Technical Report for Ego4D Long-Term Action Anticipation Challenge 2025

Qiaohui Chu[1][2], Haoyu Zhang[1][2], Yisen Feng[1], Meng Liu[3], Weili Guan[1],
Yaowei Wang[1][2], Liqiang Nie[1]

[1]Harbin Institute of Technology (Shenzhen)   [2]Pengcheng Laboratory
[3]Shandong Jianzhu University

{qiaohuichu8599, zhang.hy.2019, yisenfeng.hit, mengliu.sdu, honeyguan, nieliqiang}@gmail.com;
wangyw@pcl.ac.cn

Abstract

In this report, we present a novel three-stage framework developed for the Ego4D Long-Term Action Anticipation (LTA) task. Inspired by recent advances in foundation models, our method consists of three stages: feature extraction, action recognition, and long-term action anticipation. First, visual features are extracted using a high-performance visual encoder. The features are then fed into a Transformer to predict verbs and nouns, with a verb–noun co-occurrence matrix incorporated to enhance recognition accuracy. Finally, the predicted verb–noun pairs are formatted as textual prompts and input into a fine-tuned large language model (LLM) to anticipate future action sequences. Our framework achieves first place in this challenge at CVPR 2025, establishing a new state-of-the-art in long-term action prediction. Our code will be released at https://github.com/CorrineQiu/Ego4D-LTA-Challenge-2025.

1. Introduction

Egocentric video understanding has emerged as a critical research direction, driven by the rapid proliferation of large-scale egocentric video datasets. Unlike traditional third-person videos, egocentric videos capture everyday activities directly from the wearer's perspective, inherently embedding the actor's attention, actions, and subjective intentions [20]. This unique viewpoint closely aligns egocentric research with the goals of Embodied AI, where understanding and predicting human behavior from the actor's perspective is essential. However, due to frequent head or body movements and limited environmental context, egocentric videos present distinct and significant analytical challenges, inspiring various complex research problems [2, 3, 16, 18, 19, 21].

A representative task in this area is Long-Term Action Anticipation (LTA) from the Ego4D benchmark[4], which aims to predict an actor's future behavior from observed video sequences, typically represented as sequences of verb–noun pairs. Compared to third-person LTA tasks, egocentric LTA has greater potential to directly facilitate natural and effective human–computer interaction by predicting future actions from the actor's own viewpoint.

Recently, notable breakthroughs in visual perception and vision–language understanding [5, 9, 10, 15–17], coupled with advancements in temporal sensitivity of large language models (LLMs)[14], have significantly enhanced the models' capabilities in video content understanding and representation. Research in egocentric LTA has gradually shifted from small-scale predictive models to methods based on LLMs. Palm [6] initially demonstrated the feasibility of using Llama2-7B [13] for action anticipation, while AntGPT improved upon Palm by fine-tuning Llama2-7B, achieving notably better performance. However, as highlighted in both Palm and AntGPT, the accuracy of the initial action recognition models significantly affects the final prediction quality, and existing recognition models still have considerable room for improvement.

To address this issue, we propose an efficient three-stage framework comprising (1) feature extraction, (2) action recognition, and (3) long-term action anticipation. By leveraging a high-performing visual encoder for feature extraction, incorporating hand–object interaction cues, and integrating a verb–noun co-occurrence matrix, our framework improves the accuracy of the action recognition stage. These enhancements collectively strengthen the reliability and precision of the subsequent long-term action anticipation. Empirical validation demonstrates the effectiveness of our method. Specifically, in the Ego4D Long-Term Action Anticipation Challenge, our approach surpasses all other competitors on the public leaderboard, achieving first place in the challenge.
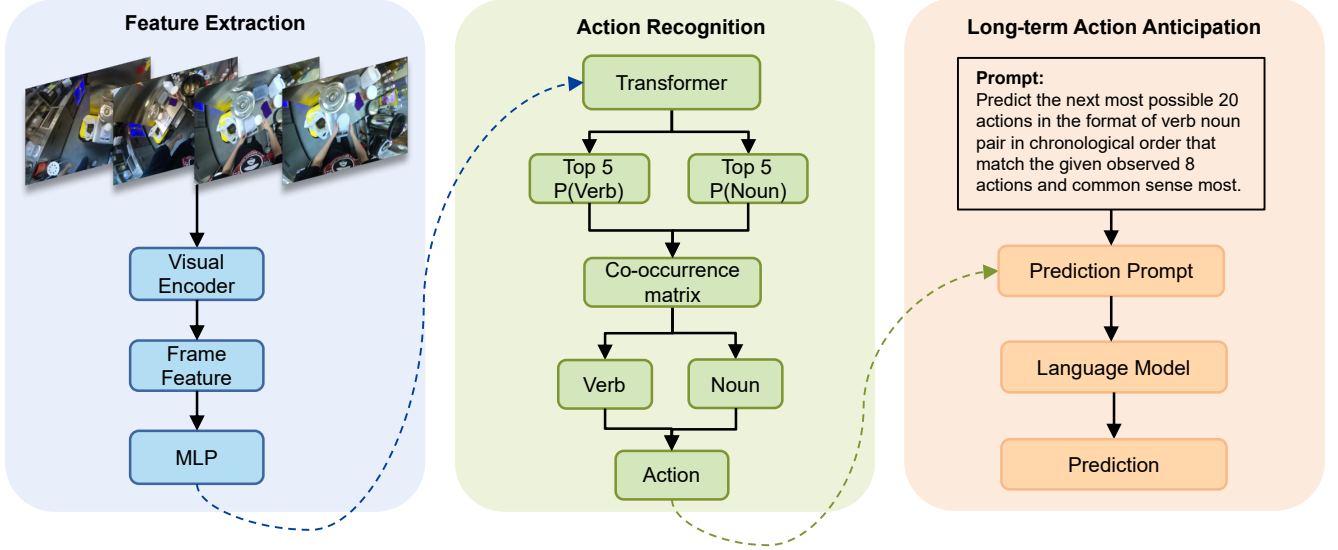
Figure 1. An illustration of our two-stage pipeline.

2. Methodology

Our proposed method, as illustrated in the Figure 1, is composed of three main stages: feature extraction, action recognition, and long-term action anticipation.

2.1. Feature Extraction

In the feature extraction stage, we adopt the general framework introduced by AntGPT [22], uniformly sampling four frames from each video segment that corresponds to an action. After evaluating two visual encoders, EgoVLP [8] and EgoVideo-V [11], we select EgoVideo-V due to its superior performance for extracting visual features. To enhance the quality of input features, we incorporate a segmentation-based method [12] to extract visual information from hand–object interaction regions. Both the hand–object interaction regions and the original video frames are fed into the visual encoder to obtain their respective features, which are then fused using a lightweight MLP. This enriched input helps improve the downstream action recognition accuracy.

2.2. Action Recognition

In the action recognition stage, visual features are fed into a Transformer model to predict candidate verbs and nouns, from which we obtain the top five predictions for both verbs and nouns. Inspired by the verb-noun co-occurrence matrix proposed in QueryMamba [24], we further integrate a similar co-occurrence matrix into our recognition model to enhance prediction accuracy. Specifically, we multiply the probability distributions of the top-five verbs and nouns output from the Transformer model by a normalized co-occurrence matrix. Each top-five verb generates five noun candidates with the highest co-occurrence prob-

abilities, forming the top-five verb–noun pairs. Similarly, each top-five noun generates five verb candidates, forming another set of top-five noun–verb pairs. From these ten candidate verb–noun pairs, we select the pair with the highest overall probability as the final prediction of the action recognition stage.

2.3. Long-term Action Anticipation

In the action anticipation stage, existing studies typically convert visual observation histories into textual form using an action recognition model, and then utilize LLMs for inference, as demonstrated by methods such as Palm, AntGPT, and EgoVideo. Notably, AntGPT and EgoVideo achieved improved results by fine-tuning Llama2-7B and Vicuna-7B [23], respectively. Following a similar approach, we feed the textual representations of visual observation histories obtained from the action recognition stage into a fine-tuned Llama2-7B model, predicting a sequence of 20 future actions.

3. Experiment

3.1. Performance Comparison

The performance on the LTA task is evaluated using the Edit Distance (ED) metric, which measures the average number of insertions, deletions, substitutions, and transpositions between the predicted sequences and the ground truth sequences. Table 1 summarizes results from existing LLM-based methods and the public leaderboard on the Ego4D v2 test set. As shown in the table, our proposed framework achieves the best (ranked first) ED scores for both nouns and overall actions. In terms of verb ED, our method ranks second, with only a slight difference of 0.06% compared to the

Table 1. Performance comparison of existing work and the top six teams on the public leaderboard. The best results are shown in bold.

Method	Rank	LTA ED↓		
		Verb	Noun	Action
VideoLLM [1]	–	0.7210	0.7250	0.9210
Palm (Challenge) [6]	–	0.6956	0.6506	0.8856
AntGPT [22]	–	0.6503	0.6498	0.8770
PALM (Paper) [7]	–	0.6471	0.6117	0.8503
Charlie's (AntGPT)	6	0.6531	0.6446	0.8748
HAI-PUI	5	0.6362	0.6258	0.8663
Doggeee (w/o finetuning)	4	0.6663	0.6240	0.8650
BigMac (mtp)	3	**0.6340**	0.6395	0.8649
123456ABCD (newdict)(EgoVideo) [11]	2	0.6354	0.6367	0.8504
Ours (iLearn2.O)	1	0.6346	**0.5986**	**0.8493**

top-ranked method. These results clearly demonstrate the superior performance of our framework compared to existing approaches.

3.2. Ablation Study

Table 2 shows ablation results for different visual encoders and inference LLMs. The results show that our approach consistently outperforms baseline methods in recognition accuracy, regardless of whether EgoVLP or EgoVideo-V is used as the backbone encoder. Furthermore, we observe that the advanced EgoVideo-V achieves the highest action recognition accuracy, reaching 22.15%.

For the long-term action anticipation, we utilize a fine-tuned Llama2-7B as our anticipation model. Despite the action recognition accuracy achieved by our model being slightly lower than the highest accuracy from the fine-tuned EgoVideo-V, our overall LTA predictions clearly outperform all baselines. By combining improvements across all three stages, we achieve superior overall performance on the LTA task, demonstrating the positive contributions of each component to the final prediction accuracy.

3.3. Case Study

Figure 2 illustrates both successful and failed examples from our action recognition model. In the failed action recognition case, we notice the model struggles to differentiate between semantically related nouns or verbs with varying levels of granularity. For example, the model incorrectly identifies "screw" as "metal"; although a screw is indeed a type of metal, the difference in granularity leads to incorrect predictions. Addressing this limitation in future work may involve methods to enhance the model's ability to recognize actions at different semantic granularities.

Figure 3 illustrates an example from our action anticipation model. For this failed example, the model demonstrates primarily insufficient predictive capability for future scenarios, resulting in repetitive sequences closely related to the
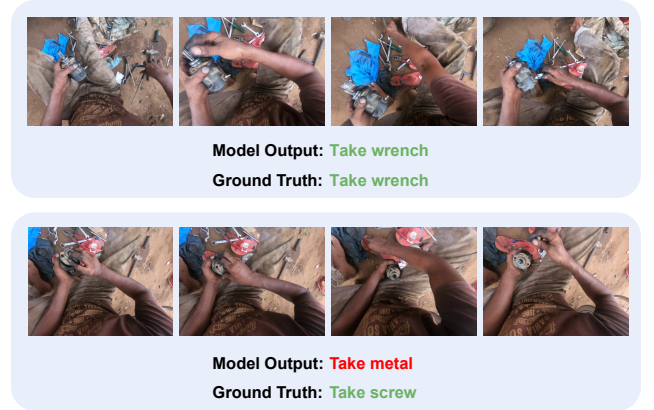

Model Output: Take wrench
Ground Truth: Take wrench

Model Output: Take metal
Ground Truth: Take screw

Figure 2. Successful and failed cases of Action Recognition model.

Prompt:

Predict the next most possible 20 actions in the format of verb noun pair in chronological order that match the given observed 8 actions and common sense most. Below is the observed 8 actions.
Observed actions: take screw, put screw, take screw, put screw, attach screw, repair rod, put screwdriver, take wrench
Prediction:

Model Output

Take **wrench**, **put** wrench, **take wrench**, put wrench, take **wrench**, put wrench, **take wrench**, **put wrench**, **take** wrench, put wrench, **take wrench**, put wrench, take wrench, put wrench, take wrench, put wrench, take wrench, put wrench, take wrench, put wrench

Ground Truth

Put **wrench**, **put** drill, **take wrench**, repair energy, put **wrench**, repair energy, **take wrench**, **put wrench**, **take** screw, repair energy, **take wrench**, repair energy, repair energy, take phone, put cement, apply cement, apply cement, wipe floor, dip sponge, dip sponge
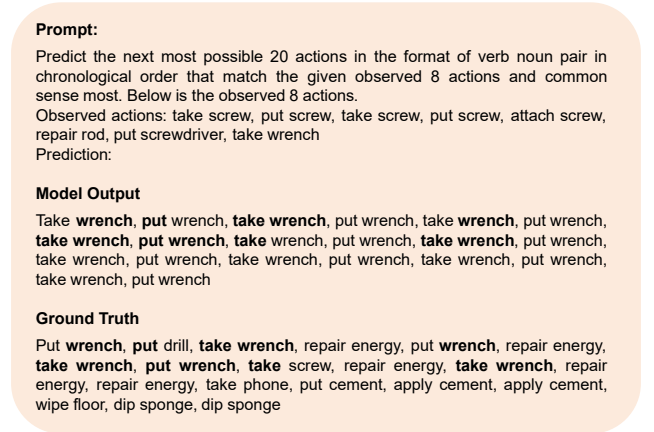
Figure 3. An example from our action anticipation model.

final observed actions. For instance, future actions could involve "take phone," "put cement," and "apply cement," but our model failed to anticipate these tasks, leading to repet-

Table 2. Performance comparison of our method and baselines. † indicates using a sliding-window strategy; ‡ indicates the model is finetuned on the Ego4D dataset. In the AR table, bolded results indicate performance using a zero-shot visual encoder without applying the sliding window strategy. In the LTA table, bolded results represent the best-performing scores.

Method	LLM	Visual Encoder	AR Acc↑			LTA ED↓		
			Verb	Noun	Action	Verb	Noun	Action
PALM [7]	LLaMA2-7B [13]	EgoVLP [8]	32.87	38.26	15.18	–	–	–
		EgoVLP†	40.32	45.53	20.63	0.6471	0.6117	0.8503
Egovideo [11]	Vicuna-7B [23]	EgoVideo-V‡	43.65	52.21	27.64	0.6354	0.6367	0.8504
Ours (EgoVLP)	LLaMA2-7B	EgoVLP	30.90	40.28	15.36	0.6491	0.6429	0.8718
Ours (EgoVideo-V)	LLaMA2-7B	EgoVideo-V	**38.45**	**51.50**	**22.15**	**0.6346**	**0.5986**	**0.8493**

itive or incorrect predictions. To mitigate this, we plan to incorporate task-intent recognition modules in future work, providing the model with clearer intent-driven guidance for more accurate predictions.

4. Conclusion

We introduce a three-stage enhanced framework for the LTA task, which substantially improves performance in both the action recognition and long-term action anticipation. Experimental results demonstrate that our method achieves outstanding performance, securing first place in the CVPR 2025 Ego4D Long-Term Action Anticipation Challenge and surpassing previous state-of-the-art approaches.

References

[1] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 3

[2] Yisen Feng, Haoyu Zhang, Yuquan Xie, Zaijing Li, Meng Liu, and Liqiang Nie. Objectnlq@ ego4d episodic memory challenge 2024. *arXiv preprint arXiv:2406.15778*, 2024. 1

[3] Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, and Liqiang Nie. Object-shot enhanced grounding network for egocentric video. *arXiv preprint arXiv:2505.04270*, 2025. 1

[4] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1

[5] Weili Guan, Xuemeng Song, Haoyu Zhang, Meng Liu, Chung-Hsing Yeh, and Xiaojun Chang. Bi-directional heterogeneous graph hashing towards efficient outfit recommendation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 268–276, 2022. 1

[6] Daoji Huang, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. *arXiv preprint arXiv:2306.16545*, 2023. 1, 3

[7] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *European Conference on Computer Vision*, pages 140–158. Springer, 2024. 3, 4

[8] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 2, 4

[9] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 15–24, 2018. 1

[10] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851, 2018. 1

[11] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 2, 3, 4

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 4

[14] Yunxiao Wang, Meng Liu, Rui Shao, Haoyu Zhang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Liqiang Nie. Time: Temporal-sensitive multi-dimensional instruction tuning and benchmarking for video-llms. *arXiv preprint arXiv:2503.09994*, 2025. 1

[15] Haoyu Zhang, Meng Liu, Zan Gao, Xiaoqiang Lei, Yinglong Wang, and Liqiang Nie. Multimodal dialog system: Relational graph-based context-aware question understanding. In *Proceedings of the 29th ACM international conference on multimedia*, pages 695–703, 2021. 1

[16] Haoyu Zhang, Meng Liu, Yuhong Li, Ming Yan, Zan Gao, Xiaojun Chang, and Liqiang Nie. Attribute-guided collaborative learning for partial person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14144–14160, 2023. 1

[17] Haoyu Zhang, Meng Liu, Yaowei Wang, Da Cao, Weili Guan, and Liqiang Nie. Uncovering hidden connections: Iterative tracking and reasoning for video-grounded dialog. *arXiv preprint arXiv:2310.07259*, 2023. 1

[18] Haoyu Zhang, Meng Liu, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Multi-factor adaptive vision selection for egocentric video question answering. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59310–59328. PMLR, 2024. 1

[19] Haoyu Zhang, Yuquan Xie, Yisen Feng, Zaijing Li, Meng Liu, and Liqiang Nie. Hcqa@ ego4d egoschema challenge 2024. *arXiv preprint arXiv:2406.15771*, 2024. 1

[20] Haoyu Zhang, Qiaohui Chu, Meng Liu, Yunxiao Wang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. *arXiv preprint arXiv:2503.09143*, 2025. 1

[21] Haoyu Zhang, Yisen Feng, Qiaohui Chu, Meng Liu, Weili Guan, Yaowei Wang, and Liqiang Nie. Hcqa-1.5@ ego4d egoschema challenge 2025. *arXiv preprint arXiv:2505.20644*, 2025. 1

[22] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023. 2, 3

[23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 2, 4

[24] Zeyun Zhong, Manuel Martin, Frederik Diederichs, and Juergen Beyerer. Querymamba: A mamba-based encoder-decoder architecture with a statistical verb-noun interaction module for video action forecasting@ ego4d long-term action anticipation challenge 2024. *arXiv preprint arXiv:2407.04184*, 2024. 2