Retrieval-Augmented Egocentric Video Captioning

Supplementary Material

Contents

1. Experiment Details

1.1. Datasets & Evaluation Metrics

In this section, we detail each dataset and corresponding evaluation metric.

Epic-Kitchen-100. Epic-Kitchen [2] is an egocentric video recognition benchmark, containing 100 hours of cooking videos, with captions, for example, put hot pepper in pan, open fridge. We use 9668 video clips with an average duration of 3.7 seconds in the test set. The evaluation task is multiple instance retrieval. Given a query video, the model is required to rank the candidate captions, *i.e.*, captions with higher similarity are more semantically related to the query video. The evaluation metric is mean Average Precision (mAP) and normalised Discounted Cumulative Gain (nDCG) [4], both evaluate the video-text matching ability of the model. Please refer to [2] for the full definition of nDCG.

EgoMCQ. Egocentric Multiple Choice Question is introduced in [7] to evaluate the model's text-video retrieval ability. The task is formulated as a multiple choice question problem, where the model is required to select the matched video from five candidate videos, given a query text description. The authors in [7] considered two settings, *i.e.,* inter-video and intra-video. The difference is that whether the candidate videos are sourced from the same video. The intra-video setting is more challenging as all five candidate samples come from the same video with similar environment. We take the full EgoMCQ data, including 39k questions covering 198k narrations sourced from Ego4d [3]. Among them, 61.5% questions belong to the inter-video setting, and the intra-video setting holds the remaining 38.5% questions. The evaluation metric is the Top-1 Accuracy.

SummMCQ. Egocentric Summary Multiple Choice Question is first introduced in [1] to evaluate the model's long-form retrieval ability. SummMCQ shares the same multiple choice question formulation as EgoMCQ. The difference between them is that SummMCQ adopts a text summary as the query and the candidates are long videos. The text summaries along with the start/end timestamps are given in the Ego4d dataset [3]. HierVL [1] did not release their constructed evaluation set, thereby we curate the test data ourselves. In specific, our SummMCQ benchmark consists of 1614 samples, and the videos have an average duration of 181.4 seconds. Each sample has a text summary as query, and the candidate set contains one matched video and four randomly sampled long-form videos. Similar to EgoMCQ, Top-1 Accuracy is adopted as the evaluation metric.

YouCook2. YouCook2 is an exocentric instructional video dataset [17], containing 2000 long untrimmed videos from 89 cooking recipes. The narrations describe the main action of the actor, *e.g., combine lemon juice sumac garlic salt and oil in a bowl, place chicken on a plate or tray and season generously with mixed spices*. We evaluate our model's video-text retrieval ability on the validation set of YouCook2 based on two settings, *i.e.,* **YouCook2-Clip** and **YouCook2-Video**. YouCook2-Clip has 3350 video clips with an average duration of 19.7 seconds, that targets at evaluating model's short-term video-text retrieval ability. Given a video clip, the model aims to select the associated caption from all 3350 captions. Different from YouCook2-Clip, YouCook2-Video assesses the long-term retrieval ability of the model. Following [8], we construct this benchmark by concatenating the captions of all short-term segments in a given long video and formulate this as a video-paragraph retrieval task, resulting in 436 test samples with an average duration of 212.4 seconds. For both YouCook2-Clip and YouCook2-Video, we adopt Top-1/Top-5/Top-10 Recall (R@1/R@5/R@10) as evaluation metrics.

CharadesEgo. CharadesEgo [13] is an action recognition dataset containing egocentric and exocentric videos that cover 157 daily activity classes. Prior works [14, 16] adopted this dataset to evaluate the model's action recognition ability. In this work, we leverage the paired ego-exo videos in CharadesEgo to evaluate our model's cross-view retrieval ability. To collect such paired videos, each actor first performs the action with the video being recorded from the third-person view, and the same action is performed again with a head-mounted phone camera to collect the paired video of the first-person view. We filter the dataset by only keeping the videos with high quality (quality score = 7) and high video-text relevance (relevance score = 7), where the scores are provided in the dataset. This results in a total of 145 ego-exo video pairs with an average duration of 22.7 seconds. Top-1/Top-5/Top-10 Recall (R@1/R@5/R@10)

Method	#Params	GFLOPs	Memory	ROUGE-L	CIDEr
LaViLA-GPT2-XL [78]	2345	585	18.6	51.4	103.0
Ours Cap-MPT-1shot	2440	714	23.5	61.2	197.3

Table 1. Efficiency and performance comparison. We measure the Parameters(MB), GFLOPs and Memory(GB) on an A100 GPU.

are adopted as evaluation metrics.

EgoLearner. EgoLearner is a self-collected dataset, containing 72 hours of first-person cooking videos recorded in different kitchens via Pupil Invisible Glasses [5]. We also source 19 hours of third-person instructional videos from the Internet, depicting the procedure of cooking the same dishes. Different from prior ego-exo datasets [11, 12], where the egocentric and exocentric videos are recorded synchronously in time, the ego- and exo-videos in EgoLearner are weakly aligned, *i.e.,* cooking the same dish. The videos are annotated with fine-grained captions with start/end timestamps, highlighting the actor's action and interacting objects, *e.g. Press down on the onion just cut and chop it into small pieces.*. Each short video clip is obtained by cropping from the original video based on the timestamps.

To evaluate the model's cross-view video-video retrieval ability, we curate the EgoLearner-MCQ benchmark by the following steps: (1) We manually select 1952 pairs of ego-exo video clips performing the same action *e.g. put the plate on the table*. (2) Among 1952 ego-exo video pairs, 1089 pairs are selected for the Ego2Exo MCQ setting, where the query is an egocentric video and the five candidates are exocentric videos. The remaining 862 pairs are used for the Exo2Ego setting. (3) For each test case, apart from the correct paired video clip, We randomly sample four video clips in the same video as the negative samples, *i.e.* the challenging intra-video setting as adopted in EgoMCQ [7].

To benchmark the model's retrieval-augmented captioning ability, we simply adopt 1089 egocentric video clips in the Ego2Exo setting as the test set for captioning, along with their fine-grained caption annotations. All the exocentric videos in the test set are considered as candidate videos. We adopt BELU-4, METEOR, ROUGE-L, and CIDER as the evaluation metrics. **All the data will be released to the community.**

1.2. Inference Cost

In Table 1, we show the inference cost and captioning performance of the models. Despite comparable network parameters with LaViLA-GPT2-XL [16] (2440 MB vs 2345 MB), our retrieval-augmented captioning model has more FLOPs and memory consumption due to the additional processing (1-shot) in-context exocentric video and narration. By leveraging the retrieved samples, our model is capable of generating captions that are closely related to the content of the egocentric video.

2. Visualisation Results

2.1. Caption Refinement

In our caption refinement, we design the prompt for LLM as shown in Fig. 1. The prompt consists of (1) a **system message** at the beginning; (2) **Human Message & Rules**, containing a brief instruction of the task and a list of ten detailed requirements for the caption refinement; (3) **Prompt Example Input & Output**, demonstrating the example input/output format. We further constrain the matching between input and output sentences by specifying '*The output must contain M sentences begin with each timestamp*', where M is the number of input sentences.; (4) **The current input** ASR transcripts organised in the desired input format.

Fig. 2 illustrates the examples of refined captions vs. the original ASR transcripts on HowTo100M [9]. We list six samples in total, covering multiple scenarios, including cooking, hand-crafting, and car & repairing. As seen in Fig. 2, the ASR transcripts are transformed into shorter, but more informative captions in the descriptive style. Irrelevant words and phrases such as '*I'm gonna*', '*and then*', '*sort of*', '*so now we're going to*' are removed. In the 3rd case (Video_ID = –2EwlPH81M), the 4th refined caption further adds the phrase '*to the sauce*' when adding pickled peppers, and the 5th refined caption replaces '*this*' in the ASR transcript with the specific object, *i.e.,'green rice'*. Both cases show that our refined captions fully leverage the context information among ASR transcripts and manages to infer the unclear reference or missing details in the caption, enabling better visual-textual alignment.

We further list two failed cases as shown in Fig. 3. In the first case, the LLM outputs captions with fixed patterns such as '*Explain that*' or '*Mention that*'. This usually occurs in the videos where the actors read the instructions of the tools without introducing specific human actions. The second case (Video_ID=–C16CkNhzU) shows a temporal shift problem. Both the 1st and the 2nd ASR transcripts describe the ingredients that will be used, while LLM summarises them together, as shown in the blue characters. The 2nd refined caption, however, describes the content in the 3rd ASR transcript, causing the temporal shift problem. Despite we constrain the output with prompt '*The meaning of each sentence should exactly match the input sentence with the same timestamp.*' in the rule, the LLM still generates undesired output. A possible solution would be post-processing the refined captions by comparing sentence similarity between the original ASR transcript and the refined caption, or leveraging video-text similarity [6, 10, 15] between the video clip and the refined caption, to filter and re-generate the undesired captions. We leave this to our future work.

2.2. Ego-Exo Pair

Fig. 4 illustrates the constructed pseudo ego-exo pairs from Ego4d [3] and HowTo100M [9], followed by the ground-truth ego-exo pairs in EgoLearner.

For the constructed ego-exo pairs, most of them have similar nouns and verbs in the captions, indicating that relevant actions are performed by actors in Ego4d and HowTo100M, respectively. We observe that even the same object, *e.g., avocado*, has distinct visual appearance in first-person and third-person perspectives. Unlike prior joint ego-exo datasets [12, 13], where the egocentric and exocentric videos are recorded in the same environment, our constructed pairs cover a wide range of daily activities. The asynchronous nature makes our data more practical and scalable in real world scenarios.

Additionally, we visualise the ground-truth ego-exo videos in EgoLearner paired by human annotators. In contrast to CharadesEgo [13], the ego- and exo-videos in EgoLearner are collected in an asynchronous manner from different environment. These two datasets, together, serve as benchmarks to evaluate EgoInstructor's cross-view retrieval ability comprehensively.

2.3. Retrieval-Augmented Captioning

Fig. 5 demonstrates more retrieval-augmented captioning results on Ego4d cooking and EgoLearner. Generally, our EgoInstructor is capable of capturing objects that are hard to identify, *e.g.,* spaghetti and nylon. These objects sometimes occur in the textual description of retrieved exocentric videos, serving as explicit cues for egocentric video caption generation. The GPT4-V model can capture the main action and the general context precisely, but it fails to describe the interacting objects accurately. A failure case (the last one in Fig. 5) occurs when our captioning model struggles to comprehensively generate captions for scenarios in EgoLearner where the caption involves multiple actions and interacting objects. Compared with our few-shot setting, fine-tuning our model on EgoLearner dataset may alleviate this issue.

System Message

This is a conversation between human and an AI assistant.

Human Message & Rules

Human:

You are given the automatically recognized speech (ASR) transcripts from a long, instructional video. Please summarise each ASR into one sentence. You must follow ten rules:

1. Each sentence should contain less than 50 words.
2. Keep only the main action, verbs, nouns in the sentence in the present time.
3. Each sentence should begin with the timestamp.
4. The meaning of each sentence should exactly match the input sentence with the same timestamp.
5. Infer the related action when the input ASR does not contain enough information, such as 'hi' or 'thank you'.
6. Infer the coreference in the sentence from the context, replace 'it' or 'them' with specific objects.
7. Some simple introductions to the related object can be added, if possible.
8. Some simple explanation (why) of the action is performed, if possible.
9. The output should have exactly the same number of sentences as the input ASRs.
10. The generation ends with the special token <\END>.

Prompt Example Input

<ASR>

<3s~4s> Hi, these are what we have today.

<10s~12s> i have a little cookie sheet to catch the extra glaze and you're just going to take one of these little warm donuts and you're just gonna plop them in here and i'm gonna let you do this",

<15s~16s> just flip them over with the spoon this smells really good and then just scoop them up and plop them right back on here',

<23s~24s> so we've let these guys sit for about 15 to 20 minutes just so they can harden",

The output must contain 4 sentences begin with each timestamp.

Prompt Example Output

Assistant

<3s~4s> Introduce the ingredients to the audience.

<10s~12s> Take the donuts and plop the donuts here.

<15s~16s> Flip the donuts over with the spoon to make sure the donuts are well cooked on both sides, then scoop the donuts up and plop the donuts back.

<23s~24s> The donuts take 15 to 20 minutes to harden. <\END>

Current Input

<ASR>

<8s~10s> Today, I'm gonna share it with you all of our secrets for making cheesesteak sandwiches

…

<118s~121s> Don't forget to click the subscribe button.

The output must contain M sentences begin with each timestamp.

Assistant

Figure 1. Full prompt for context-aware caption refinement, consisting of four parts: (1) System message; (2) Human message and a list of ten rules; (3) Prompt input and output; (4) User input. Words in red are only for illustration, which are not used in practice.

Figure 2. Examples of refined captions (coloured) vs. the original ASR transcripts in HowTo100M.

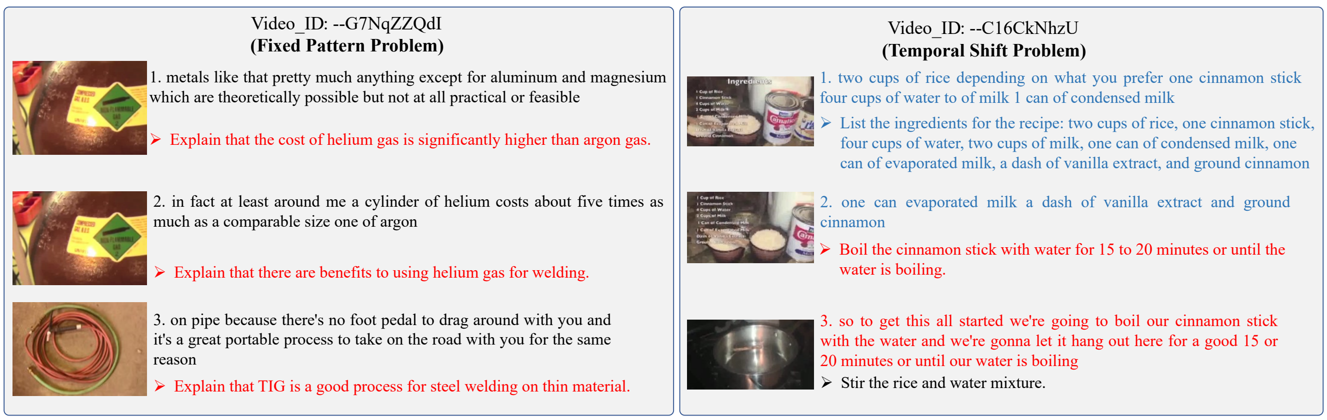


Figure 3. Examples of failed refined captions. The problems include (1) fixed caption pattern (left), *i.e.*, the caption always starts with specific pattern like '*Explain that*' and (2) temporal shift problem (right), *i.e.*, the 2nd refined caption corresponds to the 3rd ASR transcript.

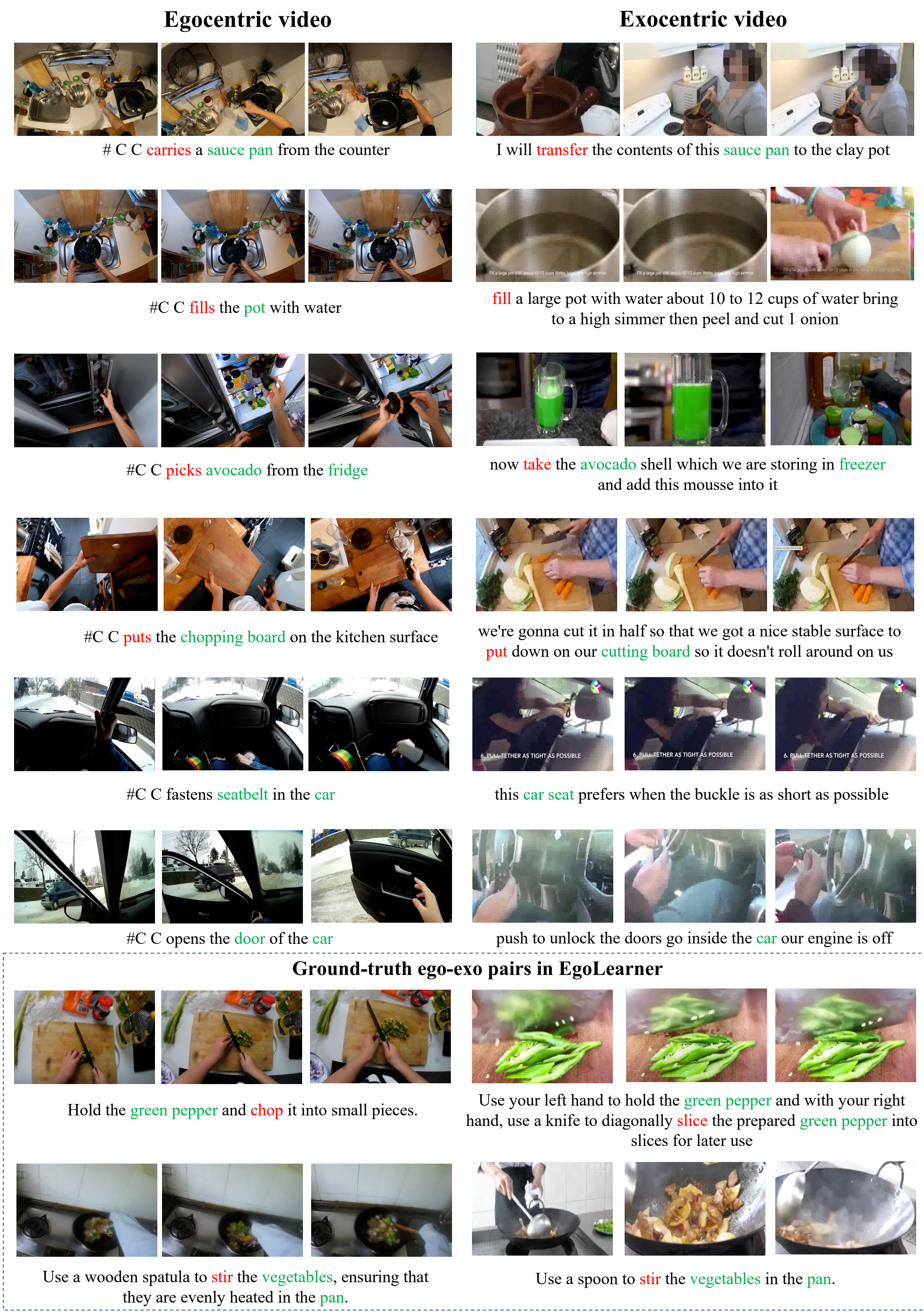


Figure 4. Examples of the pseudo ego-exo pairs constructed from Ego4d and HowTo100M. The last two rows illustrate the ground-truth ego-exo pairs in EgoLearner. We highlight the common verbs (red) and nouns (green) in ego- and exo-videos.

Ego4d Cooking

GT Caption: #C C takes a packet from the fridge.

GPT4-V: a person's hand interacting with items in an open refrigerator drawer, either placing items inside or retrieving them

Ours (w/o exo) : #C C takes out the soy sauce bottle from the refrigerator and open the lid.
Ours (w/ exo): #C C picks a container from the refrigerator.

GT Caption: #C C puts the pepper into the nylon.

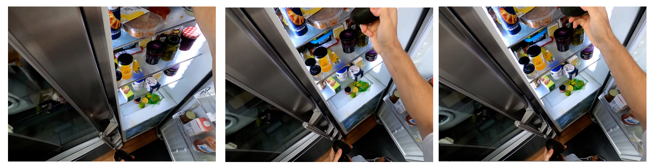
GPT4-V: a person preparing food by slicing cheese on a cutting board in a kitchen.

Ours (w/o exo) : Pick up a small piece of ginger from the bag.
Ours (w/ exo): #C C puts the pepper in the nylon.

GT Caption: #C C stirs the spaghetti in the pot with the chopsticks.

GPT4-V: a person cooking, specifically stirring food in a pan on a stovetop.

Ours (w/o exo) : #Hold the chopsticks and stir the soup in the pot.
Ours (w/ exo): #C C stirs the spaghetti in the pot with the chopstick.



GT Caption: #C C presses the avocado to see if it's ripe.

GPT4-V: a person reaching into a refrigerator to either place an item inside or take something out.

Ours (w/o exo) : #C C opens the refrigerator, while the right hand takes an eggplant.
Ours (w/ exo): #C C checks the avocados.

GT Caption: #C C picks up two potatoes from a basket tray with her left hand.

GPT4-V: a person cooking at a stovetop, specifically turning or stirring food in a pot or pan.

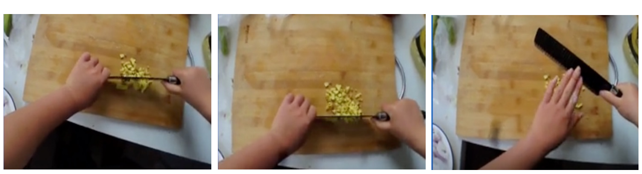
Ours (w/o exo) : #C C put the lid on the table.
Ours (w/ exo): #C C picks the potato from the tray on the cabinet with her left hand.

GT Caption: #C C takes out a yoghurt cup from the tray with her right hand.

GPT4-V: a person reaching into a refrigerator, likely to retrieve or store an item.

Ours (w/o exo) : #C C takes out an egg from the refrigerator.
Ours (w/ exo): #C C takes out a cup from the tray with her right hand.

EgoLearner



GT Caption: Use the knife to hold and cut the ginger into slices.

GPT4-V: a person chopping or mincing ingredients on a cutting board, likely as a step in preparing a meal.

Ours (w/o exo) : Chop it up into little cubes.
Ours (w/ exo): Press the ginger with the knife on the chopping board.

GT Caption: Press the tomato, hold the knife, slice the tomato into pieces, turn it around, and cut the tomato pieces into cubes.

GPT4-V: a person slicing a tomato and preparing other vegetables on a cutting board, likely for a recipe or meal.

Ours (w/o exo) : Take the ginger be careful because it's really hot.
Ours (w/ exo): Press the tomato on the chopping board with the knife with both hands.

Figure 5. Examples of retrieval-augmented captioning results of our EgoInstructor on Ego4d cooking and EgoLearner.

References

[1] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 1

[2] 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

[3] 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. 1, 3

[4] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. 1

[5] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014. 2

[6] 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. 2

[7] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 1, 2

[8] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *Proceedings of the European Conference on Computer Vision*, pages 413–430. Springer, 2022. 1

[9] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2, 3

[10] 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*, pages 8748–8763. PMLR, 2021. 2

[11] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 2

[12] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2, 3

[13] 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, 3

[14] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3307–3317, 2023. 1

[15] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. 2

[16] 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

[17] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1