Bridging the Gap: Adapting Video Language Models for Egocentric Understanding

Final Update

Group 4: Michael Cruz, Christopher Lee, Saurabh Aggarwal, Robert Martin

Problem Statement and Objectives

Problem Statement

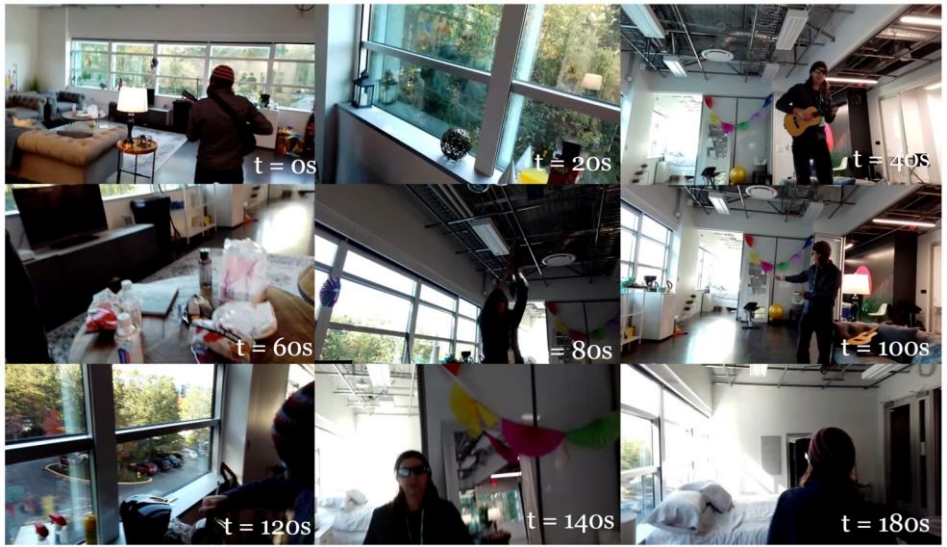
- EgoCentric data has been overlooked by researchers
- Overlooking EgoCentric may have negatively impacted existing methods
- Utilizing EgoCentric data could extend capabilities of existing methods

Objectives

- Benchmark existing LVLM’s on egocentric data (EgoSchema dataset)
- Perform fine-tuning on egocentric data (Ego4D dataset)

EgoCentric Video Data

- Video from first person perspective
- Potential Uses
  - AR/VR
  - Law Enforcement
  - Activity Recognition
  - Memory Enhancement
  - Navigation and Guidance



EgoSchema Dataset

- Over 5000 human curated multiple choice question pairs
- Subset of 500 pairs provided with answers

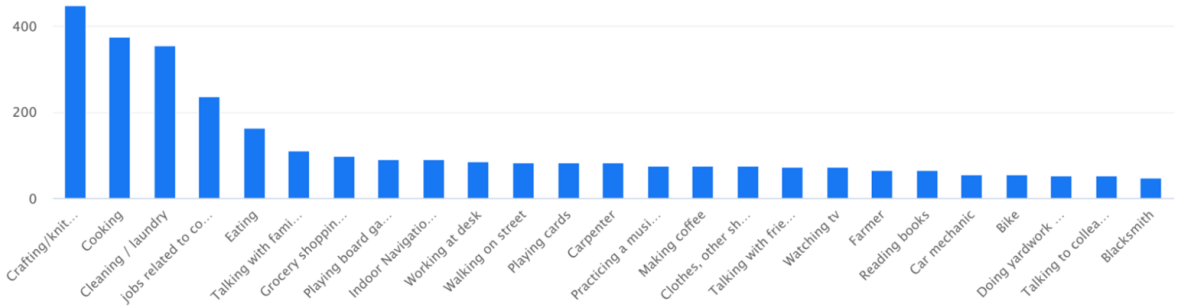
? What is the overarching behavior of C and the man in the video?

1. C teaches the man game rules but the man seems distracted and is not paying attention
2. The man teaches C how to play the card game while organizing the deck for future games
3. C and the man are playing a card game while keeping track of it in a notebook
4. C shows the man how to properly shuffle cards while the man plays them
5. The man shows C a new card game while C takes notes for future reference

Full Video Link: youtu.be/Tp4q5GeHVMY

Ego4D Dataset

- Data Types
  - Full Scale Videos
  - Clips
  - Annotations
  - Visualization Data
  - Video Components
  - Features



Experiments

Benchmarking on EgoSchema

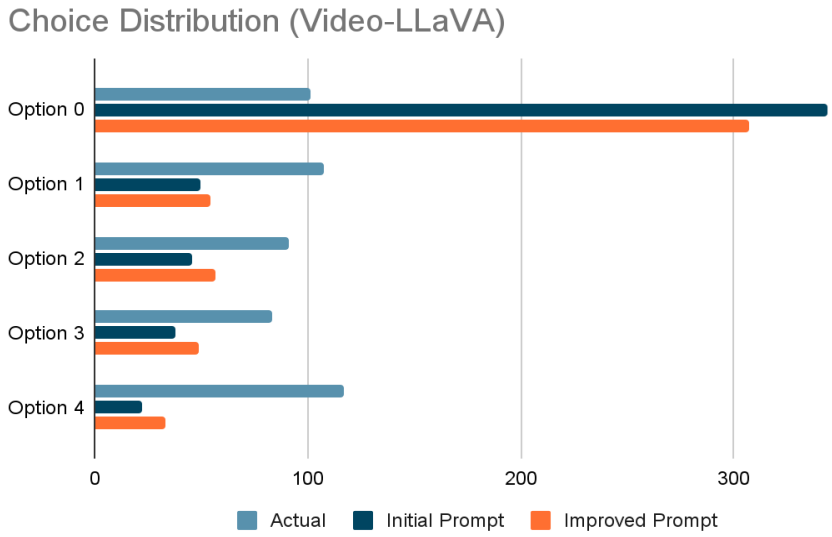
- Benchmark on the 500 multiple choice question pairs with answers
- Record the accuracy achieved by LVLMs

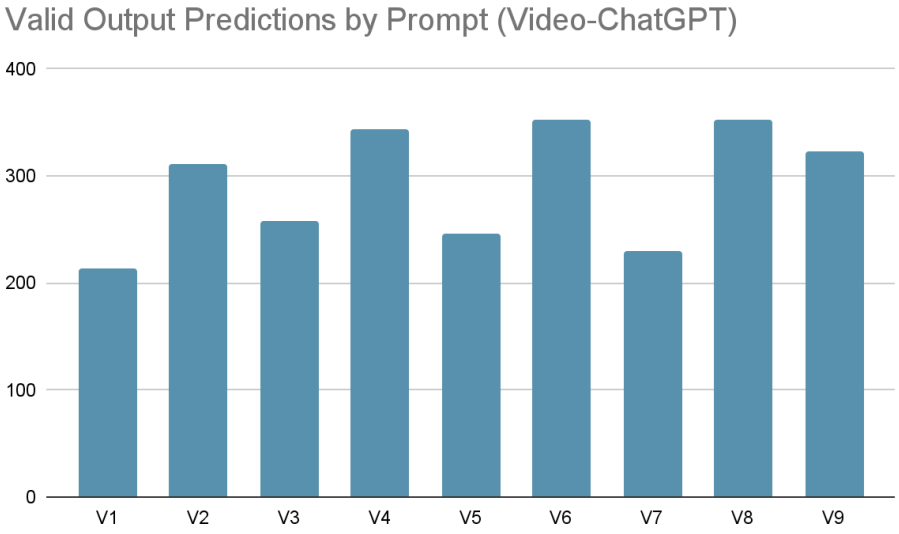
Fine-tuning on Ego4D

- Fine-tune on the full scale videos
- Create fine-tuning data using the Ego4D narration annotations

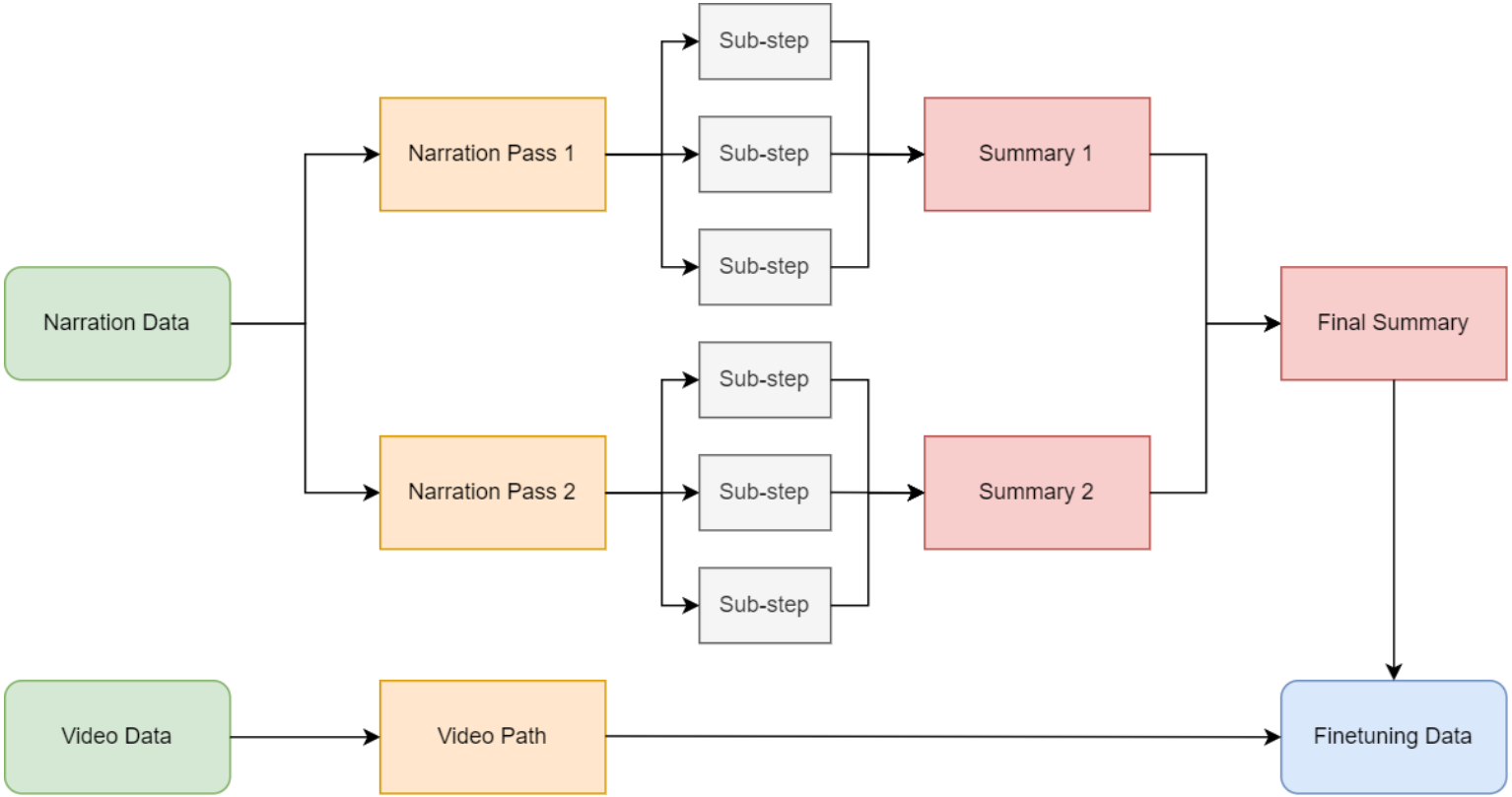
Prompt Engineering for Benchmarking

- General Formatting
- Combating Choice Bias
- Optimizing for Desired Outputs





Finetuning Data Generation



Experimental Results

- Achieved commendable results on EgoSchema
- Finetuning results were not fully completed

Method	Accuracy
VIOLET	19.9
mPLUG-Owl	31.1
InternVideo	32.1
InternVideo2-6B	41.1
Video-ChatGPT	27.6
Video-LLaVA	37.4
Gemini 1.5 (1st Frame)	54.3
Gemini 1.5 (16 frames)	**64.5**
Gemini 1.5 (150 frames)	63.6

Limitations

- Uniform selection of frames for tuning
- Reliance on narration annotations from Ego4D
- Commendable performance in zero-shot multiple choice VQA
  - Bias is choice selection
  - Difficulty generating valid outputs

Conclusion

- Extended LVLM’s for Long-Context Video Question Answering (EgoSchema)
- Developed a fine-tuning scheme using Ego4D narration annotations

- Obtained commendable benchmark results
- Made significant progress towards completing fine-tuning

Future Work

- Narration-based frame selection (nonuniform)
- Exploring fine-tuning for specific Ego4D benchmark tasks
- Creating improved narration annotations for Ego4d using a different method

References

[1] Lin, Bin, et al. "Video-llava: Learning united visual representation by alignment before projection." arXiv preprint arXiv:2311.10122 (2023).
[2] Maaz, Muhammad, et al. "Video-chatgpt: Towards detailed video understanding via large vision and language models." arXiv preprint arXiv:2306.05424 (2023).
[3] Shen, Junxiao, John Dudley, and Per Ola Kristensson. "Encode-Store-Retrieve: Enhancing Memory Augmentation through Language-Encoded Egocentric Perception." arXiv preprint arXiv:2308.05822 (2023).
[4] Pramanick, Shraman, et al. "Egovlpv2: Egocentric videolanguage pre-training with fusion in the backbone." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[5] Qinghong Lin, Kevin, et al. "Egocentric Video-Language Pretraining." arXiv e-prints (2022): arXiv-2206.
[6] Jia, Baoxiong, et al. "Egotaskqa: Understanding human tasks in egocentric videos." Advances in Neural Information Processing Systems 35 (2022): 3343-3360.
[7] Xue, Zihui, et al. "Egocentric video task translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[8] Tang, Yunlong, et al. "Video understanding with large language models: A survey." arXiv preprint arXiv:2312.17432 (2023).
[9] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.
[10] "Pre-Annotations Narrations." Annotation Guidelines Ego4D, ego4d-data.org/docs/data/annotationguidelines/#pre-annotations-narrations.
[11] "DoMSEV Dataset." Papers with Code, paperswithcode.com/dataset/domsev.
[12] "Ego4D Dataset." Papers with Code, paperswithcode.com/dataset/ego4d.
[13] "EOAD (Egocentric Outdoor Activity Dataset)." Zenodo, zenodo.org/records/7738719.
[14] Wang, Yi, et al. "InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding." https://doi.org/10.48550/arXiv.2403.15377, (2024).
[15] Reid, Machel, et al. 'Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context'. http://arxiv.org/abs/2403.05530. arXiv, (2024). [16] Mangalam, Karttikeya, Raiymbek Akshulakov, and Jitendra Malik. "Egoschema: A diagnostic benchmark for very long-form video language understanding." Advances in Neural Information Processing Systems 36 (2024).

Questions?