Vision Language Models Are Few-Shot Audio Spectrogram Classifiers

Satvik Dixit **Laurie M. Heller** **Chris Donahue**
Carnegie Mellon University
{satvikdixit, laurieheller, chrisdonahue}@cmu.edu

Abstract

We demonstrate that vision language models (VLMs) are capable of recognizing the content in audio recordings when given corresponding spectrogram images. Specifically, we instruct VLMs to perform audio classification tasks in a few-shot setting by prompting them to classify a spectrogram image given example spectrogram images of each class. By carefully designing the spectrogram image representation and selecting good few-shot examples, we show that GPT-4o can achieve 59.00% cross-validated accuracy on the ESC-10 environmental sound classification dataset. Moreover, we demonstrate that VLMs currently outperform the only available commercial audio language model with audio understanding capabilities (Gemini-1.5) on the equivalent audio classification task (59.00% vs. 49.62%), and even perform slightly better than human experts on visual spectrogram classification (73.75% vs. 72.50% on first fold). We envision two potential use cases for these findings: (1) combining the spectrogram and language understanding capabilities of VLMs for audio caption augmentation, and (2) posing visual spectrogram classification as a challenge task for VLMs.

1 Introduction

Vision-language models (VLMs) have emerged as a powerful paradigm for multimodal artificial intelligence, capable of jointly processing and reasoning over visual and textual information [8, 3]. By integrating computer vision and natural language processing capabilities, VLMs have demonstrated remarkable capabilities across a wide range of tasks, from image captioning and visual question answering to object detection and scene understanding [20, 13]. The success of VLMs in these domains suggests that their generalized visual reasoning abilities could be leveraged for tasks in other modalities, supposing those modalities can be reasonably represented as images.

Here we investigate whether VLMs might be able to perform audio classification when presented with audio in the form of *spectrograms*. Spectrograms, which visually represent the frequency content of audio signals over time, have long been used as input for audio classification tasks [15, 4, 12, 16, 11]. These time-frequency representations capture essential acoustic features that are often more informative than raw waveforms for many audio analysis tasks. Recent commercial VLMs, such as GPT-4o [1], Claude-3.5 Sonnet [2] and Gemini-1.5 [18], have demonstrated impressive zero-shot and few-shot capabilities across various visual tasks [19]. These models are trained on image-text data and accordingly have seen spectrograms and associated text during pre-training. However, their potential for processing spectrograms and classifying audio content remains unexplored to the best of our knowledge.

We propose a novel task called *visual spectrogram classification* (VSC) which tasks models with recognizing the content in audio recordings from their equivalent visual spectrograms. We benchmark VLMs and attempt to measure human expert performance on this task. Our approach involves

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

carefully designing the spectrogram image representation and the few-shot prompting strategy to enable the VLM to reason about audio content from visual patterns.

We conduct comprehensive experiments to evaluate the zero-shot and few-shot performance of various commercial VLMs on this task. We conduct ablation studies to optimize spectrogram hyperparameters, exploring variations in amplitude scale, frequency axis, colormap, and spectrogram style transformations. We also experiment with different numbers and types of examples for few-shot learning. Finally, we compare the few-shot performance of VLMs with the performance of human experts on VSC and commercial audio language models on audio classification (AC).

Demonstrating that VLMs can comprehend sounds through spectrograms opens up a range of potential applications. One of our aims, for instance, is to improve the quality of captions in audio captioning datasets which often contain vague or inaccurate descriptions [7]. While large language models (LLMs) are capable of generating captions, they are prone to hallucination. Incorporating audio information would help the multimodal LLM to ground the captions but it is not immediately clear how to do so. Current audio language models (ALMs) like Pengi [5], SALMONN [17] and GAMA [10] allow audio as input but lag behind commercial VLMs in terms of language understanding abilities, as evidenced by the size of their underlying LLMs (Pengi uses GPT-2 as the language model component with 124M parameters compared to the hundreds of billions now common in state-of-the-art models [9]). If we could provide both the vague caption and the audio information to the VLM as a spectrogram, we may be able to get captions that are both grounded in audio as well as sufficiently descriptive.

Our research contributes to cross-modal learning and VLM adaptability to novel tasks. By demonstrating the effectiveness of VLMs in processing spectrograms for audio classification, we highlight the potential for these models to bridge the gap between visual and auditory domains. VSC serves as a new benchmark task for VLMs as a test of their audio spectrogram understanding capabilities.

The primary contributions of this work are:

- Proposing a novel task, Visual Spectrogram Classification (VSC), which demonstrates VLMs' capability to classify audio content using spectrogram images.
- Conducting ablation studies to find optimal spectrogram hyperparameters for VSC.
- Comparing the performance of latest models from the GPT-4, Claude and Gemini series on the VSC task in zero-shot and few-shot settings.
- Comparing VLM performance with human experts on the VSC task and commercial audio language models on the audio classification task.

2 Tasks & Methods

2.1 Task definition for visual spectrogram classification (VSC)

Visual spectrogram classification (VSC) is a novel task that involves classifying audio content based on the spectrogram representations. In zero-shot settings, the model analyzes the spectrogram and selects the most likely audio class. For few-shot settings, the model is provided with example spectrograms for each sound class to guide the classification process.

2.2 Default spectrogram extraction hyperparameters

We establish default parameters for spectrogram extraction to ensure consistency. Audio files are resampled to 22,050 Hz, with Short-time Fourier transform (STFT) computed using a 2,048-sample window size and 512-sample hop length. Both frequency and amplitude scales are logarithmic (as is common in audio research), using the 'viridis' colormap. To improve clarity, we added axis labels and removed the colormap scale from the spectrogram images.

2.3 Prompting VLMs to perform VSC

For zero-shot experiments, we input the spectrogram and a text prompt listing all classes, instructing the model to select the most likely class. Few-shot experiments employ in-context learning [6], providing one example spectrogram per class alongside the test spectrogram.

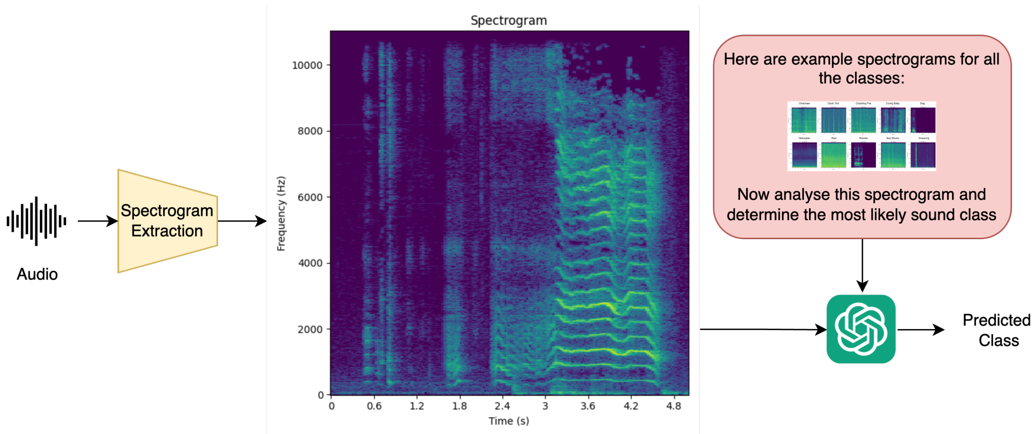

Figure 1: Experimental setup of the visual spectrogram classification task in the few-shot setting

3 Experiments & Results

3.1 Dataset and Models

Dataset: We utilize the ESC-10 dataset, a subset of ESC-50 [14], comprising 400 5-second environmental sound recordings across 10 classes. The dataset is divided into five folds, each containing 80 audio clips (8 per class). Figure 2 illustrates example spectrograms for each class.

Models: Our experiments employ six state-of-the-art VLMs:

- GPT-4: GPT-4o (most powerful) and GPT-4o-mini (lightweight)
- Claude: Claude-3.5 Sonnet (latest) and Claude-3 Opus (computation-intensive)
- Gemini: Gemini-1.5 Pro (most powerful) and Gemini-1.5 Flash (lightweight)

All models were accessed via API endpoints using the latest versions as of September 2024

3.2 Comparing zero-shot performance of VLMs on the visual spectrogram classification task

We evaluated the zero-shot VSC performance of all the VLMs on the first fold of ESC-10 using the default spectrogram hyperparameters. The first column in Table 1 shows the zero-shot VSC performance. GPT-4o outperformed other state-of-the-art models in the zero-shot setting.

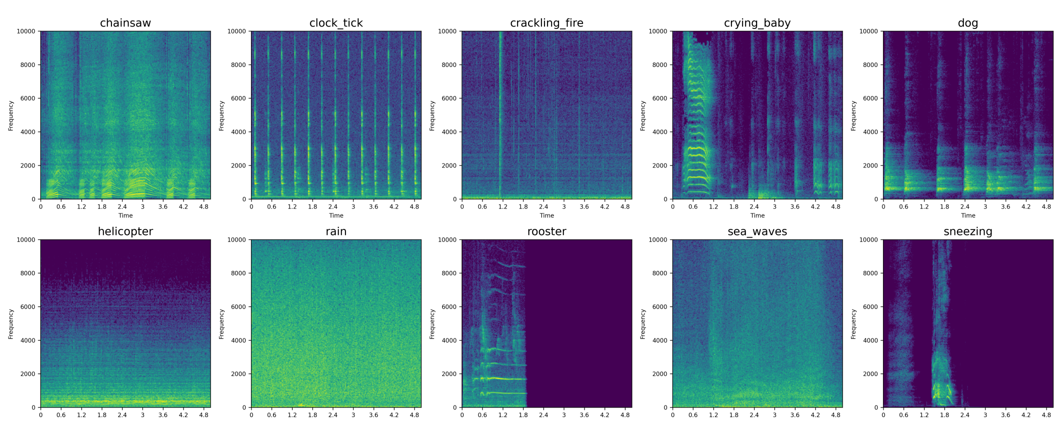

Figure 2: Example audio spectrograms for each class in the ESC-10 dataset

Table 1: Zero-shot and few-shot performance of VLMs (on VSC) and ALMs (on AC) on the first fold of the ESC-10 dataset. Zero-shot performance with tuned parameters (see appendix) is also shown. * indicates that the hyperparameters were tuned on GPT-4o, so it has an advantage over other models.

Modality	Model	Zero-shot (default parameters) (%)	Zero-shot (tuned parameters) (%)	Few-shot (%)
Vision	GPT-4o	27.50	35.00*	70.00*
	GPT-4o mini	21.50	22.50	38.75
	Claude-3.5 Sonnet	21.50	22.50	56.25
	Claude-3 Opus	11.25	12.50	15.00
	Gemini-1.5 Flash	16.25	17.50	10.81
	Gemini-1.5 Pro	18.75	18.42	20.31
Audio	Gemini-1.5 Flash (Audio)	27.50	27.50	40.26
	Gemini-1.5 Pro (Audio)	38.36	38.36	44.78

3.3 Few-shot performance of VLMs on the visual spectrogram classification task

We investigated GPT-4o's few-shot learning capabilities on the first fold of ESC-10. Initial experiments used randomly selected examples for each class. We then optimized example quality by applying K-means clustering to mel spectrograms from the remaining folds, selecting spectrograms closest to cluster centroids for 10-shot, 20-shot, and 30-shot tasks. We also explored clustering on amplitude spectrograms and manual example selection.

Table 1 (third column) shows few-shot performance with randomly chosen examples. All models demonstrated significant accuracy improvements with examples provided. GPT-4o's 10-shot accuracy increased to 70% from 35% in zero-shot. Table 2 illustrates the impact of example selection methods on GPT-4o VSC performance. Cluster-based and handpicked examples improved 10-shot performance. Optimal performance (76.25%) was achieved with 2 examples per class using K-means clustering on Mel spectrograms. Interestingly, 3 examples per class slightly decreased accuracy.

Table 2: Few-shot VSC performance of GPT-4o on ESC-10 (first fold) with different example selection methods

# per class	Selection method	Feature	Accuracy (%)
1	Random	Mel	70.00
1	Hand-picked	Mel	75.00
1	K-means (k=3)	Mel	73.75
1	K-means (k=3)	Amp	60.00
2	K-means (k=3)	Mel	76.25
3	K-means (k=3)	Mel	70.00

Table 3: Few-shot VSC performance of human experts on the ESC-10 dataset (first fold)

Expert	Accuracy (%)
Expert 1	67.5
Expert 2	57.5
Expert 3	71.25
Ensembled	72.5

3.4 Human expert evaluations

To benchmark human performance on the VSC task, we conducted evaluations with three experts (professors and graduate students from audio or speech research labs). Each expert classified 80 spectrograms from the first fold of ESC-10, using the same 10-shot mel spectrogram examples provided to the VLM 10-shot mel spectrogram clustering scenario for consistency.

The accuracies of the experts on the 10-shot VSC task on the first fold of ESC-10 are shown in table 3. The mean inter-annotator agreement (Cohen's Kappa Score) is 0.53. Notably, GPT-4o in the same scenario achieved a slightly higher accuracy (73.75%) than the ensembled human expert performance (72.5%), demonstrating the model's potential to match and even exceed human expert level performance on this task. The confusion matrices for GPT-4o and ensembled human expert predictions, shown in Figure 3, reveal that both approaches struggled more with distinguishing similar spectrograms such as 'dog' and 'sneezing' or 'helicopter', 'rain' and 'sea waves'.

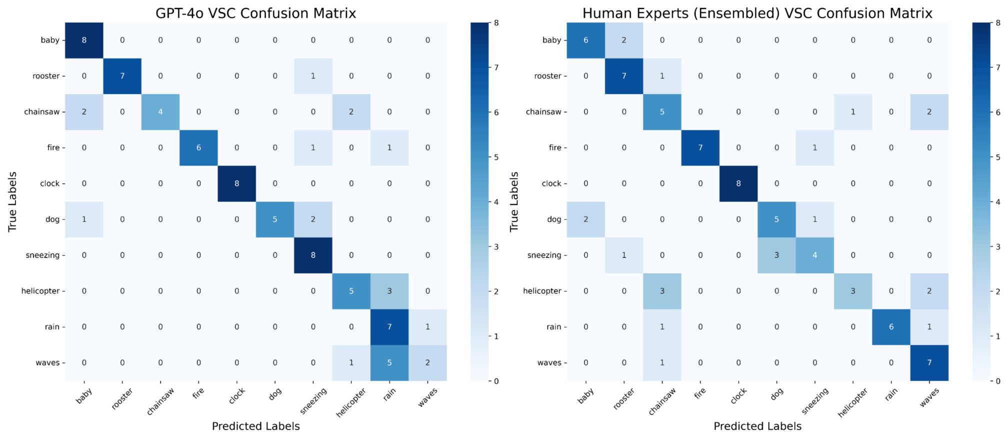

Figure 3: Confusion matrices for GPT-4o (left) and the ensembled human expert predictions (right)

3.5 Cross-validated task performance

To assess the robustness of our approach, we conduct cross-validated VSC performance of GPT-4o and AC performance evaluations of Gemini-1.5 pro on the full ESC-10 dataset. To evaluate GPT-4o on a more challenging task, we use a custom subset of the ESC-50 dataset. ESC-50 is a more comprehensive dataset with 2000 audio files across 50 classes. We create a subset of 100 randomly selected audio files (2 audio files per class) from the first of the five folds for VSC evaluation.

The accuracies of the GPT-4o and Gemini-1.5 pro models on the ESC-10 dataset are 59.00% and 49.62% respectively. This highlights the competitive performance of GPT-4o on VSC, even when compared to commercial audio language models. The 50-shot VSC accuracy of GPT-4o on the customized subset of the ESC-50 dataset is 14% which shows its poor generalization capability when the number of classes increases.

4 Limitations

Firstly, current VLMs do not yet match the performance of traditional audio classification methods. This performance gap is likely because VLMs are primarily trained on natural images and text, not spectrograms. Moreover, this task is difficult even for human experts.

Secondly, our experiments reveal a significant drop in classification accuracy as the number of audio classes increases (59% accuracy on the ESC-10 dataset (10 classes) to 14% accuracy on the ESC-50 dataset (50 classes)). This decline suggests that VLMs struggle to differentiate between a large number of audio classes; the visual patterns in spectrograms may become less distinctive or more confusing to the model as the number of classes or the similarities between them grows.

It is important to note, however, that as VLMs continue to evolve and become larger in size and potentially encounter more spectrograms in their training data, we anticipate improvements in their visual spectrogram classification capabilities.

5 Conclusion

In this study, we propose the visual spectrogram classification (VSC) task and evaluate vision language models such as GPT-4o in zero-shot and few-shot settings. Our experiments demonstrate that VLMs can interpret the ability to classify audio spectrograms, with few-shot learning greatly enhancing performance. We show that GPT-4o performs better than commercial multimodal models with audio understanding, and can even outperform human experts on this task. As VLMs continue to evolve, we envision expanded applications in audio classification and understanding.

6 Acknowledgements

This work was generously supported by Sony via the Sony Research Award Program (RAP).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Model card, Anthropic, March 2024.

[3] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024.

[4] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification, 2016.

[5] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks, 2024.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.

[7] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740, 2019.

[8] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In *International Joint Conference on Artificial Intelligence*, 2022.

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities, 2024.

[11] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2017.

[13] Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models, 2024.

[14] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

[15] Karol J. Piczak. Environmental sound classification with convolutional neural networks. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2015.

[16] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. *ArXiv*, abs/1711.02520, 2017.

[17] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024.

[18] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[19] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.

[20] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Appendix A: Spectrogram hyperparameters ablation study

We varied key parameters to generate diverse spectrograms from the same audio files:

- Amplitude scale: logarithmic or linear
- Colorscale: viridis or magma
- Frequency scale: logarithmic or linear
- Spectrogram style: raw amplitude magnitude, mel spectrogram, or mel-frequency cepstral coefficients (MFCCs)
- Visual elements: presence or absence of colorbar and labels
- Image resolution: standard or low (adjusted via GPT-4o's 'detail' parameter)

Some variations (e.g., amplitude and frequency scales) are common in spectrogram-based classifiers, while others (e.g., colorscale, text labels) are more specific to VLMs. The aim was to assess how these variations in the spectrogram representation affected the model's ability to classify audio signals. Table 4 presents the results of these experiments, showing the zero-shot VSC performance of GPT-4o on the ESC-10 dataset (first-fold) with different spectrogram hyperparameters. Notably, amplitude spectrograms with a linear frequency axis yielded the highest classification accuracy (35%), suggesting that these settings better capture relevant audio features. The zero-shot performance on tuned parameters of all models is shown in the second column of the Table 1.

Table 4: Zero-shot VSC performance of GPT-4o on ESC-10 (first fold) with various hyperparameters

Parameter	Accuracy (%)
Default parameters	27.50
Linear frequency axis	35.00
Linear amplitude scale	30.00
Remove labels	26.25
Show colorbar	23.75
Magma colormap	25.00
Mel spectrogram	25.00
MFCCs	13.75
Low resolution	20.00

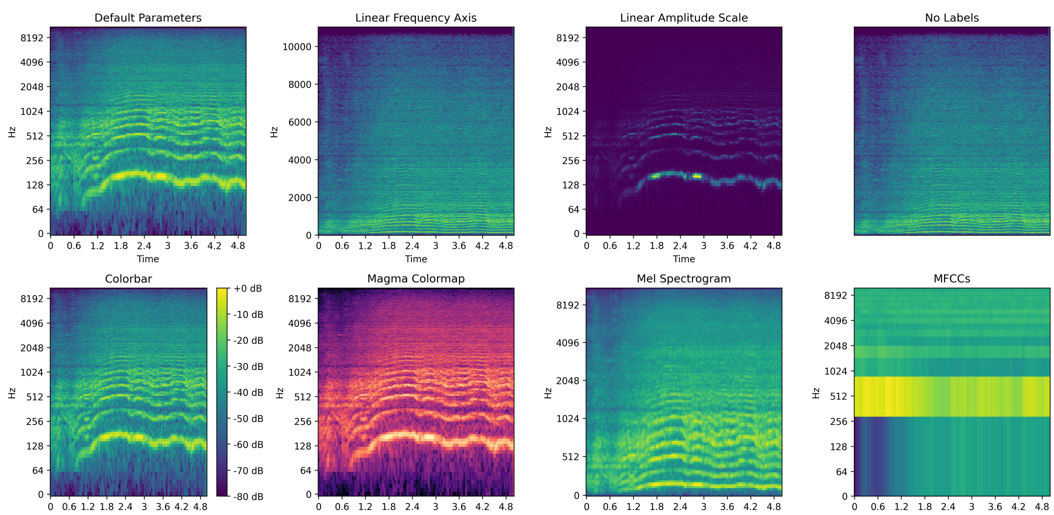

Figure 4: Example spectrograms for the same audio using different configurations described in the hyperparameter ablation section

Appendix B: Prompts for VSC in zero-shot and few-shot settings

Figures 5 and 6 show the zero-shot and few-shot prompt templates for the visual spectrogram classification task.

```
ZERO_SHOT_PROMPT = """
{
    "role": "system",
    "content": "You are a helpful assistant with expertise in
       recognizing patterns and identifying classes based on
       visual representations of audio data."
},
{
    "role": "user",
    "content": [
        {
            "type": "text",
            "text": "Your task is to analyze a spectrogram,
               which is a visual representation of the
               frequency spectrum of sound over time, and
               determine the most likely sound class from a
               given list of possibilities. Analyze the
               spectrogram image, considering factors such as
               frequency patterns, intensity, and time
               variations. Focus solely on the patterns
               presented in the spectrogram. Do not let any
               assumptions about common sounds or
               environmental settings influence your decision.
                Here are the classes: ['dog', 'chainsaw', '
               crackling_fire', 'helicopter', 'rain', '
               crying_baby', 'clock_tick', 'sneezing', '
               rooster', 'sea_waves']. Your response must
               always contain the exact name of the class only
               . For example, if you believe the spectrogram
               matches best with rain, your response would be
               rain. Here is the spectrogram:"
        },
        {
            "type": "image_url",
            "image_url": {
                "url": "data:image/png;base64,{image}"
            }
        }
    ]
}
    """
```

Figure 5: The template for prompting the VLM for zero-shot VSC. {image} is replaced by the spectrogram image to be classified.

```
FEW_SHOT_PROMPT = """
{
    "role": "system",
    "content": "You are a helpful assistant with expertise in recognizing patterns and identifying classes based on visual representations of audio data."
},
{
    "role": "user",
    "content": [
        {
            "type": "text",
            "text": "Your task is to analyze spectrograms, which are visual representations of the frequency spectrum of sound over time, and determine the most likely sound class for a given spectrogram.\nHere are examples of spectrograms for different sound classes:"
        },
        {
            "type": "examples",
            "content": [
                {
                    "type": "text",
                    "text": "Spectrogram for {category-i}:"
                },
                {
                    "type": "image_url",
                    "image_url": {"url" "data:image/png;base64,{example-image-i}"}
                }
            ]
        },
        {
            "type": "text",
            "text": "\nNow, given a new spectrogram, analyze it considering factors such as frequency patterns, intensity, and time variations. Focus solely on the patterns presented in the spectrogram. Do not let any assumptions about common sounds or environmental settings influence your decision.\nYour task is to determine which of the example classes the new spectrogram most closely resembles. Your response must contain only the exact name of the class.\nHere is the new spectrogram to classify:"
        },
        {
            "type": "image_url",
            "image_url": {
                "url": "data:image/png;base64,{image}"
            }
        }
    ]
}
    """
```

Figure 6: The template for prompting the VLM for few-shot VSC. {category-i} and {example-image-i} are replaced by the name and example spectrogram for the i-th class where i will take values from 1 to n for n-shot classification and {image} is replaced by the spectrogram image to be classified.