EgoLink: Egocentric Language-Vision Interactive Network Knowledge Challenge

Yueying Feng[1,2], Bohan Yu[1], Renhe Sun[1], Zitong Wang[1], Chang Yao[2], Jian Liu[1], Jingyuan Chen[2]
[1]Ant Group, China [2]Zhejiang University, China
{ybh441692, sunrenhe.srh}@antgroup.com yueyingf@zju.edu.cn

1 Abstract

The **Ego**centric **L**anguage-Vision **I**nteractive **N**etwork **K**nowledge(**EgoLink**) Challenge redefines the cognitive boundaries of embodied agents in social contexts. While Embodied AI ultimately aims to perceive and interact from an egocentric perspective, current research predominantly emphasizes physical navigation while neglecting deep social understanding. EgoLink introduces a large-scale, real-world egocentric benchmark that employs a multi-dimensional **Multiple-Choice Question(MCQ)** format to evaluate models' reasoning capabilities across emotions, causal logic, and behavioral intents in human interactions. This challenge bridges the gap between perception and social cognition, advancing Embodied AI toward socially-aware general intelligence.

2 Importance and Research Impact

As AR/VR devices, AI glasses, and social robots become increasingly ubiquitous, understanding human-centric social interactions has emerged as a central challenge in multimedia research. While current Multimodal Large Language Models (MLLMs) excel at scene description, they struggle with long-range causal reasoning and fine-grained emotional understanding in egocentric contexts. EgoLink addresses these gaps through a structured MCQ-based benchmark integrating multimodal fusion, video understanding, and social commonsense reasoning. By providing rigorous evaluation of social reasoning capabilities, EgoLink will accelerate advancements in multimodal memory and temporal causal inference—foundational research for developing trustworthy, proactive embodied AI systems with broad societal and industrial impact.

3 Research Tasks

The **EgoLink Challenge** features a single integrated track where models demonstrate social intelligence by answering **multi-dimensional video MCQs** across four key cognitive dimensions:

- **Emotional Perception & Temporal Localization:** Identifying emotion categories and their precise temporal boundaries within egocentric video streams
- **Social Causal Reasoning:** Analyzing triggers behind emotions and explaining causal mechanisms driving social actions or reactions
- **Behavioral Intent Prediction:** Inferring subsequent intentions and potential social goals of interlocutors based on conversational context
- **Egocentric Semantic Summarization:** Selecting the most accurate high-level description that best captures the essence of social exchanges

4 Dataset and Documentation

4.1 Dataset Foundation and Construction

The EgoLink Challenge builds upon the ***E*^{3} (Exploring Embodied Emotion)** dataset (Lin, Feng et al., NeurIPS 2024 [1]), a pioneering large-scale egocentric video benchmark for embodied emotion understanding. While E^3 provides foundational egocentric video data with emotion annotations, EgoLink transforms this resource into a comprehensive social reasoning benchmark through systematic re-annotation and task reformulation.

4.1.1 Data Source and Scale

- **Base Dataset:** E^3 egocentric video corpus, featuring real-world social interactions captured from first-person perspectives (`https://exploring-embodied-emotion-official.github.io`).
- **Video Collection:** Approximately 20,000+ egocentric clips spanning diverse social scenarios
- **Total Duration:** Over 70 hours of annotated egocentric footage
- **Scenario Coverage:** Daily conversations, workplace meetings, social gatherings, educational settings, service interactions, and family activities
- **Participant Diversity:** 500+ unique individuals across varied demographics, ages, and cultural backgrounds

4.1.2 Dataset Organization and Design Philosophy

The EgoLink Challenge is designed to cultivate deep reasoning capabilities in embodied AI. Rather than providing massive training data that encourages pattern matching, we provide a two-tier supervision strategy: foundational E^3 emotion annotations for basic perception, and 5,000 carefully designed MCQs for reasoning development. Our design follows three core principles: (1) **Limited Supervision**—constrained MCQ training (5,000 samples) encourages generalizable reasoning over pattern memorization; (2) **Transfer Learning**—models leverage E^3's emotion understanding and pretrained knowledge to bridge perception and social reasoning; (3) **Cognitive Leap**—MCQs test causal inference, temporal generalization, intent prediction, and multimodal synthesis beyond basic emotion recognition.

Dataset Splits The EgoLink dataset is partitioned into four subsets to ensure robust training and evaluation:

- **Training Set:** Built upon the E^3 emotion recognition split, it encompasses over 70 hours of egocentric video with extensive metadata. It is augmented with approximately 5,000 MCQs covering emotional perception, social causality, intent prediction, and semantic summarization.
- **Validation Set:** Systematically derived from E^3 validation samples, this split maintains a balanced distribution across all targeted cognitive dimensions.
- **Test Set:** Highly curated MCQs with complex edge cases and compositional tasks. A hidden subset including out-of-distribution scenarios is reserved for final ranking to prevent overfitting and ensure real-world transferability.

4.2 Documentation and Resources

4.2.1 Project Resources

- **Official Website:** `https://egolink-challenge.github.io` (to be launched)
  Serves as the central hub for dataset downloads, live leaderboards, documentation, and FAQs.
- **GitHub Repository:** `https://github.com/egolink/challenge2026` (to be launched)
  Provides the official codebase, including:
  - Data loading, preprocessing pipelines, and evaluation toolkits.
  - Baseline model implementations and training recipes.
  - Submission format validators and example code.

4.2.2 Data Access and Licensing

- **License:** Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- **Access:** Free registration with institutional email verification
- **Usage Requirements:**
  - *Attribution:* All publications must cite the E^3 dataset (Lin, Feng et al., NeurIPS 2024)
  - *Non-Commercial:* Restricted to research and educational purposes only
  - *ShareAlike:* Derivative works must be distributed under the same license

4.2.3 Technical Support

- **Discussion Forum:** GitHub Discussions for community Q&A and troubleshooting
- **Documentation:** Comprehensive manual covering dataset construction, annotation protocols, evaluation metrics, submission specifications, and troubleshooting guidance

4.3 Ethical Considerations

We follow the principles of privacy protection and copyright respect. We obtain explicit consent from video owners and ensure that they understand the use of personal information (*e.g.*, portrait, voice). We comply with the data protection regulations of *YouTube*, *Douyin*, and *BiliBili*, providing transparency to participants and guaranteeing their right to withdraw consent. We collect only necessary data and ensure strict data usage boundaries to support research while protecting privacy.

4.4 Future Extensions

The EgoLink dataset is designed for sustained growth to support emerging egocentric AI research directions:

- **Longitudinal & Temporal Enhancement:** Follow-up recordings tracking evolving social relationships; frame-level annotations enabling fine-grained temporal modeling
- **Embodied Action Grounding:** Annotations of hand gestures, body movements, and physical interactions linking social cognition with embodied actions visible from egocentric viewpoints
- **Advanced Dialogue Annotation:** Fine-grained dialogue act labels, pragmatic functions, conversational strategies, and turn-taking patterns from the ego-agent's perspective

5 Evaluation Criteria

The evaluation framework ensures objectivity by utilizing a MCQ format to eliminate subjective bias. Performance is primarily measured by **Overall Top-1 Accuracy**, supplemented by dimension-specific breakdowns that offer granular insights into model proficiency across emotion, causality, intent, and summarization.

6 Commitment to Maintenance

The organizing team commits to maintaining the official website, leaderboard, and evaluation infrastructure through 2029, ensuring at least three years of continuous dataset access and technical support. By providing regular updates and community engagement through workshops and forums, we ensure EgoLink serves as a sustainable benchmark for egocentric social reasoning research.

7 Publicity and Collaboration

We will collaborate closely with ACMMM 2026 organizers to maximize challenge visibility and participation:

- **Publicity Channels:** Academic mailing lists, social media campaigns, and official ACM Multimedia channels
- **Workshop:** Dedicated Grand Challenge workshop during ACMMM 2026 to showcase winning solutions, present technical insights, and foster community dialogue
- **Community Building:** Establishment of a sustained research community around egocentric social intelligence

8 Contact Information

- **Yueying Feng**, Ant Group & Zhejiang University, `yueyingf@zju.edu.cn`
  Role: Dataset management, annotation protocols, and technical evaluation
- **Bohan Yu**, Ant Group, `ybh441692@antgroup.com`
  Role: Publicity, community engagement, and conference liaison
- **Renhe Sun**, Ant Group, `sunrenhe.srh@antgroup.com`
  Role: Technical infrastructure, leaderboard maintenance, and platform support

References

[1] Wang Lin, Yueying Feng, Wenkang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. E^3: Exploring Embodied Emotion Through A Large-Scale Egocentric Video Dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, Datasets and Benchmarks Track, volume 37, 2024.