MISAR: A Multimodal Instructional System with Augmented Reality

Jing Bi[*], Nguyen Manh Nguyen[*], Ali Vosoughi[*], and Chenliang Xu

University of Rochester, Rochester, NY

{jing.bi@, nguyen.nguyen@, mvosough@ur., chenliang.xu@}rochester.edu

Abstract

Augmented reality (AR) requires the seamless integration of visual, auditory, and linguistic channels for optimized human-computer interaction. While auditory and visual inputs facilitate real-time and contextual user guidance, the potential of large language models (LLMs) in this landscape remains largely untapped. Our study introduces an innovative method harnessing LLMs to assimilate information from visual, auditory, and contextual modalities. Focusing on the unique challenge of task performance quantification in AR, we utilize egocentric video, speech, and context analysis. The integration of LLMs facilitates enhanced state estimation, marking a step towards more adaptive AR systems. Code, dataset, and demo will be available at https://github.com/nguyennm1024/misar.

1. Introduction

The effective design of augmented reality (AR) systems hinges on multimodal interfaces, blending visual, auditory, and linguistic elements to optimize human-computer interactions. When traditional text-based methods (e.g keyboard input) are not feasible or efficient for communication, auditory communication becomes vital, enabling real-time assistance even when visual information is limited. Similarly, visual data interpretation enhances AR system functionality by analyzing the user's environment, allowing for contextual support and proactive guidance. These auditory and visual capabilities together [11] contribute to a more predictive and responsive AR system, thereby elevating the efficiency and accuracy of task completion.

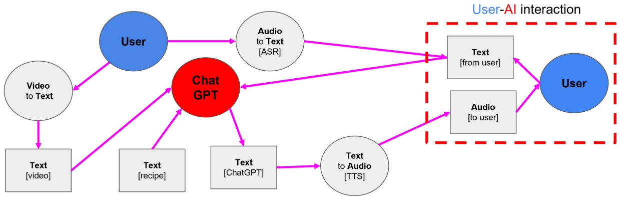


Figure 1. Architecture of the proposed multimodal integration model, highlighting GPT3.5-Turbo (*aka* ChatGPT) as a central component. The model seamlessly integrates video-to-text, text-to-speech, audio-speech recognition, and GPT3.5-Turbo. However, errors may be induced during the video-to-text conversion process, and the model does not encompass understanding of environmental audio

The evolution of large language models (LLMs) has elevated language from a basic communication tool to a critical modality in computational reasoning and decision-making, integrating elements of context interpretation and inferential reasoning into the computational process [18, 5, 9, 6]. While significant research has been dedicated to converting various sensory modalities into textual formats for applications such as video narration and complex dialog systems [18, 16, 8, 5], these studies have yet to robustly address the intricate challenges specific to the augmented reality (AR) environment.

Building on previous research, we introduce a novel method that integrates multiple modalities—visual, auditory, and contextual—using language models for reasoning and human interaction [18, 5, 9, 6]. Specifically, we address the unexplored problem of applying language models to AR for task performance quantification through egocentric video, speech, and context analysis. Our proposed architecture, illustrated in Fig. 1, positions GPT3.5-Turbo as the central component, aggregating information from diverse sources including visual cues, user and recipe inputs, and auditory data. These inputs inform GPT3.5-Turbo's contextually relevant responses and enable accurate state estimation of the user. The goal is to leverage the reasoning capabilities of language models to fill existing gaps in state estimation, moving towards a fully intelligent system.

In summary, our work makes two main contributions.

[*] Equal contribution.

Approved for public release; distribution is unlimited.

This work has been supported by the Defense Advance Research Projects Agency (DARPA) under Contract HR00112220003. The content of the information does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

First, we introduce a comprehensive approach that uses AI to assist people in working more accurately and efficiently. Second, we show that using Large Language Models (LLMs) to enhance video captions improves the quality of these descriptions, enabling effective scene understanding without training with domain specific data. These contributions significantly advance the field of intelligent computational systems.

2. Related Works

Our work encompasses multiple modalities, including speech, video, and natural language. In this section we highlight the related work to build the system that we aim.

Large Language Models: Large Language Models (LLMs), a remarkable advancement in artificial intelligence (AI), have instigated transformative changes across multiple disciplines, including computer vision [17, 12], natural language processing [13, 10], and speech processing [15, 4]. Their implications extend to augmented and virtual reality (AR/VR), where they have the potential to serve as critical components. LLMs excel in language comprehension, reasoning, and response generation, rendering them effective decision-making cores in complex systems. In the proposed framework detailed in this paper, we employ GPT-3.5 [3] from GPT3.5-Turbo to serve as the decision-making ”brain” of the system. This highlights the pivotal role that LLMs, particularly GPT-3.5, can play in enhancing the capabilities of AR/VR systems.

Video Captioning: The task of generating natural language descriptions from video, encompassing both third-person and egocentric perspectives, has garnered significant attention within the research community. Recent studies have focused on learning video embeddings that are cross-attended with generative LLMs like GPT-2 [14], aiming to bridge the domain gap between textual and visual representations. In the present work, we employ a pretrained model known as LaViLa [18], for the specific task of video captioning.

Speech Processing: Speech is the most commonly used method of interaction among humans, so is that for the human-computer interaction also it can play a role. In this research we use google API for both speech to text (ASR) and text to speech (TTS) which was found in [1, 2].

3. Method

Our system integrates ASR technology to transcribe verbal inputs into textual format, which is synergistically coupled with audiovisual data streams captured by the Hololens2 camera. This enables the system to acquire a comprehensive perspective of the user's visual field. The video footage procured from the camera is subsequently processed to generate summarizations and textual descriptions. LLMs serve as the computational epicenter of the system, fulfilling dual roles: they not only answer the user questions but also augment the fidelity of captioning in video descriptions, thereby enhancing the system's overall capability in scene understanding tasks.

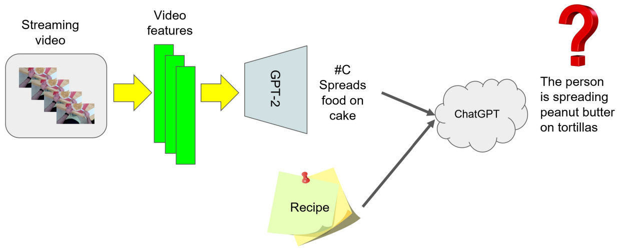


Figure 2. Architecture of the proposed model showcasing the integration of GPT3.5-Turbo (ChatGPT) into the workflow. By incorporating GPT3.5-Turbo, we enhance the quality of generated text descriptions and establish a seamless means of communication between instructional input, users, and the video text.

3.1. Recipe-based Instructional Dialog

In this section, we will present the system's conversation feature. This feature is focused on the two most frequently asked questions of the support system, including asking about steps and asking about how to fix it when you get mistakes.

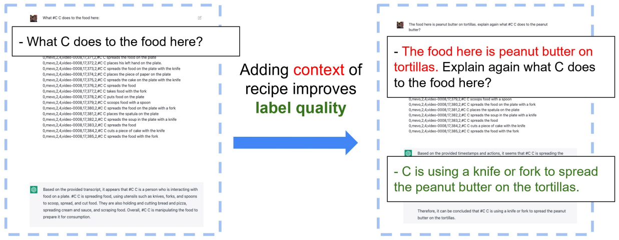
Asking for Step Instructions: As for the question type about the steps to perform a particular task, in our demo we experimented with the question types about what is the next step, what is the previous step. There is also a question form asking how each step is. These types of questions require the system to be able to understand the context of the conversation, understand the recipe, and understand the current state of the user.

Asking for Fixing Mistakes Errors are an inevitable aspect of human performance, particularly for individuals who lack prior experience in executing specific tasks. Consequently, the capacity for an assistive system to facilitate error correction becomes critically important. Providing solutions to rectify these mistakes necessitates that the system possess both generalized knowledge across multiple domains as well as contextual understanding specific to the user's current situation. This dual requirement underscores the need for a nuanced and adaptive system capable of offering effective support for error mitigation.

3.2. Enhancing Video Captions

By integrating GPT3.5-Turbo into the workflow, we can improve the generated text descriptions and establish an effective means of communication between instructional input, users, and the video text. The architecture of the proposed model is shown in Fig. 2.

The integration of GPT3.5-Turbo as a central communi-



What #C C does to the food here:

- What C does to the food here?

```
0,mevo,2,4,video-0008,17,371,2,#C C spreads the food on the plate
0,mevo,2,4,video-0008,17,372,2,#C C places his left hand on the plate.
0,mevo,2,4,video-0008,17,373,2,#C C spreads the food on the plate with the knife
0,mevo,2,4,video-0008,17,374,2,#C C places the piece of paper on the plate
0,mevo,2,4,video-0008,17,375,2,#C C spreads the cake on the plate with the knife
0,mevo,2,4,video-0008,17,376,2,#C C spreads the food
0,mevo,2,4,video-0008,17,377,2,#C C takes food with the fork
0,mevo,2,4,video-0008,17,378,2,#C C puts food on the plate
0,mevo,2,4,video-0008,17,379,2,#C C scoops food with a spoon
0,mevo,2,4,video-0008,17,380,2,#C C spreads the food on the plate with a fork
0,mevo,2,4,video-0008,17,381,2,#C C places the spatula on the plate
0,mevo,2,4,video-0008,17,382,2,#C C spreads the soup in the plate with a knife
0,mevo,2,4,video-0008,17,383,2,#C C spreads the food
0,mevo,2,4,video-0008,17,384,2,#C C cuts a piece of cake with the knife
0,mevo,2,4,video-0008,17,385,2,#C C spreads the food with the fork
```

Based on the provided transcript, it appears that #C C is a person who is interacting with food on a plate. #C C is spreading food, using utensils such as knives, forks, and spoons to scoop, spread, and cut food. They are also holding and cutting bread and pizza, spreading cream and sauce, and scraping food. Overall, #C C is manipulating the food to prepare it for consumption.

The food here is peanut butter on tortillas. explain again what #C C does to the peanut butter?

- The food here is peanut butter on tortillas. Explain again what C does to the food here?

```
0,mevo,2,4,video-0008,17,379,2,#C C scoops food with a spoon
0,mevo,2,4,video-0008,17,380,2,#C C spreads the food on the plate with a fork
0,mevo,2,4,video-0008,17,381,2,#C C places the spatula on the plate
0,mevo,2,4,video-0008,17,382,2,#C C spreads the soup in the plate with a knife
0,mevo,2,4,video-0008,17,383,2,#C C spreads the food
0,mevo,2,4,video-0008,17,384,2,#C C cuts a piece of cake with the knife
0,mevo,2,4,video-0008,17,385,2,#C C spreads the food with the fork
```

Based on the provided timestamps and actions, it seems that #C C is spreading the

- C is using a knife or fork to spread the peanut butter on the tortillas.

Therefore, it can be concluded that #C C is using a knife or fork to spread the peanut butter on the tortillas.

Figure 3. The impact of leveraging larger models such as GPT3.5-Turbo on enhancing the quality of textual labels within the given context. Leveraging larger LLMs and inclusion of contextual information such as recipes or instructions significantly improve the accuracy and richness of generated textual labels, leading to a better understanding and interpretation of the content."

cation point allows for a seamless interaction between users and the video content. Users can provide instructions or ask questions related to specific video segments, and GPT3.5-Turbo responds with informative and contextually relevant explanations. This dynamic interaction not only improves the accuracy and richness of the generated video-to-text labels but also empowers users to actively engage with the content and obtain valuable insights. As shown in Fig. 3, leveraging larger models like GPT3.5-Turbo allows for substantial improvements in the quality of textual labels within the given context. The application of these advanced models enhances the accuracy and richness of generated textual labels, ultimately enhancing the overall understanding and interpretation of the content.

3.3. An User-friendly Interactive System

In the comprehensive system we propose, users can effortlessly engage via AR glasses, facilitating a streamlined human-computer interaction predominantly through speech. The interactive experience is designed to emulate collaboration with a seasoned expert, thereby elevating the user's overall task performance. Moreover, the system serves as a cost-effective alternative to traditional methods of advanced human training, obviating the need for the hiring of specialized instructors. This approach offers a scalable solution for enhancing human performance while reducing associated training costs.

4. How to scale up with multimodal LLMs guidance?

In our video-to-text pipeline, we utilized the LaViLa model [18], as depicted in Fig. 4. Pre-trained on the Ego4D dataset [7], the model is designed to transfer this expertise effectively to Perceptually-enabled Task Guidance (PTG) data, thereby achieving accurate and coherent text generation from video inputs.

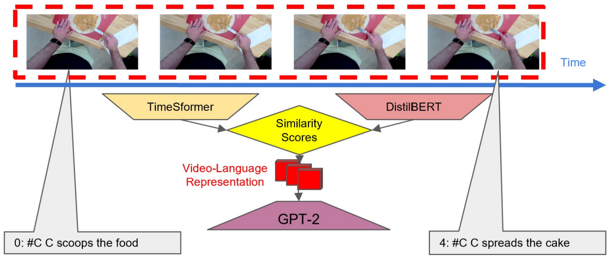
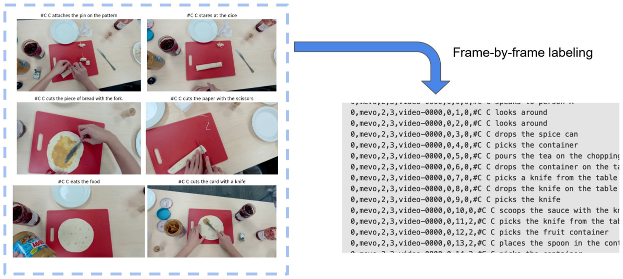


Figure 4. Converting visual content to descriptive text with Video-Conditioned Text Generation using LLMs.


#C C attaches the pin on the pattern
#C C stares at the dice
#C C cuts the piece of bread with the fork.
#C C cuts the paper with the scissors
#C C eats the food
#C C cuts the card with a knife
Frame-by-frame labeling

```
0,mevo,2,3,video-0000,0,1,0,#C C looks around
0,mevo,2,3,video-0000,0,2,0,#C C looks around
0,mevo,2,3,video-0000,0,3,0,#C C drops the spice can
0,mevo,2,3,video-0000,0,4,0,#C C picks the container
0,mevo,2,3,video-0000,0,5,0,#C C pours the tea on the chopping
0,mevo,2,3,video-0000,0,6,0,#C C drops the container on the ta
0,mevo,2,3,video-0000,0,7,0,#C C picks a knife from the table
0,mevo,2,3,video-0000,0,8,0,#C C drops the knife on the table
0,mevo,2,3,video-0000,0,9,0,#C C picks the knife
0,mevo,2,3,video-0000,0,10,0,#C C scoops the sauce with the kn
0,mevo,2,3,video-0000,0,11,2,#C C picks the knife from the tab
0,mevo,2,3,video-0000,0,12,2,#C C picks the fruit container
0,mevo,2,3,video-0000,0,13,2,#C C places the spoon in the cont
```

Figure 5. Automating frame-by-frame labeling and annotation of video files through video-to-text conversion models. This scalable and versatile approach enables efficient natural language labeling across diverse domains. However, it is important to note that the generated text may contain noise, lack crucial details and environmental dynamics, or even produce unrelated annotations.

4.1. Scaling the Natural Language Labeling of Videos

We explore the application of video-to-text conversion as a means of automating the frame-by-frame labeling and annotation process for video files. In this approach, we leverage video Transformers and GPT-2 [14] decoder with cross-modal attention for video understanding, visual-linguistic attention, and natural language generation to streamline the arduous task of labeling videos with textual descriptions. An example of the automated annotation of the videos is shown in Fig. 5. By leveraging pre-trained models, video-to-text conversion enables efficient analysis of the visual content, resulting in the generation of descriptive labels for each frame in the video. Moreover, with its ability to handle large-scale video datasets, video-to-text conversion presents a scalable solution for effectively scaling up the natural language labeling of videos, opening doors to more extensive applications in various domains. Despite the benefits, having language decoders such as GPT-2 makes it vulnerable to noisy label generation, which sometimes is not related to the scene, lacks sufficient information, attends to wrong or less important details, or is entirely unrelated/wrong text. This encourages us to use an even larger LLM to help polish the results of the GPT-2 here.

4.2. Multimodal LLMs Guidance and User-AI Interaction

The proposed model offers a comprehensive integration of various modalities, including video-to-text, text-to-speech, audio-speech recognition, and GPT3.5-Turbo. This multimodal approach ensures that all forms of input, regardless of their original format, are converted into text representations. By unifying the data into a common textual format, the model facilitates seamless integration and interaction across different modalities.

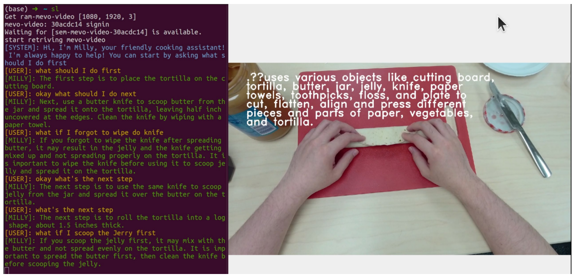
By combining the capabilities of video-to-text, text-to-speech, audio-speech recognition, and GPT3.5-Turbo, the proposed model enables a conversational and interactive experience. Users can provide instructions, queries, or engage in natural language conversations with the system, which responds accordingly based on the fused textual representations. This holistic approach enhances the model's ability to understand and interpret diverse input types, enabling more comprehensive and effective communication between users and the system.

5. Experiments

Data: We collated in-house data on an infrequently documented pinwheel recipe. With the appropriate legal clearances, we utilized a helmet with video and audio functionalities in our lab to record participants making the recipe, ensuring hands were also captured (unlike HoloLens2, which does not record hands, leading us to exclude the use of HL2SS). The resulting videos have varying duration, and they come annotated with the 13 steps of the method. A subset of this data is shared on our GitHub repository.

Video Conversion with LaViLa: For each frame, we produce a sentence, given the video's frame rate of 30 fps. Considering that the quickest human response and action takes around 250 ms, generating captions for every 8 frames aligns with real-time performance needs in physical augmented reality. This frequency matches the swiftest observable human movement.

Results and Discussion: In Tab. 1, the results show that the text after being polished with LLMs was significantly more relevant to the context under consideration. From the trained video to text model with general data, we have made the generated text more contextual, thereby opening the potential for reducing the domain gap between pretrained models and domain specific data without finetuning is required. From there, it shows that our proposed method can be easily adapted to many different domains easily. Tab. 2 illustrates the impact of recipe types on the quality of generated captions. The data reveal that medium-length recipe descriptions yield captions most closely aligned with the intended context. This outcome is logical; brief descriptions lack sufficient information for LLMs to accurately capture context, while overly extensive descriptions introduce noise that compromises caption quality. We demonstrate the efficacy of our dialog system through a live demo. The system is user-friendly and capable of providing immediate responses to user inquiries. During the evaluation phase, the system was subjected to a variety of question types, including those that describe requisite actions for task completion as well as troubleshooting procedures for errors. A screenshot of the demo is shown in Fig. 6. The demonstration video can be accessed at our Github.



```
(base) ➜ ~ sl
Get ram-mevo-video [1080, 1920, 3]
mevo-video: 30acdc14 signin
Waiting for [sem-mevo-video-30acdc14] is available.
start retriving mevo-video
[SYSTEM]: Hi, I'm Milly, your friendly cooking assistant!
 I'm always happy to help! You can start by asking what s
hould I do first
[USER]: what should I do first
[MILLY]: The first step is to place the tortilla on the c
utting board.
[USER]: okay what should I do next
[MILLY]: Next, use a butter knife to scoop butter from th
e jar and spread it onto the tortilla, leaving half inch
uncovered at the edges. Clean the knife by wiping with a
paper towel.
[USER]: what if I forgot to wipe do knife
[MILLY]: If you forgot to wipe the knife after spreading
butter, it may result in the jelly and the knife getting
mixed up and not spreading properly on the tortilla. It i
s important to wipe the knife before using it to scoop je
lly and spread it on the tortilla.
[USER]: okay what's the next step
[MILLY]: The next step is to use the same knife to scoop
jelly from the jar and spread it over the butter on the t
ortilla.
[USER]: what's the next step
[MILLY]: The next step is to roll the tortilla into a log
 shape, about 1.5 inches thick.
[USER]: what if I scoop the Jerry first
[MILLY]: If you scoop the jelly first, it may mix with th
e butter and not spread evenly on the tortilla. It is imp
ortant to spread the butter first, then clean the knife b
efore scooping the jelly.
```

.??uses various objects like cutting board, tortilla, butter, jar, jelly, knife, paper towels, toothpicks, floss, and plate to cut, flatten, align and press different pieces and parts of paper, vegetables, and tortilla.

Figure 6. Screenshot of the demo showcasing the dialog system. Full video is available in our GitHub repository.

Reference recipe	LaViLa [18]	Ours
2-Words Text	0.817	**0.845**
Full sentence	0.841	**0.854**
Descriptive sentences	0.834	**0.848**

Table 1. The numbers indicates the ability of the proposed model to classify the step based on natural language descriptions. The LaViLa method has been compared to our method. The recipe texts used to compute similarity are referred to at three different levels: *2-Words text*, *full sentence*, and *descriptive sentences*. Examples are *"spread butter"*, *"Evenly spread butter on the tortilla"*, and *"Gently spread the scooped butter evenly over the entire surface of the tortilla, leaving a small margin around the edges to allow for spreading when you roll it"* , respectively.

Step	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
#Samples	**11**	**16**	**773**	**152**	**523**	**153**	**387**	**466**	**407**	**338**	**355**	**1752**	**338**
LaViLa - short	0.830	0.810	0.825	0.810	0.849	0.823	0.822	0.802	0.838	0.807	0.797	0.805	0.839
LaViLa - medium	0.823	0.840	0.849	0.837	0.842	0.840	0.847	0.845	0.847	0.840	0.835	0.835	0.844
LaViLa - long	0.822	0.830	0.834	0.833	0.845	0.834	0.838	0.838	0.835	0.834	0.836	0.831	0.827
Ours - short	0.843	0.863	0.846	0.829	0.849	0.841	0.841	0.842	0.850	0.845	0.811	0.846	**0.873**
Ours - medium	**0.845**	**0.865**	**0.853**	**0.858**	**0.847**	**0.858**	**0.849**	**0.862**	**0.854**	**0.848**	**0.840**	**0.858**	0.859
Ours - long	0.842	0.847	0.846	0.850	0.847	0.855	**0.847**	0.847	0.851	0.844	0.842	0.851	0.841

Table 2. The values represent step-wise similarity scores between the generated narratives and the reference recipe descriptions. The table specifies the number of samples for each step, all originating from a single video. Here, "short" corresponds to a 2-word text, "medium" is a full sentence, and "long" refers to detailed sentences as seen in 1.

6. Conclusion

In summary, this paper introduces a system designed to enhance both human's accuracy and efficiency in task performance. Additionally, we advocate for the incorporation of LLMs to improve the quality of video captions. This enhancement opens the way for more effective text-based scene understanding. Our methodology's effectiveness is substantiated through both experimental outcomes and system demo.

References

[1] https://cloud.google.com/speech-to-text. 2

[2] https://cloud.google.com/text-to-speech. 2

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2

[4] Ciprian Chelba, Daniel M. Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. Large scale language modeling in automatic speech recognition. *ArXiv*, abs/1210.8440, 2012. 2

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[6] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2021. 3

[8] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 1

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[10] Stephen MacNeil, Andrew Tran, Dan Mogil, Seth Bernstein, Erin Ross, and Ziheng Huang. Generating diverse code explanations using the gpt-3 large language model. *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2*, 2022. 2

[11] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *ArXiv*, abs/2209.09634, 2022. 1

[12] Nguyen Manh Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, and Minh Hoai. Dictionary-guided scene text recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7379–7388, 2021. 2

[13] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Haiquan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. In *International Conference on Learning Representations*, 2022. 2

[14] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2, 3

[15] Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Ryan Padfield, James Qin, Daniel Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle D. Tadmor, Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovi’c, Damien Vincent, Jiahui Yu, Yongqiang Wang, Victoria Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukás Zilka, and Christian Havnø Frank. Audiopalm: A large language model that can speak and listen. *ArXiv*, abs/2306.12925, 2023. 2

[16] Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. Unveiling cross modality bias in visual question answering: A causal view with possible worlds vqa. *arXiv preprint arXiv:2305.19664*, 2023. 1

[17] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *ArXiv*, abs/2305.18279, 2023. 2

[18] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 1, 2, 3, 4