Zero-shot audio captioning with audio-language model guidance and audio context keywords

Leonard Salewski **Stefan Fauth** **A. Sophia Koepke** **Zeynep Akata**

University of Tübingen, Tübingen AI Center

Abstract

Zero-shot audio captioning aims at automatically generating descriptive textual captions for audio content without prior training for this task. Different from speech recognition which translates audio content that contains spoken language into text, audio captioning is commonly concerned with ambient sounds, or sounds produced by a human performing an action. Inspired by zero-shot image captioning methods, we propose ZerAuCap, a novel framework for summarising such general audio signals in a text caption without requiring task-specific training. In particular, our framework exploits a pre-trained large language model (LLM) for generating the text which is guided by a pre-trained audio-language model to produce captions that describe the audio content. Additionally, we use audio context keywords that prompt the language model to generate text that is broadly relevant to sounds. Our proposed framework achieves state-of-the-art results in zero-shot audio captioning on the AudioCaps and Clotho datasets. Our code is available at `https://github.com/ExplainableML/ZerAuCap`.

1 Introduction

The growing demand for seamless human-computer interaction has increased the interest in translating audio and visual information into text. As a result, improving frameworks that automatically caption audio and visual information has gained attention. This has resulted in systems for various applications that try to improve accessibility for visually impaired or hearing impaired people [27, 13, 14].

Audio captioning aims at translating general audio signals into textual descriptions. This is different from speech recognition which transforms speech into written text. Common audio captioning methods rely on supervised learning, requiring extensive training data that spans across various audio categories. This requirement limits the adaptability of trained models to new audio contexts which is crucial as the amount and diversity of newly created audio content continue to grow.

Concurrently to our work, [32] introduced the zero-shot audio captioning task which tackles the aforementioned challenges and limitations. Their proposed approach seeks to generate textual descriptions for audio content by using guidance through an audio-text matching score that determines if the generated caption describes audible content. They use this information to iteratively optimize the key-value pairs of the LLM that generates the caption. Different to this, our proposed ZerAuCap framework employs a much simpler approach: We first determine a set of keywords that captures the audible events of the given soundusing a pre-trained audio-text matching network. These audio context keywords guide the text generation by being part of the input prompt for an LLM that is tasked to write a caption. Inspired by [31, 34] that proposed a visually-guided zero-shot text generation framework using vision-text matching information, our framework uses guidance from audio-text matching scores combined with the language model's word probability to select the next word out of a list of candidate words proposed by the LLM. Our approach produces better results at lower computational cost than [32].

37th Conference on Neural Information Processing Systems (NeurIPS 2023) - ML for Audio Workshop.

Our ZerAuCap framework achieves state-of-the-art results on the audio captioning benchmark AudioCaps [15], significantly outperforming [32]. In addition, we set a new benchmark for zero-shot audio captioning on the Clotho dataset [7], outperforming our baselines.

To summarize, we make the following contributions: (1) We propose a novel framework for zero-shot audio captioning that uses a two-fold guidance approach to steer the language generation of a LLM. (2) Our framework achieves state-of-the-art results for zero-shot audio captioning on the AudioCaps [15] and Clotho [7] datasets. (3) We show that both guiding approaches contribute to the final performance of our approach.

2 Related Work

Audio captioning. The audio captioning task consists of generating a textual description for a given sound [6]. This is closely related to audio-text retrieval which aims at learning to map audio and textual representations to a joint embedding space, allowing retrieval of a matching caption for a given audio snippet and vice versa [28, 41, 33, 17, 23, 40]. Audio captioning and audio-text retrieval have been popularized by the introduction of audio-text datasets, such as Clotho [7], AudioCaps [15], Audio Caption [39], and WavCaps [25]. In particular, the recurring DCASE audio captioning challenge [42] uses the Clotho dataset for benchmarking. Many frameworks have tackled automatic audio captioning using supervision from audio-text pairs in the AudioCaps and Clotho datasets [18, 43, 8, 24, 22, 16, 11, 4, 3]. Different from those works, we consider the zero-shot audio captioning setting that goes beyond zero-shot audio classification [12] and that does not make use of task-specific training on audio captioning data. Instead, we leverage pre-trained audio-language models [26] to guide the caption generation process.

Zero-shot captioning. Recently, several zero-shot image captioning methods were proposed [34, 36, 35, 44, 38] which use CLIP [30] to guide text generation with an LLM. A first line of work optimizes hidden activations in the language model to adapt the predicted tokens towards the captioning target [36, 35]. In contrast, our approach avoids this costly step and directly chooses the next token based on a fitness function. A second line of work chooses the next token by introducing guidance in the decoding process [44, 34]. [44] conditions the LLM on CLIP-detected class names and then ranks multiple sampled captions based on CLIP similarity, whilst [34] builds a sentence token by token and selects the next token based on its CLIP similarity to the image. In contrast, [31] masks the image with the attention patterns of a VQA model and uses the same model for guiding the translation of its attention patterns into natural language. A third line of work finetunes language models in an unsupervised way to understand CLIP text embeddings which at test time are replaced by CLIP image embeddings [38]. [19] uses the same pre-training approach and proposes a simple approach to close the modality gap in a training-free manner. Unlike the previous two approaches, our model does not rely on finetuning. Concurrent to our work, and most closely related, [32] introduced the zero-shot audio captioning task. Different from their work we do not optimize the hidden states of the LLM, but choose the next token according to language model plausibility and audio-relevancy.

3 Method

In this section, we explain how our ZerAuCap framework (**Zer**o-shot **Au**dio **Cap**tioning) automatically generates audio captions for audio clips without task-specific training, i.e. in a zero-shot manner. We use an audio-language model, pre-trained with a contrastive training objective, for two distinct purposes: *zero-shot keyword selection* and *audio-relevancy guiding*. First, we select a set of top-l keywords that have high similarity to the input audio clip. These keywords are composed of very short natural language descriptions of various different audible effects, e.g. sounds produced by humans or other living beings, or environmental sounds. We then provide the top keywords to the LLM to condition its subsequent generation of an audio caption on the audible concepts contained within the audio clip. This setup exploits the world knowledge and the capability of the LLM to generate plausible sentences. We also use the audio-language model to guide the token-by-token generation from the LLM. To do this, we evaluate the match of the already generated text sequence extended by the current candidate tokens with the audio clip. We then choose a token that has a high similarity to the given audio clip, whilst still being considered a likely next token by the LLM.

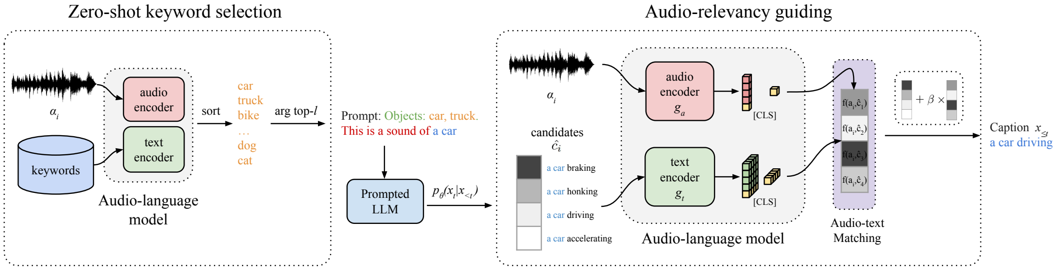

Figure 1: Our ZerAuCap framework generates audio captions without task-specific training (zero-shot). First, we determine a set of l relevant keywords that match the audio clip. An LLM is prompted with these keywords and yields the probability distribution over the next token. We evaluate the match of these candidate tokens with the input audio and choose the next token according to a weighted sum of the LLM prediction and the audio-text matching quantified by a pre-trained audio-language model.

Zero-shot keyword selection. Given a list of keywords K, we use the pre-trained audio and text encoders of the audio-language model (g_a and g_t) to encode the individual keywords k and the audio clip α_i. We consider the cosine similarity of the keywords and audio clip embeddings and select the top-l keywords K^* according to:

$$K^* = \underset{k \in K}{\arg\operatorname{top}\text{-}l}\ \operatorname{CosSim}\left(g_a(\alpha_i), g_t(k)\right) = \underset{k \in K}{\arg\operatorname{top}\text{-}l}\ \frac{g_a(\alpha_i) \cdot g_t(k)}{\|g_a(\alpha_i)\|_2 \cdot \|g_t(k)\|_2}. \tag{1}$$

We compose a prompt for the pre-trained LLM which was used to generate audible candidate tokens. Commonly, LLMs are trained to autoregressively model the probability of the next token conditioned on the previous tokens. We exploit this capability and build a prompt which consists of a *keyword prompt* "OBJECTS: ", the top-l keywords K^* and a *default prompt* "THIS IS A SOUND OF". Then, the probability of the next generated token is:

$$p_\theta(x_t \mid \underbrace{x_0, \ldots, x_{b-1}}_{\text{keyword prompt}}, \underbrace{x_b, \ldots, x_{d-1}}_{\text{top-}l\text{ keywords } K^*}, \underbrace{x_d, \ldots, x_{h-1}}_{\text{default prompt}}, \underbrace{x_h, \ldots, x_{t-1}}_{\text{autoregressive modelling}}). \tag{2}$$

Audio-relevancy guiding. After prompting the LLM, we re-weigh the LLM's top predicted tokens using audio-relevancy guiding. First, the LLM predicts the probability distribution over all tokens of the LLM's vocabulary. To accelerate the predictions, we then select the top-m predicted tokens, which we call *candidate* tokens c_i. As the LLM empirically assigns very small probabilities to all tokens except the top-m, the impact of the ignored tokens on the final weighted sum in the audio-relevancy guiding would have been negligible anyway.

We rank all candidate tokens c_i by matching the currently generated sentence extended by the candidate tokens (which we will call $\hat{c}$) with the audio clip. This measures how well the generated text matches the audio clip. The similarity $f(g_a(\alpha_i), \hat{c}_i)$ of a candidate sequence $\hat{c}_i$ with respect to the embedded audio clip $g_a(\alpha_i)$ is determined by first computing the cosine similarity $\operatorname{CosSim}(\cdot,\cdot)$ of all possible audio-text matches,

$$f(g_a(\alpha_i), \hat{c}_i) = \frac{e^{\kappa \cdot \operatorname{CosSim}(g_a(\alpha_i), g_t(\hat{c}_i))}}{\sum_{j \in 1,\ldots,k} e^{\kappa \cdot \operatorname{CosSim}(g_a(\alpha_i), g_t(\hat{c}_j))}}, \tag{3}$$

with a scalar temperature κ. We select the next token based on a weighted sum of the probabilities assigned to each of the candidate tokens by the LLM as well as the audio-text similarity. For each time step t, the next token x_t is selected according to:

$$x_t = \underset{i \in 1,\ldots,m}{\arg\max} \left\{ p_\theta\left(c_i \mid x_0, \ldots, x_{<t}\right) + \beta \cdot f\left(g_a(\alpha_i), \hat{c}_i\right) \right\}, \tag{4}$$

where β is a scalar weighting factor. We append the selected token to the prompt and iterate this process until the language model generates a token containing a period.

4 Experiments

Experimental setup. We adapted the 1.3B parameter version of OPT [45] as the LLM for all experiments unless stated otherwise. For both guiding strategies (*zero-shot keyword selection* and

		AudioCaps [15]						**Clotho** [7]					
Setting	**Framework** ↓	B4	M	RL	C	S	Sr	B4	M	RL	C	S	Sr
Zero-shot	No audio (baseline)	0.0	4.1	17.8	0.1	0.0	0.0	0.0	3.8	16.6	0.2	0.1	0.2
	Shaharabany et al. [32]	**9.8**	8.6	8.2	9.2	-	-	-	-	-	-	-	-
	ZerAuCap $_{\text{OPT}}$ (ours)	6.8	**12.3**	**33.1**	**28.1**	**8.6**	**18.3**	**2.9**	**9.4**	**25.4**	**14.0**	**5.3**	**9.7**
Supervised	HYU [5] (ensemble)	-	-	-	-	-	-	20.2	19.7	42.2	54.1	14.6	34.3
	HTSAT-BART [26]	28.3	25.0	50.7	78.7	18.2	48.5	16.8	18.4	38.3	46.2	13.3	29.7

Table 1: Zero-shot audio captioning results on the AudioCaps [15] and Clotho [7] datasets.

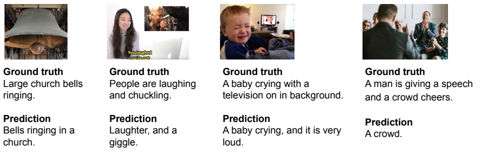
You laughed you're out
Ground truth
Large church bells ringing.
Prediction
Bells ringing in a church.
Ground truth
People are laughing and chuckling.
Prediction
Laughter, and a giggle.
Ground truth
A baby crying with a television on in background.
Prediction
A baby crying, and it is very loud.
Ground truth
A man is giving a speech and a crowd cheers.
Prediction
A crowd.

Figure 2: Qualitative example results for ZerAuCap. Video frames are only for illustration.

	AudioCaps [15]						**Clotho** [7]					
Ablation ↓	B4	M	RL	C	S	Sr	B4	M	RL	C	S	Sr
No keywords	0.0	5.9	20.7	2.5	2.4	2.4	0.6	5.4	18.1	1.4	1.2	1.3
No guiding with audio-relevancy	6.3	1.2	32.8	28.1	8.2	18.1	2.2	9.1	25.1	12.5	5.0	8.8
ZerAuCap $_{\text{GPT2}}$	1.7	9.4	23.5	11.6	6.8	9.2	1.7	9.4	23.5	11.6	6.8	9.2
ZerAuCap $_{\text{OPT}}$ (ours)	**6.8**	**12.3**	**33.1**	**28.1**	**8.6**	**18.3**	**2.9**	**9.4**	**25.4**	**14.0**	**5.3**	**9.7**

Table 2: ZerAuCap model ablations on AudioCaps [15] and Clotho [7].

audio-relevancy guiding), we used the pre-trained WavCaps [26] audio-language model to measure the audio-text matching. The list of audible keywords is derived from the publicly available AudioSet class list [9]. We separated the tags for classes that contained more than one class, resulting in 614 audio keywords. We set the number of selected keywords to $l = 2$, the scalar weighting factor to $\beta = 0.5$, and the number of candidates to $m = 45$, using the validation set of AudioCaps [15]. We evaluate the quality of our captions on the respective test sets of AudioCaps and Clotho [7] using standard natural language generation metrics BLEU [29] (B), METEOR [2] (M), ROUGE-L [20] (RL), CIDEr [37] (C), SPICE [1] (S), and SPIDER [21] (Sr).

Quantitative results. We present quantitative results for zero-shot audio captioning in Table 1. ZerAuCap outperforms [32] by a wide margin on most metrics, setting a new state of the art on AudioCaps [15]. Additionally, it outperforms our baseline that does not have access to any audio input on both datasets. As randomly generated texts already might have some overlap with the ground-truth captions, our baseline yields non-zero scores. Interestingly, on AudioCaps the baseline significantly outperforms the recently proposed zero-shot audio captioning model [32] in terms of the ROUGE-L score. Our model always outperforms the baseline and [32] for all metrics but BLEU-4. This indicates that the recall of 4-grams is higher with [32]. However, this value may be skewed due to the very short captions generated by [32]. We additionally report results from supervised methods that represent an upper bound. This comparison indicates that our model already captures significant portions of the audible content and converts them into plausible captions.

Qualitative results. In Figure 2 we can observe that our generated audio captions capture the audible events from given audio clips across a large variety of settings. For the failure case (right), our model does not provide a very detailed caption and thus fails to capture the full content of the audio clip.

Ablations. We ablate the core components of our approach in Table 2. Specifically, we study the effects of our guiding components and of using different LLMs. Using *no keywords* drastically reduces performance across all metrics. This indicates that it is important to condition the LLM on audible concepts. When using *no guiding with audio-relevancy* (i.e. $\beta = 0$), performance degrades. Combined, this shows that both guiding techniques contribute to the overall performance of ZerAuCap. Since [32] uses GPT-2 as their base LLM, we also run our model with GPT-2 and still outperform their approach, even though their audio-language model [10] is stronger than ours.

5 Conclusion

We introduced a zero-shot audio captioning framework that converts audio clips into textual captions using a twofold audio-based guiding approach without any training. Our proposed ZerAuCap framework achieves state-of-the-art results on the AudioCaps and Clotho benchmarks. Furthermore, our results show that keyword-based guiding is highly beneficial for obtaining better audio captions.

6 Acknowledgements

The authors thank IMPRS-IS for supporting Leonard Salewski. This work was partially funded by the BMBF Tübingen AI Center (FKZ: 01IS18039A), DFG (EXC number 2064/1 – Project number 390727645), and ERC (853489-DEXIM).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005.

[3] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

[4] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023.

[5] Jae-Heung Cho, Yoon-Ah Park, Jaewon Kim, and Joon-Hyuk Chang. Hyu submission for the dcase 2023 task 6a: automated audio captioning model using al-mixgen and synonyms substitution. Technical report, DCASE2023 Challenge, 2023.

[6] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. Automated audio captioning with recurrent neural networks. In *IEEE WASPAA*, 2017.

[7] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, 2020.

[8] Ayşegül Özkaya Eren and Mustafa Sert. Audio captioning based on combined audio and semantic embeddings. In *IEEE ISM*, 2020.

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *CVPR*, 2023.

[11] Félix Gontier, Romain Serizel, and Christophe Cerisara. Automated audio captioning by fine-tuning bart with audioset tags. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2021.

[12] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas R. Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2021.

[13] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *CVPR*, 2023.

[14] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. Autoad ii: The sequel-who, when, and what in movie audio description. In *ICCV*, 2023.

[15] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proc. NACCL*, 2019.

[16] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, Jaeheon Sim, Jinwoo Lee, and Kyogu Lee. Exploring train and test-time augmentations for audio-language learning. *arXiv preprint arXiv:2210.17143*, 2023.

[17] A. Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 2022.

[18] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *arXiv preprint arXiv:2012.07331*, 2020.

[19] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *ICLR*, 2023.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

[21] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 2017.

[22] Xubo Liu, Qiushi Huang, Xinhao Mei, Tom Ko, H Lilian Tang, Mark D Plumbley, and Wenwu Wang. Cl4ac: A contrastive loss for audio captioning. *arXiv preprint arXiv:2107.09990*, 2021.

[23] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. Audio-text retrieval in context. In *ICASSP*, 2022.

[24] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Audio captioning transformer. *arXiv preprint arXiv:2107.09817*, 2021.

[25] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023.

[26] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, MarkD . Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023.

[27] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP*, 2021.

[28] Andreea-Maria Oncescu, A. Sophia Koepke, Joao Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. In *INTERSPEECH*, 2021.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method and for automatic and evaluation of machine and translation. In *ACL*, 2002.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[31] Leonard Salewski, A. Sophia Koepke, Hendrik P.A. Lensch, and Zeynep Akata. Zero-shot translation of attention patterns in vqa models to natural language. In *DAGM GCPR*, 2023.

[32] Tal Shaharabany, Ariel Shaulov, and Lior Wolf. Zero-shot audio captioning via audibility guidance. *arXiv preprint arXiv:2309.03884*, 2023.

[33] Malcolm Slaney. Semantic-audio retrieval. In *ICASSP*, 2002.

[34] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv: 2205.02655*, 2022.

[35] Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint:2207.11100*, 2022.

[36] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, 2022.

[37] Ramakrishna Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[38] Junyan Wang, Yi Zhang, Ming Yan, Ji Chao Zhang, and Jitao Sang. Zero-shot image captioning by anchor-augmented vision-language space alignment. *arXiv preprint arXiv:2211.07275*, 2022.

[39] Mengyue Wu, Heinrich Dinkel, and Kai Yu. Audio caption: Listen and tell. In *ICASSP*, 2019.

[40] Yusong Wu, K. Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint arXiv:2211.06687*, 2022.

[41] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

[42] Huang Xie, Felix Gontier, Samuel Lipping, Konstantinos Drossos, Tuomas Virtanen, and Romain Serizel. DCASE2022 challenge task 6: Automated audio captioning, 2022. URL `http://dcase.community/challenge2022/task-automatic-audio-captioning`.

[43] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Zeyu Xie, and Kai Yu. Investigating local and global information for automated audio captioning with transfer learning. *arXiv preprint arXiv:2102.11457*, 2021.

[44] Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, 2023.

[45] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.