Generating Accurate and Diverse Audio Captions through Variational Autoencoder Framework

Yiming Zhang, Ruoyi Du, Zheng-Hua Tan, *Senior Member, IEEE*,
Wenwu Wang, *Senior Member, IEEE*, Zhanyu Ma, *Senior Member, IEEE*

***Abstract*—Generating both diverse and accurate descriptions is an essential goal in the audio captioning task. Traditional methods mainly focus on improving the accuracy of the generated captions but ignore their diversity. In contrast, recent methods have considered generating diverse captions for a given audio clip, but with the potential trade-off in caption accuracy. In this work, we propose a new diverse audio captioning method based on a variational autoencoder structure, dubbed AC-VAE, aiming to achieve a better trade-off between the diversity and accuracy of the generated captions. To improve diversity, AC-VAE learns the latent word distribution at each location based on contextual information. To uphold accuracy, AC-VAE incorporates an autoregressive prior module and a global constraint module, which enable precise modeling of word distribution and encourage semantic consistency of captions at the sentence level. We evaluate the proposed AC-VAE on the Clotho dataset. Experimental results show that AC-VAE achieves a better trade-off between diversity and accuracy compared to the state-of-the-art methods. The code is publicly available at https://github.com/XinMing0411/AC-VAE**

***Index Terms*—Diverse audio captioning, variational autoencoder, diverse caption generation**

I. Introduction

AUDIO captioning (AC) is a multimodal task aiming at generating natural language descriptions for audio clips [1]–[5].AC aims to obtain human-like descriptions by summarizing the audio events and scenes within an audio clip, and the semantic relationships between them [6], [7].

Thanks to the success of the DCASE Challenges [8], significant progress on AC has been made recently. Most existing AC methods [9]–[19] use maximum likelihood estimation (MLE) or reinforcement learning (RL) to improve the accuracy of the generated captions, by measuring the similarity between the generated captions and human-annotated ground-truth (GT). As an example, Table I shows the captions generated by different methods that describe the same audio clip. While the MLE and RL methods describe the audio clip accurately, the generated captions are general and lack diversity. In contrast, due to the ambiguity of audio and the complexity of linguistic representations, people describe the content of audio in greater diversity, e.g., with richer vocabulary, more flexible grammatical structures.

TABLE I
CAPTIONS GENERATED BY HUMAN AND TRADITIONAL METHODS.

Model	Generated Text
MLE	it is *raining* and the *rain* is *hitting* the ground. it is *raining* and the *rain* is *hitting* the pavement. it is *raining* and the *rain* is *pouring* down.
RL	a *fire* is *crackling* and a in a. a *fire* is *crackling* and a in the. a *fire* is *crackling* on a roof and a.
GT	a *fire* lowly *crackles* and *pops* on occasion. the *fire* *snaps* and *crackles* as the *log* begins to *burn down*. the *raindrop* *hitting* the ground is *crackling*.

To address the above issues and to generate captions that are more naturally aligned with human perception, Mei *et al.* [20], [21] and Xu *et al.* [22] proposed diverse audio captioning methods based on generative models. They utilized conditional generative adversarial networks (cGANs) with a combination of multiple discriminators, operating at the coarse-grained sentence level to ensure the quality of generated captions. This leads to captions of good diversity, but at certain cost of captioning accuracy.

In this work, we leverage the capability of the variational autoencoder (VAE) [23] architecture and propose AC-VAE to achieve accurate and diverse audio captioning. Compared to GAN-based methods, which can only implicitly model at the coarse-grained sentence level, AC-VAE explicitly models the semantic distribution of each word. During generation, it randomly samples from the distribution to ensure the diversity of the generated captions in a more fine-grained way. Also, to ensure accuracy, we propose the autoregressive prior module and the global constraint module to maintain semantic coherence between the generated sentences and the ground-truth captions. In addition, the previous metrics fail to measure the trade-off between accuracy and diversity, and only evaluate them individually. To address these issues, we propose the diversity-accuracy equilibrium score (DAES) to evaluate the overall accuracy-diversity performance.

Our contributions are threefold:

1) We present a novel VAE-based diverse audio captioning method, dubbed AC-VAE, to generate accurate and diverse audio captions.
2) We introduce a new metric, dubbed DAES, designed to assess the accuracy-diversity trade-off of AC models.
3) Comparative experiments were conducted on the Clotho

Y. Zhang, R. Du, and Z. Ma are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: {zhangyiming, duruoyi, mazhanyu}@bupt.edu.cn.

Z.-H. Tan is with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark. E-mail: zt@es.aau.dk.

W. Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, United Kingdom. E-mail: w.wang@surrey.ac.uk.

This work was supported in part by Beijing Natural Science Foundation Project No. Z200002, and in part by National Natural Science Foundation of China (NSFC) No. 62225601, U23B2052, and in part by Youth Innovative Research Team of BUPT No. 2023QNTD02, and in part by the scholarship from China Scholarship Council No. 202306470064.

(Corresponding author: Zhanyu Ma)

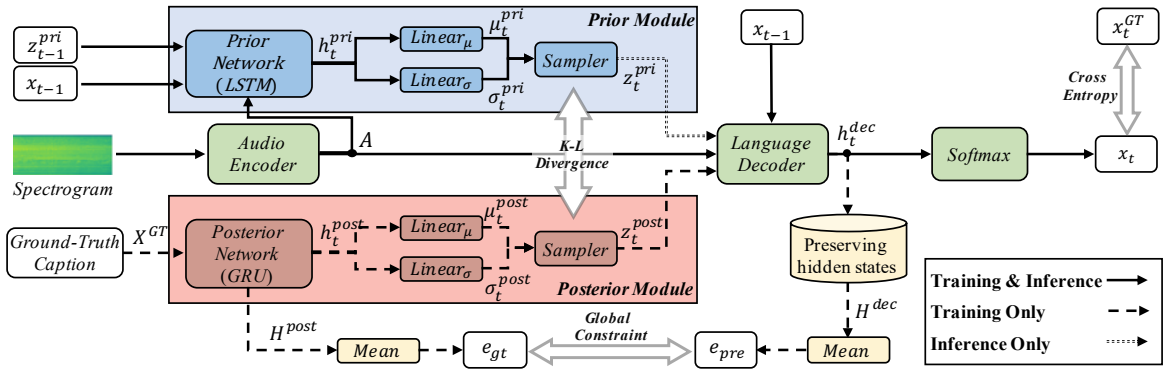

Fig. 1. The architecture of our proposed AC-VAE. The posterior module models the word latent distribution at each position with contextual information of Ground-Truth caption, the autoregressive prior module models the present word distribution based on the generated words, and the global constraint is utilized to ensure the semantic consistency of the generated captions.

dataset [2], revealing that the AC-VAE method outperforms SoTA methods [20], [21] in 6 out of 8 metrics.

II. Proposed Method

A. Overview of the Proposed Method

We formulate the AC task as the problem of estimating the posterior probability of the caption $\boldsymbol{X} = (x_1, \ldots, x_T)$ for a given acoustic feature $\boldsymbol{A}$, denoted as $p(\boldsymbol{X}|\boldsymbol{A})$, and T is the number of words. Instead of directly modeling $p(\boldsymbol{X}|\boldsymbol{A})$, we introduce a latent variable $\boldsymbol{Z} = (z_1, \ldots, z_T)$, which captures the distribution of words at each position t. Therefore, $p(\boldsymbol{X}|\boldsymbol{A})$ can be modelled as variational structure:

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A}) = \prod_{t=1}^{T} p(x_t|x_{\leq t-1}, z_{\leq t}, \boldsymbol{A}). \tag{1}$$

During Training: As shown in Fig. 1, at each position t, the ground-truth caption $\boldsymbol{X}^{GT}$ is fed into the posterior module to obtain the posterior latent z_t^{post}:

$$z_t^{post} \sim q(z_t^{post}|\boldsymbol{X}^{GT}), \tag{2}$$

where $q(z_t^{post}|\boldsymbol{X}^{GT})$ represents the posterior module for modeling the posterior distribution.

The prior latent z_t^{pri} is modeled based on the acoustic feature $\boldsymbol{A}$, the prior latent of the previous position $z_{\leq t-1}^{pri}$, and the generated text $x_{\leq t-1}$ as

$$z_t^{pri} \sim p(z_t^{pri}|z_{\leq t-1}^{pri}, x_{\leq t-1}, \boldsymbol{A}), \tag{3}$$

where $p(z_t^{pri}|z_{\leq t-1}^{pri}, x_{\leq t-1}, \boldsymbol{A})$ is modeled by the prior module.

Then the language decoder utilizes $z_{\leq t}^{post}$ and $\boldsymbol{A}$ to predict the current word x_t as

$$x_t \sim p(x_t|z_{\leq t}^{post}, x_{\leq t-1}, \boldsymbol{A}). \tag{4}$$

According to [23], [24], the variational structure is optimized by maximizing the evidence lower bound (ELBO), and thus the loss function $\mathcal{L}_t^{ELBO}$ for each position t is:

$$\begin{aligned}\mathcal{L}_t^{ELBO} =& KL[q(z_t^{post}|\boldsymbol{X}^{GT})||p(z_t^{pri}|z_{\leq t-1}^{pri}, x_{\leq t-1}, \boldsymbol{A})] \\ &- \mathbb{E}_{q(z_t^{post}|\boldsymbol{X}^{GT})}[\log p(x_t|z_{\leq t}^{post}, x_{\leq t-1}, \boldsymbol{A})],\end{aligned} \tag{5}$$

where the first term is the KL divergence constraint and the second is the cross-entropy reconstruction term.

In addition, we use the global embedding of the target caption e_{gt} to constrain the global embedding of the predicted sentence e_{pre}, thus encouraging them to be semantically consistent at the sentence level. The global constraint loss $\mathcal{L}_{global}$ is $\|e_{pre} - e_{gt}\|^2$.

Therefore, the total loss function $\mathcal{L}$ is

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t^{ELBO} + \alpha \cdot \mathcal{L}_{global}, \tag{6}$$

where the hyperparameter α is set to 0.03 in our experiments. **During Inference:** Since target captions are not available during inference, the posterior module is excluded. The prior latent z_t^{pri} replaces the posterior latent z_t^{post} to generate the current word. Therefore, Equation (4) is reformulated as

$$x_t \sim p(x_t|z_{\leq t}^{pri}, x_{\leq t-1}, \boldsymbol{A}), \tag{7}$$

B. The Caption Posterior Module

We empirically use a bi-directional GRU network and the linear layers to model the posterior distribution of the latent variable $\boldsymbol{Z}^{post} = (z_1^{post}, \ldots, z_T^{post})$.

Specifically, $\boldsymbol{X}^{GT}$ is fed into the posterior module to get the hidden states $\boldsymbol{H}^{post} = (h_1^{post}, \ldots, h_T^{post})$, and the hidden state h_t^{post} is the t-th element of $\boldsymbol{H}^{post}$. Then, h_t^{post} is fed into the linear layers (Linear_μ and Linear_σ), respectively, to obtain the mean μ_t^{post} and standard deviation σ_t^{post} at position t as

$$\mu_t^{post} = \text{Linear}_\mu(h_t^{post}), \tag{8}$$

$$\sigma_t^{post} = \exp(0.5 * \text{Linear}_\sigma(h_t^{post})), \tag{9}$$

where exp is an exponential function and the output of Linear_σ is the logarithmic variance for optimization purposes. Finally, we sample from the posterior distribution to generate z_t^{post}:

$$z_t^{post} = \mu_t^{post} + \epsilon * \sigma_t^{post}, \tag{10}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is a random variable.

C. The Autoregressive Prior Module

Unlike other VAE-based methods [25], [26], we empirically design an autoregressive prior module to model the prior variables $\boldsymbol{Z}^{pri} = (z_1^{pri}, \ldots, z_T^{pri})$, as shown in Fig. 1.

The prior module includes the LSTM network and the linear layers. Given the prior latent z_{t-1}^{pri}, the generated word x_{t-1}, and $\boldsymbol{A}$, the prior network generates the hidden state at the current position h_t^{pri}:

$$a_t^{pri} = \text{Softmax}\left(\boldsymbol{V}^{pri} \cdot \tanh(\boldsymbol{W}^{pri} \cdot [\boldsymbol{A}; x_{t-1}])\right) \cdot \boldsymbol{A}, \tag{11}$$

$$h_t^{pri} = \text{LSTM}\left(\left[x_{t-1}; z_{t-1}^{pri}; a_t^{pri}\right], h_{t-1}^{pri}\right), \tag{12}$$

where a_t^{pri} is the context audio embedding, which aggregates $\boldsymbol{A}$ using the attention mechanism, $\boldsymbol{V}^{pri}$ and $\boldsymbol{W}^{pri}$ are learnable weights, $[\cdot;\cdot]$ is the concatenation operation, and Softmax and $\tanh$ are the activation functions.Similar to the posterior module, the hidden state h_t^{pri} is then used to generate the prior latent z_t^{pri}.

D. Audio Encoder and Language Decoder

In this work, we use the pre-trained 10-layer convolutional neural network (CNN10) [27] as the audio encoder to extract the audio feature $\boldsymbol{A}$. For the language decoder, a single-layer GRU [28] is used to estimate the word probabilities $p(x_t|z_t, x_{t-1}, \boldsymbol{A})$ at the current position t as

$$a_t^{dec} = \text{Softmax}\left(\boldsymbol{V}^{dec} \cdot \tanh(\boldsymbol{W}^{dec} \cdot [\boldsymbol{A}; x_{t-1}])\right) \cdot \boldsymbol{A}, \quad (13)$$

$$h_t^{dec} = \text{GRU}\left(\left[x_{t-1}; z_t; a_t^{dec}\right], h_{t-1}^{dec}\right), \quad (14)$$

where a_t^{dec} is the context audio embedding of the decoder, $\boldsymbol{V}^{dec}$ and $\boldsymbol{W}^{dec}$ are learnable weights, z_t represents z_t^{post} during training and z_t^{pri} during inference, and h_t^{dec} is the hidden state of the decoder at the current position t. Following a linear layer and the Softmax function, the output word probabilities are obtained.

E. The Global Constraint

As shown in Equation (5), the loss function $\mathcal{L}_t^{ELBO}$ is used for optimizing the model with constraints only on local information captured at the word level. To improve the quality of the generated text of the model, sentence-level semantic constraints are applied.

We preserve the decoder hidden state h_t^{dec} at each position and obtain all the decoder hidden states $\boldsymbol{H}^{dec}$. Then, the global mean pooling is applied on $\boldsymbol{H}^{post}$ and $\boldsymbol{H}^{dec}$ to generate global embeddings e_{gt} and e_{pre}, respectively,

$$e_{gt} = \text{Mean}(\boldsymbol{H}^{post}), e_{pre} = \text{Mean}(\boldsymbol{H}^{dec}), \quad (15)$$

where $\text{Mean}(\cdot)$ represents the global mean pooling, and then e_{gt} and e_{pre} are used in Equation (6) for training the model.

III. Experiment Setup

A. Dataset

We conduct the experiments on the Clotho dataset [2] which is the official benchmark in the DCASE Challenge for the AC task. Each audio clip has five captions. The annotator only uses the audio signals for annotation without additional signal.

B. Implementation Details

We extract a 64-dimensional log-Mel spectrogram from each audio clip as the input to the audio encoder, where the window shift and size are 1024 and 2048 sampling points. The model is trained by Adam optimizer [29] with an initial learning rate of 5×10^{-4}. The warm-up strategy is used to increase the learning rate linearly in the first five epochs. The model is trained for 15 epochs. During inference, we generate five captions using a beam search with a beam size 5.

C. Evaluation Metrics

Similar to [20], [21], we employ accuracy metrics (BLEU_4, CIDEr, and SPIDEr) and diversity metrics (Vocab, mBLEU_4, Div-1, and Div-2) to quantify model performance. Inspired by the F-Score, we devise the DAES to measure the overall accuracy-diversity performance by combining accuracy and diversity metrics sets, normalized relative to human performance:

$$DAES = \frac{2}{3/\sum_{a \in Acc} Norm(a) + 4/\sum_{d \in Div} Norm(d)}. \quad (16)$$

Here, Acc and Div represent the accuracy and diversity metrics sets, respectively. $Norm(\cdot)$ represents the normalization of model performance relative to human performance, typically falling within the range of 0 to 1. It is important to note that for mBLEU_4, smaller values indicate better performance, so the normalization involves dividing the human reference score by the model's score.

Hence the higher DAES, the higher the overall performance.

IV. Results and Analysis

A. Comparison with Baseline Models

The baseline MLE method is a traditional method, which is trained by the cross-entropy loss and shares a similar architecture of encoder and decoder with AC-VAE.

As shown in Table II, the last row represents the human performance, which is calculated by treating one of the five human-annotated captions as the predicted caption, with the remaining four captions serving as references. Subsequently, scores for all five captions are computed in parallel and averaged. The baseline MLE achieves better results in accuracy metrics but performs worse in diversity and DAES metrics. This result is expected, since as previously discussed, MLE encourages the appearance of commonly used n-grams in the target caption, and these accuracy metrics primarily measure the degree of n-grams matching. Compared to the baseline, our proposed method shows a decrease of 0.037 in accuracy metrics (e.g., CIDEr). However, there is a substantial increase in the diversity metric (e.g., Div-2), rising from 0.233 to 0.574, indicating enhanced diversity in the generated captions.

B. Comparison to State-of-the-Art

We compare our proposed AC-VAE with other methods that have achieved state-of-the-art performance in diverse audio captioning tasks using the cGAN framework [20], [21]. These methods generate different captions for the same audio clip by utilizing different noises. As compared to the aforementioned cGAN methods, our proposed AC-VAE achieves better performance in six out of all eight metrics, and especially in the accuracy metrics, our method is significantly superior. The experimental results demonstrate that our method can generate more accurate caption descriptions of the audio content while maintaining diversity. This is crucial for ensuring the faithfulness of the captions generated about the audio clip.

TABLE II
EXPERIMENTAL RESULTS OF THE PROPOSED METHOD COMPARED WITH BASELINE MODEL AND OTHERS' METHODS

Model		$BLEU_4$↑	CIDEr↑	SPIDEr↑	Vocab↑	$mBLEU_4$↓	Div-1↑	Div-2↑	DAES↑
Baseline	MLE	0.153	0.382	0.254	613	0.951	0.212	0.233	0.360
Other works	cGAN [21]	0.119	0.291	0.198	897	**0.432**	**0.423**	0.559	0.448
	cGAN [20]	0.122	0.295	0.199	818	0.615	0.335	0.442	0.411
AC-VAE		**0.130**	**0.345**	**0.230**	**899**	0.442	0.417	**0.574**	**0.488**
Human		0.321	0.901	0.566	3516	0.321	0.561	0.724	1.000

TABLE III
ABLATION STUDIES OF DIFFERENT MODULES

	WD	LA	GC	CIDEr↑	Div-2↑	DAES↑
a.				0.314	0.359	0.382
b.	✓			0.304	0.561	0.445
c.			✓	-	-	-
d.	✓	✓		0.327	0.579	0.476
e.	✓		✓	0.335	0.564	0.478
f.	✓	✓	✓	0.345	0.574	0.488

C. Ablation Studies

In this section, we conduct an ablation study to investigate the contribution of the mechanisms in our proposed method. The results are summarized in Table III. "WD" is the word-level distributional constraint mechanism, "LA" is the learnable autoregressive prior module, and the "GC" represents the global constraint module.

Row *a* in Table III represents the vanilla conditional VAE-based AC method, which models the distribution at the sentence level, like the GAN-based methods [20], [21]. Row *b* means that the word distribution at each position follows a standard normal distribution. From the experimental results, we can find that Row *b* achieves significant improvement in diversity and DAES metrics compared to Row *a*. This is because the captions generated by sampling on fine-grained word distributions can be more diverse compared to the vanilla cVAE modeling distributions at the coarse-grained sentence level. Row *c* shows that the model fails to converge when only the global constraint module is used.

Comparing Row *b* with Rows *d* and *e*, respectively, we find that the autoregressive prior module can improve the model performance on almost all metrics because it accurately models word distributions with additional audio information and semantic information of the generated words. The global constraint module provides additional sentence-level information from the ground-truth captions as an extra constraint target. Using the above modules, AC-VAE (in Row *f*) achieves the best overall performance.

D. How Diversity Is Derived?

To examine the influence of latent variables on caption generation and understand the factors contributing to the achieved diversity, we conduct an analysis for the latent variables. As shown in Fig. 2, the $[\mu - 2\sigma, \mu + 2\sigma]$ interval of the latent variable distributions at different locations ($t = 2, 4, 6$) are sampled uniformly. It can be found that the latent variable distributions at each location can facilitate the generation of multiple words, and each word corresponds to a different distribution interval. In the inference stage, the latent variables generated by random sampling cause the model to generate different words at the same word location and affect the distribution of latent variables at subsequent word locations.

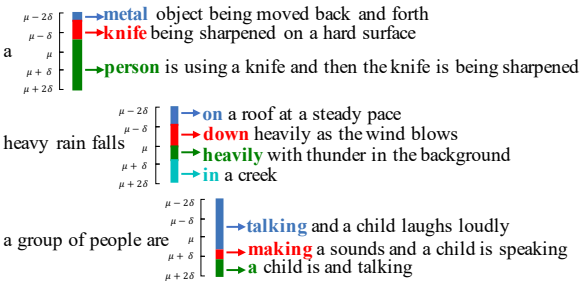

Fig. 2. The corresponding words of the latent space for three different examples of AC-VAE. Different colors represent different words.

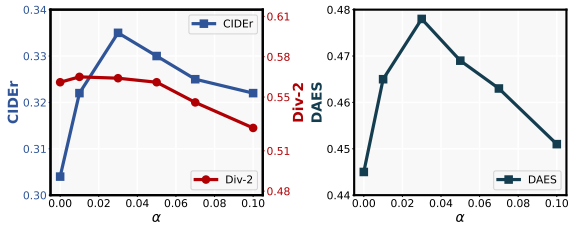

Fig. 3. The effect of hyperparameter α on performance metrics. The left figure shows the experimental results of CIDEr and Div-2, and the right figure shows the experimental results of DAES.

E. The Effect of Different Hyperparameter α

Fig. 3 shows the results of varying the hyperparameter α. We can find that when α increases, the accuracy (CIDEr) is further improved, but the impact on the diversity (Div-2) is little, meaning that with the introduction of global information, the generated captions can describe the audio contents more accurately. When α continues to increase, the performance of the model decreases significantly, and the overall performance of the model, DAES, is highest when α is 0.03.

V. CONCLUSION

We have presented a new variational framework-based audio captioning method called AC-VAE. To improve diversity, we model the semantic distribution at the word position level in a more fine-grained way than at the sentence level. To ensure accuracy, we introduce the autoregressive prior module and the global constraint module to maintain semantic coherence between the generated sentences and the ground-truth captions. Experimental results show that our proposed model can generate captions with better semantic consistency and comparable diversity as compared with SOTA methods.

REFERENCES

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

[2] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[3] Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito, "A Transformer-Based Audio Captioning Model with Keyword Estimation," in *Proc. Interspeech 2020*, 2020, pp. 1977–1981.

[4] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, pp. 1–18, 2022.

[5] X. Xu, Z. Xie, M. Wu, and K. Yu, "Beyond the status quo: A contemporary survey of advances and challenges in audio captioning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[6] X. Xu, H. Dinkel, M. Wu, and K. Yu, "Audio caption in a car setting with a sentence-level loss," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[7] Y. Zhang, H. Yu, R. Du, Z.-H. Tan, W. Wang, Z. Ma, and Y. Dong, "Actual: Audio captioning with caption feature space regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[8] S. Lipping, K. Drossos, and T. Virtanen, "Crowdsourcing a dataset of audio captions," *arXiv preprint arXiv:1907.09238*, 2019.

[9] M. Kim, K. Sung-Bin, and T.-H. Oh, "Prefix tuning for automated audio captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] E. Labbé, T. Pellegrini, and J. Pinquier, "Irit-ups dcase 2022 task6a system: stochastic decoding methods for audio captioning," DCASE Challenge, Tech. Rep, Tech. Rep., 2022.

[11] Z. Ye, H. Wang, D. Yang, and Y. Zou, "Improving the performance of automated audio captioning via integrating the acoustic and semantic information," *arXiv preprint arXiv:2110.06100*, 2021.

[12] X. Liu, Q. Huang, X. Mei, H. Liu, Q. Kong, J. Sun, S. Li, T. Ko, Y. Zhang, L. H. Tang *et al.*, "Visually-aware audio captioning with adaptive audio-visual attention," *arXiv preprint arXiv:2210.16428*, 2022.

[13] A. Koh, X. Fuzhao, and C. E. Siong, "Automated audio captioning using transfer learning and reconstruction latent space similarity regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7722–7726.

[14] X. Xu, H. Dinkel, M. Wu, and K. Yu, "A CRNN-GRU based reinforcement learning approach to audio captioning." in *DCASE*, 2020, pp. 225–229.

[15] X. Mei, Q. Huang, X. Liu, G. Chen, J. Wu, Y. Wu, J. Zhao, S. Li, T. Ko, H. L. Tang *et al.*, "An encoder-decoder based audio captioning system with transfer and reinforcement learning," *arXiv preprint arXiv:2108.02752*, 2021.

[16] X. Liu, Q. Huang, X. Mei, T. Ko, H. L. Tang, M. D. Plumbley, and W. Wang, "Cl4ac: A contrastive loss for audio captioning," *arXiv preprint arXiv:2107.09990*, 2021.

[17] C. Chen, N. Hou, Y. Hu, H. Zou, X. Qi, and E. S. Chng, "Interactive audio-text representation for automated audio captioning with contrastive learning," *arXiv preprint arXiv:2203.15526*, 2022.

[18] F. Xiao, J. Guan, H. Lan, Q. Zhu, and W. Wang, "Local information assisted attention-free decoder for audio captioning," *IEEE Signal Processing Letters*, vol. 29, pp. 1604–1608, 2022.

[19] Q. Han, W. Yuan, D. Liu, X. Li, and Z. Yang, "Automated audio captioning with weakly supervised pre-training and word selection methods." in *DCASE*, 2021.

[20] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Diverse audio captioning via adversarial training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8882–8886.

[21] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Towards generating diverse audio captions via adversarial training," *arXiv preprint arXiv:2212.02033*, 2022.

[22] X. Xu, M. Wu, and K. Yu, "Diversity-controllable and accurate audio captioning based on neural condition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 971–975.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations, ICLR*, 2014.

[24] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[25] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] J. Xu, B. Liu, Y. Zhou, M. Liu, R. Yao, and Z. Shao, "Diverse image captioning via conditional variational autoencoder and dual contrastive learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 1, pp. 1–16, 2023.

[27] 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.

[28] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.