TOWARDS DIALOGUE MODELING BEYOND TEXT

Tongzi Wu[⋆†], [*] *Yuhao Zhou* [⋆], *Wang Ling* [⋆], *Hojin Yang* [⋆], *Joana Veloso* [⋆]
Lin Sun [⋆], *Ruixin Huang* [⋆], *Norberto Guimaraes* [⋆], *Scott Sanner* [†]

[⋆] Talka AI, Toronto, ON, Canada [†] University of Toronto, Toronto, ON, Canada

ABSTRACT

In this paper, we model aspects of communication beyond the words that are said. Specifically, we aim to detect interruptions and active listening events, which are important elements in any dialogue. We build a dataset with fine-grained annotations for each category and train multimodal models that take into account all channels in a digital conversation, that is, the video, the audio, and the text. Our experiments show that multimodality is a necessary component in modeling the complexity of the non-textual components of the conversation as different artifacts require different modalities to capture effectively.

***Index Terms*—** Multimodality, Speech Recognition, Video Processing, Machine Learning, Dialogue

1. INTRODUCTION

A large body of research has been applied in modeling dialogue focusing on spoken or edited text [1], or in detecting events that are orthogonal to the dialogue [2]. However, a large part of the communication occurs through *non-verbal* artifacts [3]. Thereby, the usage of text-only representations of dialogues poses an incomplete view of the information being transmitted. An obvious example is the fact that a simple nod is equivalent to speaking the words "Yes" or "I agree", but without a textual footprint. In fact, most natural dialogues have an overwhelming amount of non-textual events, such as *non-verbal* active listening [4], where the listener directly interacts with the speaker by nodding or smiling. Another frequent non-textual cue is interruptions [5], where the flow of the speaker's discourse is disrupted, often resulting in overlapping discourse [6]. These are not only difficult to map into text due to the overlapping discourse [7] but also not always detectable in text, for instance, when the speaker intends to say "I am going to buy a carpet for my new house" and is interrupted after the word "carpet", the resulting sentence "I am going to buy a carpet" is still valid.

Thus, with the goal of supporting the field of dialogue towards incorporating *non-verbal* events, we build a dataset of fine-grained classification of interruption and active listening events (Section 2). Then, we present a multimodal model that uses textual, audio and visual features to detect these events (Section 3). Our experimental results show that both visual and auditory channels are needed for learning to classify the whole range of events (Section 4). Finally, we tie our work with previous work in Section 5 and conclude in Section 6.

[*] Work done during an internship at Talka AI.

2. DATASET

Our dataset is composed of interruptions and active listening events with fine-grained subcategorizations over 1109 Zoom meetings spanning 801.8 hours of video.

2.1. Interruptions

Interruptions break the current speaker's discourse. Following [6, 8], we subcategorize interruptions into *competitive* and *cooperative* interruptions. *Competitive* interruptions attempt to shift the current speaker's discourse and public's attention towards a different object, whereas *cooperative* interruptions support the speaker and topic being relayed.

2.2. Active Listening

Active listening is a technique used by the listener to improve mutual understanding. While certain aspects of active listening cannot be fully observed, such as not being distracted by unrelated thoughts and paying attention to the speaker's body language, we focus on detecting cues that are observable. In our Zoom recordings, we observe that active listeners frequently give encouraging verbal cues, such as, "That's right" and "uh huh" but also non-verbal cues, such as nodding and smiling. Thus, we subcategorize active listening events as *verbal* and *non-verbal*.

2.3. Annotation Scheme

Each event is annotated as a tuple (t_1, t_2, l), where we annotate the interval $[t_1, t_2]$ milliseconds with label $l \in \mathcal{L}$. Here, $\mathcal{L}$ is composed of the set of possible labels, namely, *cooperative* and *competitive* for interruptions and *verbal* and *non-verbal* for active listening. Finally, different types of events can overlap, for instance, nodding and smiling can occur concurrently.

2.4. Annotation Quality

The annotation is done by a team of annotators. Each annotator went through a screening test before making annotations in the dataset. We then trained each annotator on the descriptions of the events with examples. 200 videos were randomly selected to test for the agreement scores across different annotators, and the agreement score is above 90%.

2.5. Transcription Quality

We used an in-house transcription pipeline that includes speech recognition, diarization, punctuation, inverse-text-normalization (ITN), and capitalization. For speech recognition, our evaluation of a range of data sources (phone calls, podcasts, etc) shows that our

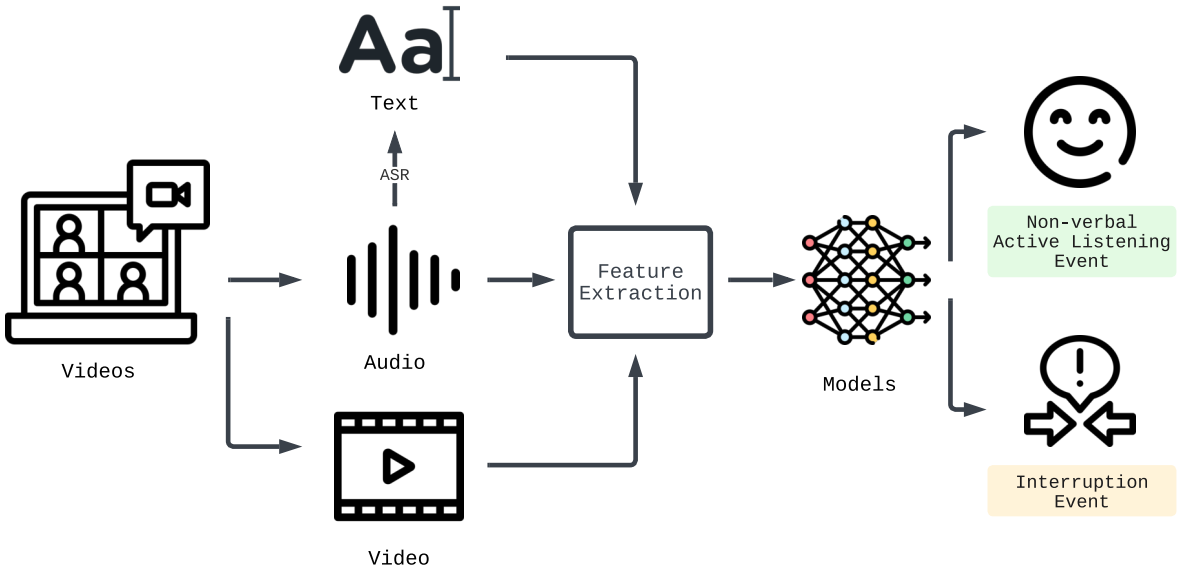

Fig. 1. Event Prediction Model on video data. Videos are extracted, into three separate channels: visual, acoustic and textual. Features are then extracted for each of the channels. A neural network is then used to aggregate the features for event classification.

in-house transcription service has a compelling word-error-rate of 10.9 whereas AWS Transcription Service is 14.3. We believe the quality of the transcription is sufficient for its downstream tasks.

2.6. Statistics

After annotating all videos, we were able to obtain 23633 interruption annotations (16682 *cooperative* and 6951 *competitive* interruptions). Furthermore, we obtained 68040 *verbal* and 4438 *non-verbal* active listening annotations. It is important to also report that more than 99% of the *non-verbal* active listening annotations relate to one of the participants either smiling or nodding. Furthermore, the number of annotations for *non-verbal* active listening is substantially lower, since most videos used the full-screen mode for recording where only the speaker is visible.

3. MODEL

We devise a model that predicts given a five-second video x, where each event $\mathcal{L}$ is present within x (Figure 1). Thus, the model generates a binary vector $\mathbf{y} = y_1, \ldots, y_{|\mathcal{L}|}$, where each index i confirms the presence or absence of a different event.

3.1. Data Pre-processing

For a k-second video, we extract k five-second segments. These correspond to extracting a moving window starting from $[-4, 1]$ seconds til $[k-4, k]$ seconds, with an interval of one second between windows. Each window is treated as a datapoint (x, y) with labels $\mathbf{y}$. We treat a given label y_i as positive if there is an annotation (t_1, t_2, l), where $[t_1, t_2]$ overlaps with x.

For training, we extract all segments from our inventory of annotated videos. Inference can be performed by iterating through the five-second segments of a given video and predicting the starting point of a given event in the last second of the positively labeled event, and the ending point as the starting time of the first segment where a negative prediction is obtained. For instance, a smile event starting from 5 to 7 seconds is predicted if the segments $[1, 5], [2, 6], [3, 7], [4, 8], [5, 9], [6, 10]$ are labelled positively and $[7, 11]$ is labelled negatively.

3.2. Features Extraction

From each segment x, we extract textual, audio and visual features. This step allows us to tackle the lack of labeled training data by using pre-trained models that have been trained on large amounts of unlabeled data. Furthermore, as text, audio and video have different input formats, this step unifies all modalities into a single vector representation.

Audio Features We followed [9] and extract the audio features from the 5-second segment with wav2vec 2.0 large model [10] pre-trained on [11]. Thus, each audio vector $\mathbf{x}_a$ extracted from x is processed into a sequence of vectors $\mathbf{e}_{\text{a}} = \phi_{\text{wav2vec2}}(x_{\text{a}}), \mathbf{e}_{\text{a}} \in \mathbb{R}^{T_a \times D_a}$, where T_a is the time sequence of the latent 1D-convolutional features, and D_a is the dimension of each audio feature.

Text Features Transcriptions of the videos with word-level timestamps are generated with a finetuned wav2vec 2.0 base model on scrapped YouTube captions. Overlapping texts, denoted as x_t, are extracted based on the start and end time of the 5-second segment.

We then embed the text using the Open Pre-trained Transformer Language Models (OPT) [12]. Each text feature is denoted as $\mathbf{e}_{\text{t}} = \phi_{\text{OPT}}(x_{\text{t}}), \mathbf{e}_{\text{t}} \in \mathbb{R}^{D_t \times T_t}$. Similar to audio features, T_t and D_t are the

lengths and the dimension of the text feature.

Visual Features For visual features, we use the visual encoder CLIP [13] and obtain the visual embedding from the sub-sampled frames in videos $\mathbf{e}_{\text{v}} = \phi_{\text{CLIP}}(\mathbf{x}_{\text{i}}), \mathbf{e}_{\text{v}} \in \mathbb{R}^{D_v \times T_v}$.

3.3. Transformer-based Multi-modal Network

The Transformer model is tested in many studies to be effective when dealing with multiple types of input. We design the second model in our paper based on the 1-D convolutional neural network and classic Transformer encoder.

The audio feature extracted in 3.2 is first pushed through four layers of 1-D convolution neural networks and transformed into audio input denoted as $\mathbf{u}_{\text{a}}^{'} \in \mathbb{R}^{T_a^{'} \times D_a^{'}}$. Then each of the $\mathbf{u}_{\text{a}}^{'}, \mathbf{u}_{\text{t}}, \mathbf{u}_{\text{v}}$ is separately passed to a positional encoding layer followed by a Transformer encoder. The output of the last layer in each Transformer encoder module, denoted by $\mathbf{z}_{\text{a}}, \mathbf{z}_{\text{t}}, \mathbf{z}_{\text{v}}$, are then concatenated in the sequence dimension as $\mathbf{z}$. Finally, $\mathbf{z}$ is input to several linear layers to produce the event possibility.

3.4. MLPMixer-based Multi-modal Network

We design our multi-modal networks based on the MLPMixer model [14], and enhance its performance by applying normalization techniques over different modalities. The advantage of the MLP-Mixer mainly lies in the combination of channel-mixing operations and token-mixing operations, which serves as a substitute for the self-attention mechanisms. While it is originally designed for visual tasks where typically only one image is involved, it is shown in footnote[1] that we are inspired by this multi-modal solution and apply a similar architecture in our design.

With the textual, audio and visual features obtained from 3.2 denoted as $\mathbf{e}_{\text{a}}$, $\mathbf{e}_{\text{t}}$ and $\mathbf{e}_{\text{v}}$, we first apply a dropout layer followed by a layer normalization on each kind of representation for a fair distribution among different modalities, and then concatenate these features to form the input $\mathbf{X} \in \mathbb{R}^{\mathbf{c} \times \mathbf{l}}$, where $\mathbf{c} = \mathbf{c}_a = \mathbf{c}_t = \mathbf{c}_v$ and $\mathbf{l} = \mathbf{l}_a + \mathbf{l}_t + \mathbf{l}_v$. The input $\mathbf{X}$ is then passed to several Mixer layers, which remain the same structure as described in [14], except for a dropout layer placed after the first linear layer in each MLP unit. The output of Mixer layers goes through a global average pooling and a linear classifier to produce predicted logits.

4. EXPERIMENTS

We now perform experiments to establish the multimodal baselines for the dataset we propose and analyze the multimodal aspects of interruptions and active listening.

4.1. Setup

Dataset Splits There are four events $\mathcal{L}$ in our dataset: *Competitive* interruptions, *Cooperative* interruptions, *verbal* and *non-verbal* active listening. To avoid overfitting to particular speakers, we divide the train, dev and test sets, so that videos do not overlap in different sets. We conduct experiments on each kind of event individually, with identical settings: we assigned 80% of the data for training, 10% of the data to the dev set and another 10% to the test set.

Models For the MLPMixer, the input sequence is padded as $T_a = 250$, $T_t = 70$ and $T_v = 12$, while the dimension of each feature is $D = 512$. The MLPMixer model is designed to have 24 Mixer layers, and each Mixer contains the token-mixing MLP dimension as $D_S = 512$ and channel-mixing MLP dimension as $D_C = 1024$. The Transformer model consists of four 1-D convolutional neural networks which are used to process audio input solely, a positional encoding layer combined with a Transformer encoder module applied for each modality, followed by three linear layers. The 1-D convolutional neural networks are with output feature lengths of 512, 128, 128 and 256 respectively. Each positional encoding layer is with a dropout rate of 0.05, followed by a Transformer encoder module, which contains 2 layers of 8 heads Transformer encoder with a dropout rate of 0.1, and finally produces a vector $\mathbf{z}_{\text{modality}} \in \mathbb{R}^{D_{\text{modality}}}$, modality $\in \{\text{audio}, \text{text}, \text{visual}\}$. The output dimensions of the final four linear layers with ReLU activation are defined as 1024, 512, 512, and 2, where the final layer feeds into a softmax layer to produce the probability of an event.

[1] https://pytorch.org/blog/how-disney-improved-activity-recognition-with-multimodal-approaches-with-pytorch/

4.2. Event Detection Results

Accuracy results obtained from our models are reported in Table 1. We can observe that using audio features only (row *Audio*), we can perform well on both interruption and *verbal* active listening events. As expected, using only audio features is not sufficient to predict *non-verbal* active listening events. For *non-verbal* active listening events, visual features are needed (row *Video*). We can also observe that combining the modalities yields the best overall accuracy across different types of events. Additionally, the MLPMixer (row *MLP-Mixer*) that combines different modalities over time outperforms the Transformer for three of the four events (row *Transformer*).

	Interruptions		*Active Listening*	
	Comp	**Coop**	**Verbal**	**Non-Verbal**
Unimodal				
TEXT	73.50	76.08	66.52	67.36
AUDIO	87.26	89.63	89.10	58.20
VIDEO	54.59	59.40	53.06	89.32
Multimodal				
TRANSFORMER	87.26	89.48	89.62	**89.53**
MLPMIXER	**87.74**	**90.18**	**90.05**	89.32

Table 1. Accuracy results on different events using a single modality or multiple modalities. *Competitive* and *cooperative* interruptions are results are illustrated in columns *Comp* and *Coop* and the aspects of verbal and non-verbal active listening events are illustrated in columns *Verbal* and *Non-Verbal*.

It is important to mention that none of the optimal results are obtained using the text input, which shows that other modalities are needed to capture different aspects of communication. Table 2 illustrates some examples with the highest probability for each different event on the development set.

We can see that text can be used to some extent to detect verbal active listening and both interruption events. It can do so by detecting some particular constructions, where the construction of the speaker's sentence is broken (e.g. "how we're gonna no, no, no, no"). However, we observed that in many instances, the ASR does not capture short interruptions, such as "yeah", likely due to low language modeling scores. This suggests that annotations of interruptions and active listening can be used to augment ASR results in dialogues.

Surprisingly, text can be used to some extent to predict non-verbal active listening. We can observe in Table 2 that these correspond to positive remarks with a high chance that the listener is performing active listening (e.g. "really appreciate it. this has been super helpful.").

Non-verbal active listening examples
$(s1)$yeah, of course. um, cool. $(s2)$no, all good.
$(s1)$super helpful. um, cool.
$(s1)$perfect. well, awesome. I'm glad that you came to you.
$(s1)$really appreciate it. this has been super helpful.
Verbal active listening examples
$(s1)$yeah, yeah. $(s2)$ I'm in London. $(s1)$well, yeah. $(s2)$whereabouts are you? $(s1)$north
$(s1)$it really it looks like I'm just a movie set now. $(s2)$yeah. $(s1)$it's all thing. $(s2)$yeah. $(s1)$hopefully it won't
$(s1)$be what we call a contributor. $(s2)$ um, right. $(s1)$and then anyone else your customers are technically contributors
$(s1)$early next week. and $(s2)$ great, $(s1)$ some of that. $(s1)$good. excellent. great. $(s2)$ well,
Competitive interruption examples
$(s1)$ like solitude or not solid. $(s2)$ excuse me. you ever heard a snowbird?,
$(s1)$ I just want to let you know that. $(s2)$ but if you're not a one.
$(s1)$not going 45 min on how we're gonna $(s2)$ no, no, no, no, no, no, no, no. absolutely not. let me ask you,
$(s1)$ and.$(s2)$sorry. let me rephrase. okay. Wednesday,
Cooperative interruption examples
$(s1)$ I see. cool. $(s2)$okay. so it seems like the
$(s1)$ lost track of what we were doing here so join yo, you can correct me from sake. $(s2)$ I think, you know, I think what johnny was trying to say
$(s1)$ there's a status change. $(s2)$ so yeah, that makes sense. and I mean, we're quite a small team.
$(s1)$go for a pint after this. $(s2)$right. where are you based? $(s1)$I'm in. I'm

Table 2. Examples of correct predictions for both verbal and non-verbal active listening, as well as competitive and cooperative interruptions. Markers $s1$ and $s2$ are used to identify different speakers.

5. RELATED WORK

5.1. Self-supervised Learning

Self-supervised learning emerged as a crucial paradigm in recent years. The learning process breaks down into the pre-training phase which learns general-purpose features from unlabeled data and the fine-tuning phase in which the model is applied to labeled data. This has demonstrated impressive results in natural language [15, 16], vision [17, 18, 19], and audio [20, 10]. We make use of the released pre-trained models and finetuned them on conversation-specific features.

5.2. Deep Multimodal Learning

Deep learning algorithms have unlocked avenues in learning from heterogeneous modalities of data. Different fields of research combine multiple modalities for a variety of tasks, including representation learning [21, 22, 23], robotics [24], cross-modality prediction and retrieval [25, 26], etc. Our work relates closely with the field of video understanding [27, 28, 29]. However, the aforementioned works predominantly rely on visual cues in objects and motions. Most of the tasks are centered around actions and procedures, whereas our work focuses on the conversations between people, where features of visual, acoustic, and textual modalities are critical.

5.3. Conversation Dataset

There have been studies on learning the intrinsic features from conversations. Datasets are collected to classify smiles and eyelid positions [30], classify emotions from conversations [31, 32], detect communication-critical events within meetings [4, 33]. We focus on conversational features that rely on temporal features in the form of a video. This greatly extends previous datasets which mostly 1. contain a single modality of audio, or text; and 2. have a limited amount of annotated data.

6. CONCLUSION

We built a dataset of interruptions and active listening in a conversational setup. These events play a critical role in conversations and are not reflected in a transcript stored in text. We propose a multimodal neural network that uses pre-trained embeddings for each modality, which are combined into an established MLPMixer model and a Transformer model for optimal results. Results indicate that the visual and audio channels play a critical role in the detection of interruptions and active listening events. Finally, the combination of these cues is more effective using MLPMixer, which merges the different modalities progressively.

The data presented in this paper will be made available at **https://www.talka.ai**.

7. REFERENCES

[1] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang, "You impress me: Dialogue generation via mutual persona perception," in *ACL*, 2020.

[2] Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea, "Conversational transfer learning for emotion recognition," 2020.

[3] Barbara J. Grosz and Candace L. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, 1986.

[4] Harry Weger Jr., Gina Castle Bell, Elizabeth M. Minei, and Melissa C. Robinson, "The relative effectiveness of active listening in initial interactions," *International Journal of Listening*, 2014.

[5] John Local and Peter French, "Prosodic features and the management of interruptions," in *Intonation in Discourse*. 1986.

[6] Khiet P Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee.," in *Interspeech*, 2013, pp. 1404–1408.

[7] Anshuman Tripathi, Han Lu, and Hasim Sak, "End-to-end multi-talker overlapping speech recognition," in *ICASSP 2020-*

2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6129–6133.

[8] Li-chiung Yang, "Interruptions and intonation," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 3, pp. 1872–1875.

[9] Henry Zhou, Alexei Baevski, and Michael Auli, "A comparison of discrete latent variable models for speech representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3050–3054.

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

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

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al., "MLP-Mixer: An all-MLP architecture for vision," *NeurIPS*, 2021.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, 2018.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," *CoRR*, 2019.

[18] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord, "Data-efficient image recognition with contrastive predictive coding," *CoRR*, 2019.

[19] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui, "Spatiotemporal contrastive video representation learning," *CoRR*, 2020.

[20] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour, "AudioLM: a language modeling approach to audio generation," 2022.

[21] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency, "Multiscale benchmarks for multimodal representation learning," *CoRR*, 2021.

[22] Andrew Owens and Alexei A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *CoRR*, 2018.

[23] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi, "Through-wall human mesh recovery using radio signals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10113–10122.

[24] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman, "Audio-visual embodied navigation," *CoRR*, 2019.

[25] Yan Zeng, Xinsong Zhang, and Hang Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," *CoRR*, 2021.

[26] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "Visual question answering," *CoRR*, 2015.

[27] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi, "MERLOT Reserve: Neural script knowledge through vision and language and sound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16375–16387.

[28] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees G. M. Snoek, "Audio-adaptive activity recognition across video domains," 2022.

[29] Yuhao Zhou, Makarand Tapaswi, and Sanja Fidler, "Now you shake me: Towards automatic 4D cinema," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7425–7434.

[30] Wen-Ming Cao, Ning Li, et al., "Face recognition based on manifold learning and renyi entropy," *Journal of Intelligent Learning Systems and Applications*, vol. 2, no. 1, pp. 49–53, 2010.

[31] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, 2019.

[32] Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu, "A large-scale Chinese personalized and emotional dialogue dataset for conversational AI," 2022.

[33] Chung-Hsien Wu, Wei-Bin Liang, and Jui-Feng Yeh, "Interruption point detection of spontaneous speech using inter-syllable boundary-based prosodic features," *ACM Trans. Asian Lang. Inf. Process.*, vol. 10, 2011.