

# GazeNLQ @ Ego4D Natural Language Queries Challenge 2025

Wei-Cheng Lin<sup>1\*</sup>, Chih-Ming Lien<sup>1\*</sup>, Chen Lo<sup>1</sup>, Chia-Hung Yeh<sup>1, 2</sup>

<sup>1</sup>National Taiwan Normal University <sup>2</sup>National Sun Yat-sen University

{linwc510, lien1119, chyeh, clo20}@ntnu.edu.tw

## Abstract

*This report presents our solution to the Ego4D Natural Language Queries (NLQ) Challenge at CVPR 2025. Egocentric video captures the scene from the wearer’s perspective, where gaze serves as a key non-verbal communication cue that reflects visual attention and offer insights into human intention and cognition. Motivated by this, we propose a novel approach, GazeNLQ, which leverages gaze to retrieve video segments that match given natural language queries. Specifically, we introduce a contrastive learning-based pre-training strategy for gaze estimation directly from video. The estimated gaze is used to augment video representations within proposed model, thereby enhancing localization accuracy. Experimental results show that GazeNLQ achieves R1@IoU0.3 and R1@IoU0.5 scores of 27.82 and 18.68, respectively. Our code is available at <https://github.com/stevenlin510/GazeNLQ>.*

## 1. Introduction

The goal of the Ego4D [6] Natural Language Queries (NLQ) challenge is to temporally localize the segment of egocentric video that corresponds to a given natural language query. Existing approaches generally fall into two categories: pretraining a foundation model to learn transferable representations suitable for various downstream tasks [2, 9, 11, 12], or developing specialized grounding model tailored to the NLQ task[4, 7, 10].

Pretraining foundation models on large-scale dataset has yielded impressive results on numerous downstream tasks. For instance, InternVideo [2] explores three types of feature extractors as backbone and fine-tunes them on the Ego4D training set. EgoVLP [9] constructs a large-scale egocentric training dataset and adapts video-text contrastive learning to explore representations. EgoVideo [11] enhances training data quality by filtering and selecting samples from multiple existing datasets, leveraging video-text contrastive learning for model training. Alternatively, task-specific models

such as GroundNLQ [7] adopt a two-stage pretraining strategy framework and introduces a multi-modal multi-scale grounding module that enables early fusion of video and text features. ObjectNLQ [4] enhances video representation by incorporating object-level information extracted through an object detection model.

Despite these advancements, most methods focus on visual and textual modalities, with limited exploration of auxiliary sensor data such as head motion or gaze signals in egocentric video understanding. Recently, EgoDistill [15] demonstrated the utility of head motion signals captured by the inertial measurement unit (IMU) of a head-mounted camera to facilitate efficient egocentric video understanding. Given that IMU data has been shown to improve classification accuracy in egocentric action recognition, it raises the question of whether the characteristics of egocentric video can similarly benefit egocentric video-language grounding. In egocentric video, gaze aligns closely with the camera wearer’s field of view, serving as a natural and informative cue for providing valuable information about visual attention, cognitive process, and underlying intentions. Understanding gaze behavior is essential for many applications, including cognitive science and psychology, human-robot interaction, and virtual and augmented reality. Recognizing the central role of gaze in revealing attention and intention in egocentric contexts, we aim to leverage this cue to advance understanding in egocentric video analysis. Therefore, We propose GazeNLQ, a novel framework that incorporates gaze to enhance natural language grounding in egocentric videos. We use a contrastive learning strategy to train the gaze estimator, which predicts gaze directly from video. The estimated gaze is then used to augment video features, leading to promising results on the NLQ task.

## 2. Method

This section presents GazeNLQ detailing the multi-modal feature representation and proposed model architecture.

### 2.1. Multi-modal Feature Representation

**Text and Video Representation.** Following GroundNLQ [7], We extract textual token representations using the

---

\*equal contributions.

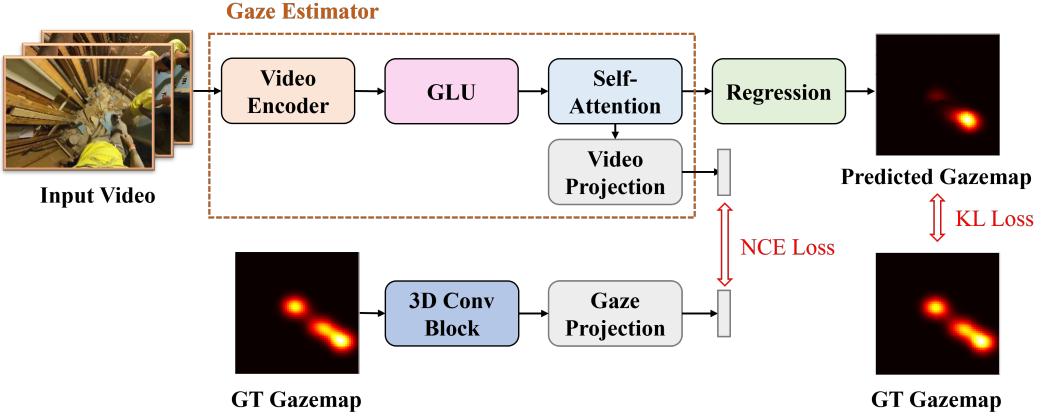


Figure 1. The proposed training framework for gaze estimator using contrastive learning.

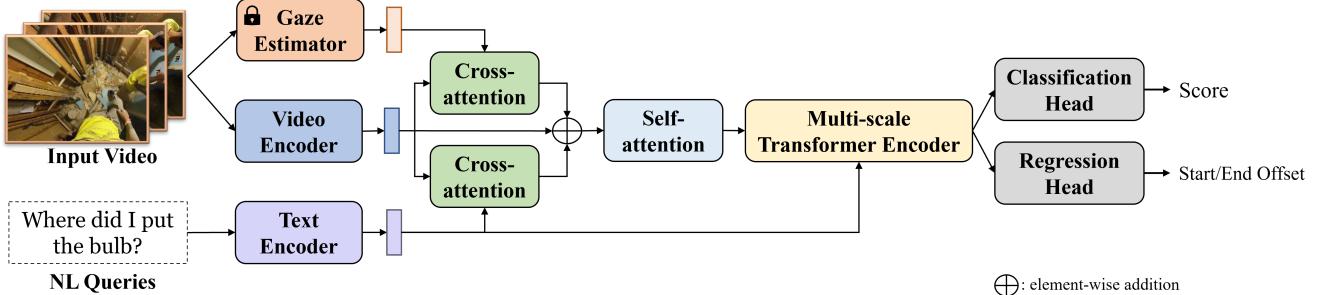


Figure 2. The proposed model for video temporal grounding.

CLIP [13] text encoder and construct video representation by concatenating features from InterVideo [2] and EgoVLP [9].

**Gaze Representation.** Since gaze annotations are not available for all video in the NLQ dataset, we train a gaze estimator using only the annotated data. The gaze estimator directly estimates the gaze from video. Our approach utilizes the dual-branch structure and contrastive learning for training, as illustrated in Fig. 1. The architecture includes a video encoder, 5 Gated Linear Unit (GLU) layers, an self-attention layer, and a video projection head. The video features are first extracted from Omnivore [5] video encoder, then processed through the GLU and an attention layer before being projected into an aligned gaze embedding space. For the gaze branch, we follow the preprocessing procedure provided by [8] to generate the gaze map for each frame from the raw gaze data. These gaze maps are processed through a 3D convolution block and a gaze projection head to produce corresponding gaze embeddings. To align video embeddings with gaze embeddings, we employ

the contrastive loss:

$$\mathcal{L}_{\text{NCE}} = \sum_i -\log \frac{\exp(v_i \cdot g_+ / \tau)}{\sum_j \exp(v_i \cdot g_j / \tau)}, \quad (1)$$

where  $v_i$  is video embedding,  $g_+$  is the positive gaze embedding,  $g_j$  is negative samples and  $\tau$  is a temperature hyperparameter. Additionally, a regression module predicts the gaze map  $G_{\text{pred}}$ , which is compared to the ground truth  $G_{\text{GT}}$  using the KL divergence loss:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(G_{\text{GT}} \| G_{\text{pred}}), \quad (2)$$

where the  $D_{\text{KL}}$  denote the KL divergence between  $G_{\text{GT}}$  and  $G_{\text{pred}}$ . The total loss  $\mathcal{L}_{\text{gaze}}$  is defined as:

$$\mathcal{L}_{\text{gaze}} = \mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{KL}}. \quad (3)$$

## 2.2. Model Architecture

The overall architecture of the proposed method is illustrated in Fig. 2. The framework extracts gaze, video, and textual embeddings using a gaze estimator, a video encoder, and a text encoder, respectively. Two cross-attention modules are then employed to align and integrate the gaze and video embeddings. The resulting

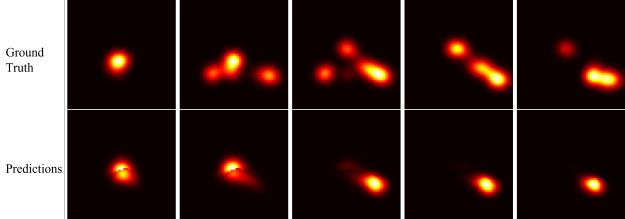


Figure 3. Visualization of gaze estimation. The top row shows the ground-truth gaze heatmaps, while the bottom row shows the predicted heatmaps.

embeddings are combined via element-wise addition and further refined using a self-attention. Next, we leverages the multi-scale transformer encoder architecture introduced in [7] to enhance the modeling of hierarchical and temporal dependencies. Final predictions are produced by the classification head, which scores each interval in the feature pyramid, and a regression head, which estimate the boundary distances from the interval, similar to the approach described in [7]. Model training employs the binary classification loss  $\mathcal{L}_{cls}$  and Intersection over Union (IoU) regression loss  $\mathcal{L}_{reg}$ . The total loss of video temporal grounding  $\mathcal{L}_{localization}$  is defined as:

$$\mathcal{L}_{localization} = \mathcal{L}_{cls} + \mathcal{L}_{reg}. \quad (4)$$

### 3. Experiment

#### 3.1. Implementation Details

**Gaze Estimator Training.** We begin by pretraining a gaze estimator using features extracted from the pretrained Omnivore model [5]. Omnivore processes video segments with a window size of 32 frames and a stride of 16 frames, yielding a single feature vector per temporal window. To align the ground-truth supervision with this temporal resolution, we average the corresponding gaze heatmaps over each 32-frame segment at a resolution of  $64 \times 64$ , as illustrated in Fig. 3. For contrastive learning, both video and gaze representations are projected into embedding size of 384, which aligns with the dimensionality of the subsequent finetuning stage. The gaze estimator is trained using a learning rate of  $1 \times 10^{-3}$  with a batch size of 16.

**Grounding Model Finetuning.** We adopt the GroundNLQ architecture [7] and initialize it with pretrained weights from a model trained on narration data [14] to establish a strong starting point. Following pretraining of the gaze estimator, we incorporate it into the GroundNLQ pipeline for end-to-end finetuning. Additionally, we investigate a model variant called GazeNLQ\* that employs negative gaze embedding, directing the model’s attention to regions outside the gaze area. During this phase, we freeze the gaze

Table 1. Performance comparison on NLQ *test* split.

Method	Test Private			
	R1@0.3	R1@0.5	R5@0.3	R5@0.5
GroundNLQ [7]	24.50	17.31	40.46	29.17
GroundNLQ <sup>†</sup> [7]	25.67	18.18	42.05	29.80
ObjectNLQ <sup>†</sup> [4]	27.02	19.28	43.66	30.87
GroundVQA [3]	26.67	17.63	39.94	27.70
EgoVideo [11]	25.07	17.31	40.88	29.67
EgoVideo <sup>†</sup> [11]	28.05	19.31	44.16	31.37
GazeNLQ	25.24	17.58	39.99	30.24
GazeNLQ*	25.45	17.48	40.23	29.87
<b>GazeNLQ<sup>†</sup></b>	<b>27.82</b>	<b>18.68</b>	<b>43.53</b>	<b>30.97</b>

<sup>†</sup>Ensemble results

Table 2. Performance Comparison on NLQ *val* split.

Method	Validation			
	R1@0.3	R1@0.5	R5@0.3	R5@0.5
GroundNLQ [7]	26.98	18.83	53.56	40.00
GroundVQA [3]	29.70	-	-	-
EgoVideo [11]	28.65	19.73	53.30	40.42
GazeNLQ	26.98	17.88	52.50	39.54
GazeNLQ*	27.22	18.08	52.61	39.63

estimator’s weights and train the combined model for ten epochs, incorporating a warm-up period of four epochs. For the finetuning process, we utilize a learning rate of  $2.5 \times 10^{-5}$  and a batch size of 8. All experiments are conducted using a single NVIDIA RTX 4090 GPU. During inference, we apply Soft-NMS [1] to merge overlapping moment predictions, optimizing the final localization outputs.

**Ensemble.** We combines predictions from GroundVQA [3], which followed the strategy by EgoVideo [11]. GroundVQA incorporates the question-answering data into the video grounding task by using the large language model.

#### 3.2. Performance Comparison

Tab. 1 reports the comparison results on the NLQ *test* split. Our ensemble approach achieves an R1@0.3 score of 27.82 and an R1@0.5 score of 18.68, demonstrating competitive performance. Notably, the variant incorporating negative gaze embeddings slightly outperforms the standard (positive) gaze formulation. This is an interesting finding that we plan to explore further in future work to understand its implications and potential for enhancing grounding performance.

Tab. 2 presents results on the NLQ *val* split without ensembling. While our method improves the R1@0.3 score compared to GroundNLQ, it results in a slight decrease in the R1@0.5 score. This indicates that our approach is more

Table 3. Ablation study of whether freeze the weights of Gaze Estimator on NLQ *val* split.

Weights	Validation			
	R1@0.3	R1@0.5	R5@0.3	R5@0.5
Unfreeze	26.98	18.10	51.49	38.60
Freeze	27.22	18.08	52.61	39.63

effective at retrieving relevant segments within a relaxed temporal threshold but less accurate under stricter alignment constraints.

### 3.3. Ablation Study

We conducted an ablation study to evaluate whether freezing the weights of the pretrained gaze estimation module affects grounding performance in Tab. 3. Interestingly, freezing the gaze model’s weights results in better performance compared to finetuning. Since the gaze estimator is trained on a relatively small dataset and may not generalize well when finetuned jointly with the grounding model, which is trained on a large-scale narration dataset.

### 3.4. Case Analysis

Fig. 4a shows successful examples in NLQ, where our model accurately locates the target of the text description. However, the failure examples are presented in Fig. 4b. In the top figure, the error arises from an imprecise temporal boundary—GazeNLQ captures only the first half of the ground truth event (“chop the vegetables”), indicating difficulty in handling long-duration actions. In the bottom figure, the model fails due to a misunderstanding of the object involved in the activity. However, we believe the ground truth annotation may not accurately reflect the subject, as the action of placing the bulb was performed by someone other than the camera wearer.

### 3.5. Discussion

This study represents an early stage in our research on egocentric video grounding with gaze, and remains a room for future improvement. First, there exists a feature discrepancy between the gaze training stage and the video grounding stage due to the use of different video feature extractors. The video features for gaze estimator are from Omnivore [5], while the grounding stage employs [2] video features. This mismatch may hinder the seamless transfer of learned representations, potentially impacting grounding performance.

Second, gaze information serves as a strong spatial prior during the gaze estimation phase, capturing precise locations of visual attention. However, in the grounding stage, the video features lack explicit spatial information. This non-spatial feature structure limits the ability to directly

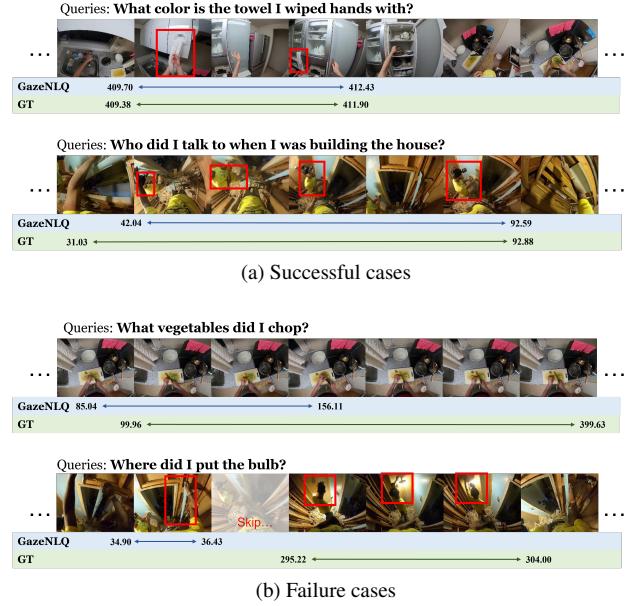


Figure 4. Four examples of GazeNLQ on NLQ *val* split: two successful cases (a) and two failure cases (b).

leverage the spatial cues provided by gaze tokens, necessitating additional processing or fusion strategies to align gaze with video features, which may introduce inefficiencies or loss of spatial detail.

Third, our approach relies on finetuning a pretrained GroundNLQ model rather than training from scratch using narration data. This finetuning strategy may constrain the model’s ability to fully adapt to the nuances of our dataset, particularly in integrating gaze information with text queries. Training from scratch with narration data could potentially yield a more robust model but was not pursued due to resource and time constraints at this stage.

## 4. Conclusion

This report presents GazeNLQ, our proposed method for the Ego4D natural language queries challenge at CVPR 2025. GazeNLQ employs a contrastive learning-based pretraining strategy for gaze estimation, which is a core component of the overall framework. The incorporation of estimated gaze into the video representation enhances the model’s ability to localize relevant content in response to natural language queries, as demonstrated by experimental results. These improvements highlight the promise of leveraging gaze to advance egocentric video understanding. Future work will focus on developing consistent feature extractors across stages, incorporating spatial information in grounding features, and exploring training from scratch to enhance model adaptability.

## References

- [1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms — improving object detection with one line of code. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5562–5570, 2017. 3
- [2] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint:2211.09529*, 2022. 1, 2, 4
- [3] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, pages 12934–12943, 2024. 3
- [4] 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
- [5] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022. 2, 3, 4
- [6] 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 *CVPR*, pages 18995–19012, 2022. 1
- [7] Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. Groundnlq@ ego4d natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023. 1, 3
- [8] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation and beyond. *IJCV*, pages 1–18, 2023. 2
- [9] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wen-zhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, pages 7575–7586, 2022. 1, 2
- [10] Naiyuan Liu, Xiaohan Wang, Xiaobo Li, Yi Yang, and Yuet-ing Zhuang. Reler@ zju-alibaba submission to the ego4d natural language queries challenge 2022. *arXiv preprint arXiv:2207.00383*, 2022. 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. 1, 3
- [12] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*, pages 5285–5297, 2023. 1
- [13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning*, pages 8748–8763, 2021. 2
- [14] Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *Computer Vision and Pattern Recognition (CVPR), 2023 IEEE Conference on*. IEEE, 2023. 3
- [15] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. In *NeurIPS*, 2023. 1