ERQA+: An Enhanced Benchmark on Embodied Reasoning

BAAI *FlagEval* Team

TL;DR: We introduce `ERQA+` **, a new embodied reasoning benchmark** that complements ERQA (from Gemini Robotics Team) in multiple aspects:

- **Egocentric scenes:** `ERQA+` is specifically adapted to embodied perspectives by focusing on egocentric scenes.
- **Reduced contamination:** `ERQA+` was completely newly created by manual annotation over recent robotic videos, rather than reusing samples from earlier VQA datasets or benchmarks.
- **Extended taxonomy:** `ERQA+` has been collected with a more fine-grained taxonomy to cover additional embodied reasoning skills.
- **Question types:** `ERQA+` comprises multiple question types, and also increases the number of options for those formed as multiple-choice questions.
- **Difficulty:** the creation of `ERQA+` includes multiple stages of data filtering to exclude trivial or easy samples.

Abstract

Embodied reasoning capabilities from modern vision-language models are crucial for embodied intelligence. We introduce `ERQA+` , a new benchmark for embodied reasoning that addressed a few limitations from ERQA, an earlier benchmark that directly motivates this work. We attach links to the benchmark, evaluation code, and other potential updates on this website: https://flageval-baai.github.io/ERQA-Plus-page/

1 Introduction

1.1 Embodied intelligence in the LLM era

The rapid development of large language models (LLMs) and vision-language models (VLMs) (OpenAI, 2025b; Gemini Team, 2025) has brought a paradigm shift in embodied intelligence, enabling robotic agents to perceive, reason, and act in the physical world with better generalization. Modern VLMs have demonstrated remarkable general-purpose capabilities in visual recognition and semantic understanding due to pre-training on large-scale image and text data. Recent advances in embodied AI either directly include explicit VLM components to borrow their strong generalization power for perception, reasoning, and planning (Gemini Robotics Team, 2025b;a; Cheang et al., 2025), or derive Vision-Language-Action (VLA) models from VLMs to translate high-level language instructions directly into precise, context-aware physical actions (Kim et al., 2024; Black et al., 2024; Physical Intelligence, 2025; Shukor et al., 2025).

That said, key properties relevant to embodied intelligence remain challenging, including but not limited to spatial reasoning, task planning, navigation, etc. An earlier benchmark RoboVQA (Sermanet et al., 2024) focuses on interactive, long-horizon scenarios, but the unbounded output space makes it very difficult to faithfully evaluate. To measure progress specifically in *embodied reasoning*, Gemini Robotics Team (2025b) introduce a benchmark named *Embodied Reasoning Question Answering*, known as ERQA in short. ERQA specifically focuses on capabilities likely required by an embodied agent that perceives and interacts with the physical world. Figure 1 (top) shows three examples from ERQA. They may require a VLM to recognize or track objects across different video frames, or to infer the spatial relationship with other objects in the scene.

1.2 Why `ERQA+`

ERQA makes a useful start on benchmarking embodied reasoning. However, it has clear limitations given that it is not serving as an ordinary general-purpose visual question-answering (VQA) benchmark. Specifically:

1. ERQA is formed by 400 problems in a simple multiple-choice question answering (MCQA) format, with each problem having no more than four options only. Many questions only involve as few as two options, typically Yes-No questions, making random guess easier.

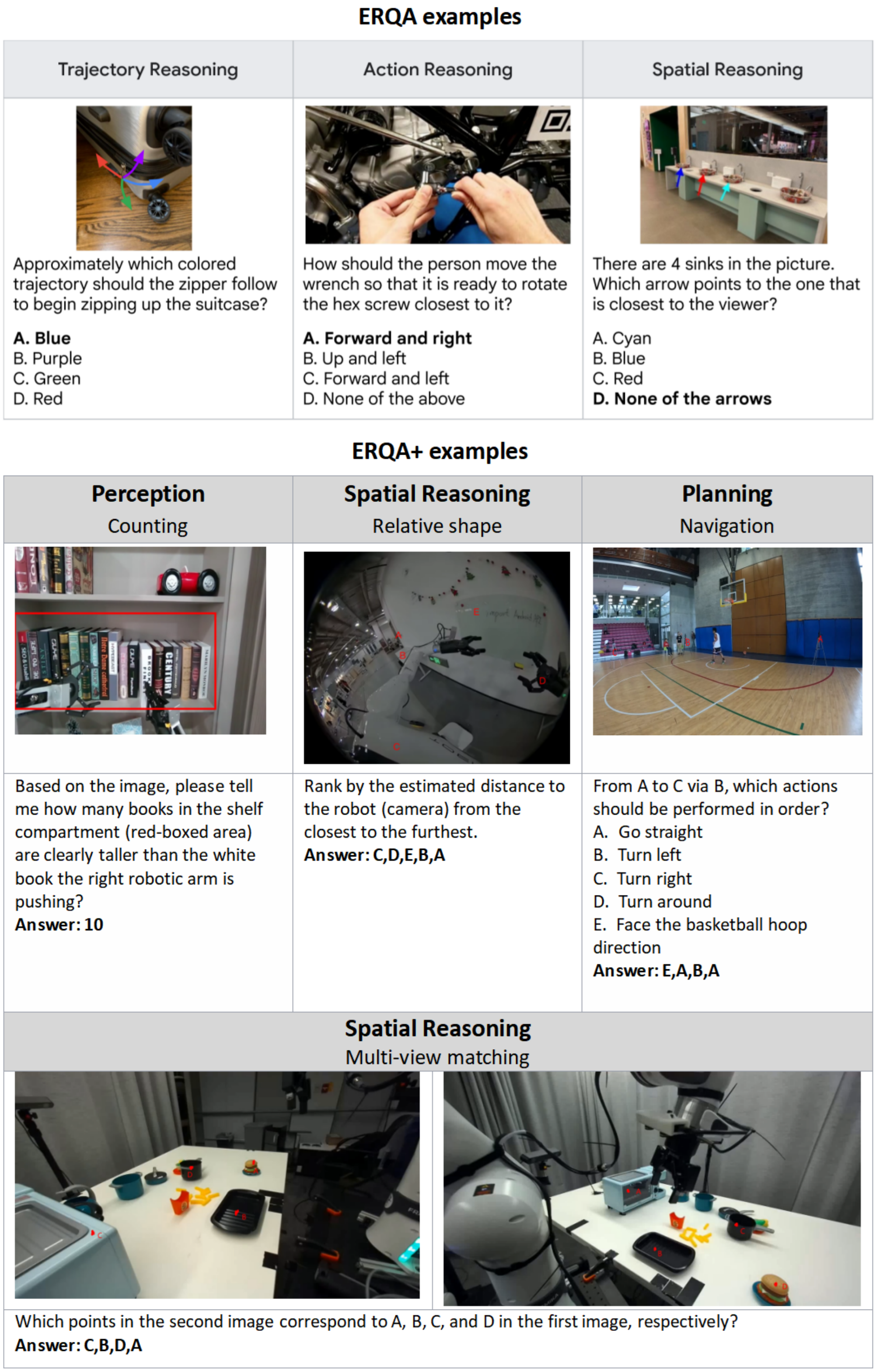

Figure 1: Examples from the ERQA (above; all are multiple-choice questions with four or less options) and ERQA+ (below; with more diverse and more difficult problem types).

2. ERQA is mostly derived from image datasets covering general scenes (e.g., street view), with only a few of them taken from an embodied or ego-centric viewpoint. Using samples or exact images from much earlier datasets also leaves potential risk of evaluation data contamination.
3. ERQA is guided by a rather coarse-grained taxonomy, with categories including: spatial reasoning, action reasoning, trajectory reasoning, state estimation, task reasoning, multi-view reasoning, pointing, and other.

To address this need, we introduce `ERQA+` , our newly collected benchmark that inherits the original intention of ERQA, but with some new properties that make it more relevant to embodied reasoning and more challenging:

- The annotation of `ERQA+` is based on manually selected video frame(s), using videos sampled from a diverse set of recent robotic datasets that guarantee an egocentric viewpoint that naturally appears as input to an embodied agent.
- `ERQA+` is collected with the guidance from a more detailed and crafted taxonomy of embodied reasoning (Section 2.1).
- `ERQA+` by design contains new types of prompts or questions, while disallowing multiple-choice questions that have very few options to choose from, making it difficult for guessing.
- The collection process of `ERQA+` introduces new mechanisms for difficulty filtering, leaving out the easiest questions that are relatively trivial to measure embodied reasoning on more modern VLMs.

As shown in Figure 1 (examples on the bottom half), the examples have richer visual scenes but still captured in egocentric viewpoints. The problems are generally more difficult for random guessing as the solution space becomes combinatorially larger.

2 The `ERQA+` Benchmark

2.1 Taxonomy

The collection of `ERQA+` is guided by a slightly more detailed taxonomy, with four coarse categories and 11 fine-grained categories:

- **Spatial reasoning:** covering a wide range of capabilities that needs 2D or 3D relationship understanding: with subcategories *Dynamic* for reasoning on potential movements or traces, *multi-view matching* from simultaneous multiple viewpoints, understanding of *relative shape*, *relative distance*, or *relative direction* among a number of objects or places
- **Perception:** measuring how good a model is at *object and scene recognition*, *state and activity understanding*, and a combinatorial variant of *counting*
- **Planning:** including *goal decomposition* and *navigation*, reflecting reasoning capabilities in different stages
- **Prediction:** with only one subcategory (*future prediction*) measuring capabilities in prediction future events, mostly counterfactual

2.2 Benchmark Statistics

The current release of `ERQA+` contains 800 QA pairs, distributed according to Figure 2. The types of questions also pass beyond the simplest one-in-four or one-in-two multiple-choice questions, with additional types introduced such as sorting, matching, integer-valued counting, composite-judgment that sequentially combines multiple smaller yes-no questions, and also a few open-form questions. The statistics are shown in Table 1. During evaluation, different question types will also be attached with different formatting prompt template, specified in Appendix B.1.

In the next section, we will briefly describe how `ERQA+` has been collected and post-processed.

Statistics	Number
Total Questions	**800**
Multiple-choice	283 (35.4%)
Sorting	214 (26.8%)
Matching	101 (12.6%)
Counting	121 (15.1%)
Composite-judgment	64 (8.0%)
Open	17 (2.1%)

Table 1: Question type of ERQA+.

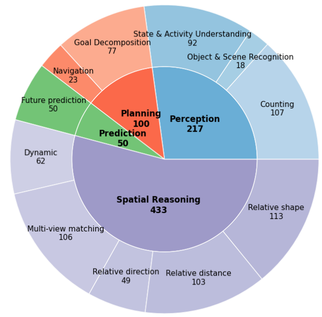

Figure 2: Distribution of task categories.

3 Data Collection

Our benchmark is annotated on manually selected frames from egocentric videos, followed by multiple stages of quality review and filtering. We briefly illustrate the entire process in Figure 3, and describes more details in the following subsections.

3.1 Data sources of egocentric videos

We sample videos from 8 publicly available embodied datasets: AgiBot World (AgiBot-World-Contributors et al., 2025), DROID (Khazatsky et al., 2024), Ego-Exo4D (Grauman et al., 2024), EgoLife (Yang et al., 2025), EgoDex (Hoque et al., 2025), RoboMind (Wu et al., 2024), BDRBench-20 (Li et al., 2024), and Egocentric-10K (AI, 2025). These datasets cover a wide range of environments such as houses, offices, kitchens, supermarkets, and factories. The videos feature diverse activities such as cooking, cleaning, repairing, and object manipulation, providing a rich source of embodied interactions for our benchmark.

3.2 Question annotation

All question-answer pairs are manually labeled to ensure their quality and correctness. The construction of each QA pair usually follows a uniform sequence of steps:

1. Determine the category of the question and select a certain frame or several frames from the video that are suitable for the question.

2. Write the questions and answers, and make necessary annotations on the selected frame(s).

3. Test the question using six small MLLMs to ensure that it is not a simple visual recognition or can be answered without pictures.

The questions are categorized into the following types: multiple-choice questions, ordering questions, matching questions, true/false questions, and numerical problems. Each type of question contains one or more subcategories to evaluate different aspects of the model's capabilities. More details about the annotations and interfaces are presented in the Appendix A.

In the third step, we chose six small MLLMs, phi-4-multimodal-instruct (Microsoft, 2025), qwen2.5-vl-32b-instruct (Qwen Team, 2025), qwen-2.5-vl-7b-instruct (Qwen Team, 2025), glm-4.5v (GLM-V Team, 2025), mistral-small-3.2-24b-instruct (Mistral AI Team, 2025) and gpt-4.1-mini (OpenAI, 2025a) for preliminary detection of the questions. Through the responses of these six models, we ensure that the answers to all questions must be derived through reasoning based on the content of the images. In order to make the data more challenging at the same time, we require that at least two of these six models give incorrect answers. After the data collection is completed, there will be two rounds of review to recheck the category, expression, correctness of the answers and difficulty.

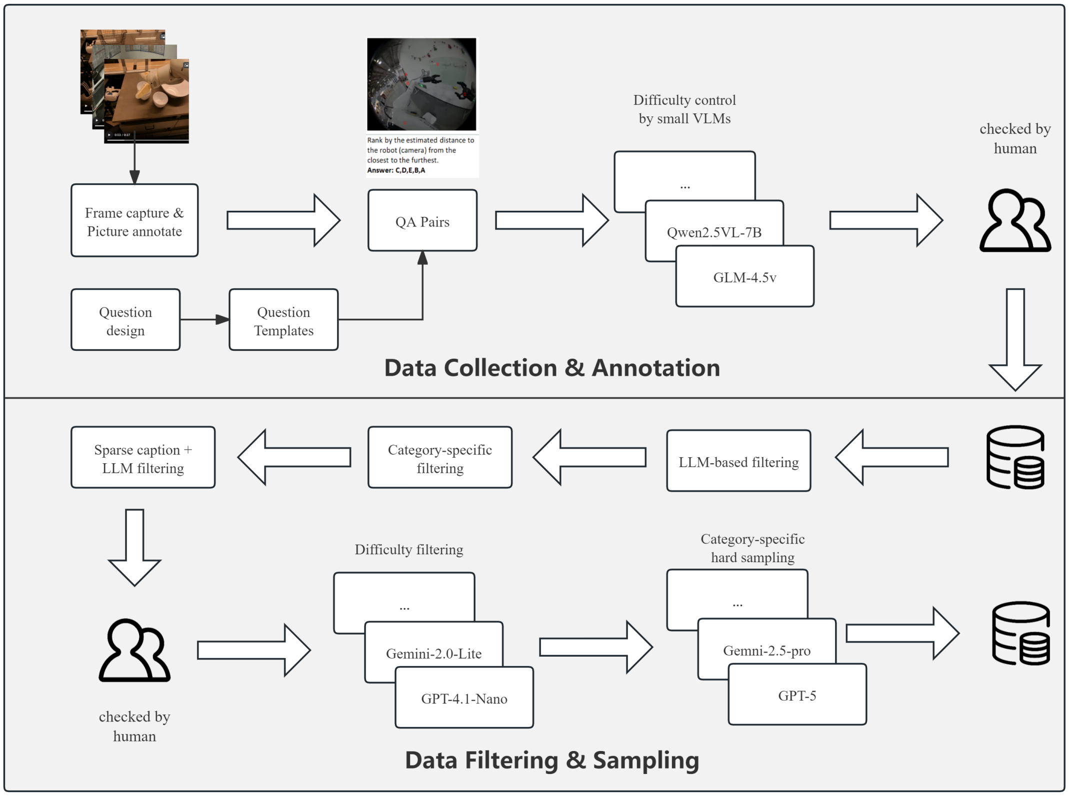

Figure 3: The completion process of data collection, annotation, filtering and sampling.

3.3 Filtering

After multiple rounds of quality review, we also conduct additional filtering schemes in this work to exclude samples that might have been too easy or relatively trivial, henceforth possessing less value for the purpose of evaluation.

LLM-based filtering Similar to many earlier VQA benchmarks, many questions from the initial set we collect might be correctly answered based on just commonsense knowledge encoded in text, even without the need to check the image. We prompt a strong LLM (GPT-5-mini) with the textual questions only, and exclude those that can be correctly guessed.

Sparse caption + LLM filtering Some questions left out from the first step may still be trivial or technically easy as the only dependency would be a simple object recognition. We adopt an "LLM plus perception" filter that first transcribes significant objects or scenes using a strong VLM (Gemini 2.5 Pro), and then attach them with the textual question.

Category-specific filtering We design specific filtering schemes for a few categories. For instance, we found that the target answer for initial questions collected under the category of *counting* heavily bias towards zero or one. For such questions, we only keep those that are sufficiently challenging.

Difficulty filtering We exclude questions that can be answered by all of the smaller and weaker models we evaluated earlier in preliminary trials, including GPT-4.1-nano, Gemini-2.0-Lite, Qwen3-VL-8B.

Category-specific hard sampling The final filtering stage involves targeted down-sampling for several categories that might have overrepresented in the data. We specifically reduce the number of *counting* problems, *multi-view matching* problems, and *relative shape* ordering problems, as we have collected more than intended proportions of those subsets probably due to the relatively easier annotation cost.

4 Results

4.1 Evaluated models

We evaluate the most typical VLMs on ERQA+ , including the frontier proprietary models like Gemini 2.5 Pro (more recently Gemini 3 Pro) and the GPT-5 family, open-weight models like Qwen3-VL-235B-A22B, along with a few smaller VLMs such as Qwen3-VL-8B, Gemma3-12B, and Phi-4-multimodal.

4.2 Experimental results

We show results from all evaluated VLMs on our ERQA+ benchmark in Table 2. Compared with ERQA where the metrics are rather close among different models, our new ERQA+ benchmark clearly enlarges the gap between smaller models and larger models. The overall metrics are 10%-25% lower, showing that ERQA+ is also more challenging.

Model	ERQA+					ERQA
	Perception	Planning	Prediction	Spatial Reasoning	All	All
Gemini-3-Pro-preview	53.5	49.0	74.0	59.1	57.3	66.0
Gemini-2.5-Pro	33.6	37.0	54.0	33.3	35.1	57.3
Gemini-2.5-Flash	32.3	22.0	48.0	26.6	28.9	53.3
GPT-5	42.4	38.0	56.0	46.7	45.0	59.3
GPT-5-Mini	34.1	22.0	40.0	39.3	35.8	53.8
Qwen3-VL-235b-a22b	33.2	34.0	48.0	32.8	34.0	49.5
GPT-5-Nano	18.9	12.0	18.0	21.2	19.3	45.8
Qwen3-VL-8B	23.5	14.0	32.0	21.0	21.5	41.8
Gemma3-12b	15.7	8.0	14.0	13.9	13.6	36.8
Phi-4-Multimodal	18.4	11.0	14.0	14.3	15.0	36.0

Table 2: Performance comparison of various vision-language models on ERQA and ERQA+.

Frontier models such as Gemini 3 Pro and GPT-5 are showing stronger performance on many aspects of embodied reasoning. As shown in Table 3, Gemini 3 Pro (Preview) is showing very good performance on multi-view matching, state & activity understanding, and future prediction.

L1 Category	L2 Category	Correct	Total	Accuracy (%)
Overall	Total	458	800	57.25
Perception	Total	116	217	53.50
Perception	Counting	45	107	42.10
Perception	State & Activity Understanding	60	92	65.20
Perception	Object & Scene Recognition	11	18	61.10
Planning	Total	49	100	49.00
Planning	Goal Decomposition	41	77	53.20
Planning	Navigation	8	23	34.80
Spatial Reasoning	Total	256	433	59.10
Spatial Reasoning	Dynamic	39	62	62.90
Spatial Reasoning	Relative distance	39	103	37.90
Spatial Reasoning	Relative direction	27	49	55.10
Spatial Reasoning	Multi-view matching	88	106	83.00
Spatial Reasoning	Relative shape	63	113	55.80
Prediction	Total	37	50	74.00
Prediction	Future prediction	37	50	75.60

Table 3: Detailed results on Gemini-3-Pro-Preview

Some other findings:

Larger VLMs are stronger in recognition and prediction Models at the frontier, such as GPT-5 and the very recently announced Gemini 3 Pro (Preview), are showing superior capabilities across multiple dimensions of embodied reasoning. On the other side of the table, it is shown to be very difficult for smaller VLMs to accurately identify the objects marked on the picture. Figure 4 shows an example on which the smaller GPT-5-Nano model makes many errors on perception or relational understanding, while GPT-5 has nailed it.

Question	GPT 5 Reasoning Trace	GPT 5 Nano Reasoning Trace
  In the first image, which points in the second image correspond to A, B, C, and D respectively? Answer: D,C,B,A	... In the first image, point A is on the left side near the window, with B on the couch backrest, C by the suitcase wheel, and D on the top right corner of the bag. So, point D corresponds to the window location in the second image. So, I've determined that the first image's point A corresponds to point D in the second image. For point B, which is on the couch in the first image, it corresponds to point C in the second image. Point C from the first image is near the wheel, but I can't find a specific corresponding dot on a wheel in the second image. Point B in the second image is near the top left corner of the bag, potentially close to the wheel area. Lastly, point D in the first image aligns with point A in the second. I'll conclude that the mapping is as follows: A corresponds to D, B to C, C to B, and D to A. So, the final answer is: Final Answer: D, C, B, A.	... I need to determine the correct mapping of points A, B, C, and D across both images. In the first image, A is near the left window frame, while D is at the couch, yet in the second image, D is by the window frame on the right. It's a bit confusing! When I look closely at the positions, it seems that A might correspond to B in the second image, while D might align in other ways. I think I need to solidify the mappings—perhaps in order: D, A, B, C for clarity? In the first image, D is at the back of the sofa, while in the second image, there's also a D label in the same area. But the task requires me to specify how A, B, C, and D correspond between the two images. Notably, I see C near the suitcase wheel in the first image, while B appears near the left side of the suitcase in the second. After analyzing the positions, my best guess for the final answer is: Final Answer: B, A, D, C.

Figure 4: Example on the superiority of larger models (GPT-5) versus a smaller VLM (GPT-5-Nano)

Relative spatial relationships remains a challenge We notice that even the strongest models are still making errors on spatial understanding, especially those formatted with our new question types (e.g., sorting). As one may notice in Figure 5, models are still giving superficial claims about relative distance judging from the 2D image of a 3D scene. Identifying the 3D relative spatial relationships of objects on 2D images remains a considerable challenge in the current reasoning paradigm.

Question	Gemini 3 Reasoning Trace	GPT 5 Reasoning Trace
 Which point among ABCDE is closest to the left gripper? Answer: B	...... **Point A** is positioned just above the main body of the left gripper. The vertical distance on the 2D image plane between the top of the left gripper and Point A is relatively small. **Point E** is much higher up and further away on the wall. **Point B** is further to the right and deeper into the background than Point A. **Points C and D** are on the right side of the image, much closer to the right gripper than the left one. **Conclusion**: Comparing the proximity of all labeled points to the left gripper, Point A is the nearest both in terms of 2D pixel distance and physical location within the scene's layout (being on the left side behind the arm). Final Answer: A	... Comparing the 2D distances from the left gripper to each labeled point: - A is on the plant pot near the left side, close to the left gripper. - B and C are higher and more central/right, farther away. - D is near the right chair, closest to the right gripper. - E is far upper-left. Thus, A is the closest to the left gripper. Final Answer: A

Figure 5: An example that none of the evaluated models can answer correctly.

5 Conclusion

We introduce ERQA+ , a new benchmark on embodied reasoning that addresses several limitations from another earlier benchmark ERQA. Our evaluation results show magnified gaps between smaller models and larger models, while also showing a bigger room for improvement in embodied reasoning. We hope this resource could help future development of embodied vision-language models with improved spatial reasoning, planning, and event understanding capabilities as targets.

Acknowledgment

We thank all members in our *FlagEval* team for earlier contribution to a smaller round of initial trial annotation.

References

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL https://arxiv.org/abs/2503.06669.

Build AI. Egocentric-10k, 2025. URL https://huggingface.co/datasets/builddotai/Egocentric-10K.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. GR-3 technical report, 2025. URL https://arxiv.org/abs/2507.15493.

Gemini Robotics Team. Gemini Robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer, 2025a. URL https://arxiv.org/abs/2510.03342.

Gemini Robotics Team. Gemini Robotics: Bringing ai into the physical world, 2025b. URL https://arxiv.org/abs/2503.20020.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2025. URL https://arxiv.org/abs/2505.11709.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ZMnD6QZAE6.

Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024.

Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Mistral AI Team. Mistral Small 3, Jan 2025. URL https://mistral.ai/news/mistral-small-3.

OpenAI. Introducing GPT-4.1 in the API, April 2025a. URL https://openai.com/index/gpt-4-1/.

OpenAI. Introducing GPT-5, August 2025b. URL https://openai.com/index/introducing-gpt-5/.

Physical Intelligence. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

Qwen Team. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. RoboVQA: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652, 2024. doi: 10.1109/ICRA57147.2024.10610216.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics, 2025. URL https://arxiv.org/abs/2506.01844.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant, 2025. URL https://arxiv.org/abs/2503.03803.

A Annotation Details

We include the annotation guideline and the interfaces of relevant tools in this section.

A.1 Annotation guideline

The annotation guideline contains detailed guidance on the specific steps from loading a robotic video to the end result of annotated questions and video frames. During the annotation process, an additional interface is also provided to instantly test a candidate data sample on weaker or smaller VLMs to get a sense of potential question difficulty.

Specific guidelines:

1. **Video upload:** The video annotation tool operates as a browser-based frontend utility (See Figure 6). For initial configuration, users are required to input a username in the upper-right corner of the interface. This identifier is persisted locally, thereby eliminating the need for repetitive input in subsequent sessions. To import one or more target videos, users can either click the 'Upload Video' button or directly drag and drop the files into the designated gray area. To identify specific frames for annotation, users can navigate the video by dragging the progress bar or pressing the spacebar to toggle playback.
2. **Annotation:** Annotation is performed on paused frames using the left-side panel. Users first specify the target category (question type and sub-class). The tool supports four distinct annotation modes: point, rectangle, arrow, and text label, each allowing for adjustments in color and line width. Furthermore, the tool supports error correction via undo operations or by clearing the entire canvas. Subsequently, users are required to input the corresponding prompt and answer for the current frame. Pressing the 'S' key saves the annotation, which is then immediately displayed in the annotation log on the right-hand panel. For scenarios involving multi-frame prompts, the text input (question and answer) must be provided only on the initial frame of the sequence.
3. **Preliminary trial:** By clicking the central 'Multi-model QA' button, users are redirected to a dedicated frontend interface. Here, users copy the text prompt and drag the corresponding image from the annotation log into the input field. Clicking the 'Run Six Models' button triggers concurrent inference across six multimodal large models. Finally, clicking 'Download JSON' exports the results into a JSON file, which serves as the requisite input for the tool's 'JSON File' module.
4. Expected result: Upon completion of all annotation tasks, clicking the 'Finish All Annotations' button in the bottom-left corner triggers the download of a ZIP archive, which include "uploaded_data.json" (the raw user input), "annotation_summary.json" (aggregated annotation results), an "images directory" (housing all relevant image files), and a "README.md" file.

A.2 Interfaces

Figure 6 shows the screenshot of the interface for video frame extraction and question annotation. Figure 7 show the interface for preliminary trials on smaller VLMs during annotation.

B Evaluation Details

B.1 Evaluation prompts

To simplify answer parsing, we add specific prompt suffices after different types of problems:

- Open-form: ***Finalize your output with:** 'Final Answer:', followed by a string representing the correct answer.*
- Multiple-choice questions: ***Finalize your output with:** 'Final Answer:', followed by a letter or a comma-separated list of letters which means the option of correct answer. **Format example:** 'Final Answer: A'.*
- Matching: ***Finalize your output with:** 'Final Answer:', followed by a comma-separated list. **Format example:** 'Final Answer: A, B, C, D, E'.*
- Ordering: ***Finalize your output with:** 'Final Answer:', followed by a comma-separated list. **Format example:** 'Final Answer: A, B, C, D, E'.*
- Counting: ***Finalize your output with:** 'Final Answer:', followed by a number. **Format example:** 'Final Answer: 123'.*
- Composite-judgment: ***Finalize your output with:** 'Final Answer:', followed by a comma-separated list composed of 0 or 1.**Format example:**'Final Answer:[1,1,0]'.*

For more details, please check our GitHub repository referred in the project website.

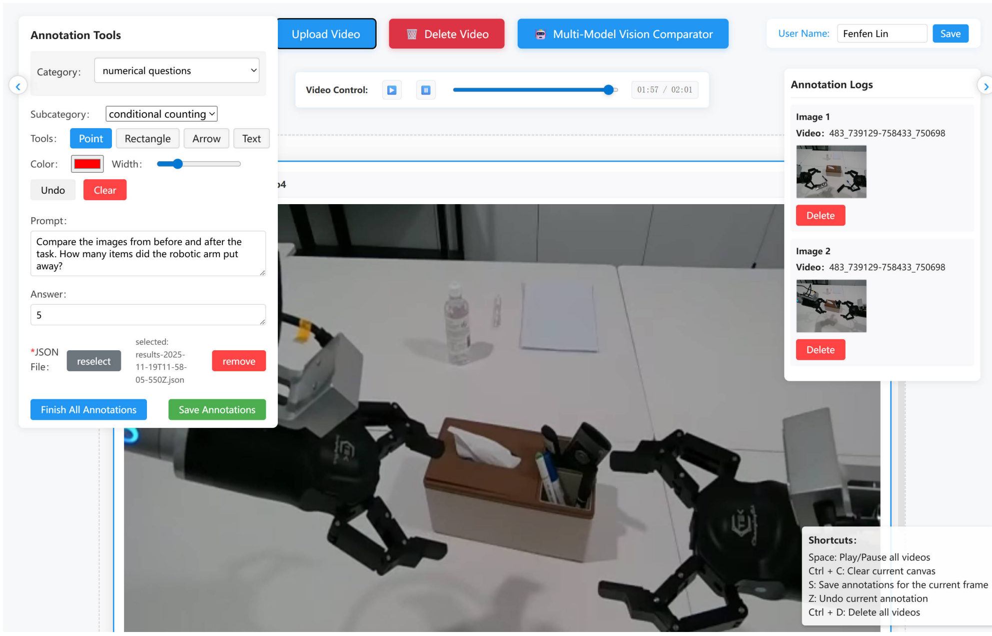

Figure 6: The interface of the annotation tool

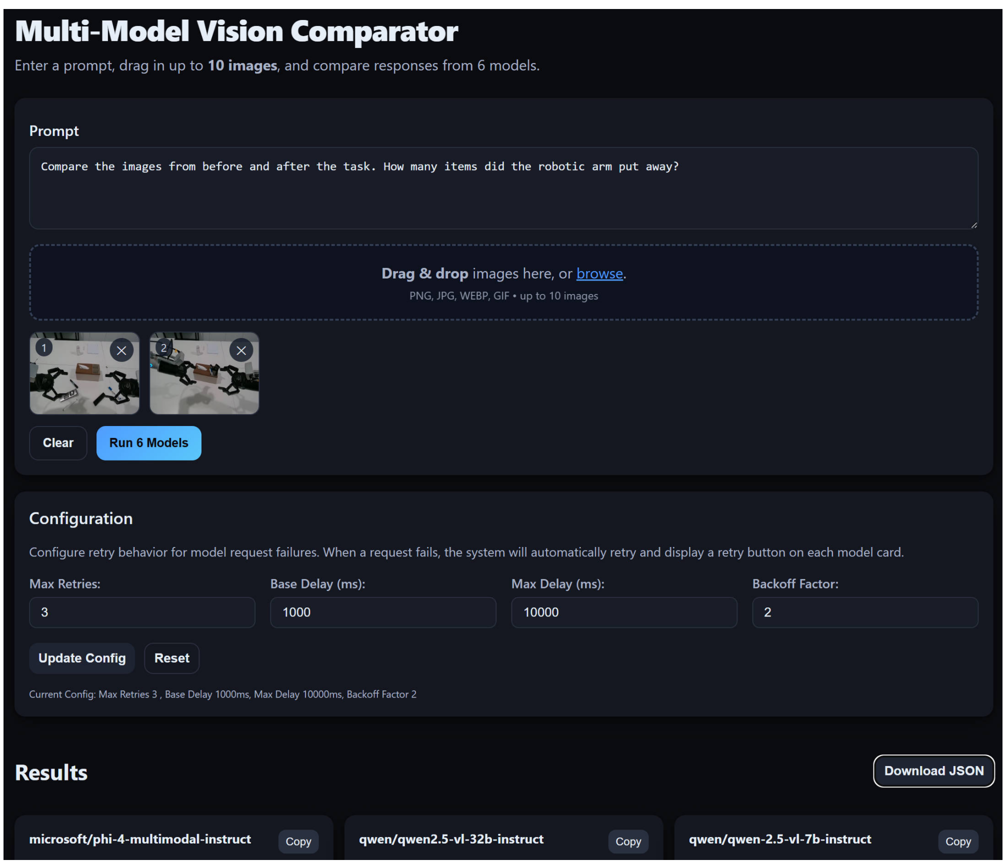

Figure 7: The interface of the tool for preliminary MLLM trials to examine the difficulty of samples during data annotation