ALANAVLM: A Multimodal Embodied AI Foundation Model for Egocentric Video Understanding

Anonymous ACL submission

Abstract

AI personal assistants deployed via robots or wearables require embodied understanding to collaborate with humans effectively. However, current Vision-Language Models (VLMs) primarily focus on third-person view videos, neglecting the richness of egocentric perceptual experience. To address this gap, we propose three key contributions. First, we introduce the Egocentric Video Understanding Dataset (EVUD) for training VLMs on video captioning and question answering tasks specific to egocentric videos. Second, we present ALANAVLM, a 7B parameter VLM trained using parameter-efficient methods on EVUD. Finally, we evaluate ALANAVLM's capabilities on OpenEQA, a challenging benchmark for embodied video question answering. Our model achieves state-of-the-art performance, outperforming open-source models including strong Socratic models using GPT-4 as a planner by 3.6%. Additionally, we outperform Claude 3 and Gemini Pro Vision 1.0 and showcase competitive results compared to Gemini Pro 1.5 and GPT-4V, even surpassing the latter in spatial reasoning. This research paves the way for building efficient VLMs that can be deployed in robots or wearables, leveraging embodied video understanding to collaborate seamlessly with humans in everyday tasks, contributing to the next-generation of Embodied AI[1].

1 Introduction

Embodied cognition posits that our understanding of the world is fundamentally shaped by our physical bodies and their interaction with the environment (Johnson, 2015). Humans leverage this embodied understanding to intuitively grasp physical tasks, anticipate actions, and communicate effectively through nonverbal cues. For robots and AI systems to become true collaborators, they too must develop a similar understanding. Egocentric understanding of video data also has key applications in areas such as wearable computing, VR and AR, and video game technology.

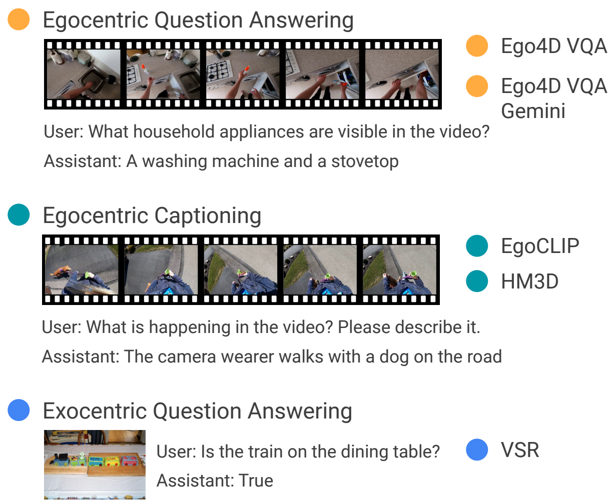


Figure 1: **Egocentric Video Understanding Dataset (EVUD)**: a collection of egocentric video caption generation and video question-answering tasks that can be used for instruction-tuning video-based VLMs.

In the scenario of an embedded artificial personal assistant, e.g., smart glasses (or a wearable camera for blind and partially sighted people), that can support the user in providing responses to visual queries, we want to build AI systems that can understand videos of the user's activities and of their visual-spatial environment. For this task, it becomes essential that the model is able to receive as input a sequence of frames before generating an answer. Recently, by leveraging pre-trained powerful Large Language Models (LLMs), Vision-Language Models (VLMs) have been proposed by using adapters that fuse representations generated by visual experts with textual tokens that can be manipulated by text-only language models (e.g., Liu et al., 2024). Following this approach, VLMs have been extended to handle video understanding tasks

[1]Code available <link removed for review>

as well (e.g., Maaz et al., 2023). However, most of these models have been developed using datasets that include a majority of third-person view videos only ignoring the importance of modelling egocentric videos (e.g., Xu et al., 2017; Caba Heilbron et al., 2015; Maaz et al., 2023). As demonstrated by Grauman et al. (2023), modelling both perspectives is challenging, and dedicated data creation efforts are required to distil this capability into VLMs.

In this paper, we provide a recipe for building VLMs that can solve tasks involving egocentric videos by extending existing video-based VLMs which are trained only on third-person view videos. Concretely, we present three main contributions: 1) we introduce the **Egocentric Video Understanding Dataset (EVUD)**, a collection of egocentric video caption generation and video question-answering tasks that can be used for instruction-tuning video-based VLMs, which underwent a rigorous human evaluation, 2) we leverage parameter-efficient training to extend existing VLMs and **train ALANAVLM** using a limited computational budget; 3) we extensively **evaluate different model variants on OpenEQA** (Majumdar et al., 2024), a challenging real-world benchmark for embodied video question-answering, and achieve state-of-the-art results compared to similarly-sized open-source models and competitive performance with much larger, proprietary variants. We also conducted rigorous human evaluation and quality control of a large portion of EVUD and elicited an error analysis on our system outputs that we hope will inform the next generation of egocentric video-based VLMs.

2 EVUD: Egocentric Video Understanding Dataset

We developed the **Egocentric Video Understanding Dataset (EVUD)** to train VLMs for egocentric video question-answering tasks. This dataset includes 29,477 examples and its components are described below (see Figure 2 for an overview).

2.1 Ego4D VQA

We consider the Ego4D collection as a high-quality source of egocentric videos that were collected in diverse settings with different types of cameras (Grauman et al., 2021). Specifically, from the Ego4D NLQ training set, we gathered 13,849 annotated clips extracted from 933 videos (see Appendix A). Then, we filtered questions having corresponding human-annotated answers which resulted in 1,137 question-answer (QA) pairs, all of which were added to EVUD.

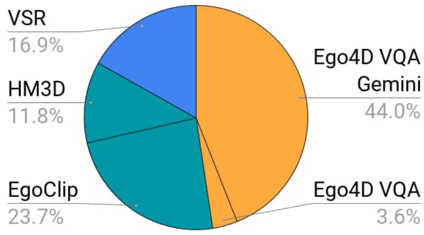


Figure 2: EVUD is built ensuring that the majority of examples focus on visual question answering (Ego4D VQA, Ego4D VQA Gemini and VSR), as well as image captioning (HM3D and EgoClip).

2.2 Ego4D VQA Gemini

Inspired by LLM-based approaches for generating training data (e.g., Li et al., 2023; Wang et al., 2022) and state-of-the-art performance of large multimodal language models, we prompted Gemini Pro 1.5 (Gemini Team, 2024) in a zero-shot multimodal fashion to produce a dataset consisting of 96K question and answer pairs requiring video understanding. These QA pairs belong to seven different categories corresponding to those specified in the OpenEQA episodic-memory question answering dataset (Majumdar et al., 2024): object recognition, attribute recognition, object state recognition, object localization, spatial reasoning, functional reasoning, and world knowledge.

We gathered all 13,849 clips from Ego4D NLQ (see Section 2.1) and used them as reference for the following data generation process. Each extracted clip was queried with Gemini Pro 1.5 using the VertexAI API [2] using zero-shot multimodal prompting with default settings (see Appendix B for the prompt definition).

In total, 13,789 of the clips successfully passed the Gemini Pro 1.5 filters, corresponding to 99.6% of the input clips. Of these clips, 100% of the outputs were successfully parsed to extract the seven *(category, question, answer)* tuples, resulting in an overall dataset of 96,523 egocentric video QA pairs (see Appendix C for a summary and Appendix D for examples of generated data). To use this data for training, the QA pairs were formatted into a series of QA turns. In EVUD, we used 12,978 clips between 2 and 60 seconds in length (corresponding to 90,846 QA dialogues).

[2]https://cloud.google.com/vertex-ai

2.2.1 Ego4D VQA Gemini Dataset Evaluation

To evaluate the quality of the generated data, we took a random set of 200 clips (corresponding to 1,400 examples) and one of the authors determined whether the questions, categories, and answers were relevant and correct, following the human evaluation schema of the Self-Instruct dataset (Wang et al., 2022; details in Appendix E).

Gemini demonstrated a strong ability to generate appropriate questions tailored to the specified categories and visual context, achieving an overall rate of 87.1% for appropriate questions and 95.6% for appropriate categories. However, it performed considerably worse (58.9%) in generating correct and acceptable answers. Additionally, the model's proficiency varied across categories, especially with regards to answer correctness (see Figure 5 in Appendix E). For object localization, spatial reasoning, and object recognition, fewer than 50% of the answers were deemed correct and acceptable.

In cases where the model-generated answer was found to be incorrect and/or unacceptable, the gold standard answer was also annotated. These gold standard answers were integrated into EVUD by replacing the model-generated answers for those questions. In this way, 575 examples were updated to human gold annotated answers and 825 model-generated were found to be satisfactory.

2.3 VSR

In order to distil fine-grained visual understanding skills into AlanaVLM, we use the Visual Spatial Reasoning (VSR) dataset (Liu et al., 2023) as a source of data for generating polar VQA pairs. In particular, for each example in the training set, we give the statement to a language model (*Llama-3 8B*, AI@Meta, 2024), and prompt it [3] to transform the statement into the corresponding question. Then, we use the truth value associated with the statement to generate an answer, randomly selecting "True" or "Yes" for positive answers, or "False" or "No" for negative answers. This results in 7,680 examples that are part of EVUD.

2.4 EgoClip Captioning

To further improve AlanaVLM's visual grounding ability, we also included a portion of the 3.8M EgoClip video-caption pairs (Lin et al., 2022). To build our captioning dataset, we sample only clips whose length is between 2 and 60 seconds resulting in 7,000 clips. We then convert the abstracted language in the original captions into natural language prompts using rules (see Appendix G). We used all 7,000 clips with associated captions in EVUD.

[3]We use Llama-3 via Ollama and report the prompt we used in Appendix F.

2.5 HM3D Captioning

The OpenEQA benchmark is composed of two different settings: ScanNet scenes which are very photorealistic (Dai et al., 2017), and HM3D scenes which contain many visual artefacts (Ramakrishnan et al., 2021). Considering that most video-based VLMs are trained on videos recorded in real-world settings, there is a mismatch with HM3D videos. Therefore, inspired by Ehsani et al. (2023), we use the Habitat simulator (Savva et al., 2019) to generate the shortest paths to specific objects relevant to the OpenEQA benchmark. Specifically, we first extract all the noun phrases from the OpenEQA benchmark using spaCy (Honnibal et al., 2020) to get our candidate set of objects $\mathcal{O}$. Then, for each *training* scene in HM3D, we spawn the agent in a random location and create the shortest paths to all the objects in the current scene which are also in $\mathcal{O}$. Given these shortest paths, we create 3,475 short videos with associated captions generated using a fixed set of prompts similar to EgoClip (examples in Appendix G) and used them all in EVUD.

3 Model Training

We build AlanaVLM by fine-tuning Chat-UniVi (Jin et al., 2023) — a vision & language foundation model equipped with video understanding capabilities — on EVUD. This fine-tuning step is essential for injecting the *egocentric video understanding* skills that are unique to AlanaVLM. We decide to build our model starting from Chat-UniVi for several reasons. First, it is an open-source model whose code and weights are publicly available. Second, it is designed for handling language, images, and videos taking an arbitrary number of frames into account. Third, it outperforms other open-source vision and language foundation models in classic video understanding tasks.

As follows, we describe the fine-tuning recipe that we used to build our model trying to preserve the original capabilities that were distilled during the instruction tuning stage. We mitigate the forgetting of previously learned skills by leveraging rehearsal (Robins, 1995), which consists in the retraining of the model on a small percentage of the previously learned information as the model is trained on new information. We fine-tune our model using Low-Rank Adaptation (LoRa; Hu et al., 2021), which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. In all our experiments, we fine-tune ALANAVLM on rehearsal data and EVUD. See Appendix H for training details.

4 Evaluation & Results

For our evaluation, we use the episodic memory use case of the OpenEQA benchmark (Majumdar et al., 2024). OpenEQA uses GPT-4 to rank the appropriateness of the generated answers concerning the ground-truth answers. To favour reproducibility, we use the highly capable open-weight model LLama-3 70B (AI@Meta, 2024).

Model	SN	HM3D	All
GPT-4 (text-only)*	32.5	35.5	33.5
GPT-4V (50f)*	57.4	51.3	55.3
Claude 3 (20f)*	n/a	n/a	36.3
Gemini 1.0 Pro V. (15f)*	n/a	n/a	44.9
Gemini 1.5 Flash (50f)	**74.0**	**69.7**	**72.5**
Gemini 1.5 Pro (50f)	66.9	61.0	64.9
Chat-UniVi (text-only)	43.4	32.4	39.7
Chat-UniVi (50f)	43.4	40.4	42.3
AlanaVLM (50f)	**47.8**	**44.8**	**46.7**

Table 1: Results on OpenEQA comparing AlanaVLM against other VLMs (with *n*f indicating the number of frames) on ScanNet (SN), HM3D, and all instances. (*): Results taken from Majumdar et al. (2024).

We compare several ablations to derive AlanaVLM's best configuration and we report additional details in Appendix I and Table 4. In this section, AlanaVLM is the best-performing model which is trained on Ego4D VQA, Ego4D VQA Gemini, VSR, and EgoClip. Table 1 shows the overall performance of AlanaVLM with respect to other VLMs on ScanNet (SN), HM3D, and all OpenEQA instances (All). AlanaVLM outperforms the base model Chat-UniVi by 4.4%. Despite having only 7B parameters and being fine-tuned with LoRa, AlanaVLM outperforms Gemini 1.0 Pro Vision and Claude 3 and its performance is comparable with all other larger VLMs except GPT-4V and the Gemini 1.5 models. However, AlanaVLM outperforms GPT-4V on spatial questions (Table 4). We do not attempt full fine-tuning to provide a more cost-effective solution; it is reasonable to expect even better results after this stage

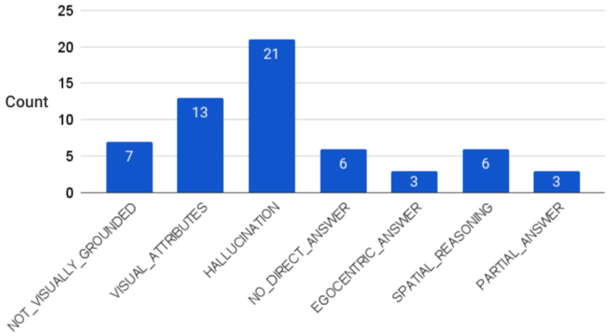


Figure 3: Human error analysis performed on 98 QA pairs on OpenEQA.

is completed (cf. Table E in Jin et al. 2023).

We notice that most models perform better on SN than on HM3D, probably because of its higher quality. Adding HM3D to the ALANAVLM training doesn't help either, maybe because its descriptions are not fine-grained enough. Finally, we note that the most recent version of Gemini surpasses all other frontier models in this task presumably due to its ability to encode higher-resolution video frames leveraging its 1M context length.

Error analysis To gain further insights into AlanaVLM predictions, we perform an error analysis based on 98 QA pairs and derive a categorization of the errors. We find that 60% of answers are incorrect. We notice that the Pearson correlation between human and LLM ratings was 0.76. Moreover, in roughly 7% of cases, the LLM said that the answers were wrong even though humans noticed that both answers applied. As shown in Figure 3, most errors concern visual attributes (e.g., object colours) or hallucinations (e.g., missing objects). Additionally, we find that ALANAVLM struggles with spatial reasoning which is required to understand the relationships between objects (error category = `SPATIAL_REASONING`). We note that in a few cases ALANAVLM generates answers that are not aligned with the camera wearer's egocentric point of view (error category = `EGOCENTRIC_ANSWER`). This highlights the need for more robust visual encoders for VLMs that can capture fine-grained details of the visual scenes when trained with egocentric vision perception (Pantazopoulos et al., 2023). Finally, we also highlight the problem of current VLMs being overpowered by the original LLM probability distribution which produces not only hallucinations but answers that are not visually grounded or which indirectly answer the question (Guan et al., 2024).

Limitations

In this paper, we present a training recipe for designing and training VLMs that can perform visual question answering in an embodied setting specifically when receiving a video stream. When designing our training recipe, we made sure that fundamental tasks such as captioning and question answering are well represented in our dataset mixture because they somehow elicit different visual grounding capabilities. To the best of our knowledge, this is the first paper that describes a training recipe for building VLMs able to *generate* responses about egocentric videos.

Despite its strengths, this paper has some limitations that we acknowledge in this section: 1) ALANAVLM is trained using LoRa therefore it is not fully leveraging the training on EVUD to the full extent as demonstrated by Jin et al. (2023); 2) to avoid potential overfitting and to facilitate fast training times, EVUD includes roughly 39K instances; this is somehow unconventional compared to current training regimes involving millions of examples. However, we don't consider this as a downside of our training recipe because most of the generated datasets in our mixture can be easily scaled up allowing one to further boost performance; and 3) as shown by our quality control evaluation, the Ego4D VQA Gemini data had an accuracy of 58.9% for the generated answers. Relying on frontier models to generate training data inherently has a disadvantage in that the generated training data is only as good as the capability of those models. We ameliorated a small batch via our human control step, but in future, advances in frontier models (e.g. the performance improvement we saw in Gemini 1.5 Flash) may result in more robust vision-language training datasets.

Finally, it is important to note that, despite its competitive performance on this benchmark, ALANAVLM still has several important limitations in terms of its visual understanding capabilities based on the careful human error analysis that we performed. Particularly, most of the errors can be considered as visual hallucinations of objects that either are not present in the scene or that are more prominent than the target object. Additionally, more research is required to understand how to design visual resamplers that are able to generate more fine-grained visual representations for the LLM which do not discard important visual attributes and spatial information—another major bottleneck for ALANAVLM as well as proprietary models such as GPT-4V, and in general of many current VLMs as demonstrated by Pantazopoulos et al. (2024).

Ethics Statement

Egocentric video understanding with VLMs presents a powerful new approach to analyzing first-person videos. However, this capability raises significant ethical considerations that must be addressed.

It is important to prioritize user privacy by ensuring informed consent is obtained for all video data collection. All our datasets are derived from academic benchmarks in which anonymization techniques are employed wherever possible to minimize the risk of identifying individuals within the videos. For instance, we have used Ego4D which has strict policies about the usage of such data.

Another important consideration is the potential for bias in VLM development, particularly if trained on imbalanced datasets. When building EVUD, we made sure to cover diverse and representative datasets during training including both image, first-person videos, and third-person videos. However, we acknowledge that this has to be proven improved when considering the deployment of this ALANAVLM in the real world. For instance, in household settings like OpenEQA, it is important to make sure that the model is trained on culturally relevant objects without favouring western-centric object distributions (Liu et al., 2021).

References

AI@Meta. 2024. Llama 3 model card.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839.

Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. 2023. Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. *arXiv preprint arXiv:2312.02976*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2023. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*.

Kristen Grauman et al. 2021. Ego4D: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.

Mark Johnson. 2015. Embodied understanding. *Frontiers in psychology*, 6.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. MIMIC-IT: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

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. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. 2024. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

Georgios Pantazopoulos, Malvina Nikandrou, Amit Parekh, Bhathiya Hemanthage, Arash Eshghi, Ioannis Konstas, Verena Rieser, Oliver Lemon, and Alessandro Suglia. 2023. Multitask multimodal prompted training for interactive embodied task completion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 768–789.

Georgios Pantazopoulos, Alessandro Suglia, Oliver Lemon, and Arash Eshghi. 2024. Lost in space: Probing fine-grained spatial understanding in vision and language resamplers. *arXiv preprint arXiv:2404.13594*.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. 2021.

Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Yutaro Yamada, Yingtian Tang, Yoyo Zhang, and Ilker Yildirim. 2022. When are lemons purple? The concept association bias of CLIP. *arXiv preprint arXiv:2212.12043*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

A Ego4D Preprocessing

We designed a preprocessing step to utilize the Ego4D NLQ videos. Specifically, 13,849 Ego4D NLQ clips were extracted by slicing the original 933 NLQ training set videos from $int(min(0, clip\_start))$ to $int(max(video\_length, clip\_end))$ for each clip (Grauman et al., 2021). The mean clip length is 12.1 seconds, with a min of 1.0 seconds and max of 481.0 seconds. See Figure 4 for a distribution of the lengths for the 13,355 clips of ≤ 60 seconds length.

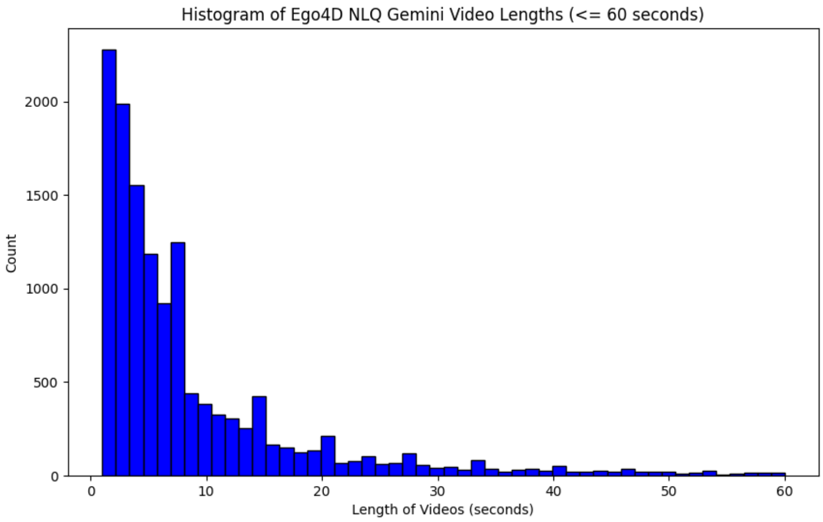


Figure 4: Length distribution of Ego4D NLQ clips.

B Gemini Pro 1.5 Prompt

To generate the Ego4D VQA Gemini dataset we adapted the prompt from MIMIC-IT. We updated it to the video modality and to obtain a question and answer pair for each of the seven OpenEQA task categories (Li et al., 2023; Majumdar et al., 2024):

Ego4D VQA Gemini prompt

```
[INPUT VIDEO]
```

You are an intelligent embodied agent that can answer questions. You will be shown a video that was collected from a single location.

Your task is to generate a question for each of the following categories: object recognition, attribute recognition, object state recognition, object localisation, spatial reasoning, functional reasoning, world knowledge.

Ask diverse questions and give corresponding short answers. Include questions asking about the visual content of the video. The questions you posed can include the actions and behaviors of people or objects in the video, the chronological order of events, and causal relationships. Only include questions that have definite answers. Do not ask any questions that cannot be answered confidently.

Don't use headers. You should use the following format for each category:

Category: <category>
Question: <question>
Short answer: <answer>

Assistant:

C Ego4D VQA Gemini Dataset Diversity

Summary of question and QA pair counts generated by Gemini are shown in Table 2 for each category. Overall, a diverse set of questions were generated with 42,357 unique questions. When considering the union of questions and answers, 76,438 unique QA pairs were generated.

Category	Unique count	
	Questions	QA pairs
Object recognition	5,565	12,497
Object state recognition	4,729	7,627
Object localization	5,304	12,411
Attribute recognition	4,417	8,796
Spatial reasoning	8,968	12,364
Functional reasoning	5,656	10,831
World knowledge	7,718	11,912
Total	**42,357**	**76,438**

Table 2: Counts of unique questions and QA pairs in the Ego4D VQA Gemini dataset

D Ego4D VQA Gemini Dataset Examples

D.1 Domestic kitchen scene

Example of high quality generations in all categories

Category: Object recognition
Question: What household appliances are visible in the video?
Short answer: A washing machine and a stovetop

Category: Attribute recognition
Question: What color is the countertop?
Short answer: Gray

Category: Object state recognition
Question: Is the cabinet door open or closed at the start of the video?
Short answer: Closed

Category: Object localisation
Question: Where is the orange bottle placed?
Short answer: Inside the cabinet, to the left of the cleaning supplies

Category: Spatial reasoning
Question: Is the sink located to the left or right of the stovetop?
Short answer: Right

Category: Functional reasoning
Question: What is the purpose of the cabinet under the sink?
Short answer: To store cleaning supplies and other household items

Category: World knowledge
Question: What type of cleaner is typically used to clean a stovetop?
Short answer: Degreaser or multi-surface cleaner

D.2 Nighttime street scene

Example of high quality generations in all but one category, with an incorrect answer for spatial reasoning

Category: object recognition
Question: What kind of animal is in the video?
Short answer: A cat

Category: attribute recognition
Question: What color is the cat?
Short answer: Black and white

Category: object state recognition
Question: Is the phone on or off?
Short answer: On

Category: object localisation
Question: Where is the life preserver?
Short answer: Attached to the railing

Category: spatial reasoning
Question: Is the cat closer to the person holding the phone or the person walking?
Short answer: The person holding the phone

Category: functional reasoning
Question: What is the phone being used for?
Short answer: It is not clear what the phone is being used for

Category: world knowledge
Question: What time of day is it?
Short answer: Night

D.3 Shop counter scene

Example of high quality generations in all but one category, with changing object state resulting in a partially correct answer for object state recognition

Category: object recognition
Question: What items are on display behind the cashier?
Short answer: cigarettes, cigars, chewing tobacco

Category: attribute recognition
Question: What color is the cash register?
Short answer: black

Category: object state recognition
Question: Is the cashier's mouth open or closed?
Short answer: open

Category: object localisation
Question: Where is the display of lighters located?
Short answer: to the left of the cashier

Category: spatial reasoning
Question: What is closer to the camera, the lighters or the cigarettes?
Short answer: the lighters

Category: functional reasoning
Question: What is the purpose of the cash register?
Short answer: to process transactions

Category: world knowledge
Question: What is the legal age to purchase tobacco products in the United States?
Short answer: 21 years old

D.4 Outdoors dog scene

Example of hallucinated objects in questions, incorrect or over-confident answers for most categories

Category: Object recognition
Question: What breed of dog is in the video?
Short answer: There is a Bernese Mountain dog, a white Pitbull, and a black and white Border Collie.

Category: Attribute recognition
Question: What color is the dog's collar?
Short answer: Red

Category: Object state recognition
Question: Is the sun shining?
Short answer: Yes

Category: Object localisation
Question: Where is the person standing?
Short answer: At the bottom of a hill

Category: Spatial reasoning
Question: Are the dogs uphill or downhill from the person?
Short answer: Uphill

Category: Functional reasoning
Question: What is the leash used for?
Short answer: To keep the dog from running away

Category: World knowledge
Question: What season is it?
Short answer: It is most likely early spring

D.5 Garden landscaping scene

Example of non-visually grounded answers for object recognition, object state recognition and functional reasoning categories

Category: object recognition
Question: What type of equipment is being used to dig the hole?
Short answer: A mini excavator.

Category: attribute recognition
Question: What color is the handle of the shovel?
Short answer: Brown.

Category: object state recognition
Question: Is the hole completely dug?
Short answer: No, it is still being dug.

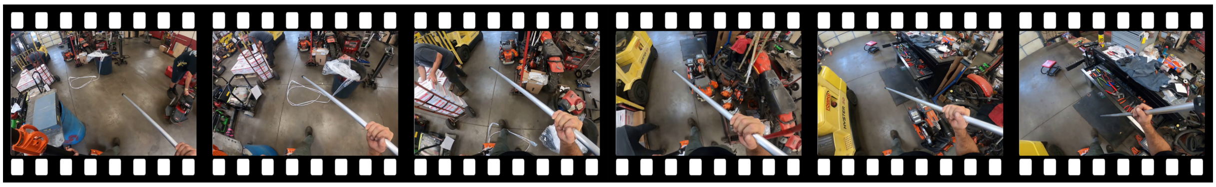
Category: object localisation
Question: Where is the red ribbon located?
Short answer: Next to the hole that is being dug.

Category: spatial reasoning
Question: Is the shovel closer to the mini excavator or the hole?
Short answer: The hole.

Category: functional reasoning
Question: What is the purpose of the red ribbon?
Short answer: It is likely marking the location of underground utilities.

Category: world knowledge
Question: What type of professional is likely operating the mini excavator?
Short answer: A landscaper or construction worker.

D.6 Workshop scene



Example of hallucinated objects in questions and answers for object recognition, object state recognition, object localisation and spatial reasoning categories

Category: object recognition
Question: What types of equipment are visible in the workshop?
Short answer: The workshop contains various equipment, including chainsaws, leaf blowers, a forklift, hand tools, and a workbench.

Category: attribute recognition
Question: What color is the forklift in the video?
Short answer: Yellow

Category: object state recognition
Question: Are the chainsaws in the video new or used?
Short answer: It is difficult to tell from the video alone whether the chainsaws are new or used.

Category: object localisation
Question: Where is the workbench located?
Short answer: The workbench is located on the right side of the video, against the wall.

Category: spatial reasoning
Question: How many chainsaws are to the left of the red toolbox?
Short answer: There are five chainsaws to the left of the red toolbox.

Category: functional reasoning
Question: What is the purpose of the extension pole shown in the video?
Short answer: The extension pole is likely used to reach high places, possibly for tasks like pruning trees or cleaning gutters.

Category: world knowledge
Question: What types of businesses typically use forklifts for their operations?
Short answer: Forklifts are commonly used in warehouses, distribution centers, construction sites, and other businesses that handle heavy materials.

E Human Evaluation of Gemini Pro 1.5 Generated Training Data

Following the human evaluation schema described for the Self-Instruct dataset (Wang et al., 2022), an author of the present work evaluated the 1,400 examples as follows:

1. Is the question appropriate for the clip?
2. Is the question appropriate for the category?
3. Is the answer correct and acceptable for the clip and question?

E.1 Results

Results of the human evaluation are shown in Figure 5. The rates of appropriate generated questions ranged from 79.0% for the functional reasoning category to 95.0% for the object recognition category. For assigning the questions to appropriate categories, Gemini performed favourably, with a range of 86.0% for object recognition to 99.5% for functional reasoning and object localization. Gemini performed markedly worse with generating correct and acceptable answers for the clips, ranging from 36.5% for object localization to 83.5% for world knowledge. The superior performance in the world knowledge category could be due to the advantages of relying on the language model's encoded knowledge, without the need to refer to the visual context of the scene. In addition, VLMs have an observed concept association bias and weakness in compositional understanding, with tasks such as spatial reasoning being especially prone to errors (Yamada et al., 2022; Yuksekgonul et al., 2022).

Issues with the generated questions and answers often included hallucinated objects, non-visually grounded answers and changing camera angles resulting in partially correct answers. See Appendix D for examples of generated *(category, question, answer)* tuples.

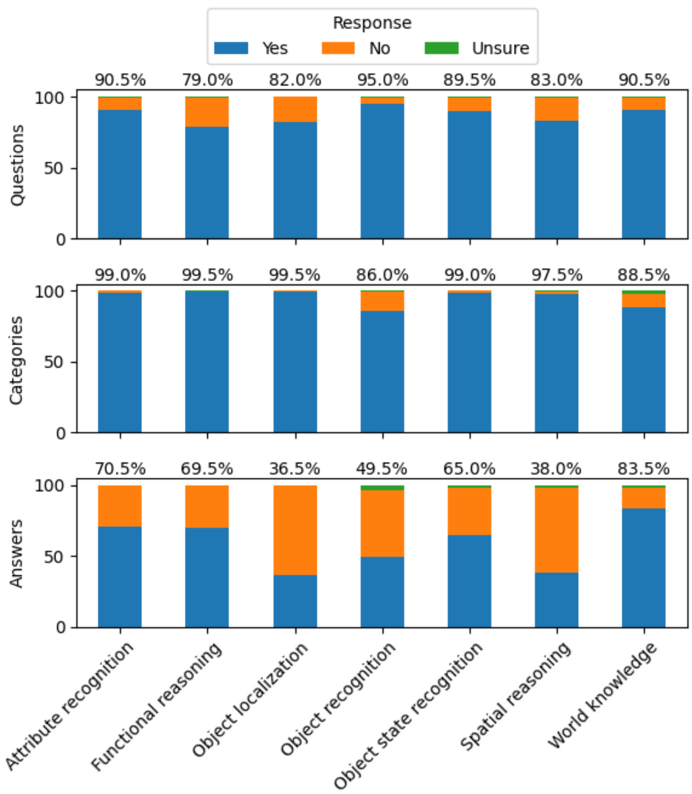


Figure 5: Results of human evaluation on 1,400 examples. The percentage of appropriate question, appropriate category, and correct answer are shown on a per category basis. Text labels show the percentage of questions/categories/answers in each category found to be appropriate and/or correct.

E.2 Inter-Annotator Agreement

In addition, 10% of the 200 clips were randomly chosen for evaluation by another human expert and used to estimate inter-annotator agreement by calculating Cohen's Kappa. Agreement between annotators was found to be fair for question and category appropriateness, with scores of 0.210 and 0.232 respectively. For answer correctness and acceptability, agreement was found to be moderate, with a score of 0.403. Although the scores show an agreement between experts, they also indicate the difficulty of evaluating generated questions and answers, with clips often changing state, for example, via object movement and multiple camera angles resulting in changing spatial relationships.

F VSR Prompt

VSR prompt

Generate a polar question from the following statement about a picture. Keep as many words as you can of the statement in the question and do not add unnecessary words. Always generate just the question. Do not include any explanations.
Statement: `<statement>`

G EgoClip Preprocessing

Given the original EgoClip dataset, we preprocess it using specific rules to convert it into a more natural caption. Specifically, we first sample a prompt from a list of predefined prompts (see below) and then apply conversion rules to the original caption. Specifically, following the Ego4D guidelines [4], we replace "#C" with "the camera wearer", "#O" with "another person", and "#UNSURE" with "something". Finally, we delete the prefix "Summary" when included.

EgoClip prompt instructions

```
[
"Can you please provide a brief description of the video?",
"Describe the content of the video.",
"What is happening in the video? Please describe it.",
"Can you summarize the key events or actions in the video?",
"Describe the visual elements and any notable features in the video.",
"Provide a narrative description of the video.",
"What's in the video?",
"What can you see in this video?",
"What's happening in the video?",
"What is the main focus of the video?"
]
```

H Training Details

We build our rehearsal dataset composed of previously learned examples starting from the Chat-UniVi instruction tuning dataset, which includes instances from LLaVa (Liu et al., 2024; composed of NLP and COCO examples), MIMIC-IT (Li et al., 2023), and Video-ChatGPT (Maaz et al., 2023). Since we wanted ALANAVLM to forget language skills as little as possible, and to have good video understanding capabilities, we adapted the distribution of previously learned examples, giving slightly less emphasis to the text and image instances, and much more emphasis to the Video-ChatGPT instances. In particular, we bring the percentage of LLaVa NLP instances, LLaVa COCO instances, MIMIC-IT instances, and Video-ChatGPT from 5%, 82%, 13%, and 25% to 10%, 20%, 50%, and 20%. Given the instances

[4] https://ego4d-data.org/docs/data/annotation-guidelines/#narrations

Subset	Sampled instances	Sampling percentage
NLP	1000	10
COCO	2000	20
VideoChat	5000	50
MIMIC	2000	20
Total	**10000**	**100**

Table 3: Proportion of data used for vision+language rehearsal during our fine-tuning stage. Data are derived from several data sources used for Chat-UniVi instruction-tuning (Jin et al., 2023).

resulting from the changed distribution, we sample 1% from each subset in order to build the rehearsal data leveraged in our experiments which is composed of 10,000 instances (see Table 3).

Following best practices in using LoRa[5], we employ the Adam optimizer with a learning rate equal to 3e-4 to fine-tune for one epoch and we set the rank R equal to 64 and value of α equal to 128.

H.1 Computational Experiments

ALANAVLM is a 7B parameter model trained using A10 NVIDIA GPUs available in AWS. Each training run lasted approximately 8 hours on a single GPU thanks to LoRA. Running all the configurations of ALANAVLM required an overall computational budget of 80 GPU/hours.

I Extended Evaluation & Results

I.1 Response Generation

I.1.1 ChatUniVi Variants

To generate the ChatUniVi responses, we use the default parameters, i.e., we set the temperature of the model to 0.2 and use beam search with a single beam. For processing the input videos, we consider two approaches. For the first approach, we process the videos by sampling frames at a rate of 1 frame per second. Then, if there are more than a maximum of 100 frames, we resample 100 frames uniformly from the sampled set. For the second approach, we sample 50 frames uniformly at the original frame rate of the video, if there are more than 50 frames. Otherwise, we use all the available frames. Frames are also resized to 224×224 pixels, as per the original model resolution (Jin et al., 2023).

I.1.2 Gemini 1.5 Variants

We also evaluate two variants of Gemini 1.5, i.e., the Pro and Flash variants. For this, we used a similar protocol as the one used for the evaluation of Gemini 1.0 Pro Vision on the OpenEQA benchmark (Majumdar et al., 2024). We accessed these models through the Vertex AI API[6]. The prompt was constructed by concatenating the prompt and the frames as follows:

Gemini prompt

You are an intelligent question answering agent. I will ask you questions about an indoor space and you must provide an answer.
You will be shown a set of images that have been collected from a single location.
Given a user query, you must output ‘text‘ to answer to the question asked by the user.
<FRAME> ... <FRAME>
User Query: question

By submitting the frames instead of the video, we could control the number of frames that were sent to the model. In the results section, we show two sets of results. One set of results used the full frame sizes and the other set corresponds to sending frames resized to 224×224 pixels.

[5] https://lightning.ai/pages/community/lora-insights/
[6] https://cloud.google.com/vertex-ai

I.2 Evaluation Protocol

To evaluate the models in this paper, we follow an evaluation protocol inspired by the OpenEQA benchmark (Majumdar et al., 2024). Concretely, we submit the prompt below to a Llama3 70B model (AI@Meta, 2024) through the together.ai API[7]. There is another variant for examples that include "extra answers", which follows a similar format, but the model is also prompted to check the extra answers to make an assessment as to whether the generated response answers the given question. As illustrated in the prompt, the Llama3 model is prompted to give a score between 1 and 5 depending on how well the generated response matches any of the ground-truth answers. Once we have obtained the scores for all the samples in the dataset, we normalise them and compute their mean and bootstrapped standard error.

Llama3 prompt

You are an AI assistant who will help me to evaluate the response given the question and the correct answer.
To mark a response, you should output a single integer between 1 and 5 (including 1, 5).
5 means that the response perfectly matches the answer.
1 means that the response is completely different from the answer.

Example 1:
Question: Is it overcast?
Answer: no
Response: yes
Your mark: 1

Example 2:
Question: Who is standing at the table?
Answer: woman
Response: Jessica
Your mark: 3

Example 3:
Question: Are there drapes to the right of the bed?
Answer: yes
Response: yes
Your mark: 5

Your Turn:
Question: question
Answer: answer
Response: prediction

I.3 Results

Table 4 shows the results per category, per subset, and for all instances of blind models, VLMs, and AlanaVLM's ablations. When it comes to ablations, we evaluated different mixtures of the EVUD to verify the impact of different data sources on the overall performance in the OpenEQA benchmark. Additionally, we also experimented with different numbers of video frames. Following the OpenEQA evaluation protocol, we use bootstrapping to estimate standard deviations associated with the different model configurations.

[7] https://www.together.ai/

J Error Analysis of AlanaVLM's performance

The subset of examples used for the human evaluation of AlanaVLM's performance has been obtained through stratified sampling based on question categories for each dataset. Since we have seven question types per category and two subsets (ScanNet and HM3D), we obtained 98 examples. The mistakes made by AlanaVLM were pointed out by two authors of the present work who provided ratings and categorised the errors according to special categories that were created in a bottom-up fashion. To compute the percentage of correct answers according to humans, we counted the number of times where human ratings were ≥ 4 and LLM ratings were ≤ 2.

K Data and Model Release Details

We will release both the EVUD dataset as well as the trained checkpoints that were produced in the context of this paper alongside their predictions for the OpenEQA benchmark. We plan to release the model checkpoints and code under MIT license. On the other hand, we will release the EVUD under CC BY 4.0. All these artefacts will be released on Huggingface Hub upon acceptance.

Model	Object Recognition	Object State Recognition	Object Localisation	Attribute Recognition	Spatial Understanding	Functional Reasoning	World Knowledge	SN	HM3D	All
Blind LLMs										
GPT-4*	15.4	51	20.3	31.5	31.4	52.2	34.2	32.5 ± 1.2	35.5 ± 1.7	33.5 ± 1.0
Chat-UniVi (text-only)	33.1 ± 2.6	55.5 ± 3.0	24.0 ± 2.4	29.2 ± 2.8	38.2 ± 2.9	51.4 ± 2.7	48.9 ± 2.9	43.4 ± 1.3	32.4 ± 1.7	39.7 ± 1.1
Proprietary Multi-Frame VLMs										
GPT-4V (50f)*	51.4	57.7	53.3	65.2	42.6	63.8	52.3	57.4 ± 1.3	51.3 ± 1.8	55.3 ± 1.1
Claude 3 (20f)*	37.0	45.5	13.1	39.2	37.0	37.9	47.3	n/a	n/a	36.3 ± 1.1
Gemini 1.0 Pro V. (15f)*	41.5	56.9	33.3	41.9	37.6	52.2	52.1	n/a	n/a	44.9 ± 1.1
Gemini 1.5 Flash (50f)	**73.6** ± 2.6	**76.0** ± 2.6	61.4 ± 2.3	**81.8** ± 2.2	56.7 ± 3.0	**78.3** ± 2.2	**81.1** ± 2.2	**74.0** ± 1.1	**69.7** ± 1.7	**72.5** ± 0.9
Gemini 1.5 Flash (50f - 224 x 224)	71.0 ± 2.7	75.5 ± 2.6	**62.8** ± 2.4	80.8 ± 2.2	55.9 ± 3.0	76.8 ± 2.2	74.1 ± 2.6	71.9 ± 1.2	69.1 ± 1.7	71.0 ± 1.0
Gemini 1.5 Pro (50f)	73.1 ± 2.6	60.9 ± 2.7	56.3 ± 2.5	74.4 ± 2.4	**59.4** ± 3.0	63.5 ± 2.7	67.6 ± 2.7	66.9 ± 1.2	61.0 ± 1.8	64.9 ± 1.0
Gemini 1.5 Pro (50f 224 x 224)	69.0 ± 2.7	61.4 ± 2.7	53.0 ± 2.6	69.7 ± 2.5	55.6 ± 3.0	61.5 ± 2.7	63.5 ± 2.8	64.3 ± 1.3	57.1 ± 1.8	61.9 ± 1.0
Open-Source Multi-Frame VLMs										
Chat-UniVi	28.9 ± 2.6	57.1 ± 3.0	23.8 ± 2.4	35.6 ± 2.9	37.4 ± 2.9	59.1 ± 2.7	52.7 ± 2.9	42.6 ± 1.3	39.8 ± 1.9	41.7 ± 1.1
Chat-UniVi (50f)	33.8 ± 2.7	45.1 ± 2.8	27.9 ± 2.5	33.4 ± 2.9	44.5 ± 3.0	**63.8** ± 2.6	52.0 ± 3.0	43.4 ± 1.3	40.4 ± 1.8	42.3 ± 1.1
Chat-UniVi (Rehearsal)	32.3 ± 2.6	55.5 ± 3.0	26.8 ± 2.4	38.0 ± 3.0	43.3 ± 3.0	57.5 ± 2.7	**58.3** ± 2.9	45.7 ± 1.4	40.8 ± 1.9	44.0 ± 1.1
Chat-UniVi (Rehearsal) (50f)	36.1 ± 2.7	44.5 ± 2.7	27.9 ± 2.4	35.6 ± 2.9	44.4 ± 3.0	57.1 ± 2.8	52.2 ± 2.9	42.9 ± 1.3	40.3 ± 1.8	42.0 ± 1.1
AlanaVLM (VQA-EgoClip)	30.1 ± 2.5	56.2 ± 3.1	29.0 ± 2.4	41.7 ± 3.0	**45.7** ± 3.0	61.1 ± 2.5	51.9 ± 3.0	47.0 ± 1.4	40.2 ± 1.9	44.7 ± 1.1
AlanaVLM (VQA-EgoClip) (50f)	**39.8** ± 2.7	54.9 ± 3.1	32.0 ± 2.4	42.7 ± 3.0	45.5 ± 3.0	59.6 ± 2.5	52.5 ± 2.9	47.4 ± 1.3	44.3 ± 1.9	46.3 ± 1.1
AlanaVLM (VQA-EgoClip-HM3D)	33.2 ± 2.7	56.3 ± 3.1	31.2 ± 2.5	40.2 ± 3.0	41.7 ± 3.0	61.2 ± 2.5	53.6 ± 3.0	46.8 ± 1.3	41.5 ± 1.9	45.0 ± 1.1
AlanaVLM (VQA-EgoClip-HM3D) (50f)	36.4 ± 2.7	53.9 ± 3.1	30.5 ± 2.4	44.5 ± 3.1	38.0 ± 2.9	56.8 ± 2.6	56.1 ± 3.0	45.9 ± 1.3	42.7 ± 1.8	44.8 ± 1.1
AlanaVLM (VQA-VSR-EgoClip) (50f)	32.0 ± 2.6	50.5 ± 3.1	29.3 ± 2.5	41.8 ± 3.0	42.7 ± 3.0	61.1 ± 2.6	50.7 ± 3.0	45.4 ± 1.4	40.0 ± 1.9	43.6 ± 1.1
AlanaVLM (VQA-VSR-EgoClip) (50f)	37.1 ± 2.6	57.5 ± 3.1	31.0 ± 2.5	**46.2** ± 3.1	43.4 ± 3.0	61.9 ± 2.5	52.5 ± 3.0	**47.8** ± 1.4	**44.8** ± 1.9	**46.7** ± 1.1
AlanaVLM (VQA-VSR-EgoClip-HM3D)	32.7 ± 2.7	**59.4** ± 3.0	**36.6** ± 2.6	39.2 ± 3.0	37.2 ± 2.9	61.9 ± 2.6	54.1 ± 3.0	47.2 ± 1.4	42.6 ± 1.9	45.6 ± 1.1
AlanaVLM (VQA-VSR-EgoClip-HM3D) (50f)	37.0 ± 2.6	55.4 ± 3.1	30.7 ± 2.5	43.9 ± 3.1	40.5 ± 2.9	58.6 ± 2.5	50.2 ± 2.9	46.7 ± 1.3	41.4 ± 1.8	44.9 ± 1.1

Table 4: Results per category, per subset, and for all instances of blind models, VLMs, and AlanaVLM ablations (with *n*f indicating the number of frames). Standard deviations were estimated using bootstrapping as per the OpenEQA evaluation protocol (Majumdar et al., 2024). In this table, we refer to the union of Ego4D VQA NLQ human annotated QA pairs (Section 2.1) and Ego4D VQA Gemini (Section 2.2) as VQA. (*): Results taken from Majumdar et al. (2024).