TSPE：任务特定提示集成以改进零样本音频分类

Nishit Anand[1] Ashish Seth[1] Ramani Duraiswami[1] Dinesh Manocha[1]
[1]University of Maryland, College Park, MD, USA
{nishit, aseth125, ramanid, dmanocha}@umd.edu

摘要—音频语言模型（ALMs）在零样本音频分类任务中表现出色，该任务通过利用描述性自然语言提示，在测试时对之前未见过的音频片段进行分类。我们引入了 *TSPE*（任务特定提示集成），这是一种简单、无需训练的硬提示方法，通过对不同音频分类任务定制提示来提升 ALMs 的零样本性能。与使用通用模板化的提示如“汽车的声音”相比，我们生成内容丰富的提示，例如“从隧道中传来的汽车声音”。具体来说，我们利用标签信息识别合适的声音属性，比如“大声”和“微弱”，以及适当的声音来源，比如“隧道”和“街道”，并将这些信息整合到 ALMs 用于音频分类的提示中。此外，为了增强音频文本对齐，我们在 TSPE 生成的任务特定提示之间进行提示集成。在 12 个多样化的音频分类数据集上评估时，TSPE 通过展现绝对提升 1.23-16.36%的表现优于原版零样本评估。

I. 介绍

最近的多模态语言模型（MLMs）进展在多种模式和任务上极大地提升了性能 [27]–[31] 。通过在大量的音频-字幕配对数据集上进行训练，这些模型获得了广泛的音频概念理解能力，使它们能够在零样本设置下对新的音频类别进行分类。这种适应性使得音语言模型（ALMs）非常适合具有多样且不熟悉声音的动态环境。

对比学习基础的音频语言模型（ALMs）如对比语言音频预训练（CLAP）[19] 学习音频和文本之间的共享表示空间。这有助于它们在诸如音频分类等任务中很好地泛化并具有良好的下游性能。近年来，文献中发表了多种音频-语言编码器（ALEs），它们在音频分类和文本到音频检索等任务上表现出色。

这些模型是在大型音频文本数据集上预先训练的，使它们能够在各种音频环境中很好地泛化。虽然它们在零样本音频分类中表现出色，但这种成功往往依赖于大量的音频文本配对或额外的微调。很少有努力专注于不进行额外训练的情况下增强类似 CLAP 模型的零样本分类能力。一些方法如 Audio Prompt Learner [25] 和 TreffAdapter [26] 能够提高性能，但需要额外的训练并引入可学习参数，这会增加时间和计算成本。此外，尽管这些方法在处理分布内任务时有效，但在处理分布外（OOD）音频分类任务时往往表现不佳。另外，当前方法的一个常见限制是依赖于通用提示，例如“< 标签 > 的声音”。我们发现这些提示并不能有效地跨不同的下游任务转移，并且通常需要适应才能有意义。例如，在像 GTZAN [6] 这样的音乐流派分类任务中，“摇滚的声音”这一提示不清楚，不能传达预期的类别，而将其修改为“摇滚的声音音乐”则可以为模型提供清晰度以在适当的上下文中理解流派。

主要贡献 为了解决这个问题，我们引入了 TSPE（任务特定提示集成），这是一种无训练的方法，可以提高 ALMs 的零样本音频分类性能。它使用下游任务和标签信息自动生成每个类别标签的任务特定提示。这是为了捕捉音频-文本对齐中的语义细微差别。然后，它不使用单一的基本提示，而是使用提示集成来学习更具语义丰富性的提示表示，这有助于更好地理解提示的丰富文本表示与音频表示之间的关联，从而提高 ALM 在下游音频分类上的性能。我们的方法无需任何微调或额外训练即可实现这一性能提升，并且其亮点在于它在分布外（OOD）数据集上也能表现出良好的音频分类性能。我们在 12 个多样化的音频分类数据集上进行了广泛实验，并展示了与基本零样本评估相比有绝对改进，改进幅度为 1.23%-16.36%。

II. 相关工作

多模态编码器在跨模态学习共享表示方面显示出巨大的潜力。基于视觉-语言模型如 CLIP [16] 的对比预训练方法，音频-语言模型（ALM）已经在音频分类中实

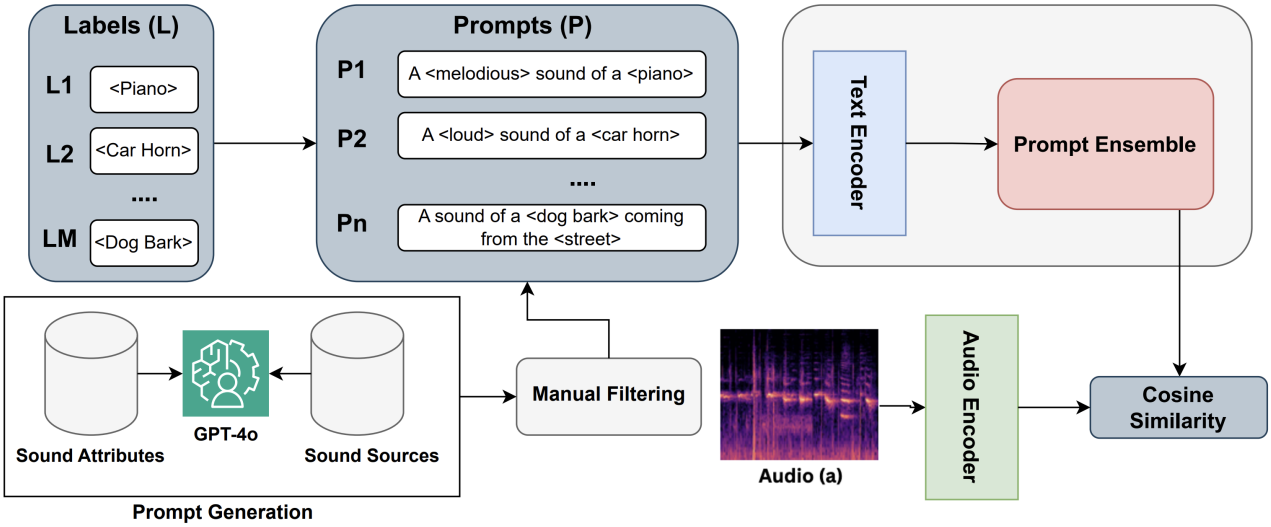

图 1. TSPE 工作流程的说明：我们从一组声音属性和来源开始，使用 GPT-4 定制一套提示模板。然后，我们手动选择与特定下游任务相关的提示，并将其传递给文本编码器以生成表示。这些表示通过提示集成进行平均。最后，我们计算音频表示与平均文本表示之间的余弦相似度。

现了最先进的零样本性能。早期的模型如 Wav2Clip [17] 和 AudioCLIP [18] 专注于将音频表示与类别标签对齐，而最近的方法如 CLAP [19] 则直接将音频映射到文本描述上，从而在零样本性能方面取得了显著的提升。尽管硬提示方法 [24] 通过使用语境丰富的提示工程 [20] 增强了视觉-语言模型的零样本能力，但对于 ALM 来说，这一技术仍然鲜有探索。相反，之前的 ALMs 依赖于计算密集型方法，例如改进对齐目标 [21] 或扩展参数和数据集 [19]。

III. 方法论

A. 分组多样化的音频标签

我们观察到，目前用于音频分类任务的提示语经常无法捕捉音频标签的多样性。例如，通用提示如“< 标签 > 的声音”在 < 标签 > 是像“摇滚”这样的音乐类型或像“海滩”这样的地点时缺乏语义连贯性。这突显了理解任务特定标签以创建更有效的提示的必要性。为了解决这个问题，我们开发了 TSPE，它利用任务和标签知识生成任务特定的硬提示。我们首先根据其分类特征（如声音类型）将标签归类到不同的组别中来实现这一目标。这些分类的具体细节如下所述：

- **乐器识别** 此类别包括来自各种音乐乐器在不同场合下演奏的声音，如歌剧、街头表演和剧院。它包含一系列的乐器，包括钢琴、吉他、铙、鼓等。与此类别相关的数据集包括京剧 [1]、Mridangam 敲击 [2]、Mridangam 定音 [3]、Nsynth 乐器 [4] 和 Nsynth 来源 [5]。
- **声场景理解** 此类别指代常见的城市声音，例如公共汽车、汽车、打桩机、钻孔机、狗叫、电梯、拥挤的街道、地铁、卡车、警报器、摩托车、空调、发动机怠速、汽车喇叭和街头音乐等产生的声音，以及日常生活中常遇到的声音，如教堂钟声、鸟鸣、鼠标点击声、办公室环境噪音、咖啡馆、有轨电车、海滩、餐厅、母鸡、公鸡、地铁站、公园和城市中心等地的声音。此类别使用的数据集包括 Cochlscene [7]、ESC50 [8]、TUT [9] 和 USD-8K（Urban Sounds）[10]。
- **音乐流派分类：** 此类别涉及将音频样本分类到不同的音乐流派中，包括古典、乡村、迪斯科、嘻哈、蓝调等。[6]。
- **冲击和紧急声音：** 此类别侧重于像爆炸、枪声和警报器这样的响亮而突然的声音。SESA 数据集常用于此任务 [11]。
- **非言语发声声音：** 此类别包括非语言的发声声音，如咳嗽、打喷嚏、清嗓子、吸鼻子、叹息和笑声。与此类别相关的数据集包括 VocalSound [12]。

B. 任务特定提示生成

我们首先为 GPT-4 提供任务类别及其标签的信息，以及声音属性的示例，如“安静”、“大声”、“静音”和“微弱”。我们还包括了声音来源的示例，比如“剧院”、“音乐会”、“房间”、“歌剧”和“街道”。然后我们请求 GPT-4 [13] 生成一份与我们的任务类别相关的 60 个声音属性和来源的列表。接下来，我们手动将这些声音属性和来源映射到每个任务类别中，确保属性和来源在特定标签中的上下文中是合适的。

对于每个任务类别，然后我们向 GPT-4 [13] 提供一种提示格式以及我们映射到该类别的属性和来源列表。我们要求它使用以下格式生成 40 个提示：

- 一个具有 < 属性的 > 声音，带有 < 标签的 >
- “来自 < 源 > 的 < 标签 > 的声音 “
- 一个 < 属性 > 声音的 < 标签 > 可以从一个 < 源 > 听到。

其中 < 标签 > 表示类别标签，< 属性 > 表示声音属性，而 < 来源 > 表示声音来源。从为每个任务类别生成的 40 个提示中，我们手动筛选出最符合任务类别要求的 20 个。我们观察到 GPT-4 [13] 产生了一些幻觉或非特性的提示，使得手动选择变得必要。

C. 硬提示和提示集成

如前所述，我们采用硬提示 [14] 而不是软提示或其他基于学习的方法，因为后者需要重新训练或微调 [15]，而我们的方法不需要训练，并且能够实现零样本改进。我们多样化的提示集通过以各种方式描述类别标签来提高下游性能，既包括它们的声学特征也包括声音来源。

我们的方法，任务特定提示集成（TSPE），显著提高了音频语言模型在音频分类 [23] 方面的零样本性能。通过为每个任务类别使用一组独特的提示，我们捕捉到了跨多种场景的声音细微差别和属性。这种方法使得语义表示更加准确，因为诸如“一段优美的 < 钢琴 > 声音”之类的提示比通用提示如“这是 < 标签 > 的声音”更好地捕捉了类别标签的细微差别。对于每个任务类别，我们生成其提示集中所有提示的文本嵌入，并对这些嵌入进行平均以增强语义表示。总之，我们首先使用硬提示创建特定于任务的提示，然后集成这些提示以提高零样本结果。

D. 将声音属性注入硬提示中

从 GPT-4 为每个任务类别生成的初始 40 个提示中，我们手动筛选出每类 20 个提示，并仔细检查声音属性是否匹配。我们确保每个提示中的声音属性与任务的类别标签有意义地相关。例如，提示“一个悦耳的的声音来自一个 < 标签 >”非常适合音乐乐器分类任务，就像“一个悦耳的的声音来自一个 < 钢琴 >”那样，在语义上是恰当的。然而，“一声悦耳的枪响 < 的声音 >”对于冲击和紧急声音来说是不匹配且不适当的。

同样，诸如“一个温柔的的声音的爆炸”这样的提示不符合影响和紧急声音，并且最好用“一个温和的的声音的一个笛子”来替换乐器分类任务。这个手动过滤步骤确保了声音属性与每个任务类别的现实特性相一致。

E. 将声音源信息注入硬提示中

在对每个任务类别由 GPT 生成的 40 个提示进行手动筛选时，我们仔细确保每个声源与类别标签自然一致。这种一致性至关重要，因为不匹配的声源会降低提示的一致性，从而减弱其捕捉正确语义背景的能力 [22]。

例如，考虑提示“可以听到从一个管弦乐队传来的小提琴的声音。”这比其他如“从一个图书馆传来小提琴的声音”或“一个动物园”，其中来源不符合小提琴的预期环境，更是一个合理的搭配。类似地，提示“来自教堂的器官的声音”与类别标签‘器官’非常吻合，提供了一个现实的场景。相比之下，选项如“来自一个机场或者一个火车站的器官的声音”则显得格格不入，降低了提示在从一系列声音中区分器官声音的效果。

IV. 实验设置

我们在五个任务类别中使用十二个不同的数据集和两个最先进的音频语言模型（ALMs），即由微软开发并于 2022 年和 2023 年分别发布的 MS-CLAP’22 和 MS-CLAP’23 [19]。对于乐器识别任务，我们在包含各种乐器的多个音频数据集上测试了我们的模型。这些数据集包括京剧 [1]、Mridangam 击打 [2]、Mridangam 定音 [3]、Nsynth 乐器 [4] 和 Nsynth 来源 [5]。对于音乐流派分类任务，我们使用了包含各种流派音乐的 GTZAN 数据集 [6]，包括摇滚、嘻哈、流行和乡村。对于声景理解，我们在多个数据集上评估我们的模型，例如 Cochlscene [7]、USD-8K（城市声音）[10]、ESC50 [8] 和 TUT Urban

表 I
TSPE 在十二个数据集上的五个任务的零样本音频分类性能。最佳结果报告在 **粗体**。

Task	Dataset	MSCLAP 2023	MSCLAP 2023 (TSPE)	MSCLAP 2022	MSCLAP 2022 (TSPE)
Musical Instruments Recognition	Beijing Opera	70.33	68.22	55.50	**71.86**
	Mridangam Stroke	45.60	**51.77**	14.69	10.72
	Mridangam Tonic	20.13	**34.01**	16.52	16.20
	NSynth Instrument	**66.72**	64.70	27.53	30.40
	NSynth Source	52.24	**53.47**	38.55	34.91
Acoustic Scene Understanding	Cochlscene	**85.07**	83.98	22.77	24.76
	USD-8K	79.37	**83.72**	72.39	70.93
	ESC-50	92.85	**94.55**	77.70	75.85
	TUT	44.07	**46.51**	21.32	24.09
Music Genre Classification	GTZAN	54.49	**59.57**	22.72	19.51
Impact and Emergency Sound	SESA	65.71	65.71	66.28	**67.81**
Non-Verbal Vocalization Sound	Vocal Sound (VS)	**80.93**	78.94	47.09	61.23

Acoustic Scenes [9]。这些数据集包含多种声音，如空调声、汽车喇叭声、有轨电车和地铁站的声音。对于冲击与紧急情况声音分类，我们使用了 SESA 数据集 [11]，其中包括爆炸、枪声和警报器等紧急相关的音效。对于非言语发声分类，我们在 VocalSound 数据集 [12] 上进行测试，该数据集包括叹息、抽鼻、笑声和打喷嚏等声音。所有结果都是通过五次运行的平均值报告的。

V. 结果

表 I展示了我们的技术在两种最先进的音音频语言模型（ALMs），MSCLAP'22 和 MSCLAP'23 [19] 上的零样本音频分类（ZSAC）效果。我们提供了针对五种任务类别和十二个涵盖各种声音的音频分类数据集，将我们的特定任务提示集成方法与传统的 ZSAC 提示进行了对比。

A. 结果分析

TSPE 能够提升模型的性能，范围从 1.23% 到 16.36%，在 MSCLAP'23 和 MSCLAP'22 上所有数据集的平均改进率分别为 2.06% 和 1.89%。我们观察到，在某些数据集上应用 TSPE 后，性能下降了。一个可能的原因是提示语不够语言丰富，以适应该任务，我们可以将此作为未来的工作内容。表 II显示了 TSPE 为不同任务生成的提示示例。

B. 超参数调优

我们从 GPT-4 生成的初始 40 个提示中选择了 $K = 20$ 个提示，并进行了消融研究以确定 K 的最佳值。具体来说，我们在 VocalSound 数据集上评估了 TSPE 在 $\{5, 10, 15, 20, 25, 30\}$这些 K 值下的性能，以考察提示数量如何影响音频语言模型（ALMs）的零样本音频分类性能。结果显示，随着 K 增加至 20 时，ALM 的性能有所提升，之后则开始下降。超过 $K = 20$ 后的性能下滑可能是由于提示数量增多导致语义噪声增加所致。图 2展示了不同提示数量对音频分类中的 MSCLAP'23 效果的影响，数据集为 [12]VocalSound。

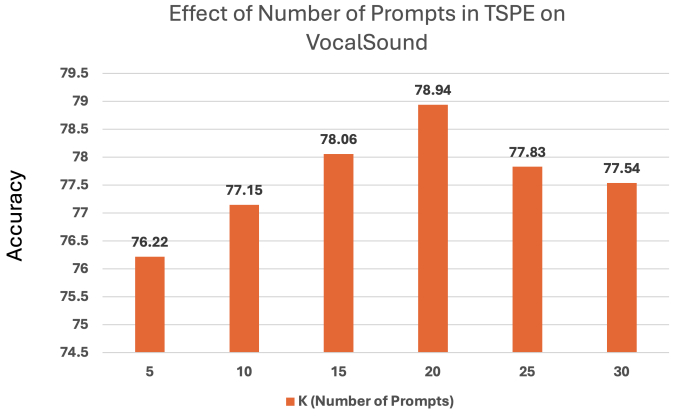

图 2. 不同数量的提示对 MSCLAP'23 在 VocalSound 数据集上的音频分类的影响

VI. 结论

在本文中，我们提出了 TSPE，这是一种旨在提升现有最先进的音频语言模型（ALMs）零样本性能的任务特定硬提示方法。与文献中常用的普通提示不同，我们的方法通过首先分析各种音频分类任务中的各个类

表 II
任务类别及其示例提示

Task	Prompt Example
Musical Instruments Recognition	The sound of a <violin> coming from an <opera> The sound of an <organ> coming from a <church>
Acoustic Scene Understanding	A <loud> sound of a <jackhammer> coming from a <street> The sound of a <bike> coming from a <road>
Music Genre Classification	The sound of <jazz> coming from a <concert hall> The sound of <rock> coming from a <room>
Impact and Emergency Sound	The sound of <gunshot> coming from a <university> A sound of an <explosion> coming from a <parking lot>
Non-Verbal Vocalization Sound	A <hushed> sound of a <cough> A sound of <laughter> coming from a <hall>

别标签来创建任务特定的提示。然后，我们识别出最相关的声学属性和来源，这些自然定义了每个类别的自然语言描述。最后，我们进行了广泛的定量和定性实验，证明 TSPE 显著优于传统的零样本评估。

VII. 限制与未来工作

1) 提示生成错误：GPT-4 生成的提示中可能出现的错误或重复措辞可能需要仔细的手动筛选。未来的工作将探索自动化质量控制方法以减少对人工监督的需求。
2) 任务特定提示中的偏差：提示的定制可能会引入任务特定的偏差到模型中。未来的工作将专注于识别和减轻这些偏差，以确保在多样化的数据集上实现稳健的表现。
3) TSPE 表示的更广泛应用：TSPE 的文本-音频表示可以增强其他任务，例如音频生成和声音事件定位。未来的研究将包括将 TSPE 扩展到这些应用。

参考文献

[1] Tian, Mi, et al. "A study of instrument-wise onset detection in Beijing opera percussion ensembles." 2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2014.

[2] Akshay Anantapadmanabhan, Ashwin Bellur, and Hema A. Murthy. 2014. Mridangam stroke dataset (1.0). 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada. Data set.

[3] Akshay Anantapadmanabhan, Ashwin Bellur, and Hema A. Murthy. 2014. Mridangam stroke dataset (1.0). 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada. Data set.

[4] Engel, Jesse, et al. "Neural audio synthesis of musical notes with wavenet autoencoders." International Conference on Machine Learning. PMLR, 2017.

[5] Engel, Jesse, et al. "Neural audio synthesis of musical notes with wavenet autoencoders." International Conference on Machine Learning. PMLR, 2017.

[6] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." IEEE Transactions on speech and audio processing 10.5 (2002): 293-302.

[7] Jeong, Il-Young, and Jeongsoo Park. "CochlScene: Acquisition of acoustic scene data using crowdsourcing." 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2022.

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

[9] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2017b. Tut acoustic scenes 2017, development dataset. Data set.

[10] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." Proceedings of the 22nd ACM international conference on Multimedia. 2014.

[11] Tito Spadini. 2019. Sound events for surveillance applications (1.0.0). Data set.

[12] Gong, Yuan, Jin Yu, and James Glass. "Vocalsound: A dataset for improving human vocal sounds recognition." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[13] Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).

[14] Jin, Woojeong, et al. "A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models." arXiv preprint arXiv:2110.08484 (2021).

[15] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.

[16] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[17] Wu, Ho-Hsiang, et al. "Wav2clip: Learning robust audio representations from clip." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[18] Guzhov, Andrey, et al. "Audioclip: Extending clip to image, text and audio." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[19] Elizalde, Benjamin, et al. "Clap learning audio concepts from natural language supervision." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[20] Sahoo, Pranab, et al. "A systematic survey of prompt engineering in large language models: Techniques and applications." arXiv preprint arXiv:2402.07927 (2024).

[21] Ghosh, Sreyan, et al. "Compa: Addressing the gap in compositional reasoning in audio-language models." arXiv preprint arXiv:2310.08753 (2023).

[22] Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[23] Zaman, Khalid, et al. "A survey of audio classification using deep learning." IEEE Access (2023).

[24] Wen, Yuxin, et al. "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery." Advances in Neural Information Processing Systems 36 (2024).

[25] Li, Yiming, Xiangdong Wang, and Hong Liu. "Audio-Free Prompt Tuning for Language-Audio Models." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

[26] Liang, Jinhua, et al. "Adapting language-audio models as few-shot audio learners." arXiv preprint arXiv:2305.17719 (2023).

[27] Wu, Jiayang, et al. "Multimodal large language models: A survey." 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023.

[28] Bai, Tianyi, et al. "A Survey of Multimodal Large Language Model from A Data-centric Perspective." arXiv preprint arXiv:2405.16640 (2024).

[29] Jin, Yizhang, et al. "Efficient multimodal large language models: A survey." arXiv preprint arXiv:2405.10739 (2024).

[30] Szot, Andrew, et al. "Grounding Multimodal Large Language Models in Actions." arXiv preprint arXiv:2406.07904 (2024).

[31] Carolan, Kilian, Laura Fennelly, and Alan F. Smeaton. "A Review of Multi-Modal Large Language and Vision Models." arXiv preprint arXiv:2404.01322 (2024).