Towards Audio-based Zero-Shot Action Recognition in Kitchen Environments

Alexander Gebhard[1,2], *Andreas Triantafyllopoulos*[1,2], *Iosif Tsangko*[1,2], *Björn W. Schuller*[1,2,3,4]

[1]CHI – Chair of Health Informatics, TUM University Hospital, Germany
[2]MCML, Munich Center for Machine Learning, Germany
[3]MDSI, Munich Data Science Institute, Germany
[4]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

***Abstract*—Human actions often generate sounds that can be recognized to infer their cause. In action recognition, *actions* can usually be broken down to a combination of *verbs* and *nouns*, of which there exist a very large number of enumerations. Contemporary datasets, like EPIC-KITCHENS, cover a wide gamut of the potential action space, but not its entirety. Arguably, the holistic characterization of human actions through the sounds they generate requires the use of zero-shot learning (ZSL). In this contribution, we explore the feasibility of ZSL for recognizing a) nouns, b) verbs, or c) actions on Epic-Kitchens. To achieve this, we use linguistic intermediation, by generating descriptions of each word corresponding to our classes using a pre-trained large language model (LLAMA-2). Our results show that human action recognition from sounds is possible in zero-shot fashion, as we consistently obtain results over chance.**

***Index Terms*—zero-shot classification, computer audition, machine learning, action recognition**

1. INTRODUCTION

Despite significant strides in the field of computer audition in recent years [1], achieving a level of auditory perception comparable to that of humans remains a considerable challenge [2]. In the case of humans interacting with their environment, achieving a comprehensive understanding of audio involves the intricate task of perceiving all actions embedded within a particular soundscape. This challenge is amplified by the fact that, in psychology, the identification of sound events by humans is deeply intertwined with the recognition of associated actions [3]. However, the multifaceted nature of human interactions with their environment gives rise to a seemingly inexhaustible array of potential action categories. This inherent complexity poses a significant challenge, making it impractical to train a model capable of recognizing every conceivable category or combination. In this regard, zero-shot learning (ZSL) proves ideal, as it generalizes findings from known classes and their combinations to new ones and thus excels in identifying categories that were previously unknown or unseen [2], [4].

As for ZSL, the majority of progress in zero-shot action recognition (ZSAR) has predominantly occurred within the field of computer vision, leveraging semantic information such as video captions and other text data [5]–[7]. Accordingly, mostly visual or audio-visual data have been investigated, not only for ZSAR but action recognition (AR) in general [8]–[11], while audio was often neglected. Some popular datasets employed for AR, such as HMDB51 [12] or UCF101 [13], do not even contain audio or only have partial audio information. As mentioned by Elizalde et al. [3], audio is an extremely important factor for people to be able to recognize actions. In this regard, verbs are often closely tied to characteristic sounds that reflect actions, interactions between objects, and occasionally the material composing the objects [3]. We therefore want to investigate an audio-based ZSAR approach in this study.

This work was partially funded from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

Similar to computer vision, it is common for audio-based ZSL approaches to leverage textual descriptions of the target classes, their features, or related information as meta information [14]–[17]. These auxiliary data can be textual descriptions of the sound classes or even the labels themselves [14], descriptions of what the target classes sound like [15], or descriptions of the musical concepts which shall be modeled in music [17]. Among the latest breakthroughs when it comes to audio-based ZSL are modifications of the CLIP approach in computer vision, exemplified by Wav2CLIP [18], AudioCLIP [19], or CLAP [16].

While ZSL has gained popularity in the audio domain, we have not come across any prior studies exploring audio-based ZSAR. This paper takes a step in that direction by carrying out initial investigations. For this purpose, we employ the EPIC-KITCHENS dataset [20], [21], which contains egocentric videos of people interacting with objects in their home kitchen environments. We chose this dataset due to the non-scripted daily activities, the availability of audio for all annotated actions, as well as the fact that the videos were narrated by the participants themselves afterwards. Furthermore, each action comprises a verb and a noun (e. g., "cut tomato"), enabling their separate investigation.

We also note that the recently published EPIC-SOUNDS dataset [22] also holds the potential to capture audible actions. Nevertheless, the dataset's action classes predominantly center around collisions and materials, giving rise to classifications like *"metal-only"* or *"cut / chop"*. Our objective, however, is to delve into the identification of the actual object(s) engaged in an interaction, alongside discerning the corresponding verb that describes or characterizes the action.

We primarily want to investigate if ZSAR based solely on audio is possible. To do this, we adopt artificially generated textual descriptions of how the actions sound as meta information. Thus, for each annotated action, we use LLAMA-2 [23], a large language model (LLM), to generate a corresponding textual description. This description is then adopted as auxiliary information for the ZSL process. We draw inspiration for this approach from the usage of artificially generated video captions in [6] as well as the textual sound descriptions for various bird species in [15] for ZSL. We furthermore distinguish the three scenarios of classifying 1) the verb, 2) the noun of an action, as well as 3) the action itself. In our modeling approach, we leverage recent research in audio-based ZSL and primarily utilize a standard ZSL method [14], [15]. Our objective is to conduct initial experiments rather than pursue state-of-the-art results. In doing so, we also explore the textual embeddings of the verbs and nouns and how they influence the model performance. The code repository of this work is publicly available on GitHub[1].

[1]https://github.com/CHI-TUM/epic-kitchens-zsl

Table 1: Statistics about the utilized data. The minimum (Min), maximum (Max), and average (Avg) are reported w. r. t. the **amount** (left) and the **total duration** (right) of the audio segments per class.

	$\sum$			Total Duration (in s)		
Class	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**
VERB	1	14 648	680	1.1	37 802	2 118
NOUN	2	3 576	330	1.7	9 032	716
ACTION	1	1 768	19	.3	3 688	59

2. DATA

We utilize the publicly available data from both EPIC-KITCHENS-100 [21] and EPIC-KITCHENS-55 [20] for our experiments. In particular, we adopt the training data from both datasets and split them into our own subsets in Section 3.2. As the test data does not come with annotated actions we neglect it. Based on the timestamps of the annotated actions we extract the action segments from the provided videos and convert them to .mp3 format, as we only exploit the audio stream. In total, we have 57 hours of audio, averaging 3.12 seconds per file.

The annotated actions are based on the narrations by the participants themselves. The core components of a narration are one verb and at least one noun. If multiple nouns are present in a narration, only the first one is considered for the action, as done in the original paper describing the dataset [20]. An action a_i for an audio segment i is defined as $a_i = (v_i, n_i)$ with v_i and n_i being the corresponding verb and noun class. For instance, the narration "add banana to jug" describes the action tuple (add, banana), thus comprising the verb class *add* and the noun class *banana*. Note that we omitted prepositional objects ("to jug"), same as the original authors of the data [20]. In total, our extracted subset from the EPIC-KITCHENS datasets comprises 97 verb, 287 noun, and 3 507 action categories. However, there is a considerable data imbalance regarding the amount of audio segments for each verb / noun / action class which is illustrated by Table 1.

Audio descriptions: We extract artificially generated textual descriptions as meta-information that describe the expected sound of the corresponding action. This approach is inspired by the incorporation of artificially generated video captions in [6] as well as the leveraging of textual sound descriptions in [15]. For this purpose, we employ LLAMA-2[2] [23], specifically, the instruction fine-tuned 7-billion parameter model, to generate adequate textual descriptions, representing the auditory characteristics associated with the specific actions. To achieve this, we utilize an appropriately designed prompting template, tailored for sound description tasks. Integrating the specific action as a variable in the prompting strategy, we guide the LLM in producing the final description. The utilized prompt is presented in Table 2. The following quote, belonging to the narration "add banana to jug" from above, gives an impression of these generated descriptions:

> "Adding a banana to a jug produces a distinctive 'slosh' sound, followed by a slight 'gurgling' or 'splashing' noise as the fruit settles at the bottom of the container."

3. METHODOLOGY

This section describes the employed features and how they are extracted, the utilized zero-shot classification method, and the experimental setup of this study.

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Table 2: Prompt employed with LLAMA-2 to generate textual sound descriptions, where {ACTION} is replaced by each narration from the EPIC-KITCHENS dataset.

Prompt template
<s>[INST] <<SYS>> You are a highly skilled audio engineer with expertise in accurately describing sounds from various actions in everyday life. I will provide you with a "kitchen action" and your task is to give me a short, precise and accurate description of the sound produced by the specific action in the phrase. The question is, how does this action sound like in terms of auditioning<< /SYS>> "{ACTION}". [/INST] This action sounds like: ###Response:

Audio features: We utilize audio spectrogram transformer (AST) embeddings as audio features, obtained through a state-of-the-art AST model[3] [24]. Prior to extracting the embeddings, the audio files are resampled to 16 kHz. The process involves converting each audio file into a 2D array representation, followed by temporal averaging to derive a 1D vector with a dimensionality of 768.

Text embeddings: We also apply a pre-trained Transformer-based language model to obtain representative embeddings for the LLAMA-2 text descriptions. For this purpose, we deploy SENTENCE-BERT (SBERT), an adaptation of the BERT model [25], which was proposed by Reimers and Gurevych [26] and is intended to reflect semantic similarity in generated sentence and paragraph embeddings. As BERT and SBERT showed no discernible difference in [15], we choose the latter variant. This decision is grounded in the notion that extracting the semantic meaning from our text descriptions is expected to be more feasible than handling the onomatopoeia of bird sounds in the case of [15]. We select the paraphrase-multilingual-mpnet-base-v2 model[4] from the available set of pre-trained SBERT models[5] and call the provided pooling method to yield the embedding vector of size 768.

Since we extract the SBERT embeddings for every LLAMA-2 description from Section 2, every audio file has a corresponding text embedding vector. For our ZSL approach, which will be described in Section 3.1, we require one text embedding vector for each class. Considering that a class is usually annotated to multiple audio files, we take into account all of these files and their corresponding LLAMA-2 descriptions. Since we now possess the SBERT embeddings for each of these descriptions, we can average them to obtain a singular text embedding vector for each class. That is, for each of our three use cases of classifying either the 1) verb, 2) noun, or 3) action, we create a .csv file which contains the classes together with their corresponding text embedding vector.

3.1. Model Training

The ZSL methodology implemented in this study is the one from Gebhard et al. [15], building upon the foundation laid by previous studies from Xie et al. [14] and Akata et al. [27]. They employ a compatibility function on an acoustic-semantic projection to classify the sound classes. This compatibility function is leveraged by a ranking hinge loss in their training process, with the sound class exhibiting the highest compatibility deemed correct. The aim is for the top-ranked class embeddings to provide the most accurate description of the audio sample.

Following the approaches of [15] and [14], we employ a single linear layer equipped with as many neurons as the size of the

[3]https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer

[4]https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[5]https://www.sbert.net

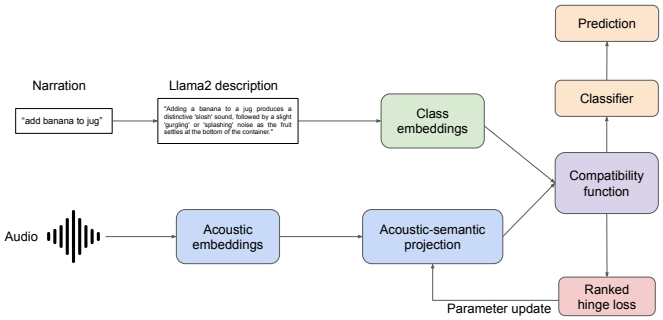


Fig. 1: Overview of the zero-shot learning pipeline adapted from [14], [15]. Audio samples are converted into acoustic embeddings using an AST model and projected into a shared semantic space. LLAMA-2-generated textual descriptions are encoded via SBERT to obtain class embeddings. The classification is based on the highest compatibility score between acoustic and class embeddings.

corresponding class embeddings to project the acoustic embeddings onto the class embeddings. Furthermore, we apply the dot product as our compatibility function. The schematic representation of our pipeline is depicted in Fig. 1.

3.2. Experimental Setup

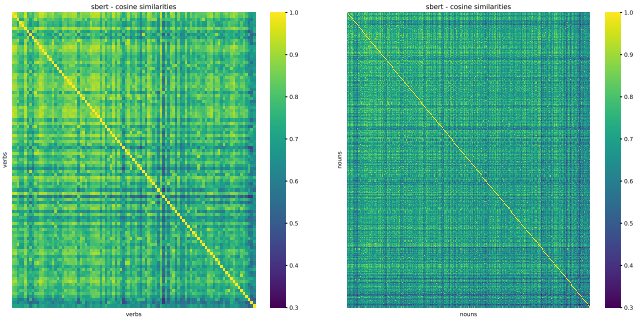
For our experiments, we employed a non-exhaustive cross-validation approach, aiming to ensure an ample amount of data for training. Consequently, we opted for an 80 – 10 – 10 split for each of the five splits. We took explicit care to ensure that these three sets of a split were mutually exclusive, meaning that no classes could appear in the other two sets. In generating the five splits, we also ensured that the development and test sets, in relation to the other four splits, consistently contained different classes, thereby avoiding any overlap.

Our study assesses the efficacy of the ZSL approach outlined in Section 3.1 when paired with the artificially generated meta-information expounded in Section 2. The experiments are executed across the five splits, and the average performance on the development/test sets is reported. We furthermore opt for three different random seeds to initialize our model and also take the mean of those.

We conduct training for a total of 30 epochs, utilizing an Adam optimizer with a learning rate of .0001 and a batch size of 16. As a compatibility function for our ranking loss, we employ the dot product, as mentioned in Section 3.1. Subsequently, the model state demonstrating optimal performance on the development set is selected for evaluation on the test set. Before analyzing the results, we conduct an exploratory data analysis on the text embeddings of the verb and noun classes to gain insights into the underlying relationships. Section 4 offers the results and the corresponding discussion.

4. RESULTS

The metrics used for the evaluation are a superset of the metrics utilized in the EPIC-KITCHENS papers [20], [21]. We use unweighted accuracy (UA), weighted accuracy (WA), and unweighted precision (UP). UA is calculated by computing the recall for each class and then averaging the results; the same is done for UP just with the precision. WA is instead computed by taking the percentage of correctly classified examples – and is thus agnostic to class imbalance. Therefore, WA can be considered as a broad overview of model performance, while UA and UP take class imbalances more into account.



(a) Cosine similarities among the verb embeddings.

(b) Cosine similarities among the noun embeddings.

Fig. 2: The pairwise cosine similarity matrices for the SBERT embeddings of the (a) verb and (b) noun classes depicted as heat maps. Each cell represents the similarity between two class descriptions. The noun embeddings exhibit clearer differentiation, suggesting stronger separability in the text embedding space.

4.1. Exploratory data analysis

For our analysis we resort to cosine similarity visualized as heatmaps and t-SNE pairwise distance visualized as scatter plot.

Cosine similarities: First, we conduct an analysis of pairwise cosine similarities within the text embeddings of the verb and noun classes. The action classes, being composed of verb and noun classes, are not considered in this analysis, allowing us to focus on the smaller components. Our aim is to gauge the strength of the textual embeddings for both the verb and noun classes, to determine which embeddings impart more distinct characteristics. Specifically, we calculate the pairwise cosine similarity between the (S)BERT embeddings of each category and every other category. This computation results in a matrix of cosine similarities w. r. t. those embeddings. Visualized in Fig. 2 as heatmaps, these matrices reveal that noun embeddings manifest more distinct representations across various classes. As a result, we anticipate the noun classification use case to relatively outperform verb classification in terms of model performance. Section 4.2 presents the results and discussions.

t-SNE distances: To further understand the relationship between the text embeddings of different classes, we apply t-SNE (t-Distributed Stochastic Neighbor Embedding) [28] to the textual embeddings for both the verbs as well as the nouns. This way, we can visualize the high-dimensional data in a lower-dimensional, 2D space[6]. Before creating the plots, we annotate the verb and noun categories to the groups specified in [21], to allow a better analysis. There are 13 verb and 21 noun groups. t-SNE has a tendency to group similar data points in the reduced-dimensional space. If embedding vectors from different classes form distinct clusters, it suggests that the original high-dimensional vectors carry information that allows for effective class separation. Knowing this and looking at the plots depicted in Fig. 3 we can, therefore, make several assumptions. For a clearer and more detailed view, we recommend inspecting the interactive plots provided in the supplementary material.

First, when looking at the t-SNE plot of the verb classes it is hard to identify some groups. The only noticeable groups are the verbs belonging to the “monitor” or the “split” group. Unfortunately, even these groups exhibit verb classes that are quite distantly related. However, there are also two small clusters in which the classes are

[6]Interactive t-SNE plots are attached as HTML-files in the supplementary material for easier visualization.

Table 3: The mean results over the development (Dev) and test (Test) sets of the five splits from Section 3.2, also averaged over three different model initialization seeds. The standard deviation (STD) for the three seeds is also presented. The displayed metrics are WA, UA, and UP). The UA-score poses the main evaluation metric. The random chance UA for the categories would be .100, .034, and .003 for the verb, noun, and action class, respectively.

	Dev			Test		
Set	**UA**	**UP**	**WA**	**UA**	**UP**	**WA**
VERB	**.179**	.144	.261	**.165**	.136	.194
STD	.005	.024	.024	.023	.040	.087
NOUN	**.065**	.061	.081	**.066**	.066	.129
STD	.008	.014	.010	.007	.012	.005
ACTION	**.027**	.031	.056	**.029**	.031	.062
STD	.001	.001	.002	.001	.001	.002

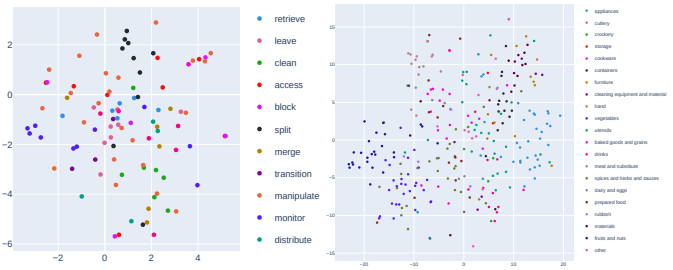
semantically connected to each other but belong to different groups. The first one is on the top right corner of the plot. It consists of four different groups comprising semantically connected verbs (knead, roll, form, unroll, stretch, fold) in the context of food preparation or cooking, especially in the process of manipulating or shaping food or dough. The other small cluster can be identified at the bottom center of the plot. The contained verbs (soak, pour, fill, wash, filter, etc.) are related through the general theme of actions associated with the handling and manipulation of liquids, particularly water.

Regarding the nouns, we can determine more distinct groups, even though there is still a lot confusion in the center of the plot. Some well recognizable groups are the nouns belonging to food categories, such as “fruits and nuts”, “meat and substitute”, or “vegetables”. In addition, other groups that are easily distinguishable are “appliances” or “materials”. Even though some groups can sometimes be mixed with other groups, such as “vegetables” and “spices and herbs and sauces”, this makes sense, since these groups are highly related in reality as well. In fact, most of the food categories are grouped close to each other or are intertwined. There are also some groups which are quite versatile in reality and have connections to many other categories, which is reflected by them being hard to group and distinguish in the plot, such as “utensils” or “rubbish”. In general, it seems as the food categories are more on the left side of the plot whereas the furniture and materials categories are more on the right side. However, there are also some noun groups which are hard to distinguish from the others, such as “utensils” or “furniture”. These analyses are another indicator for a better model performance when regarding the noun classification compared to the verb classification.

4.2. Experimental Results

The zero-shot classification (ZSC) results are depicted in Table 3 and show that each of the three uses cases 1) verb, 2) noun), and 3) action classification, mentioned in Section 1, surpassed random chance UA. We have 10 verb, 29 noun, and 251 action test classes in each of the five splits, leading to random UAs of .100, .034, and .003, respectively. As expected in Section 4.1, the noun classification could achieve better results w. r. t. relative model performance than the verbs. This suggests that the better distinction for nouns, indicated by the similarity heatmaps and t-SNE plots in Section 4.1, transferred to improved zero-shot classification performance as well.

The ZSC results w. r. t. the action classes are the best out of the three uses cases, relatively spoken, as they achieve a UA ten times higher than random guessing. It is also the most stable across the three seeds. Our interpretation in this regard is based on the fact that an action in this study consists of a verb and noun class. Accordingly, a verb can theoretically be associated with all existing noun classes and vice versa. This, in turn, means that while an action class is unique, the verb or noun of the action can also appear in other actions. For instance, the verb “add” from the action “(add, banana)” is also part of the action “(add, apple)”. Therefore, we assume that even in our zero-shot setting, where no training class is present in the test set, the information about the verb or noun class is incorporated in the text embedding of other actions and thus also trained on – i.e., there is some amount of *information leakage* between the train and the test set. This furthermore suggests that the model recognizes this partial knowledge of seen in unseen classes. This could explain the substantial performance gap observed between the action and the verb / noun ZSC. However, coming back to the main goal of this work, we achieve results better than chance for each of the three use cases, confirming the feasibility of audio-based ZSAR.



(a) t-SNE similarities of the verb groups. (b) t-SNE similarities of the noun groups.

Fig. 3: The t-SNE pairwise similarities for the SBERT embeddings of the (a) verb and (b) noun groups. While noun groups exhibit more distinct clusters (especially food-related categories), verb groups appear less separable

5. CONCLUSION

In this study we investigated the feasibility audio-based ZSAR on the EPIC-KITCHENS dataset. Textual descriptions of the action sounds were artificially generated by an LLM and leveraged as meta information. Our model surpassed random chance in all three use cases 1) verb, 2) noun, and 3) action ZSC, achieving a mean UA of .165, .066, .029 for the verb, noun, and action classes over the five test sets and three seeds. The unweighted random chance of guessing would be .100, .034, and .003 for the verb, noun, and action class, respectively. The t-SNE visualizations based on the textual embeddings revealed challenges in verb grouping, while noun categories demonstrated clearer distinctions, especially in food-related classes. In general, ZSC performed exceptionally well on the action task, attaining ten times better performance than random chance – indicating the model’s ability to recognize partial knowledge of seen in unseen classes. These findings support the potential of audio-based ZSAR and showcase promising results for future applications.

Future work can analyze a variety of other meta information, such as extracting the textual embeddings directly from LLAMA-2 or meta information from other modalities, such as image or video embeddings. Since vision has typically been the main modality [10] it would be interesting to apply it as additional information. The model performance can further be improving by a more tight integration of audio and large language models, as in the case of CLAP [16] or Pengi [29].

REFERENCES

[1] A. Triantafyllopoulos, I. Tsangko, A. Gebhard, A. Mesaros, T. Virtanen, and B. W. Schuller, "Computer audition: From task-specific machine learning to foundation models," *Proceedings of the IEEE*, vol. 113, no. 4, pp. 317–343, 2025.

[2] B. Schuller, A. Baird, A. Gebhard, S. Amiriparian, G. Keren, M. Schmitt, and N. Cummins, "New Avenues in Audio Intelligence: Towards Holistic Real-life Audio Understanding," *Trends in Hearing*, vol. 25, pp. 1–14, 11 2021.

[3] B. Elizalde, R. Revutchi, S. Das, B. Raj, I. Lane, and L. M. Heller, "Identifying actions for sound event classification," in *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2021*. IEEE, Oct. 2021.

[4] O.-B. Mercea, T. Hummel, A. S. Koepke, and Z. Akata, *Temporal and Cross-modal Attention for Audio-Visual Zero-Shot Learning*. Springer Nature Switzerland, 2022, p. 488–505.

[5] X. Xu, T. Hospedales, and S. Gong, "Semantic embedding space for zero-shot action recognition," in *Proceedings of the International Conference on Image Processing (ICIP) 2015*. IEEE, Sep. 2015.

[6] V. Estevam, R. Laroca, H. Pedrini, and D. Menotti, "Tell me what you see: A zero-shot action recognition method based on natural language descriptions," *Multimedia Tools and Applications*, vol. 83, no. 9, p. 28147–28173, Sep. 2023.

[7] S. Chen and D. Huang, "Elaborative rehearsal for zero-shot action recognition," in *Proceedings of the International Conference on Computer Vision (ICCV) 2021*. IEEE, Oct. 2021.

[8] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proceedings of the International Conference on Computer Vision (ICCV) 2019*. IEEE, Oct. 2019.

[9] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "Mict: Mixed 3d/2d convolutional tube for human action recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) 2018*. IEEE, Jun. 2018.

[10] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, p. 1366–1401, Mar. 2022.

[11] D. Ahn, S. Kim, H. Hong, and B. Chul Ko, "Star-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) 2023*. IEEE, Jan. 2023.

[12] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV) 2011*. IEEE, Nov. 2011.

[13] Z.-X. GUO, Y. YI, and H.-J. LI, "Recognizing actions from videos in the wild via adaptive feature fusion: Recognizing actions from videos in the wild via adaptive feature fusion," *Chinese Journal of Computers*, vol. 36, no. 11, p. 2330–2339, Mar. 2014.

[14] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2021.

[15] A. Gebhard, A. Triantafyllopoulos, T. Bez, L. Christ, A. Kathan, and B. W. Schuller, "Exploring Meta Information for Audio-based Zero-Shot Bird Classification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024*, IEEE. Seoul, South Korea: IEEE, 4 2024, 5 pages.

[16] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*. IEEE, 2023, pp. 1–5.

[17] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A Joint Embedding of Music Audio and Natural Language," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, Nov. 2022, pp. 559–566.

[18] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022*. IEEE, 2022, pp. 4563–4567.

[19] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022*. IEEE, 2022, pp. 976–980.

[20] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 11, pp. 4125–4141, 2021.

[21] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision (IJCV)*, vol. 130, p. 33–55, 2022.

[22] J. Huh, J. Chalk, E. Kazakos, D. Damen, and A. Zisserman, "EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, 2023.

[23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[24] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech 2021*. ISCA, 2021, pp. 571–575.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[26] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[27] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1425–1438, 2016.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[29] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 18 090–18 108.