Object-Shot Enhanced Grounding Network for Egocentric Video

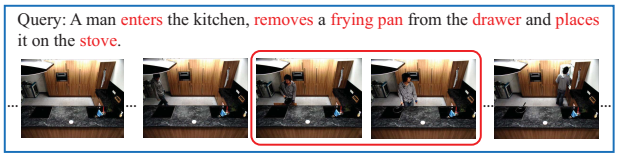
Yisen Feng[1] Haoyu Zhang[1 2] Meng Liu[3*] Weili Guan[1] Liqiang Nie[1*]
[1]Harbin Institute of Technology (Shenzhen) [2]Pengcheng Laboratory
[3]Shandong Jianzhu University
{yisenfeng.hit, zhang.hy.2019, mengliu.sdu, honeyguan, nieliqiang}@gmail.com

Abstract

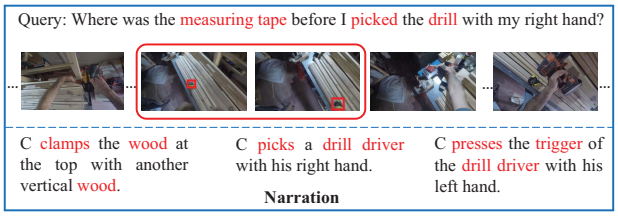
Egocentric video grounding is a crucial task for embodied intelligence applications, distinct from exocentric video moment localization. Existing methods primarily focus on the distributional differences between egocentric and exocentric videos but often neglect key characteristics of egocentric videos and the fine-grained information emphasized by question-type queries. To address these limitations, we propose OSGNet, an Object-Shot enhanced Grounding Network for egocentric video. Specifically, we extract object information from videos to enrich video representation, particularly for objects highlighted in the textual query but not directly captured in the video features. Additionally, we analyze the frequent shot movements inherent to egocentric videos, leveraging these features to extract the wearer's attention information, which enhances the model's ability to perform modality alignment. Experiments conducted on three datasets demonstrate that OSGNet achieves state-of-the-art performance, validating the effectiveness of our approach. Our code can be found at https://github.com/Yisen-Feng/OSGNet.

1. Introduction

With advancements in wearable camera technology, Ego4D [8] introduces the Natural Language Query (NLQ) task for egocentric video grounding. NLQ aims to identify the specific video moment that answers a question-type query within an untrimmed egocentric video, as shown in Figure 1(b). The dynamic and complex camera perspectives in egocentric videos [55] make video comprehension [20, 21] significantly more challenging compared to the fixed viewpoints of exocentric video grounding, as shown in Figure 1(a). Moreover, NLQ queries often focus on fine-grained details of background objects (e.g., "measuring tape" in Figure 1(b)), unlike exocentric video grounding tasks that emphasize character actions (e.g., "enters" and "removes" in Figure 1(a)), adding complexity to mining background information from video. Consequently, despite progress in exocentric video grounding, existing methods struggle with the unique challenges of NLQ. Nonetheless, NLQ enables innovative applications [53, 54], such as smart assistant systems and memory retrieval modules for autonomous robots, underscoring the urgent need to address these challenges.



(a) Exocentric video moment localization task from TACoS [38]



(b) Natural language query task and narrations from Ego4D [8]

Figure 1. Illustration of exocentric and egocentric video grounding, accompanied by annotated narrations for the egocentric video. Key verbs and nouns are highlighted in red.

Though existing methods [2, 22, 31, 34, 37] for NLQ have made significant advances in video-text pretraining, **they fail to address the issue of video features lacking the fine-grained object information needed to answer detailed queries**. As shown in Figure 1(b), egocentric video-text pretraining datasets generally provide narrations focusing on character actions involving objects. These datasets, combined with clip-narration contrastive learning, lead current video backbones to overlook fine-grained background object details in the extracted features. In contrast, NLQ emphasizes enhancing people's memory experience, often involving queries centered on background objects that are not part of active interactions, such as the "measuring tape" placed casually on the table, as illustrated in Figure 1(b).

*represents corresponding author.

Moreover, existing methods **fail to fully leverage the rich attention information embedded in egocentric videos.** Egocentric videos feature frequent camera movements, as the camera is typically worn on the head, moving with the wearer's actions. This movement implicitly encodes head motion information, signaling shifts in the wearer's attention and focus, an aspect that is crucial for NLQ but often overlooked by current methods. As shown in Figure 1(b), initially, the individual fixes the wood, then walks to pick up a drill driver, and finally uses the drill driver to drill holes in the wood. These behaviors are independent yet interconnected. Capturing these attention shift points is vital for clarifying video structure and improving video understanding.

Building upon these findings, we propose a novel **O**bject-**S**hot enhanced **G**rounding **Net**work (OSGNet) for egocentric videos, addressing the unique challenges posed by the NLQ task. **To address the challenge of insufficient fine-grained object information**, we introduce an object extraction process (see Figure 2(a)), which captures detailed object-level information. These features are then integrated using a multi-modal fusion mechanism that employs parallel cross-attention within the main branch. This strategy ensures that both visual and textual cues are effectively leveraged to enhance localization accuracy. **To exploit the dynamics of the wearer's attention**, we extract head-turning data from egocentric video, which provides insights into shifts in the wearer's focus. Based on these shifts, we segment the video into semantically distinct shots and apply contrastive learning to strengthen the model's ability to align these shots with the query. This approach enhances the model's capacity to capture and utilize attention-driven context for improved video grounding. We evaluate our model on three benchmark datasets: Ego4D-NLQ, Ego4D-Goal-Step [42], and TACoS [38]. On Ego4D-NLQ, OSGNet outperforms GroundVQA [5], achieving a 2.15% improvement in Rank@1 at IoU=0.5. On Ego4D-Goal-Step, OSGNet surpasses BayesianVSLNet [33] with a 3.65% increase in Rank@1 at IoU=0.3. On TACoS, OSGNet achieves a 3.32% improvement in Rank@1 at IoU=0.5 over SnAG [27]. These results underscore the effectiveness of our model in enhancing video grounding.

Contributions. 1) We integrate fine-grained object information into the egocentric video grounding task, thereby improving the accuracy of background object-related query localization. 2) We introduce a wearer movement-aware shot branch that leverages shot-level contrastive learning, collaborating with the main branch to further enhance egocentric video grounding performance. And 3) our OSGNet outperforms current state-of-the-art approaches across three benchmark datasets, setting a new standard for egocentric video grounding.

2. Related Work

2.1. Video Moment Localization

Video moment localization aims to determine the start and end timestamps of a video moment in response to a natural language query. Existing approaches [23, 24, 35, 43, 48, 50] can be broadly classified into proposal-based and proposal-free methods, which primarily differ in their strategies for generating candidate moments. Proposal-based methods generally follow a two-stage process involving the generation of candidate moment features and subsequently matching them with textual queries. Sliding windows have been extensively employed for generating candidates [11, 15, 16], with approaches like [7] incorporating regression to refine window boundaries. Subsequent work [4, 47, 49] enhances candidate quality by integrating text information at an early stage, while others [56, 57] increased the density of candidate moments through exhaustive enumeration.

On the other hand, proposal-free methods aim to directly generate features for individual moments or the entire video, leading to fewer features and increased computational efficiency. Some approaches [18, 28, 61] extract features for the entire video sequence, whereas others [6, 52] focus on predicting the probability of a particular moment being part of the desired query segment. The emergence of new datasets involving long video sequences [8, 41] has posed significant challenges for existing methods, which often struggle with performance degradation on longer videos. To address these challenges, [13] proposes a query-guided window selection strategy under the assumption that short windows are sufficient, while [27] presents a single-stage approach that aligns multiple video moments without making similar assumptions.

As discussed in Section 1, significant differences in the distribution of videos and queries between the exocentric video moment localization task and the egocentric video NLQ task hinder the direct applicability of existing frameworks. To overcome these challenges, our model incorporates fine-grained object-level information and employs shot-level contrastive learning, thereby enhancing the accuracy of egocentric video grounding.

2.2. Natural Language Query

The NLQ task aims to accurately localize video moments that answer natural language questions in egocentric videos. Existing approaches primarily pursue two directions: 1) Video-Text Pretraining [17, 58]. Due to the differences in feature distributions between egocentric and exocentric videos, several studies [2, 22, 31, 34] have fine-tuned video backbones using egocentric videos, improving the model's robustness in tasks involving egocentric video comprehension. 2) Data Enhancement. Ramakrishnan et al. [37] constructed a large narrative dataset of 940K video-narration

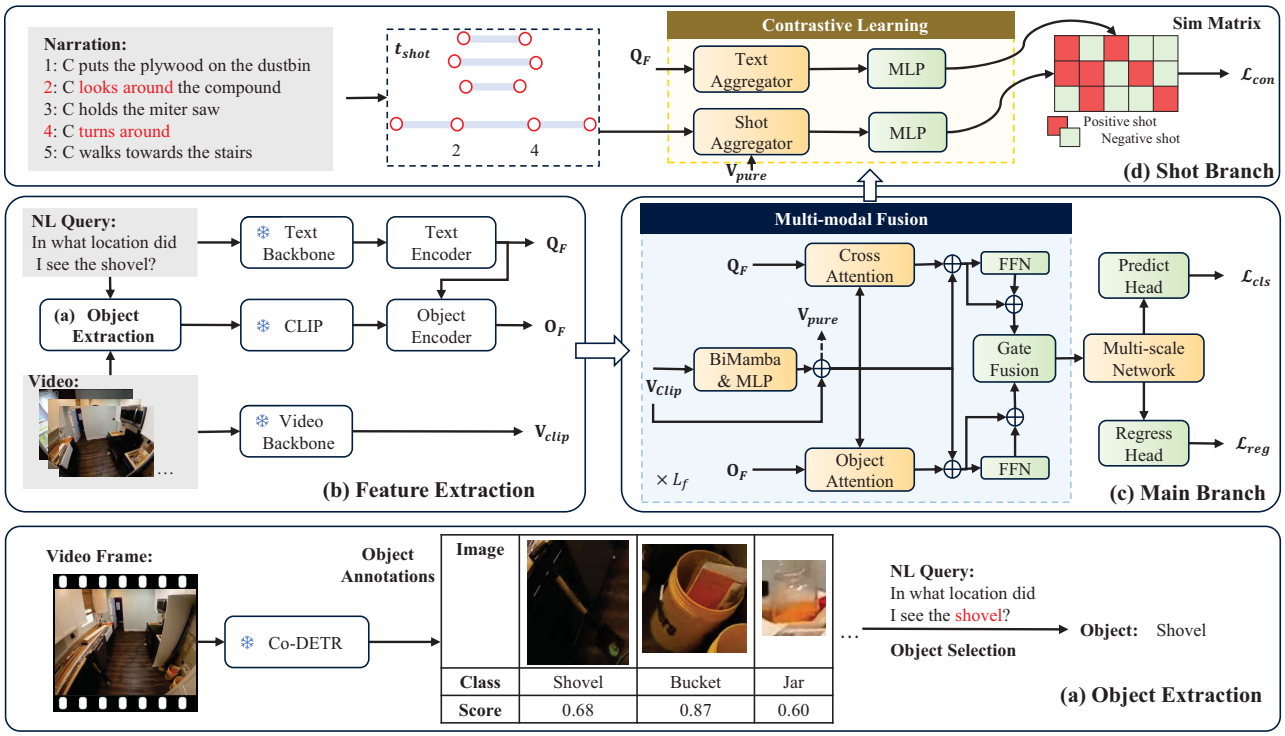


Figure 2. The framework of our OSGNet, which consists of four key components: (a) Object Extraction, which captures fine-grained object features; (b) Feature Extraction, where visual and textual cues are processed; (c) Main Branch, responsible for primary grounding tasks; and (d) Shot Branch, which leverages wearer movement dynamics and shot-level contrastive learning to improve localization accuracy.

pairs to pre-training models, followed by fine-tuning on Ego4D-NLQ, yielding substantial improvements.

However, prior methods [12, 13, 25] often treat NLQ as a general long-video localization problem [30], overlooking the need for fine-grained object information that is crucial for NLQ. Our approach addresses this gap by extracting query-relevant fine-grained object features and integrating them into the training process, thus improving localization accuracy. Additionally, existing methods tend to overlook the frequent shot transitions inherent in egocentric videos. To mitigate this, we segment egocentric videos into distinct shots and employ contrastive learning to better align the two modalities, ultimately enhancing localization performance.

3. Method

3.1. Overview

Given a video consisting of N frames, denoted as $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$, and a natural language query comprising L words, denoted as $\mathcal{Q} = \{w_1, w_2, \cdots, w_L\}$, our grounding model aims to localize the specific moment within an untrimmed egocentric video that best answers the query, represented by the timestamp $[t_s, t_e]$.

As outlined in Section 1, existing video feature representations are often insufficient for NLQ, primarily due to the absence of fine-grained object-level information. To address this limitation, we propose a novel framework that integrates object-level details into the feature representation process, significantly enhancing the localization accuracy. Our approach involves extracting features from video, text, and objects, followed by a comprehensive feature encoding process, as described in Section 3.2. The model architecture comprises two primary components: the main branch and the shot branch. The main branch (Section 3.3) combines video, text, and object features to improve the accuracy of video localization. In contrast, the shot branch (Section 3.4) focuses exclusively on video features to generate shot representations and applies contrastive learning with text inputs, enhancing the model's ability to perform modality alignment. Finally, Section 3.5 details the training and inference process of our framework.

3.2. Feature Extraction

Video Feature. We divide the video into non-overlapping clips, denoted as $\mathcal{C} = \{\mathcal{C}_i | \mathcal{C}_i = \{v_{(i-1)\times s+1}, \cdots, v_{(i-1)\times s+s}\}\}_{i=1}^{T}$, where s denotes the size of the sliding window and T denotes the number of clips in the video. Pretrained video backbones are used to extract clip-level video features, which are then projected into a feature space, yielding the video representation as

$\mathbf{V}_{clip} \in \mathbb{R}^{T\times D}$, where D denotes the dimensionality of the feature vectors.

Query Feature. Query features are extracted using a text backbone, resulting in word-level features $\mathbf{Q}_{word} \in \mathbb{R}^{L\times D_T}$, where D_T is the dimensionality of the word embeddings. To capture the relationships among words in the query, we employ a multi-layer transformer, termed the text encoder, outputting the query representation $\mathbf{Q}_F \in \mathbb{R}^{L\times D}$.

Object Feature. To effectively capture object information, we utilize the object detector Co-DETR [63], pretrained on the LVIS dataset [9], which is capable of identifying a wide range of object categories within video frames. As illustrated in Figure 2(a), we select object categories that are related to the nouns in the query and whose confidence scores exceed the threshold θ. For object representation, we encode the object categories as textual features, allowing seamless integration with the textual query. To achieve this, we employ CLIP (ViT-B/32) [36] to encode textual features of the detected object categories, resulting in $\mathbf{O}_{clip} \in \mathbb{R}^{T\times N_o\times D_o}$, where N_o is the maximum number of objects detected in a single frame[1] and D_o denotes the dimensionality of the object features.

Different objects in the same query play distinct roles in the video grounding task. For example, in the query "How many drill bit did I remove from the drill before I moved the yellow carton?", the beginning of the target moment should be related to the "drill bit" instead of the "drill", and the end of the target moment should before the "carton" appears. Therefore, we design an object encoder to refine the query-relevant object features. The object encoder is a multi-layer transformer, processing the object features in the context of the input query. To avoid confusion between object features within the same frame, we replace the conventional self-attention mechanism of transformers with a cross-attention mechanism, where object features serve as queries and query features act as keys and values. This architecture effectively encodes object-related information concerning the query, denoted as $\mathbf{O}_F \in \mathbb{R}^{T\times N_o\times D}$.

3.3. Main Branch

Our main branch consists of three core components: a multimodal fusion module to integrate features from different modalities, a multi-scale network to generate candidate moment representations at multiple temporal scales, and task-specific heads to predict the temporal offsets and confidence scores for the localized video moments.

Multi-modal Fusion. To effectively integrate fine-grained object features, we design a multi-modal fusion module that combines video, text, and object features. This module consists of multiple stacked layers, each including a bidirectional Mamba (BiMamba) block, followed by a multi-layer perceptron (MLP), cross-attention (CA) block, and Feedforward Neural Network (FFN) for query and object processing, culminating in a fusion block.

To be specific, we enhance video features using a Bi-Mamba layer [62], instead of traditional local self-attention, to better capture long-range dependencies within video data. The updated video features are computed as follows:

$$\hat{\mathbf{V}}^{(i)} = \mathbf{V}_f^{(i)} + MLP(BiMamba(\mathbf{V}_f^{(i)})), \tag{1}$$

where $\mathbf{V}_f^{(0)} = \mathbf{V}_{clip}$ and $\mathbf{V}_f^{(i)} \in \mathbb{R}^{T\times D}$ is the output of the i-th layer. Next, we apply a cross-attention block and FFN to aggregate video and query information:

$$\begin{cases} \mathbf{V}_Q^{(i)} = \hat{\mathbf{V}}^{(i)} + CA(\hat{\mathbf{V}}^{(i)}, \mathbf{Q}_F, \mathbf{Q}_F), \\ \hat{\mathbf{V}}_Q^{(i)} = \mathbf{V}_Q^{(i)} + FFN(\mathbf{V}_Q^{(i)}). \end{cases} \tag{2}$$

Similarly, we aggregate video feature $\hat{\mathbf{V}}^{(i)}$ and object features $\mathbf{O}_F$ using a parallel cross-attention block for object features. The final output $\hat{\mathbf{V}}_O^{(i)}$ is calculated as follows:

$$\begin{cases} \mathbf{V}_O^{(i)} = \hat{\mathbf{V}}^{(i)} + CA(\hat{\mathbf{V}}^{(i)}, \mathbf{O}_F, \mathbf{O}_F), \\ \hat{\mathbf{V}}_O^{(i)} = \mathbf{V}_O^{(i)} + FFN(\mathbf{V}_O^{(i)}). \end{cases} \tag{3}$$

Finally, these features are combined using a gating mechanism:

$$\begin{cases} \mathbf{A} = \sigma(MLP(\hat{\mathbf{V}}_Q^{(i)} || \hat{\mathbf{V}}_O^{(i)})), \\ \mathbf{V}_f^{(i+1)} = \mathbf{A}\cdot\hat{\mathbf{V}}_Q^{(i)} + (\mathbf{1}-\mathbf{A})\cdot\hat{\mathbf{V}}_O^{(i)}. \end{cases} \tag{4}$$

where σ is sigmoid fuction, $||$ is vector concatenation.

Multi-scale Network. The multi-scale network generates a feature pyramid to facilitate the grounding of video moments at various temporal scales. This network is a multi-layer transformer, with each layer including a 1D depth-wise convolution before the self-attention and FFN modules to enable sequence downsampling, as described in [12]. By feeding the output of the multi-modal fusion module, we can obtain the multi-scale candidate moment representations, denoted as $[\mathbf{V}_m^{(0)}, \mathbf{V}_m^{(1)}, \cdots, \mathbf{V}_m^{(L_s)}]$, where $\mathbf{V}_m^{(0)} = \mathbf{V}_f^{(L_f)}$. Here L_f and L_s are the number of layers in the multi-modal fusion and multi-scale network, respectively. The representation $\mathbf{V}_m^{(j)} \in \mathbb{R}^{T/2^j\times D}$ represents the video features at the j-th scale, with progressively reduced sequence lengths due to downsampling at each layer.

Task Heads. The task heads are responsible for decoding the multi-scale feature pyramid into final predictions for video grounding. Specifically, we use two heads: a classification head and a regression head following previous work [12, 27, 51]. Each task head is implemented using two layers of a 1D convolutional network. The classification head predicts the confidence score for each candidate

[1] If insufficient object representations are present, zero-padding is applied.

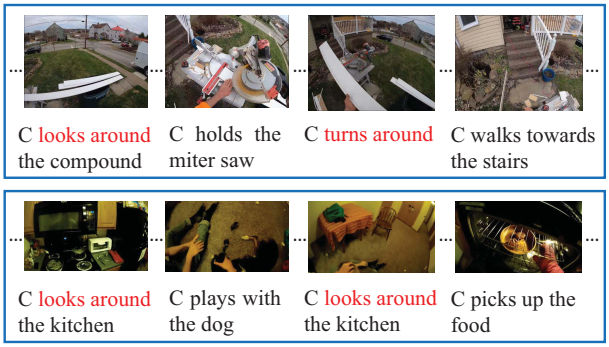


Figure 3. Illustration of captions generated by LAVILA [59] describing camera movements.

moment, while the regression head predicts the offsets of the moment boundaries relative to the anchor point.

For each feature $\mathbf{v}_k^{(j)} = \mathbf{V}_m^{(j)}[k] \in \mathbb{R}^D$, $\mathbf{v}_k^{(j)}$ represents the feature for the k-th anchor point in the j-th layer, which has $T/2^j$ anchor points with a stride of 2^j. The corresponding timestamp is computed as $t_k^{(j)} = 2^j * k$. The classification head then predicts the confidence $c_k^{(j)}$ for $\mathbf{v}_k^{(j)}$, and the regression head predicts normalized offsets $(s_k^{(j)}, e_k^{(j)})$ for $\mathbf{v}_k^{(j)}$. The predicted video moment for this anchor point is defined by the start and end boundaries: $(t_k^{(j)} - s_k^{(j)} \cdot 2^j, t_k^{(j)} + e_k^{(j)} \cdot 2^j)$.

Localization Loss. The moment localization loss, $\mathcal{L}_{ML}$, is designed to enhance model accuracy in video moment localization and is defined as follows:

$$\mathcal{L}_{ML} = \mathcal{L}_{cls} + \mathcal{L}_{reg}. \tag{5}$$

The first component $\mathcal{L}_{cls}$ is focal loss [39], which is used for classification to assess whether a proposal aligns with the query. The second component $\mathcal{L}_{reg}$ is Distance-IoU loss [60], which refines the boundaries of the localized moment by calculating the Distance-IoU between the predicted and ground truth moment boundaries. Notably, $\mathcal{L}_{reg}$ is calculated only on positive candidates.

3.4. Shot Branch

The shot branch is specifically designed to harness the attention information embedded in egocentric videos, which reflects the wearer's focus during video capture. This is achieved by segmenting the video into semantically distinct shots, with head movement serving as a key indicator for defining these boundaries. Contrastive learning is then employed to enhance the model's ability to align cross-modal features. This shot-level understanding deepens the model's comprehension of the video structure, significantly improving video grounding performance.

Shot Segmentation. As illustrated in Figure 3, the wearer's head movements, captured by the pretrained captioning model LAVILA [59], are often described in video captions with expressions such as *"looks around"* or *"turns around"*. These cues are used to define shot boundaries, denoted by timestamps $t_{shot} = \{(t_i, t_{i+1})\}_{i=1}^{N_S}$, where $t_1 = 0$ represents the start of the video, t_{N_S+1} represents the video's end, and $\{t_i\}_{i=2}^{N_S}$ corresponds to timestamps that capture head movement activities. Here N_S is the total number of shots.

Feature Aggregation. To perform contrastive learning effectively, it is crucial to extract contextual video features that represent the content of each shot independently of query and object information. This is achieved by employing multiple stacked layers, each consisting of a BiMamba block integrated with an MLP. These layers share parameters with the multi-modal fusion module, enabling us to extract self-interaction features from the video, denoted as $\mathbf{V}_{pure} \in \mathbb{R}^{T \times D}$.

To aggregate shot and query features for contrastive learning, we employ a transformer encoder with an architecture similar to the Q-Former [19]. In this setup, learnable embeddings serve as queries, while the shot and query features function as keys and values. Following this process, we apply a 1D convolution to obtain the final shot representations, denoted as $\mathbf{V}_{shot} \in \mathbb{R}^{N_S \times D}$, and text representation, denoted as $\mathbf{q}_{sent} \in \mathbb{R}^D$.

Contrastive Learning. To enhance contrastive learning, we gather all shots and their associated queries within a mini-batch, denoted as $\mathbf{Q}_{batch} = \{\mathbf{q}_{sent}^i\}_{i=1}^m \in \mathbb{R}^{m \times D}$ and $\mathbf{V}_{batch} = \{\mathbf{v}_{shot}^i\}_{i=1}^n \in \mathbb{R}^{n \times D}$, respectively, where m and n represent the number of queries and shots in the mini-batch. Next, the text and video features are projected into a joint semantic space using separate MLPs. These projected features are used for contrastive learning to enable the model to learn the alignment between video shots and natural language queries, thereby improving the accuracy of video moment localization.

To align shot features with textual queries, we apply contrastive learning using the InfoNCE loss [45], which encourages positive shot-query pairs to be closer in the embedding space while pushing apart the negative pairs. The InfoNCE loss is defined as follows:

$$\mathcal{L}_{con} = \frac{1}{m}\sum_{i=1}^{m} \frac{\sum_{(\mathbf{q}_{shot}^i, \mathbf{v}_{shot}^s) \in \mathcal{P}} exp(Sim(\mathbf{q}_{sent}^i, \mathbf{v}_{shot}^s)/\tau)}{\sum_{j=1}^{n} exp(Sim(\mathbf{q}_{sent}^i, \mathbf{v}_{shot}^j)/\tau)} + \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{(\mathbf{q}_{sent}^l, \mathbf{v}_{shot}^i) \in \mathcal{P}} exp(Sim(\mathbf{v}_{shot}^i, \mathbf{q}_{sent}^l)/\tau)}{\sum_{j=1}^{m} exp(Sim(\mathbf{v}_{shot}^i, \mathbf{q}_{sent}^j)/\tau)}, \tag{6}$$

where Sim represents the cosine similarity, $\mathcal{P}$ is the set of positive query-shot pairs where the query's ground truth intersects with the corresponding shot, and τ is the temperature coefficient.

Table 1. Performance comparison on Ego4D-NLQ. C denotes the CLIP video feature. * indicates results reproduced using the released code. † refers to results that utilize the NaQ [37] pretraining strategy.

	Method	Published	Feature	Validation				Test			
				R@1		R@5		R@1		R@5	
				0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
Ego4D-NLQ v1	InternVideo [2]	CVPRW 2022	E+I	15.64	10.17	24.78	18.30	16.45	10.06	22.95	16.10
	CONE [13]	ACL 2023	E	14.15	8.18	30.33	18.02	13.46	7.84	23.68	14.16
	SnAG* [27]	CVPR 2024	E	-	-	-	-	14.34	10.29	27.72	19.91
	NaQ† [37]	CVPR 2023	E+C	19.31	11.59	23.62	15.75	18.46	10.74	21.50	13.74
	RGNet* [10]	ECCV 2024	E	16.86	10.53	34.43	21.84	16.68	10.61	27.95	17.83
	RGNet*† [10]	ECCV 2024	E	18.66	11.72	36.37	22.43	18.21	11.69	29.80	19.06
	OSGNet		E	16.13	11.28	36.78	25.63	15.23	10.71	28.27	20.32
	OSGNet†		E	**21.97**	**15.20**	**44.61**	**32.96**	**22.13**	**15.46**	**36.54**	**25.70**
Ego4D-NLQ v2	NaQ*† [37]	CVPR 2023	E+I+C	24.10	15.03	29.90	20.85	21.70	13.64	25.12	16.33
	ASL† [40]	CVPRW 2023	E+I	22.62	15.64	46.86	32.16	24.13	15.46	34.37	23.18
	GroundNLQ† [12]	CVPRW 2023	E+I	26.98	18.83	53.56	40.00	24.50	17.31	40.46	29.17
	EgoEnv† [29]	NeurIPS 2024	E+I+C+EgoEnv	25.37	15.33	-	-	23.28	14.36	27.25	17.58
	EgoVideo† [31]	CVPRW 2024	EgoVideo	28.65	19.73	53.30	40.42	25.07	17.31	40.88	29.67
	GroundVQA*† [5]	CVPR 2024	E+I	29.68	20.23	52.17	37.83	26.67	17.63	39.94	27.70
	OSGNet†		E+I	**31.63**	**22.03**	**57.91**	**45.19**	**27.60**	**19.78**	**43.46**	**32.77**

3.5. Training and Inference

Training. Our training is divided into pretraining and fine-tuning phases. In pretraining, we use the dataset from [37], only the main branch without object cross-attention is trained with the moment localization loss. In fine-tuning, the final loss function combines both the moment localization loss and the contrastive learning loss, as:

$$\mathcal{L} = \frac{1}{C}(\mathcal{L}_{ML} + \lambda \mathcal{L}_{con}), \tag{7}$$

where C is the number of positive candidate moments (those that match the query) using momentum update, and λ is a balancing hyperparameter.

Inference. During inference, the main branch generates the start and end boundaries, along with the predicted confidence for all anchors in the pyramid. To eliminate duplicates, we apply SoftNMS [1] for deduplication.

4. Experiment

4.1. Datasets

Since our contributions primarily target the NLQ task, the Ego4D-NLQ dataset is our main choice. To further evaluate the versatility of our method, we also incorporated the Ego4D-Goal-Step dataset, another egocentric video grounding benchmark. Additionally, we used the TACoS dataset to enable more comprehensive comparisons.

Ego4D-NLQ. Ego4D-NLQ consists of two versions. Ego4D-NLQ v1 contains 1,659 videos with 11,279, 3,874, and 4,004 video-query pairs for training, validation, and testing. Due to noisy data, Ego4D-NLQ v2 was released with 2,018 videos and 13,847, 4,552, and 4,004 video-query pairs for training, validation, and testing. Both versions share the same test set.

Ego4D-Goal-Step. Ego4D-Goal-Step provides annotations for step grounding, which involves locating a video clip based on a step description. It contains 851 videos with 31,566, 7,696, and 5,540 video-query pairs for training, validation, and testing.

TACoS. TACoS is an exocentric video moment localization dataset with 127 cooking videos. The training, validation, and test sets consist of 10,146, 4,589, and 4,083 video-query pairs. For consistency with existing works, we followed the setup in [57], using 9,790, 4,436, and 4,001 video-query pairs for training, validation, and testing.

4.2. Experimental Settings

Implementation Details. We segmented videos into clips using a sliding window of 16 frames for both window size and stride. Features are extracted using EgoVLP [22] (E) and InternVideo [2] (I), then concatenated as in [12]. Text features are extracted using CLIP (ViT-B/32) [36]. For TACoS, we used C3D [44] for video features and Glove [32] for text features to ensure a fair comparison [27]. The object's confidence score threshold θ is 0.6. In pretraining, we used a batch size of 16, a learning rate of $8e^{-4}$, 4 warmup epochs, and 10 total epochs. In fine-tuning, warmup and total epochs are set to 4 and 10, respectively. Components that were not pretrained are initialized with weights from other pretrained structures. For further details, please refer to our released code.

Baselines. On Ego4D-NLQ v1, we compared our model with several strong baselines, including methods utilizing the NaQ pretraining strategy (NaQ [37] and RGNet [10])

Table 2. Performance comparison under R@1 on Ego4D-Goal-Step. * indicates results that leverage the order prior of the dataset.

Method	Validation		Test	
	0.3	0.5	0.3	0.5
EgoVideo [31]	28.02	23.66	32.99	25.92
BayesianVSLNet* [33]	18.15	8.97	35.18	20.48
OSGNet	29.61	24.94	32.77	25.50
OSGNet*	**42.61**	**35.38**	**38.83**	**30.16**

Table 3. Performance comparison on TACoS. * indicates results that utilize the E+I video feature.

Method	Published	R@1		R@5	
		0.3	0.5	0.3	0.5
VSLNet [52]	ACL 2020	29.61	24.27	-	-
2D-TAN [57]	AAAI 2020	37.29	25.32	57.81	45.04
Tri-MRF [46]	TMM 2024	52.44	41.49	76.01	63.46
DPHANet [3]	TMM 2024	47.01	34.12	-	-
MESM [26]	AAAI 2024	52.69	39.52	-	-
MRNet [14]	ACMMM 2024	55.41	38.54	77.18	64.78
SnAG [27]	CVPR 2024	56.44	44.86	81.15	70.66
OSGNet		**57.57**	**48.18**	**82.02**	**72.05**
OSGNet*		66.43	55.77	87.16	79.45

and methods without the NaQ pretraining strategy (InternVideo [2], CONE [13], SnAG [27], and RGNet [10]), using the same validation set for a fair comparison. Note that RGNet [10] and SnAG [27] employ different validation settings and have not reported performance on the Ego4D-NLQ v1 test set, we evaluated their performance in our unified setting. Further details see the *Supplementary Material*. On Ego4D-NLQ v2, we compared with strong baselines (NaQ [37], ASL [40], GroundNLQ [12], EgoEnv [29], EgoVideo [31], and GroundVQA [5]) utilizing the NaQ pretraining strategy. As GroundVQA [5] and NaQ [37] did not report performance on the Ego4D-NLQ v2 validation set, we used their published checkpoints for testing.

On Ego4D-Goal-Step, we compared with baselines, including BayesianVSLNet [33], which leverages the order prior, and EgoVideo [31], which did not use prior information. On TACoS, we compared with the classic baselines (VSLNet [52] and 2D-TAN [57]) and current baselines (Tri-MRF [46], DPHANet [3], MESM [26], MRNet [14], and SnAG [27]).

Evaluation Metric. Following previous work [13, 27], we utilized the metric Rank@m, IoU=n (R@m, n), which evaluates the percentage of queries that contain at least one correct moment among the top m retrieved predictions. A correct moment is defined as having an IoU greater than n with the ground truth.

4.3. Performance Comparison

On Ego4D-NLQ. Table 1 presents the performance comparison on the validation and the test sets of Ego4D-NLQ. On Ego4D-NLQ v2, our model outperforms all compared methods across all metrics. Specifically, compared to the strong baseline GroundVQA, our model achieves an absolute improvement of over 1.5% in R@1, 0.5 on both the validation and test sets. On Ego4D-NLQ v1, our model surpasses the baseline RGNet [10] in most metrics, including the challenging R@1, 0.5. Our carefully designed architecture results in a nearly 4.0% absolute improvement in R@1, 0.5 after pretraining. In contrast, RGNet [10] shows considerably less benefit from pretraining, likely due to its dual-branch fusion structure.

On Ego4D-Goal-Step. Table 2 presents the performance of our model compared to existing methods on the validation and test sets of Ego4D-Goal-Step. Our OSGNet outperforms all competing methods across all metrics on the validation set. Without any prior information, our model achieves an absolute improvement of approximately 1.59% in R@1, 0.3 on the validation set, compared to the strong baseline EgoVideo. When incorporating the order prior of the steps, as done by BayesianVSLNet [33], our model further improves by around 3.65% in R@1, 0.3 on the test set, surpassing BayesianVSLNet [33].

On TACoS. Table 3 presents the performance comparison of our model with existing methods on the test set of TACoS. Our OSGNet outperforms all competing methods across all metrics on the test set when using the C3D feature. Compared to the strong baseline SnAG, our model achieves an absolute improvement of approximately 3.32% in the challenging metric R@1, 0.5. Additionally, when concatenating the InternVideo [2] and EgoVLP [22] video features, along with the pretraining strategy, our model achieves 66.43% IoU=0.3 and 55.77% IoU=0.5 at R@1.

Summary. Comparing results across different datasets, we observed that our model performs lowest on Ego4D-NLQ, followed by Ego4D-Goal-Step, and highest on TACoS under the same settings. We attributed the relatively lower performance on egocentric video grounding datasets, compared to exocentric ones, to the inherent complexity of understanding egocentric video content. Furthermore, NLQ presents additional challenges due to its emphasis on grounding fine-grained objects within the background. Despite this, our model excels in R@1, 0.5 on both Ego4D-NLQ and TACoS, demonstrating the effectiveness of our proposed approach.

4.4. Ablation Study

On Model Structure. We conducted ablation studies on the validation set of Ego4D-NLQ v2 to assess the contributions of key components in our model. Specifically, we examined the impact of removing fine-grained object features by excluding the object branch and fusion block from the multi-modal encoder, as well as the effect of removing shot-level contrastive learning by eliminating the $\mathcal{L}_{con}$ term from the total loss function $\mathcal{L}$. Table 4 presents the results

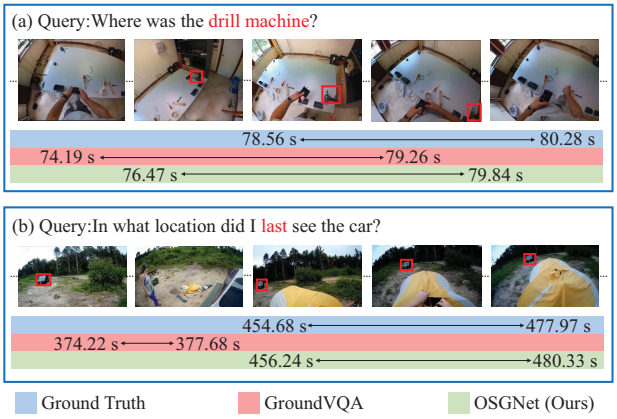


Figure 4. Qualitative comparison with GroundVQA on Ego4D-NLQ.

of these ablations. Removing $\mathcal{L}_{con}$ leads to a performance drop of around 0.4% across all R@1 metrics, indicating that the shot branch plays a crucial role in enhancing the model's ability to learn better video representations. Additionally, removing the object feature results in over 2.0% decline in the R@1 metric, highlighting the importance of object-level information for the NLQ task.

On Shot Segmentation. Table 5 displays the effect of our shot segmentation method on the validation set of Ego4D-NLQ. We hypothesized that camera movement, in addition to head rotation, may be relevant for shot segmentation. To test this, we selected movement-related verbs from the generated captions that occur more than 100 times, including "walks", "moves around", "rides", "runs", "cycles", "jogs", and "jumps". Based on these frequencies, we divided the verbs into three groups: head rotation verbs (R: "turns around", "looks around"), high-frequency movement verbs (HM: "walks", "moves around"), and other main movement verbs (RM: "rides", "runs", "cycles", "jogs", "jumps"). We first examined segmentation using only R and HM and observed that the model using R for segmentation typically performs better on R@1, which aligns with the main verbs in the captions. When combining R with HM and RM, the results are comparable to using R alone, leading us to adopt R for segmentation due to its efficiency.

On Object Features. Table 6 illustrates the impact of different types of object features on the validation set of Ego4D-NLQ. The results indicate that text-based object features outperform image-based object features by 0.65% in R@1, 0.3. We suppose that text-based object features align more effectively with the video modality.

4.5. Qualitative Analysis

We compared our model with the strong baseline GroundVQA [5] on the Ego4D-NLQ validation set through qualitative analysis. As shown in Figure 4(a), OSGNet accurately grounds the video moment where the "drill machine" is located. In contrast, GroundVQA struggles with fine-grained object localization due to the limitations of its video features, resulting in less accurate predictions. In Figure 4(b), OSGNet accurately selects the moment when the car was last seen, demonstrating stronger semantic alignment. On the other hand, while GroundVQA can identify the car's appearance in the video, it fails to predict the correct moment due to its limited video understanding.

Table 4. Ablation studies on model structure.

$\mathcal{L}_{con}$	Object	R@1		R@5	
		0.3	0.5	0.3	0.5
✗	✗	28.87	19.60	54.53	41.78
✓	✗	29.22	19.99	55.12	41.78
✗	✓	31.26	21.64	**58.19**	**45.19**
✓	✓	**31.63**	**22.03**	57.91	**45.19**

Table 5. Ablation studies on verb selection for shot segmentation.

R	HM	RM	R@1		R@5	
			0.3	0.5	0.3	0.5
✓	✓	✓	31.61	21.99	**58.52**	44.86
✓	✓	✗	31.22	21.84	56.83	44.44
✗	✓	✗	30.91	21.20	58.28	**45.41**
✓	✗	✗	**31.63**	**22.03**	57.91	45.19

Table 6. Ablation studies on the object features.

Object Features	R@1		R@5	
	0.3	0.5	0.3	0.5
Image	30.98	21.68	**58.22**	45.14
Text	**31.63**	**22.03**	57.91	**45.19**

5. Conclusion

In this paper, we propose OSGNet for egocentric video grounding. Our key improvements are twofold: 1) capturing object information to improve video representation, better aligning it with text queries; and 2) capturing shot information for contrastive learning, implicitly enhancing video representation. Through comprehensive experiments, we validate the effectiveness of these improvements.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China, No.62236003, No.62476071, No.62376140, and No.U23A20315; the Shenzhen College Stability Support Plan, No.GXWD20220817144428005; the Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions, No.2023KJ128, and the Special Fund for Taishan Scholar Project of Shandong Province.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5561–5569, 2017. 6

[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 arXiv:2211.09529*, pages 1–11, 2022. 1, 2, 6, 7

[3] Ruihan Chen, Junpeng Tan, Zhijing Yang, Xiaojun Yang, Qingyun Dai, Yongqiang Cheng, and Liang Lin. DPHANet: Discriminative parallel and hierarchical attention network for natural language video localization. *IEEE Transactions on Multimedia*, 2024. 7

[4] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8199–8206, 2019. 2

[5] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943, 2024. 2, 6, 7, 8

[6] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2448–2460, 2023. 2

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275, 2017. 2

[8] 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, 2

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 4

[10] Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. RGNet: A unified retrieval and grounding network for long videos. In *European Conference on Computer Vision*, pages 1–23, 2024. 6, 7

[11] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812, 2017. 2

[12] 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*, pages 1–5, 2023. 3, 4, 6, 7

[13] Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wk Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. CONE: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 8013–8028, 2023. 2, 3, 6, 7

[14] Jingjing Hu, Dan Guo, Kun Li, Zhan Si, Xun Yang, and Meng Wang. Maskable retentive network for video moment retrieval. In *ACM Multimedia 2024*, 2024. 7

[15] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. Video moment localization via deep cross-modal hashing. *IEEE Transactions on Image Processing*, 30:4667–4677, 2021. 2

[16] Yupeng Hu, Liqiang Nie, Meng Liu, Kun Wang, Yinglong Wang, and Xian-Sheng Hua. Coarse-to-fine semantic alignment for cross-modal moment localization. *IEEE Transactions on Image Processing*, 30:5933–5943, 2021. 2

[17] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. Victr: Video-conditioned text representations for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18547–18558, 2024. 2

[18] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022. 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 5

[20] Xiaojie Li, Shaowei He, Jianlong Wu, Yue Yu, Liqiang Nie, and Min Zhang. Mask again: Masked knowledge distillation for masked video modeling. In *Proceedings of the ACM International Conference on Multimedia*, page 2221–2232. ACM, 2023. 1

[21] Xiaojie Li, Jianlong Wu, Shaowei He, Shuo Kang, Yue Yu, Liqiang Nie, and Min Zhang. Fine-grained key-value memory enhanced predictor for video representation learning. In *Proceedings of the ACM International Conference on Multimedia*, page 2264–2274. ACM, 2023. 1

[22] 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. 1, 2, 6, 7

[23] 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. 2

[24] 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. 2

[25] Naiyuan Liu, Xiaohan Wang, Xiaobo Li, Yi Yang, and Yueting Zhuang. Reler@ zju-alibaba submission to the ego4d natural language queries challenge 2022. *arXiv preprint arXiv:2207.00383*, 2022. 3

[26] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3855–3863, 2024. 7

[27] Fangzhou Mu, Sicheng Mo, and Yin Li. SnAG: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940, 2024. 2, 4, 6, 7

[28] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 2

[29] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. EgoEnv: Human-centric environment representations from egocentric video. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

[30] Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13767–13777, 2023. 3

[31] 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*, pages 1–8, 2024. 1, 2, 6, 7

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. 6

[33] Carlos Plou, Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Ana C Murillo. CARLOR@ Ego4D step grounding challenge: Bayesian temporal-order priors for test time refinement. *arXiv preprint arXiv:2406.09575*, pages 1–4, 2024. 2, 7

[34] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 1, 2

[35] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. ChatVTG: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856, 2024. 2

[36] 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 *International Conference on Machine Learning*, 2021. 4, 6

[37] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. NaQ: Leveraging narrations as queries to supervise episodic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6694–6703. IEEE Computer Society, 2023. 1, 2, 6, 7

[38] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 1, 2

[39] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017. 5

[40] Jiayi Shao, Xiaohan Wang, Ruijie Quan, and Yi Yang. Action sensitivity learning for the ego4d episodic memory challenge 2023. *arXiv preprint arXiv:2306.09172*, 2023. 6, 7

[41] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 2

[42] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[43] Chaolei Tan, Jianhuang Lai, Wei-Shi Zheng, and Jian-Fang Hu. Siamese learning with joint alignment and regression for weakly-supervised video paragraph grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13569–13580, 2024. 2

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, Santiago, Chile, 2015. IEEE. 6

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[46] Di Wang, Xiantao Lu, Quan Wang, Yumin Tian, Bo Wan, and Lihuo He. Gist,content,target-oriented:a 3-level human-like framework for video moment retrieval. *IEEE Transactions on Multimedia*, pages 1–13, 2024. 7

[47] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2986–2994, 2021. 2

[48] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18709–18719, 2024. 2

[49] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9062–9069, 2019. 2

[50] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13901–13912, 2023. 2

[51] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 4

[52] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, 2020. 2, 7

[53] 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*, page 695–703. Association for Computing Machinery, 2021. 1

[54] 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

[55] 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

[56] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021. 2

[57] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12870–12877, 2020. 2, 6, 7

[58] Wei Zhang, Chaoqun Wan, Tongliang Liu, Xinmei Tian, Xu Shen, and Jieping Ye. Enhanced motion-text alignment for image-to-video transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18504–18515, 2024. 2

[59] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 5

[60] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12993–13000, 2020. 5

[61] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8454, 2021. 2

[62] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-First International Conference on Machine Learning*, 2024. 4

[63] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6758, 2023. 4