distribution $p(a|t)+p(t|a)$, without considering the input distribution $p(t)$ and $p(a)$, they fail to generalize well to unseen and out-of-domain data.

2.2. DiffATR: Generation Modeling

As shown in Figure 1, we approach the ATR task from a generative modeling perspective. Drawing inspiration from the diffusion models [25, 26], we employ the diffusion model as the generator. Specifically, given a query and N candidates, we aim to generate the distribution $x^{1:N}=\left\{x^i\right\}_{i=1}^N$ from Gaussian noise $\mathcal{N}(0,\mathrm{I})$. In contrast to the prior works that primarily focus on optimizing the posterior probabilities $p\left(a|t;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)+p\left(t|a;\boldsymbol{\theta_t},\boldsymbol{\theta_a}\right)$, our approach constructs the joint probabilities:

$$x^{1:N}=p(a,t|\phi)=f_\phi(a,t,\mathcal{N}(0,\mathrm{I})) \quad (3)$$

where ϕ represents the generator's parameter. Notably, the generator's learning objective is equivalent to approximating the data distribution, i.e., $\prod_{(t,a)\in\mathcal{D}}p(a,t)$, which is termed generative training [7].

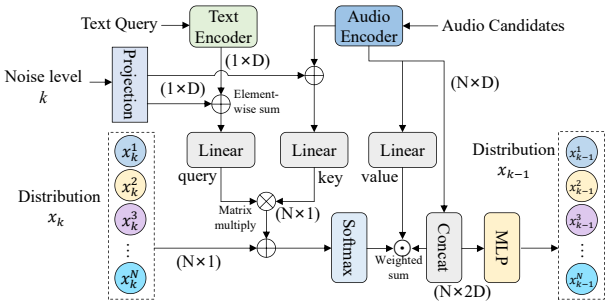


Figure 2: *Model Architecture of the Attention-based Denoising Module.*

2.2.1. Attention-based Denoising Module

Different from general audio generation tasks which mainly prioritize the authenticity and diversity of generated samples, the key to ATR task lies in mining the correspondence between the query and candidates. To this end, we introduce an attention-based denoising module to capture the relationship between query and candidates during the generation process. The overview of our DiffATR denoising module is shown in Figure 2.

For the text-to-audio retrieval task, we first project the text representation $C_t\in\mathbb{R}^D$ extracted from the text encoder into the query and the audio representation $C_a\in\mathbb{R}^{N\times D}$ from the audio encoder into key and value, where N denotes the number of audio candidates and D is the number of feature channels. The projections are represented as: $Q_t=W_{Q_t}\left(C_t+pj(k)\right)$, $K_a=W_{K_a}\left(C_a+pj(k)\right)$, $V_a=W_{V_a}\left(C_a+pj(k)\right)$, where W_{Q_t},W_{K_a} and W_{V_a} are projection matrices. The projection function $pj(\cdot)$ transforms the noise level k into D dimensional embedding. To give more weights to the audio candidates with higher joint probabilities of the previous noise level, we add the distribution x_k to the attention weight. The attention mechanism is formulated as: $E_t=\mathrm{Softmax}\left(Q_tK_a^\top+x_k\right)V_a$.

The attention module's output E_t serves as a high-level semantic embedding infused with textual query information. Then, we concatenate the audio representation C_a and the embedding E_t as the input of the denoising decoder $[C_a,E_t]\in\mathbb{R}^{N\times 2D}$. The denoising decoder is a multi-layer perceptron (MLP) including two linear layers with a Relu activation function to compute the output distribution.

Similarly, for audio-to-text retrieval, we use the projected audio representation as query and the projected text representation as key and value into our attention-based denoising module. The output distribution is computed in the same way.

2.2.2. Generation-based Optimization

From the generation viewpoint, we model the distribution $x=p(a,t|\phi)$ as the reversed diffusion process of a Markov Chain with length K. We learn the joint distribution of text and audio by iteratively denoising a variable sampled from Gaussian distribution. In a forward diffusion process $q\left(x_k|x_{k-1}\right)$, the noise sampled from Gaussian distribution is added to a ground truth data distribution x_0 at every noise level k:

$$\begin{aligned}q\left(x_k|x_{k-1}\right)&=\mathcal{N}\left(x_k;\sqrt{1-\beta_k}x_{k-1},\beta_k\mathrm{I}\right),\\ q\left(x_{1:K}|x_0\right)&=\prod_{k=1}^{K}q\left(x_k|x_{k-1}\right),\end{aligned} \quad (4)$$

Table 1: *Performance comparison of our proposed method with baseline methods on the AudioCaps dataset.*

Methods	AudioCaps					
	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ResNet38 (baseline) [4]	33.9±0.4	69.7±0.2	82.6±0.3	39.4±1.0	72.0±1.0	83.9±0.6
CNN14 (baseline) [4]	31.4±0.4	66.2±0.3	78.8±0.2	38.2±0.5	68.7±0.5	81.9±0.3
ResNet38 (Diff)	**35.6±0.3**	**71.3±0.2**	**84.1±0.3**	**41.1±0.2**	**73.4±0.1**	**85.3±0.5**
CNN14 (Diff)	**33.7±0.1**	**68.2±0.4**	**80.6±0.6**	**39.4±0.3**	**70.1±0.2**	**83.4±0.2**

Table 2: *Performance comparison of our proposed method with baseline methods on the Clotho dataset.*

Methods	Clotho					
	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ResNet38 (baseline) [4]	14.4±0.4	36.6±0.2	49.9±0.2	16.2±0.7	37.5±0.9	50.2±0.7
CNN14 (baseline) [4]	13.9±0.3	34.3±0.5	48.2±0.8	14.3±0.8	35.9±0.6	49.9±0.5
ResNet38 (Diff)	**16.2±0.4**	**37.8±0.2**	**51.6±0.2**	**17.7±0.3**	**39.1±0.3**	**51.4±0.2**
CNN14 (Diff)	**15.3±0.2**	**35.9±0.3**	**49.3±0.1**	**16.1±0.4**	**37.7±0.1**	**51.3±0.5**

where β_k decides the step size which gradually increases, the ground truth vector x_0 is defined such that only the candidate which matters the query is one, others are all zeros (e.g., $[1, 0, 0, ..., 0]$). We can sample x_k by the following formula:

$$x_k = \sqrt{\alpha_k^*} x_0 + \sqrt{1 - \alpha_k^*} \epsilon_k, \tag{5}$$

where $\alpha_k^* = \prod_{i=1}^{k} \alpha_i$ and $\alpha_k = 1 - \beta_k . \epsilon_k$ is a noise sampled from $\mathcal{N}(0, \mathrm{I})$. We follow [27] to predict the data distribution itself, i.e., $\hat{x}_0 = f_\phi(a, t, x_k)$. The training objective of the diffusion model is formulated as:

$$\mathcal{L}_G = \mathbb{E}_{(t,a)\in\mathcal{D}} \left[\mathrm{KL} \left(\hat{x}_0 | x_0 \right) \right] \tag{6}$$

This loss maximizes the likelihood of $p(a, t)$ by bringing $f_\phi(a, t, x_k)$ and x_0 closer together.

2.2.3. Discrimination-based Optimization

From the discrimination perspective, we leverage the contrastive learning method to optimize the audio/text features so that these features hold discriminative semantic information. For a text-audio pair (t, a), the similarity of the text and audio is measured by the cosine similarity of their embeddings:

$$s(t, a) = \frac{f_{\boldsymbol{\theta^t}}(t)^\top \cdot f_{\boldsymbol{\theta^a}}(a)}{\| f_{\boldsymbol{\theta^t}}(t) \|_2 \| f_{\boldsymbol{\theta^a}}(a) \|_2}. \tag{7}$$

Then, the contrastive loss can be formulated as:

$$\begin{aligned} \mathcal{L}_D = -\frac{1}{2} \mathbb{E}_{(t,a)\in\mathcal{D}} \Bigg[& \log \frac{\exp\left(s(t, a)/\tau\right)}{\sum_{a' \in \boldsymbol{A}} \exp\left(s(t, a')/\tau\right)} \\ & + \log \frac{\exp\left(s(t, a)/\tau\right)}{\sum_{t' \in \boldsymbol{T}} \exp\left(s(t', a)/\tau\right)} \Bigg], \end{aligned} \tag{8}$$

where $\boldsymbol{T}$ represents the set of text candidates. $f_{\boldsymbol{\theta^t}}(\cdot)$, $f_{\boldsymbol{\theta^a}}(\cdot)$ denote the text and audio encoders, respectively. This loss narrows the gap between semantically related text and audio in their representation space, which also helps the diffusion model to generate the joint distribution of text and audio.

Finally, the total loss for our DiffATR framework is: $\mathcal{L}_{total} = \mathcal{L}_D + \lambda \mathcal{L}_G$, which effectively combines the advantages of both generative and discriminative paradigms.

3. Experiments

3.1. Datasets

In our experiment, we leverage two publicly accessible datasets: AudioCaps [28] and Clotho [29]. The AudioCaps dataset encompasses around 50K 10-second audio clips, which are sourced from AudioSet [30], the largest dataset for audio tagging. Within this collection, the training subset comprises 49,274 audio instances, with each having a unique human-derived caption. Both validation and test subsets include 494 and 957 audio instances respectively, and each instance is paired with five human-generated captions. Clotho v2, sourced from the Freesound platform [31], features audio clips uniformly ranging from 15 to 30 seconds in length. Every instance in this dataset is coupled with five descriptive sentences. The distribution for training, validation, and testing subsets amounts to 3,839, 1,045, and 1,045 instances, respectively.

3.2. Training Details and Evaluation metrics

In our work, we follow the same pipeline in [4] to train our network, where the batch size is set to 32 for the AudioCaps dataset and 24 for the Clotho dataset. The number of feature channels is 512 and the diffusion steps are set as K = 50. λ is set as 1. We utilize BERT [32] for text encoding, while employing ResNet38 and CNN14 from Pre-trained audio neural networks (PANNs) [33] for audio encoding. We perform experiments by fine-tuning the pre-trained models. The training is divided into two phases. In the first phase, we train the feature encoder only using the contrastive loss. In the second phase, we freeze the feature encoder and only train the diffusion model, that is, we augment the cosine similarity with the similarity estimated by the diffusion model to get the final similarity. We repeat each experiment three times with different training seeds, presenting the mean and standard deviation of the metrics. We evaluate using the Recall at rank k (R@k) metric, indicating the percentage of relevant targets in the top-k ranked results. The results of R@1, R@5, and R@10 are presented.

3.3. Experimental Results

In this part, we first compare the performance of our DiffATR with previous baselines in Table 1 and Table 2. We adopt either ResNet38 or CNN14 as the audio encoder on the AudioCaps and Clotho datasets. It is clear that our DiffATR brings sig-

Table 3: *Text-to-audio and audio-to-text retrieval performances in out-of-domain retrieval settings. "Clotho → AudioCaps" indicates that "Clotho" is the source domain (the training dataset) and "AudioCaps" is the target domain (the evaluation dataset).*

Methods	Clotho → AudioCaps					
	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ResNet38 (baseline)	12.4±0.3	35.8±0.5	50.6±0.4	15.0±0.2	39.9±0.6	52.1±0.8
CNN14 (baseline)	11.5±0.6	34.9±0.5	48.8±0.3	12.2±0.9	36.4±0.8	49.9±0.4
ResNet38 (Diff)	**16.6±0.3**	**39.3±0.1**	**54.1±0.2**	**18.8±0.7**	**43.1±0.5**	**55.9±0.6**
CNN14 (Diff)	**14.9±0.8**	**38.2±1.0**	**52.6±0.3**	**15.6±0.5**	**39.6±0.8**	**53.3±0.1**

Table 4: *Text-to-audio and audio-to-text retrieval performances in out-of-domain retrieval settings. "AudioCaps → Clotho" indicates that "AudioCaps" is the source domain (the training dataset) and "Clotho" is the target domain (the evaluation dataset).*

Methods	AudioCaps → Clotho					
	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ResNet38 (baseline)	9.1±0.3	27.2±0.1	39.4±0.5	11.2±0.8	29.0±1.2	39.8±0.3
CNN14 (baseline)	8.3±0.2	25.9±0.4	37.5±0.7	9.2±0.4	26.2±0.3	38.1±0.4
ResNet38 (Diff)	**12.3±0.1**	**30.6±0.3**	**42.7±0.8**	**14.3±0.1**	**32.7±0.7**	**43.1±0.8**
CNN14 (Diff)	**11.4±0.5**	**29.3±0.6**	**41.1±0.3**	**13.1±0.2**	**29.4±0.1**	**41.9±0.4**

Table 5: *Effect of the training strategy on the AudioCaps dataset.*

Method	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
Gen	26.2	62.8	76.7	31.9	66.7	78.6
Dis	33.9	69.7	82.6	39.4	72.0	83.9
Gen + Dis	**35.6**	**71.3**	**84.1**	**41.1**	**73.4**	**85.3**

Table 6: *Effect of the diffusion steps on the AudioCaps dataset.*

Method	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
10	35.3	71.0	83.9	40.6	73.2	85.1
50	**35.6**	**71.3**	**84.1**	**41.1**	**73.4**	**85.3**
100	35.2	70.8	83.7	40.5	73.1	85.0
200	35.2	70.9	83.5	40.3	73.1	85.1

nificant boosts over different audio encoders for both datasets, which shows the effectiveness of our method.

While existing ATR methods are typically trained and evaluated on the same dataset, we also explore out-of-domain ATR for deeper insights into the adaptability and generalization of our DiffATR framework. Here, we first train our model on one dataset (the "source") and then measure its performance on another dataset (the "target") that is unseen in the training. As shown in Table 3 and Table 4, we find that our DiffATR notably surpasses the corresponding discriminative approaches in out-of-domain ATR, which strongly demonstreates that our method is more generalizable and transferrable for the unseen data than conventional discriminative models.

3.4. Ablation Study

In this part, we examine the impact of different training strategies, the number of diffusion steps, the generative loss weight, and the training batch size, utilizing ResNet38 as the audio encoder.

In Table 5, we compare different training strategies on the AudioCaps dataset. We observe that solely discriminative (Dis) training outperforms purely generative (Gen) training. The hybrid training method we present achieves the best results, confirming its capability to combine the strengths of both approaches. Table 6 studies the effect of diffusion step numbers. The results highlight that 50 steps is the sweet spot for the ATR task, surpassing the common range of 100-1000 steps used in audio generation tasks [11, 34]. We consider that this is due to the simpler probability distribution in retrieval compared to the generated audio signal distribution. Consequently, the ATR task demands fewer diffusion steps relative to generation-oriented tasks.

Table 7: *Effect of the loss weight λ on the Clotho dataset.*

λ	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
0.9	16.1	37.6	51.4	17.6	38.8	51.4
1.0	**16.2**	**37.8**	**51.6**	**17.7**	**39.1**	**51.4**
1.1	15.9	37.2	51.1	17.2	38.6	51.1

Table 8: *Effect of the training batch size on the Clotho dataset.*

batch size	Text-to-Audio			Audio-to-Text		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
16	15.8	37.3	51.2	17.4	38.7	50.8
24	**16.2**	**37.8**	**51.6**	**17.7**	**39.1**	**51.4**
32	16.1	37.6	51.9	17.6	39.3	51.2

Besides, we study the generative loss weight λ on the Clotho dataset. As shown in Table 7, the optimal value for λ is 1. This demonstrates that when the contrastive and generative loss functions contribute equally, we can get higher results as these two loss functions could act as complementary to each other. In Table 8, we can see that different training batch size settings would affect the results, but not too much, with the most favorable outcomes manifesting at a batch size of 24.

4. Conclusions

In this paper, we present DiffATR, a generative diffusion-based framework for ATR. By explicitly modeling the joint probability distribution of text and audio, DiffATR optimizes the ATR network from both a generative and discriminative perspective to handle the intrinsic limitations of the current discriminative ATR paradigm. Experiments on the AudioCaps and Clotho datasets with superior performances, justify the effectiveness of our DiffATR. Remarkably, DiffATR also excels in out-of-domain retrieval scenarios without necessitating any modifications.

5. References

[1] A. S. Koepke, A.-M. Oncescu, J. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, 2022.

[2] Y. Xin and Y. Zou, "Improving Audio-Text Retrieval via Hierarchical Cross-Modal Interaction and Auxiliary Captions," in *Proc. INTERSPEECH 2023*, 2023, pp. 341–345.

[3] Y. Xin, B. Wang, and L. Shang, "Cooperative game modeling with weighted token-level alignment for audio-text retrieval," *IEEE Signal Processing Letters*, 2023.

[4] X. Mei, X. Liu, J. Sun, M. Plumbley, and W. Wang, "On Metric Learning for Audio-Text Cross-Modal Retrieval," in *Proc. Interspeech 2022*, 2022, pp. 4142–4146.

[5] X. Cheng, W. Xu, Z. Zhu, H. Li, and Y. Zou, "Towards spoken language understanding via multi-level multi-grained contrastive learning," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 326–336.

[6] X. Cheng, Z. Zhu, Y. Li, H. Li, and Y. Zou, "Das-cl: Towards multimodal machine translation via dual-level asymmetric contrastive learning," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 337–347.

[7] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, "Generative or discriminative? getting the best of both worlds," *Bayesian statistics*, vol. 8, no. 3, pp. 3–24, 2007.

[8] C. Liang, W. Wang, J. Miao, and Y. Yang, "Gmmseg: Gaussian mixture based generative semantic segmentation models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 360–31 375, 2022.

[9] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[11] 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 *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 21 450–21 474.

[12] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.

[13] X. Cheng, Q. Dong, F. Yue, T. Ko, M. Wang, and Y. Zou, "M 3 st: Mix at three levels for speech translation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[16] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[18] X. Cheng, Z. Yao, Z. Zhu, Y. Li, H. Li, and Y. Zou, "C 2 a-slu: cross and contrastive attention for improving asr robustness in spoken language understanding," in *Proc. of INTERSPEECH*, 2023.

[19] W.-H. Li, S. Yang, Y. Wang, D. Song, and X.-Y. Li, "Multi-level similarity learning for image-text retrieval," *Information Processing & Management*, vol. 58, no. 1, p. 102432, 2021.

[20] X. Cheng, Z. Zhu, B. Cao, Q. Ye, and Y. Zou, "Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 10 495–10 505.

[21] X. Cheng, Z. Zhu, H. Li, Y. Li, X. Zhuang, and Y. Zou, "Towards multi-intent spoken language understanding via hierarchical attention and optimal transport," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 844–17 852.

[22] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, p. 100336, 2021.

[24] X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou, "Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 6492–6505.

[25] Y. Zhang, T. Zhang, D. Chen, Y. Wang, Q. Chen, X. Xie, H. Sun, W. Deng, Q. Zhang, F. Yang *et al.*, "Irgen: Generative modeling for image retrieval," *arXiv preprint arXiv:2303.10126*, 2023.

[26] P. Jin, H. Li, Z. Cheng, K. Li, X. Ji, C. Liu, L. Yuan, and J. Chen, "Diffusionret: Generative text-video retrieval with diffusion model," in *ICCV*, 2023, pp. 2470–2481.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[28] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[29] 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.

[30] 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.

[31] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[32] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.

[33] 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.

[34] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.