Egocentric Video Comprehension via Large Language Model Inner Speech

Ying Wang, Dongdong Sun, Rui Chen, Yanlai Yang, Mengye Ren
New York University
{yw3076,ds5749,rc4753,yy2694,mr3182}@nyu.edu

Abstract

The Natural Language Query (NLQ) task involves localizing a temporal window in an egocentric video that provides an answer to a posed query. These queries mimic the cognitive processes of the camera wearer involved in recalling past actions or locating misplaced items. In this work, we present a novel solution to the Ego4D NLQ challenge, inspired by the concept of "inner speech" in cognitive science. Our proposed pipeline first uses image and video captioning models to generate captions that encapsulate sufficient details from the egocentric video. The captions are then fed in a large language model (LLM) to generate coarse-grained predictions containing multiple potential response windows. Finally, a pre-trained NLQ model further filters and refines these windows. Results show that our approach outperforms the baseline NaQ++ ReLER model, suggesting a promising research direction for employing LLMs in video question-answering tasks. Code is available at https://github.com/YingWANGG/LLM-Inner-Speech.

1. Introduction

Given an egocentric video and a query, the NLQ challenge in the Ego4D Episodic Memory task requires localizing a temporal window where the answer to the query can be deduced. These queries, such as "where did I put the keys" and "how many drawers did I open," are closely related to the camera wearer's daily life experience and resemble the situations when people recall their past actions or the location of misplaced items. A myriad of methods has been explored, primarily focusing on improvements in network architecture and leveraging pre-training for better image and video features.

Recent advancements in Large Language Models (LLMs), such as OpenAI's GPT series, have demonstrated the remarkable capabilities of LLMs in their comprehension of both visual and natural language inputs. GPT-4 [17], the most advanced of this series as of our study, has shown competitive zero-shot performance across various vision and natural language learning benchmarks, thereby offering a promising paradigm shift from conventional methods.

Motivated by the success of LLMs and other foundational models in many vision and language tasks, we explore the usage of video captioning models and LLMs towards solving the NLQ task. We employ the concept of "inner speech" in formulating queries for LLMs, specifically OpenAI's GPT-4, to generate responses based on captions from egocentric videos. Inner speech refers to the internal narrative that accompanies one's cognitive processes. This concept has been extensively studied in psychology and cognitive science, with evidence suggesting it plays a crucial role in memory recall, problem-solving, and planning [1, 2, 12, 18].

One of the most challenging aspects of the Ego4D dataset is that the videos are much longer compared to prior egocentric video datasets such as Charades-Ego [21] and Epic-Kitchens [7]. Some prior works [13, 19] fail to consider the long-term temporal dependencies in long Ego4D videos due to architectural limitations. To condense the information in each video to be further processed by an LLM, we use image and video captioning models to generate a descriptive narrative that includes sufficient details about the environment, the subject's actions, and other relevant objects in the footage. This narrative acts as a simulation of inner speech when it is subsequently processed by GPT-4 to generate responses to the posed queries. Due to the restriction on the context length of GPT-4, we only require a coarse-grained prediction containing multiple intervals that potentially contain the answer at this stage. In the end, similar to [10], we use a pre-trained NLQ model to further refine the response windows.

Our proposed approach with GIT as captioning model improves the baseline model (NaQ++ ReLER [20]) by around **0.4%** on the test data and around 1% on the validation data. Utilizing the official narrations from the Ego4D dataset, our method improves the baseline model by around 1.68%, indicating the potential of our method.

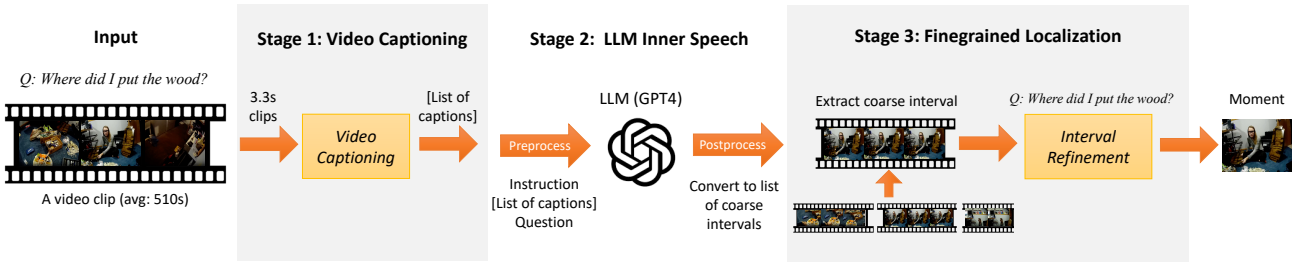

Figure 1. Pipeline of the proposed approach. Since GPT-4 cannot directly process the image or video data, we utilize an image/video captioning model to convert the input video clip into a series of captions (Stage 1). We then combine the resulting captions and the given queries and send the long text as a user prompt to GPT-4, which will generate a data frame as a response (Stage 2). We now obtain a coarse-grained range of candidate temporal windows for each query, which is shorter than the original clip. We then feed the filtered video to a base NLQ model to make the final fine-grained predictions (Stage 3).

2. Related Work

Natural Language Queries in Egocentric Videos. Much progress has been made since the introduction of the Ego4D Natural Language Queries (NLQ) task [9]. This is a challenging task due to the sparsity of annotations and the length of videos in the dataset. [9] evaluates VSLNet [26] on this task as a baseline. ReLER [16] proposes a novel multi-scale cross-modal transformer architecture, a video frame-level contrastive loss, and two data augmentation strategies. InternVideo [4] improves the quality of video features by carefully pre-training and fine-tuning a VideoMAE-L Model [22], and ensemble the features and predictions. More recently, NaQ [20] introduces a data augmentation strategy to transform video narrations into training data for the NLQ task, alleviating the problem of sparse annotation. The current state-of-the-art model, NaQ++ ReLER, is obtained by training the ReLER model with NaQ data.

Large Language Models for Multi-modal Learning. Large Language Models (LLMs) [6, 17, 23] have demonstrated an excellent ability to understand visual and natural language inputs [3]. Many prior works have explored the usage of LLMs in multi-modal learning. LlamaIndex [15] provides a data framework to integrate many different data types into a shared encoding for LLMs to ingest. PaLM-E [8] enhances PaLM [6], a powerful LLM, by adding sensor data from a robot and obtaining in a single model with general capabilities in visual, language, and robotic tasks. [28] proposes an interactive perception framework that uses an LLM to actively acquired information about the environment, reason over multi-modal information, and plan task execution. [27] proposes a chain-of-thought reasoning (CoT) prompting method for LLMs to get better performance on visual question-answering (VQA). The Socratic models [25] quantitatively evaluate the zero-shot reasoning capability of LLM on image captioning and video-to-text retrieval and demonstrate the performance is on par with current standards, while illustrative examples are shown to highlight the potential for broader multimodal applications such as egocentric video question answering and robotic perception and planning. To the best of our knowledge, our work is the first to use a general purpose chat-based LLM in the NLQ task. Similar to [25], we employ a text-based reasoning pipeline; in addition, we propose a novel pipeline for generating temporal localization outputs.

3. Methodology

Our proposed method converts egocentric videos into a series of captions and performs natural language query using LLMs. There are three stages: 1) the first stage requires a comprehensive understanding of the video context, possibly achieved by an image or a video captioning model; 2) the second stage involves an LLM such as GPT-4 effectively utilizing this context to filter the input; and 3) the last stage utilizes a pre-trained NLQ model for fine-grained answers (Figure 1). A detailed breakdown is as follows.

1) Video Captioning. First, we use a number of state-of-the-art image or video captioning models to obtain detailed narratives of given egocentric videos. Since the video data adopts a first-person view and is very different from other video datasets, we finetune existing image/video captioning models on the Ego4d Narration data to better capture the activities of the subject in the video.

Considering the redundancy of neighboring frames and the restriction on the context length, we downsample the original videos before feeding them to captioning models.

2) LLM Inner Speech. We leverage the contextual understanding and reasoning skills of LLMs to process the list of captions from the previous stage and produce a set of relevant intervals related to the natural language query. We use "*You are the person c and want to recall your memory to answer a list of questions.*" as the system prompt to give the model the context of the NLQ task. We then merge preprocessed captions and queries into a template to formulate an instructive and contextualized prompt. An example of the template is shown in the left half of Figure 2. When processing a question, LLMs can take into consideration the full

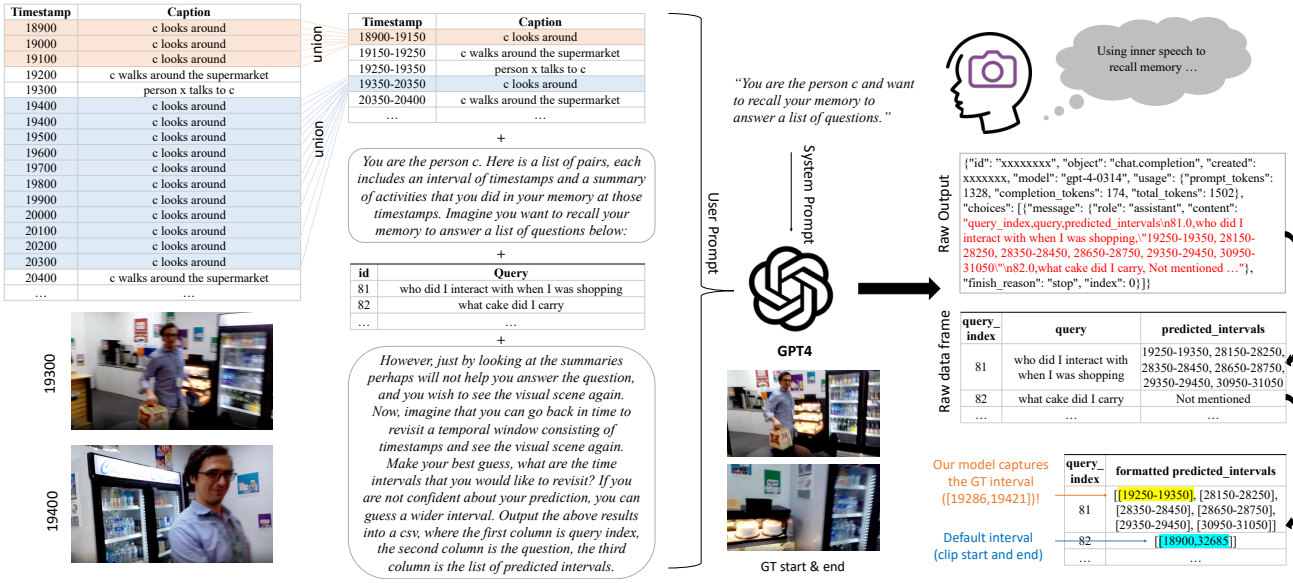

Figure 2. Example input and output of GPT-4. During prepossessing (left), we reduce the total length of captions by removing uninformative phrases such as "in the image" and union neighboring timestamps with the same caption. We then combine captions and queries using the template, encouraging GPT-4 to make reasonable guesses leveraging reasoning skills when the query is not mentioned in the captions. After gathering a response from GPT-4 (right), we convert the response string into a data frame and format the text of predicted intervals as a structured list of intervals. For null predictions, we replace them with the start and end time of the original clip.

context in the template and utilize different pieces of information to produce the most probable answer, mimicking the humans' memory recall process. For example, when asking "*Who did I interact with when I was shopping?*", GPT-4 is able to filter all captions and produce a list of intervals involving "*person x talking to c*" where *c* is the subject in the video and *x* refers to the other person.

Since it is inevitable to lose information when converting videos into texts, we particularly instruct LLMs to imagine the visual scene underlying the given captions. For example, one query asks "*What size of washer did I pick ?*" but there are no captions explicitly mentioning the washers. In this example, GPT-4 displays its capability to capture implicit information and infer based on context. GPT-4 answers "*choosing the time points where I picked items from the table or the floor, as these instances may provide more context about the objects and their locations*." By grasping nuanced relationships and dependencies within the given context, GPT-4 is able to filter out the most relevant information within the extensive video data provided.

We also perform post-processing steps on the generated response shown on the right half of Figure 2. After converting GPT4's predictions into data frames, we process the output string to a list of intervals $[(s_1, e_1), (s_2, e_2) \dots (s_n, e_n)]$. For queries without predictions, we simply use the clip start and end time (s, e) as the default interval.

3) Finegrained Localization. After obtaining a set of filtered time windows, we need to come up with a set of fine-grained timestamps. For this task, we use the current state-of-the-art NLQ model, NaQ++ ReLER [20], as the base model to refine the predictions. Since the original input, a long video clip, is filtered into a list of short candidate intervals, the NLQ model can focus on key events that occur during the specified time windows and thus make more accurate and precise predictions.

To minimize the risk of missing the target temporal window, we extend the predicted intervals by a constant window size α before feeding the filtered video into the NLQ model. Specifically, for each (s_i, e_i), the new start time is $min(s_i - \alpha, s)$ and new end time is $max(e_i + \alpha, e)$ where s and e are the start and end time of the original clip. Note that as α increases, our LLM filtering has less impact.

If the prediction for a certain query contains multiple candidate intervals, we input them separately into the localization model and record the softmax scores of the top 5 predictions for each candidate. We consider all candidates equally and select the top 5 with the highest scores among all predictions as the final output.

4. Experiments

Experiment Setup. For image and video captioning, we first use GIT [24] from HuggingFace,[1] which utilizes CLIP's vision encoder and is pretrained on 0.8 billion image-text pairs, achieving state-of-the-art performance on image/video captioning. We then finetune GIT on Ego4D Narration data to transfer the learned representations to egocentric data. We also experiment with LLAVA [14], a recently released large vision-language model that combines a vision encoder with pre-trained Vicuna [5]. We construct

[1] https://huggingface.co/docs/transformers/model_doc/git

		Validation			Test		
Model	Window	meanR@1	R@1 IoU@0.3	R@1 IoU@0.5	meanR@1	R@1 IoU@0.3	R@1 IoU@0.5
Ours (LLAVA)	240	19.13	22.15	16.11	-	-	-
NAQ++ RELER [20]	-	20.13	22.82	17.45	17.69	21.73	13.66
Ours (IGIT)	240	20.81	**24.16**	17.45	-	-	-
Ours (VGIT)	50	20.81	23.49	**18.12**	17.45	21.50	13.39
Ours (GIT++)	240	**21.14**	**24.16**	**18.12**	**18.07**	**22.02**	**14.11**
Ours (Ego4D Narr.)	50	**21.81**	**24.83**	**18.79**	-	-	-

Table 1. Performance different caption models sorted by meanR@1. IGIT=Image GIT, VGIT=Video GIT, GIT++=combining predications from IGIT+VGIT. When the window size reaches the length of the clip, the results should be exactly the same as directly applying the base model. Thus, we restrict the window size α to under 240s (around one half of the clip length). The **bold** number denotes the highest and the underlined the second highest. The last row represents using existing narrations from the Ego4D dataset, which has higher quality than machine-generated ones.

(a) **MG**: c picks the wheel; **GT**: c turns the handle of the wheel mounting machine with his left hand.

(b) **MG**: c picks a screwdriver from the floor with his left hand. **GT**: c sprays the wheel with the spray bottle in his right hand.

Figure 3. Comparison of captions generated by Image GIT (**MG**) and Ego4D Narrations (**GT**). Ground-truth captions are more accurate and contain more information.

the prompt as "*Provide a one-sentence accurate caption for the given image. The image is egocentric and the caption should focus on the person's interaction with other objects. Never make any guesses about the scene.*" to encourage the model to produce captions similar to the narrations from Ego4d. For all captioning models, we sample once every 100 frames and end up with around 120 unique captions per clip.

To save the expense of OpenAI API requests, we only use 14 clips (with 149 queries in total) from NLQ validation split as our validation dataset for all experiments. The validation and test results are summarized in Table 1. One complete run on the test dataset costs around $25 using OpenAI's GPT4 API.

Results and Discussion. We find that using Ego4D narrations gives the best performance across all metrics and the ensemble of Video GIT and Image GIT (GIT++) is the second. Since ground-truth narrations are not available in the test set, we evaluate with GIT++ on the test set and it achieves **18.07%** meanR@1 which improves the baseline model (NAQ++ RELER) by almost **0.4%**.

By inspecting the generated captions, we observe that the performance of our approach is limited by the quality of the captions. All three captioning models (Image & Video GIT, LLAVA) are biased towards predicting common objects with high frequency in training data (e.g. "screwdriver" is predicted by the captioning model but is absent from Figure 3b). When the captioning model fails to capture the target object but misclassifies something else at other timestamps, GPT-4 will confidently predict a wrong interval because it is provided with an incorrect context, leading to a performance drop compared to the base NLQ model. In contrast, an oracle that uses the Ego4D narration data, which contains fewer but more accurate annotations, outperforms all other models including baselines. This confirms that LLMs such as GPT-4 have strong reasoning skills that extend towards question answering with extra long context, and with the ever-increasing quality of captioning and reasoning delivered by LLMs, we expect that our method has the potential to further improve in the future.

We notice that most of the queries in the NLQ challenge focus on specific objects in the video while captioning models tend to focus on the surrounding context and neglect these small objects. We also experiment with adding a text-conditioned detector (MDETR [11]) to the current captioning model and applying a text-conditioned captioning model (LLAVA) by integrating queries to the prompt. However, they all suffer from bias towards false positives and do not lead to significant improvement.

5. Conclusion

Inspired by the concept of "inner speech" from cognitive science, we propose a novel framework of converting egocentric videos into textual captions and leveraging an LLM for coarse-grained filtering and a pretrained NLQ model to refine the predictions. Our proposed method improves the performance of the previous state-of-the-art model (NaQ++ ReLER) and shows potential in similar video question-answering tasks.

Acknowledgement

We thank NYU High Performance Computing for the computational resources and IT service.

References

[1] Ben Alderson-Day and Charles Fernyhough. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141(5):931, 2015. 1

[2] Barbara L Bershon. Cooperative problem solving: A link to inner speech. *Interaction in cooperative groups. The theoretical anatomy of group learning*, pages 36–48, 1992. 1

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 2

[4] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022. 2

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 3

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 1

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[10] Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wing-Kwong Chan, Chong-Wah Ngo, Zheng Shou, and Nan Duan. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *arXiv preprint arXiv:2209.10918*, 2022. 1

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. *CoRR*, abs/2104.12763, 2021. 4

[12] Jane SM Lidstone, Elizabeth Meins, and Charles Fernyhough. The roles of private speech and inner speech in planning during middle childhood: Evidence from a dual task paradigm. *Journal of Experimental Child Psychology*, 107(4):438–451, 2010. 1

[13] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining@ ego4d challenge 2022. *arXiv preprint arXiv:2207.01622*, 2022. 1

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 3

[15] Jerry Liu. LlamaIndex, 11 2022. 2

[16] Naiyuan Liu, Xiaohan Wang, Xiaobo Li, Yi Yang, and Yueting Zhuang. Reler@ zju-alibaba submission to the ego4d natural language queries challenge 2022. *arXiv preprint arXiv:2207.00383*, 2022. 2

[17] OpenAI. Gpt-4 technical report, 2023. 1, 2

[18] Marcela Perrone-Bertolotti, Lucile Rapin, J-P Lachaux, Monica Baciu, and Hélène Loevenbruck. What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, 261:220–239, 2014. 1

[19] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv e-prints*, pages arXiv–2206, 2022. 1

[20] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. *arXiv preprint arXiv:2301.00746*, 2023. 1, 2, 3, 4

[21] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018. 1

[22] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[24] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. 3

[25] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*, 2022. 2

[26] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6543–6554, Online, July 2020. Association for Computational Linguistics. 2

[27] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2

[28] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models, 2023. 2