Interactive Episodic Memory with User Feedback

Nikesh Subedi
University of Utah
nikesh.subedi@utah.edu

Loris Bazzani
University of Verona
loris.bazzani@univr.it

Ziad Al-Halah
University of Utah
ziad.al-halah@utah.edu

Abstract

In episodic memory with natural language queries (EM-NLQ), a user may ask a question (e.g., "Where did I place the mug?") that requires searching a long egocentric video, captured from the user's perspective, to find the moment that answers it. However, queries can be ambiguous or incomplete, leading to incorrect responses. Current methods ignore this key aspect and address EM-NLQ in a one-shot setup, limiting their applicability in real-world scenarios. In this work, we address this gap and introduce the Episodic Memory with Questions and Feedback task (EM-QnF). Here, the user can provide feedback on the model's initial prediction or add more information (e.g., "Before this. I'm looking for the big blue mug not the white one"), helping the model refine its predictions interactively. To this end, we collect datasets for feedback-based interaction and propose a lightweight training scheme that avoids expensive sequential optimization. We also introduce a plug-and-play Feedback ALignment Module (FALM) that enables existing EM-NLQ models to incorporate user feedback effectively. Our approach significantly improves over the state of the art on three challenging benchmarks and is better than or competitive with commercial large vision-language models while remaining efficient. Evaluation with human-generated feedback shows that it generalizes well to real-world scenarios. Project: https://nsubedi11.github.io/refocus.

1. Introduction

Episodic Memory with Natural Language Query (EM-NLQ) retrieves specific moments from a person's past visual experiences, such as wearable-camera video, using free-form text questions [7]. For example, a user might ask, "What did I put in the frying pan?" (Fig. 1), and the model must identify the exact moment in the video that answers the question. By enabling on-demand "visual recall," such systems can help users recover forgotten details, review past actions, and support embodied agents or assistive technologies in tasks such as safety checks, finding misplaced items, and retracing work steps.

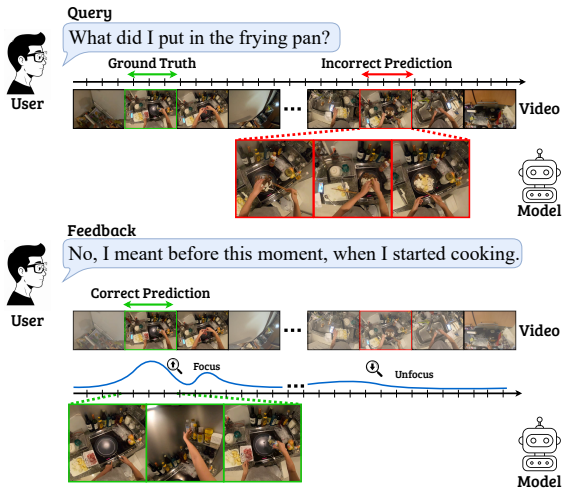

Figure 1. We introduce the interactive episodic memory with user feedback task (EM-QnF) to address ambiguous queries and model errors. Given an initial query and an incorrect model prediction (top), the user refines the query through natural language feedback, either by referring to the model's prediction or by providing additional information (middle). The model then uses the joint context of the query, prediction, and feedback to shift its focus toward relevant moments in the video that better align with the user's intent (bottom) to find the correct answer.

EM-NLQ remains very challenging. Egocentric videos are often long and untrimmed, making short answer segments difficult to find. The first-person viewpoint also introduces difficulties such as rapid head movements, motion blur, and occlusions. In addition, models must process long video histories efficiently while remaining lightweight enough for resource-constrained devices.

Recent work has focused on improving performance [5, 10, 11, 21, 23, 27], efficiency [23, 27, 32], and generalization with limited training data [31]. However, a key aspect remains underexplored: user queries are often ambiguous, and real interactions are inherently iterative. In practice, users may refine their questions and provide feedback after seeing an incorrect prediction. For example, as shown in Fig. 1, a user might respond, "No, I meant before this mo-

ment, when I started cooking." Current EM-NLQ models cannot leverage such feedback, missing the chance to refine their predictions and better match the user's intent.

Large Vision-Language Models (LVLMs) appear to be natural candidates for addressing the interactive aspect of EM-NLQ, since they build on language models designed for dialogue, instruction following, and user alignment. However, in practice, current LVLMs fall short of this promise. As our results show, fine-tuning these models for video understanding reduces their ability to respond effectively to user feedback. Additionally, their reliance on large vision-language backbones makes them highly resource-intensive and slow, which limits their practicality for fast or on-device episodic memory applications.

In this work, we take a first step toward addressing the underexplored problem of interactivity and feedback in EM-NLQ. We introduce the Episodic Memory with Questions and Feedback task (EM-QnF) and construct new datasets for feedback-based episodic memory interaction, together with a training scheme that avoids costly sequential training with feedback. We also propose ReFocus, which integrates our novel plug-and-play Feedback ALignment Module (FALM) with a variety of existing EM-NLQ models, enabling them to process and respond to user feedback efficiently and effectively. Unlike prior approaches that treat queries as fixed, one-shot inputs, ReFocus allows models to refine their predictions based on user corrections or clarifications, bringing EM-NLQ closer to the natural way humans seek information about past experiences.

Our approach significantly improves both the performance and scalability of EM-NLQ models in interactive settings. Through extensive experiments on three challenging benchmarks, our method achieves state-of-the-art performance and shows consistent gains across different EM-NLQ models and diverse evaluation settings. Finally, we validate the effectiveness and practicality of our approach through human-based feedback evaluations and comparisons with commercial multimodal LLMs, showing that our model can effectively incorporate real user feedback to produce more accurate and better aligned responses.

2. Related Works

Episodic Memory with Natural Language. This task requires localizing a response to a query within a long egocentric video. Early work adapted moment localization methods to this setting, establishing strong baselines (e.g., 2D-TAN [46], VSLNet [44]). Subsequent methods incorporated multiscale representations to handle the wide variation in response durations [11] or used object-centric features to capture a broader range of objects [5]. To address the scarcity of labeled data, some approaches leveraged more readily available supervision such as narrations [31] or adapted video-text contrastive learning to egocentric videos [21]. Other works [10, 23, 27, 32] focused on improving the efficiency and accuracy of EM-NLQ models.

However, all previous EM-NLQ methods localize the query in a one-shot manner and therefore cannot handle query ambiguity or model errors. In contrast, we propose an interactive episodic memory search setting that allows users to provide feedback to help the model refine its prediction and identify the correct response.

Large Vision-Language Models. Large Vision-Language Models (LVLMs) [1, 3, 16, 17, 22, 45] possess broad visual knowledge and excel at visual-linguistic reasoning and spatial understanding. However, these methods often struggle with time-sensitive video tasks such as temporal localization, especially in long videos. Many works [2, 13, 14, 24, 33, 43] have proposed solutions to improve the alignment between video semantics and their corresponding timestamps for such tasks. For example, TimeChat [33] introduces timestamp-aware video representations for temporal grounding, UniTime [20] proposes a multi-stage inference strategy for moment localization in long videos, and ChatVTG [30] explores training-free temporal grounding with LVLMs.

Despite strong performance on some of these tasks, these methods remain computationally expensive and, as we show in our experiments, fail to generalize well or retain their instruction-following ability when adapting to user feedback. Instead, we propose a plug-and-play feedback alignment module that enables EM-NLQ models to effectively use user feedback, leading to significant performance improvements while keeping computational cost low.

Localization with Textual Feedback. Utilizing human or synthetically generated feedback has been shown to effectively improve model capabilities across language reasoning [25, 42, 47] and visual understanding [18]. Different forms of feedback have been explored, including regions for segmentation [40], clicks for preference learning [47], and timestamps for temporal grounding [4, 20]. Among these, language feedback remains one of the richest and most user-friendly ways to refine predictions at inference time [8, 15, 26, 41] and to finetune models [42]. Closely related work also appears in localization within navigational environments [9, 37], where a locator tries to find an observer in an indoor environment, and the observer provides natural language feedback to guide the search.

However, to the best of our knowledge, textual feedback has not yet been shown to be effective for EM-NLQ. In this work, we propose an approach that leverages textual feedback to resolve localization ambiguities caused by large search spaces and imprecise user queries in egocentric videos. We further propose a feedback generation recipe that enables training at scale without expensive manual annotation, while still generalizing well to human feedback.

3. Episodic Memory with User Feedback

Our work is the first to explore the interactive aspect of EM-NLQ by enabling models to refine their predictions based on user feedback. We first formally define this new task (Sec. 3.1), then describe an effective procedure for generating feedback data for training models in this setting (Sec. 3.2). Finally, we introduce Feedback ALignment Module (FALM), a plug-and-play module that can be seamlessly integrated into existing EM-NLQ models, allowing them to process and respond to user feedback (Sec. 3.3).

3.1. Task Definition

We introduce the **Episodic Memory with Questions and Feedback** task (EM-QnF). Unlike EM-NLQ, which answers a natural language query in a one-shot manner, EM-QnF extends the task by allowing users to provide natural language feedback that guides the model toward a more accurate prediction. The goal is to enable EM-NLQ models to iteratively refine their responses based on user feedback and prior outputs.

Formally, given an egocentric video $\mathcal{V}$ and a natural language query $\mathcal{Q}$, the objective is to identify the video segment $\mathcal{R}^q \in \mathcal{V}$ that answers $\mathcal{Q}$, where $\mathcal{R} = [t_s, t_e]$ denotes the temporal window of the response defined by its start and end times. An EM-NLQ model produces an initial prediction $\mathcal{R}_1$, which may be incorrect due to query ambiguity, missing context, or model error. The user then provides feedback $\mathcal{F}_1$, a natural language statement containing additional or contrastive information relative to $\mathcal{R}_1$. The model integrates $\mathcal{F}_1$ in the context of $(\mathcal{V}, \mathcal{Q}, \mathcal{R}_1)$ to generate a refined prediction $\mathcal{R}_2 \neq \mathcal{R}_1$. This interactive process continues over multiple rounds, producing a sequence of responses $\{\mathcal{R}_1, \ldots, \mathcal{R}_n\}$ and feedbacks $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ until the final response $\mathcal{R}_n$ matches the correct answer, $\mathcal{R}_n = \mathcal{R}^q$. Without loss of generality, we refer to the current model prediction at step i as the *reference span* $\mathcal{R}^f$, which the user provides feedback on, and focus on the single-turn case in the following sections. We provide an extension to the multi-turn scenario in Sec. 3.3.

3.2. A Recipe for User Feedback Generation

To the best of our knowledge, no publicly available dataset currently supports the EM-QnF task. Collecting real user feedback is costly and time-consuming, as it requires human annotators to watch long egocentric videos and write meaningful feedback for incorrect model predictions.

To address this limitation, we propose a *synthetic feedback generation recipe* that produces useful and realistic feedback data from existing EM-NLQ datasets, effectively turning them into EM-QnF datasets. Our recipe, shown in Fig. 2, has four main steps: (1) reference span sampling, (2) caption generation for the sampled clips, (3) explanation generation for response span and (4) feedback construction.

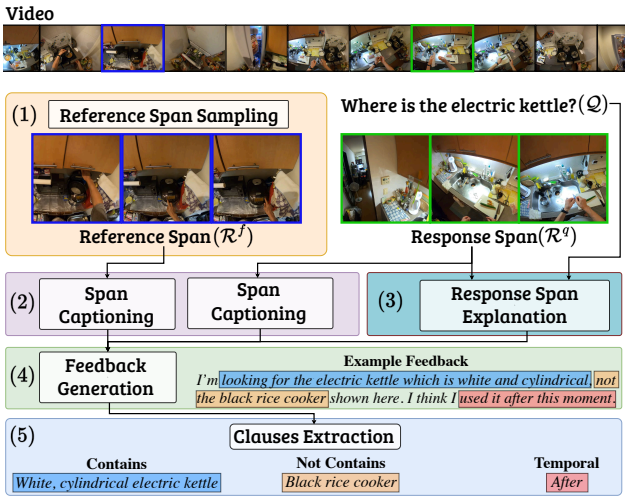

Figure 2. Our feedback generation recipe. For a query $\mathcal{Q}$ and ground-truth response span $\mathcal{R}^q$, (1) we sample a reference span $\mathcal{R}^f$, then (2) collect captions describing each span. Additionally, (3) we collect an explanation of why $\mathcal{R}^q$ answers $\mathcal{Q}$. The captions from (2) and (3) are then used to generate a feedback $\mathcal{F}$. Finally, (5) we extract three *clauses* from $\mathcal{F}$ representing different types of information, which we use to generate labels that supervise the learning of our FALM module (see Sec. 3.3).

This approach enables scalable and controlled generation of feedback examples without costly manual annotation. Furthermore, our results show that models trained on the proposed synthetic feedback data can effectively use real user feedback at inference time. Notably, synthetic feedback yields improvements comparable to those obtained with real user feedback (see Sec. 4.3), suggesting that our feedback generation recipe produces realistic feedback that aligns well with human feedback styles. Next, we describe the main steps of the proposed recipe, and provide more details in the supplementary material.

Reference Span Sampling. The goal of this step is to sample a span $\mathcal{R}^f$ that simulates an incorrect prediction for a given query $\mathcal{Q}$. An intuitive approach is to use actual model failures from an EM-NLQ model as reference spans. However, relying only on such examples limits the diversity of error types and may lead to overfitting to the behavior of a specific model. Hence, we sample two additional types of spans: (1) $\mathcal{R}^q$-similar spans, which are visually similar to the ground-truth response $\mathcal{R}^q$ (based on video feature similarity) but do not correctly answer the query ($\mathcal{R}^f \neq \mathcal{R}^q$); and (2) random spans, which are randomly sampled segments that are temporally disjoint from $\mathcal{R}^q$. We refer to spans sampled from model failures or $\mathcal{R}^q$-similar spans as *query-relevant spans*, since they may contain some information relevant to the query but are still incorrect. In contrast, we refer to random spans as *query-irrelevant spans*, since they represent completely off-target responses.

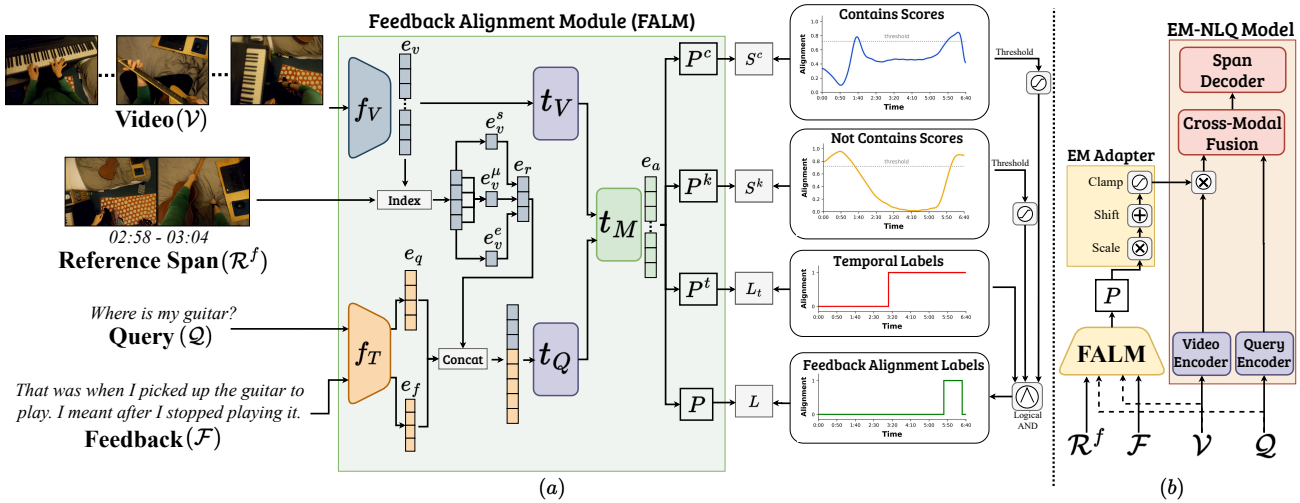

Figure 3. Our (a) Feedback ALignment Module (FALM) module and (b) its integration with an EM-NLQ model, ReFocus. FALM is trained to predict an alignment score P that indicates how well each clip aligns with the user feedback $\mathcal{F}$ in the context of the input video $\mathcal{V}$, the query $\mathcal{Q}$, and the reference span $\mathcal{R}^f$. It is then plugged into an EM-NLQ model using a lightweight adapter (b), enabling the model to leverage user feedback effectively by shifting its focus toward video clips that better match the user intent expressed in the feedback.

Response Captioning and Explanation Generation. For all ground-truth response spans $\mathcal{R}^q$ and the sampled reference spans $\mathcal{R}^f$ from previous step, we first generate textual captions describing these spans. This step reduces the computational cost of feedback generation by removing the need to process long video clips directly with large vision-language models (LVLMs), and it improves feedback quality by providing concise and relevant summaries that large language models (LLMs) can reason over effectively. For each span $\mathcal{R}_i$, we use a pre-trained LVLM to generate a textual description $\mathcal{D}_i$ that captures the visual content (e.g., objects and scenes) and actions present in the span, independent of any query $\mathcal{Q}$. Additionally, for each ground-truth span $\mathcal{R}_i^q$, we generate an explanation E_i describing why the span answers its corresponding query $\mathcal{Q}_i$. This helps us control the type of information allowed in the feedback, as explained next.

Feedback Generation. The final step generates natural language feedback $\mathcal{F}$ that guides the model from a sampled reference span $\mathcal{R}^f$ toward the correct ground-truth span $\mathcal{R}^q$. For each query $\mathcal{Q}_i$, we sample pairs $(\mathcal{R}_i^q, \mathcal{R}_j^f)$ and provide their corresponding descriptions $\mathcal{D}_i^q$, $\mathcal{D}_j^f$, the explanation E_i^q, and the relative temporal order between $\mathcal{R}_i^q$ and $\mathcal{R}_j^f$ to a reasoning-focused large language model (LLM). The LLM is prompted to generate feedback $\mathcal{F}_{i,j}$ that provides informative cues to help locate $\mathcal{R}_i^q$ without directly revealing the answer to $\mathcal{Q}_i$. To achieve this, we design prompts that encourage the feedback to include any combination of three types of information: (1) additional disambiguating details about the queried object or moment derived from $\mathcal{D}_i^q$; (2) contrastive cues highlighting differences between $\mathcal{R}_i^q$ and $\mathcal{R}_j^f$ based on $\mathcal{D}_i^q$ and $\mathcal{D}_j^f$; and (3) temporal guidance indicating whether to search before or after the current reference span $\mathcal{R}_j^f$. We further use the explanation E_i^q to instruct the LLM not to produce feedback that trivially answers the original query $\mathcal{Q}_i$.

This leads to diverse types of feedback, ranging from short phrases (e.g., "before this") that simulate impatient users to more descriptive feedback that may include different types of information based on the cues above. The feedback has an average length of 16 words, with a standard deviation of 6.8 (see Supp for the prompts and examples).

EM-QnF Samples. Following this recipe, the constructed EM-QnF dataset consists of samples of the form $(\mathcal{V}_i, \mathcal{Q}_i, \mathcal{R}_i^q, \{(\mathcal{R}_{i,j}^f, \mathcal{F}_{i,j})\})$, which can be used to train models to handle interactive feedback refinement.

3.3. FALM: Feedback ALignment Module

We introduce FALM, a plug-and-play module designed to help EM-NLQ models align with and use user feedback. Given user feedback $\mathcal{F}$, a reference span $\mathcal{R}^f$, a video $\mathcal{V}$, and a natural language query $\mathcal{Q}$, the core idea of FALM is to predict an alignment score for each clip in $\mathcal{V}$, indicating its relevance to $\mathcal{F}$.

Modern video encoders typically process long videos by dividing them into short clips and encoding each clip independently before aggregation. Formally, let $\mathcal{V} = \{C_1, \ldots, C_m\}$ denote a video segmented into m clips, where C_i is the i-th clip. FALM takes $(\mathcal{V}, \mathcal{Q}, \mathcal{R}^f, \mathcal{F})$ as

input and outputs an alignment vector $P \in [0,1]^m$, where each element P_i represents the alignment score between clip C_i and the given user feedback $\mathcal{F}$. These scores are then used to reweight the video features in existing EM-NLQ models, effectively shifting the model's attention toward clips that are most relevant to the user feedback when predicting the next response span.

FALM Architecture. As shown in Fig. 3.a, FALM encodes the video $\mathcal{V}$, query $\mathcal{Q}$, and feedback $\mathcal{F}$ using pretrained video and text encoders, f_V (ViT–1B from EgoVideo [28]) and f_T (gte-Qwen2-7B-instruct [19]), respectively. This yields feature representations $f_V(\mathcal{V}) = e_v \in \mathbb{R}^{v_t \times d}$, $f_T(\mathcal{Q}) = e_q \in \mathbb{R}^{q_t \times d}$, and $f_T(\mathcal{F}) = e_f \in \mathbb{R}^{f_t \times d}$, where d is the model dimension and (v_t, q_t, f_t) denote the number of tokens in each input sequence. We represent the reference span $\mathcal{R}^f$ by concatenating its start, end, and mean clip embeddings from the video features, $e_r = [e_v^s, e_v^e, e_v^\mu] \in \mathbb{R}^{3 \times d}$, allowing FALM to interpret the feedback $\mathcal{F}$ in the visual context of the reference span.

Next, we use a two-layer Transformer encoder (t_Q) to model interactions among $\{e_q, e_f, e_r\}$ and another two-layer Transformer encoder (t_V) to capture the overall video context from e_v. Finally, we pass the features through two Transformer decoder blocks with cross-attention (t_M) to produce video-feedback aligned embeddings e_a.

Alignment Supervision. We generate pseudo-labels to supervise the learning of FALM by indicating clip relevance based on three cues extracted from the feedback: (1) Contains: what information the correct response should contain, (2) Not Contains: what should not appear in the response, and (3) Temporal: whether to search before or after the reference span.

We prompt an LLM to extract these cues from each feedback $\mathcal{F}$ in form of short language clauses (see Fig. 2(5)). Using the EgoVideo [28] encoders, we compute <clip, clause> similarities to obtain *contains* scores S^c and *not-contains* scores S^n. To reduce noise in these similarity measures, we apply Gaussian smoothing, and min-max normalization across all clips. Finally, we invert the S^n scores to obtain $S^k = 1 - S^n$, i.e. the model should avoid clips that include information the user excluded in their feedback. To convert these scores into binary labels, we first calculate the mean and standard deviation of the scores within the correct response $\mathcal{R}^q$ [6], denoted by S_μ and S_σ for each score type, S^c and S^k. The threshold is then defined as $\delta = S_\mu - 3S_\sigma$. We assign a label of 1 to clips with scores above the threshold and 0 otherwise, resulting in binary labels L^c and L^k based on δ^c and δ^k. For temporal labels L^t, we assign 1 to clips before or after $\mathcal{R}^f$ according to the extracted temporal clause. Finally, we combine the three label types using a logical AND operation to form the final FALM labels: $L = L^c \wedge L^k \wedge L^t$. Note that not all three clauses are present in every feedback instance, and the labels are generated using only the subset of clauses extracted from each $\mathcal{F}$.

FALM Training Objective. Given encoded features e_a, four MLP heads predict *contains* (P^c) *not-contains* (P^k), *temporal* (P^t), and overall alignment (P) scores:

$$\mathcal{L} = \lambda \mathcal{L}_C(L, P) + \lambda_t \mathcal{L}_C(L^t, P^t) + \tag{1}$$

$$\lambda_c \mathcal{L}_2(S^c, P^c) + \lambda_n \mathcal{L}_2(S^k, P^k), \tag{2}$$

where $\mathcal{L}_C$ is a binary cross-entropy loss and $\mathcal{L}_2$ is $\|\cdot\|_2^2$ regression loss.

ReFocus: FALM Integration. After pretraining FALM on feedback data, we can integrate it into an EM-NLQ model by reweighting that model's video clip features using the predicted alignment scores P as shown in Fig. 3(b). In other words, FALM shifts the focus of the EM-NLQ model by emphasizing or de-emphasizing the importance of certain clips based on user feedback. To enable effective and seamless adaptation across EM-NLQ models, we introduce a lightweight EM Adapter that scales and shifts the alignment scores of FALM using two learned scalars, α and β, while fine-tuning with the EM-NLQ model: $\hat{P} = \text{clamp}(\alpha P + \beta, 0, 1)$. We denote the final approach with FALM plugged into an EM-NLQ model $\mathcal{M}$ with the adapter as ReFocus($\mathcal{M}$).

Multi-Turn Feedback Extension. While our work focuses on single-turn feedback, we propose here an extension of ReFocus to the multi-turn feedback setting. Given multiple independent feedback samples $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ for the same query $\mathcal{Q}$, we first pass each feedback to ReFocus($\mathcal{M}$) and use late fusion by averaging the output features of the Cross-Modal Encoder obtained from integrating each feedback separately, before passing the fused features to the Span Decoder, as shown in Fig. 3(b), to localize the answer. We find that this simple extension leads to significant gains without added complexity (Sec. 4.3). More advanced multi-turn modeling remains an important direction for future work, and our extension provides a strong baseline.

4. Experiments

We demonstrate the effectiveness of our approach on three challenging egocentric video datasets adapted to the EM-QnF task (Sec. 4.1), and compare it against state-of-the-art EM-NLQ and LVLM models (Sec. 4.2). We then provide an analysis of our model (Sec. 4.3), including comparisons with commercial LVLMs (Gemini-Flash) and human-generated feedback. Finally, we show the performance of our approach in multi-turn feedback scenarios.

4.1. Evaluation Setup

Datasets. We experiment with Ego4D-NLQ [7], Ego4D-GoalStep [34], and HD-EPIC [29], which are widely used

	Method	Ego4D-QnF				GoalStep-QnF				HD-EPIC-QnF			
		IoU = 0.3		IoU = 0.5		IoU = 0.3		IoU = 0.5		IoU = 0.3		IoU = 0.5	
		R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5
LVLM	TimeChat (ZS)	$1.6^{-0.2}$	N/A	$0.7^{-0.2}$	N/A	$2.3^{+0.9}$	N/A	$1.1^{+0.7}$	N/A	$0.2^{+0.0}$	N/A	$0.0^{+0.0}$	N/A
	UniTime (ZS)	$19.9^{-5.2}$	N/A	$12.3^{-3.3}$	N/A	$10.2^{-1.7}$	N/A	$6.0^{-0.8}$	N/A	$3.6^{-2.0}$	N/A	$1.3^{-1.2}$	N/A
	UniTime (FT)	$21.7^{-3.4}$	N/A	$13.3^{-2.3}$	N/A	$8.2^{-0.3}$	N/A	$5.5^{-0.1}$	N/A	$2.5^{-1.6}$	N/A	$1.0^{-0.8}$	N/A
Task-Expert	OSGNet	$29.6^{+0.4}$	$56.3^{+0.5}$	$20.5^{+0.5}$	$43.3^{+0.7}$	$30.2^{+0.6}$	$60.1^{+0.9}$	$24.7^{+0.5}$	$52.5^{+0.7}$	$14.7^{+0.3}$	$37.7^{-0.1}$	$9.6^{+0.1}$	$25.2^{+0.1}$
	ReFocus(OSGNet)	**$32.5^{+3.3}$**	**$58.3^{+3.7}$**	**$22.4^{+1.9}$**	**$45.3^{+3.3}$**	**$31.9^{+2.0}$**	**$61.0^{+2.3}$**	**$26.5^{+1.8}$**	**$53.7^{+2.4}$**	**$15.3^{+0.8}$**	**$38.3^{+1.3}$**	**$10.1^{+0.5}$**	**$25.8^{+0.9}$**
	GroundNLQ	$29.6^{+0.6}$	$56.0^{+1.0}$	$21.4^{+0.2}$	$43.0^{+0.6}$	$23.3^{+0.2}$	$53.2^{+0.3}$	$17.9^{+0.5}$	$43.7^{+0.4}$	$11.3^{+0.0}$	$33.8^{+0.9}$	$6.5^{-0.1}$	$21.1^{+0.5}$
	ReFocus(GroundNLQ)	**$33.1^{+3.3}$**	**$59.7^{+4.6}$**	**$23.7^{+2.2}$**	**$46.1^{+3.9}$**	**$26.8^{+4.9}$**	**$56.2^{+5.4}$**	**$20.3^{+3.6}$**	**$46.1^{+4.8}$**	**$15.1^{+3.0}$**	**$39.6^{+5.4}$**	**$9.1^{+2.1}$**	**$25.7^{+4.0}$**

Table 1. Model performance comparison across QnF datasets. Deltas (Δ) of feedback vs query-only performance are shown as superscripts. For LVLMs, ZS denotes zero-shot evaluation of the method, while FT is after finetuning on QnF data.

egocentric video benchmarks. While Ego4D-NLQ already provides query and response annotations, GoalStep and HD-EPIC are not designed for the NLQ task. Using NLQ templates from Ego4D-NLQ, we leverage the step descriptions from GoalStep and the narrations from HD-EPIC to generate natural language queries. We then apply our feedback generation recipe to all three datasets, resulting in question-and-feedback datasets that we refer to as Ego4D-QnF, GoalStep-QnF, and HD-EPIC-QnF. See Supp for dataset details and statistics.

Comparison with the SoTA. We compare ReFocus against several SoTA methods: LVLM-based methods like **TimeChat** [33] and **UniTime** [20], and the task expert EM-NLQ models like **GroundNLQ** [11] and **OSGNet** [5]. We evaluate UniTime and TimeChat in zero-shot setting and additionally, finetune UniTime on EM-QnF datasets as well. For task experts, we first pretrain on NaQ [31] and Ego4D NLQ [7] datasets using EgoVideo [28] as video features and gte-Qwen2-7B-instruct [19] as text features. We adapt EM-NLQ models to feedback input by concatenating the text features with reference span and feedback features to form a new query representation. Our FALM module is applied to GroundNLQ and OSGNet, which are named ReFocus(GroundNLQ) and ReFocus(OSGNet). We train all models (with and without our module) on the same QnF data for fair comparisons.

Implementation Details. Videos are divided into non-overlapping clips of 16 consecutive frames each. The clips are encoded with f_V = ViT–1B from EgoVideo [28]. For text features, we use f_T = gte-Qwen2-7B-instruct [19]. We pretrain FALM on the combined training splits of the three QnF datasets. For Refocus, we integrate FALM into the EM-NLQ model pretrained on NaQ [31] and fine-tune Refocus($\mathcal{M}$) on the combined training splits of the three QnF datasets. See Supp for full details.

Evaluation Metrics. Following the episodic memory benchmark [7], we report Recall (R1 and R5) at multiple temporal intersection-over-union (tIoU) thresholds $\{0.3, 0.5\}$ between the predicted and ground-truth spans.

4.2. Main Results

Table 1 shows the performance of our model and the baselines when evaluated with feedback. For brevity, we report results in the format X^{Δ}_{q+f}, where $\Delta = X_{q+f} - X_q$, X_{q+f} is the model performance when given both the query and the feedback, and X_q is its performance when given only the query (i.e., the initial performance before feedback). Thus, X_{q+f} shows the absolute performance after processing the feedback, while Δ shows whether, and by how much, the model benefits from the feedback.

Interestingly, the LVLM-based localization methods (Table 1, top), which incorporate an LLM in their architecture, fail to adapt to user feedback. In most metrics and across datasets, they perform worse with feedback than without it, resulting in negative Δ values. Even when fine-tuned on the new QnF data, these models do not show clear improvement in leveraging feedback, despite the stronger reasoning capabilities of their underlying LLMs, which might be expected to help them better handle such interactions.

As for EM-NLQ experts, simply training OSGNet and GroundNLQ on EM-QnF data leads to only small improvements, with $\Delta \leq 1\%$ across all metrics. Although these models are trained with EM-QnF data as well, they still largely ignore the feedback.

Across all datasets, our ReFocus consistently outperforms all other methods. Specifically, ReFocus(GroundNLQ) achieves notable gains on R1 and R5, with up to +4.9 and +5.4, respectively. We also observe consistent improvements when ReFocus is applied to OSGNet. These results demonstrate the effectiveness of our approach in leveraging user feedback to improve localization across EM-NLQ models and benchmarks. Furthermore, our ap-

Method	GoalStep-QnF IoU = 0.3 R1	GoalStep-QnF IoU = 0.3 R5	HD-EPIC-QnF IoU = 0.3 R1	HD-EPIC-QnF IoU = 0.3 R5
OSGNet	14.5$^{+0.2}$	36.7$^{+0.7}$	5.3$^{+0.4}$	16.9$^{+0.7}$
ReFocus(OSGNet)	**17.9**$^{+3.6}$	**42.0**$^{+6.8}$	**6.7**$^{+2.2}$	**18.6**$^{+4.5}$
GroundNLQ	17.7$^{+0.5}$	42.2$^{+1.6}$	6.6$^{+0.2}$	21.3$^{+0.3}$
ReFocus(GroundNLQ)	**20.7**$^{+3.7}$	**45.3**$^{+5.0}$	**8.2**$^{+1.6}$	**25.1**$^{+4.2}$

Table 2. Zero-Shot evaluation across feedback datasets when models trained on Ego4D-QnF only.

Method	Ego4D-QnF IoU = 0.3 R1	Ego4D-QnF IoU = 0.3 R5	GoalStep-QnF IoU = 0.3 R1	GoalStep-QnF IoU = 0.3 R5
Gemini-2.5-Flash	15.7$^{+1.7}$	28.7$^{+0.7}$	8.7$^{+2.7}$	16.0$^{+1.0}$
ReFocus(OSGNet)	24.0$^{+3.0}$	46.7$^{+2.7}$	21.7$^{+2.7}$	53.7$^{+3.7}$
ReFocus(GroundNLQ)	8.7$^{+8.7}$	48.0$^{+48.0}$	9.7$^{+9.7}$	54.7$^{+54.7}$

Table 3. Performance comparison between Gemini-2.5-Flash and ReFocus models on a small 100 NLQ subset where ReFocus(GroundNLQ) fails with query-only but improves with feedback. Deltas (Δ) of feedback vs query-only performance are shown as superscripts.

proach does not rely on heavyweight components such as LLMs and preserves the efficiency of the underlying task-expert model (see Supp for details).

Zero-Shot Cross Evaluation. Table 2 shows the performance on GoalStep-QnF and HD-Epic-QnF when the models trained only on Ego4D-QnF, *i.e.*, in the zero-shot setting. While competing methods show only marginal improvements in this setting, our approach generalizes much better and does not overfit to a specific type of feedback. Since GoalStep and HD-EPIC differ from Ego4D-NLQ in video content, they also likely involve different styles of feedback. Even so, our approach achieves larger Δ values on both GoalStep-QnF and HD-EPIC-QnF.

Comparison with Commercial LVLMs. We compare our approach with a commercial LVLM, Gemini-2.5-Flash, on a subset of the test sets. We sample 100 NLQs from each of the three QnF datasets where ReFocus(GroundNLQ) fails when given only the query, but where at least one user feedback example in the test set helps the model identify the correct response span. We then sample three feedback samples for each of the 100 queries, favoring query-relevant spans over irrelevant ones, resulting in 900 QnF samples across the three datasets. Table 3 shows that, despite the strong R1 performance of Gemini-2.5-Flash, it is unable to significantly improve its predictions when given feedback and a reference span. This result suggests that even commercial LVLMs, despite being trained on large-scale multi-modal data, still struggle to reason effectively over feedback in long-form videos to produce better predictions. See Supp for more details.

Method	IoU = 0.3 R1	IoU = 0.3 R5	IoU = 0.5 R1	IoU = 0.5 R5
GroundNLQ	29.56	56.42	21.63	43.71
w. FALM	**33.13**	**59.70**	**23.58**	46.26
w. FALM$_C$	31.08	57.95	22.26	44.52
w. FALM$_N$	30.89	58.03	22.38	45.02
w. FALM$_T$	32.29	59.41	23.23	**46.40**
w. FALM w/o Adapter	32.46	58.33	23.11	45.39

Table 4. Ablation of our ReFocus(GroundNLQ). Evaluated on a subset of Ego4D-QnF containing all types of FALM labels.

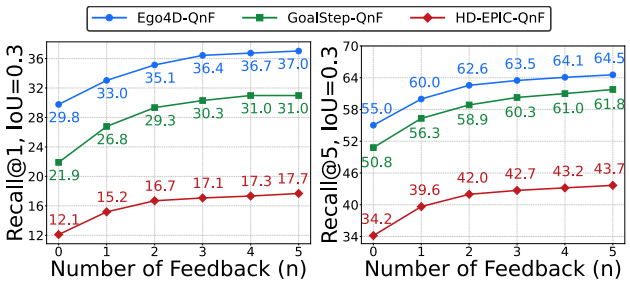

Figure 4. Multi-Turn Feedback evaluation of our ReFocus(GroundNLQ) across the three datasets.

4.3. In-depth Model Analysis

Ablations. We present an in-depth ablation study of our approach in Table 4 to examine the effect of each component in learning from feedback. We start with GroundNLQ and evaluate different components of FALM. First, we study the contribution of each supervision signal used to train FALM (Sec. 3.3). We train variants of FALM using a single supervision type: contains, not-contains, and temporal clauses, denoted as FALM$_C$, FALM$_N$, and FALM$_T$, respectively. We observe that each clause type improves the model, with the temporal clause being the most helpful. Combining all of them leads to further gains, since different feedback may contain different types of information. We also evaluate whether integrating FALM with the proposed EM adapter is beneficial. The results (last row in Table 4) show a drop in performance when the adapter is removed.

Multi-Turn Feedback. While we have focused so far on the single-turn case, we evaluate here the ability of ReFocus to handle multi-turn feedback in a zero-shot manner, following the proposed extension in Sec. 3.3. In our evaluation, we average performance over five different random samplings of n feedback per query. Figure 4 shows the performance of ReFocus(GroundNLQ) across the three datasets with multiple feedback. Interestingly, our approach

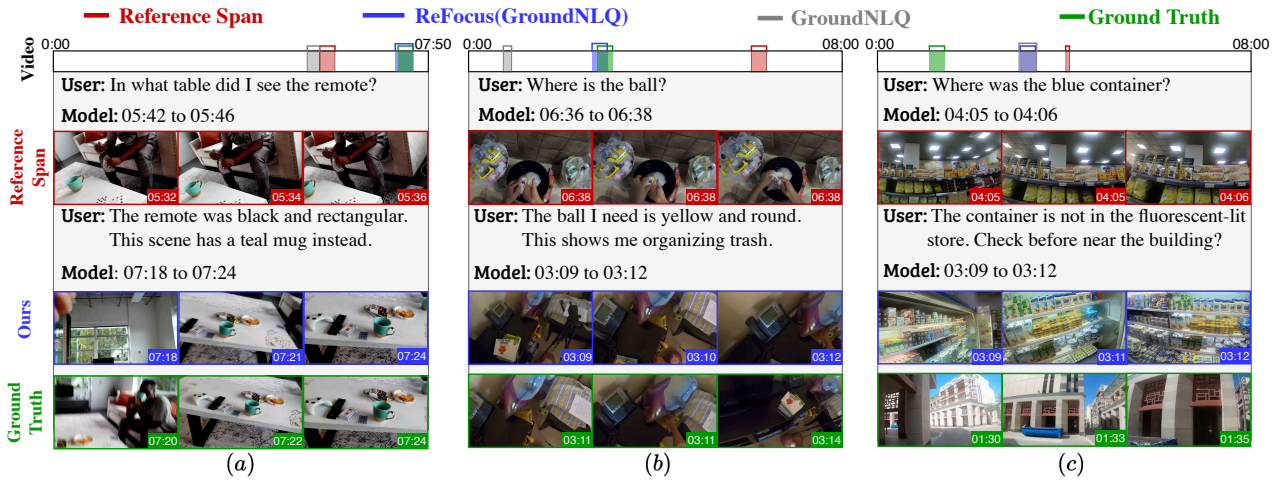

Figure 5. Qualitative results for GroundNLQ and our ReFocus(GroundNLQ) when given feedback. Examples (a) and (b) show cases where ReFocus(GroundNLQ) improves with feedback, whereas (c) shows a failure case.

	IoU = 0.3		IoU = 0.5	
Feedback Type	**ΔR1**	**ΔR5**	**ΔR1**	**ΔR5**
Generated	8.6	50.0	5.4	31.0
Human	5.8	34.4	3.4	20.4

Table 5. Comparison of our ReFocus(GroundNLQ) on generated vs human feedback on examples from Ego4d-QnF where the method fails with query-only. The reported metrics are absolute improvement in Recall when using feedback vs without.

improves as the number of feedback turns increases on all datasets, even though it was not trained in this setting. We observe substantial gains up to the third or fourth feedback turn, after which performance plateaus.

Comparison with Human Feedback. We collected feedback from human users to compare it with the quality of our generated feedback. From Ego4D-QnF, we sample unique NLQ and reference span pairs where ReFocus(GroundNLQ) fails but improves with LVLM-generated feedback. We ask 11 users to assume the role of the person wearing the camera in the egocentric video and provide feedback for the reference spans as if they were trying to recall the correct response span. Users are instructed not to answer the queries directly, but instead to guide the model toward the correct response span they are trying to recall. In total, we collect 500 unique user feedback for 271 NLQ and reference span pairs. Table 5 shows the evaluation of our approach using generated and human feedback. We observe that our approach can leverage human feedback and improve its predictions, which suggests that our proposed feedback generation recipe is effective at producing realistic and helpful feedback that generalizes to human-style inputs. Additionally, we see there is still a gap in absolute improvement between generated and human feedback, indicating room for further improvement. Methods trained with human feedback may be able to recover additional gains.

Qualitative Results and Failures. Figure 5 shows qualitative examples from the Ego4D-QnF dataset. Examples (a) and (b) show that ReFocus(GroundNLQ) improves over GroundNLQ when given additional user feedback about object attributes in contrast to the reference span. Example (c) shows an interesting failure case for our approach. The user feedback suggests that the blue container is not inside the store, but near a building, and also indicates that the model should search before the reference span. While the model correctly shifts its attention to an earlier moment, it is confused by the many blue food containers on the shelves and fails to infer that those shelves are inside the same store.

5. Conclusion

This work addresses an unexplored aspect of episodic memory search with natural language queries: the interactive nature of the task. User questions can be incomplete or ambiguous, and model predictions can be incorrect. In such cases, the model should be able to incorporate user feedback to improve its predictions. We introduce the task of interactive episodic memory with user feedback (EM-QnF), a suitable recipe for user feedback generation without manual annotations, and an effective feedback alignment module (FALM) that can be integrated into different EM-NLQ models, leading to significant improvements across multiple challenging benchmarks.

Acknowledgements We thank Santhosh Ramakrishnan for valuable feedback during the early stages of the project. This work was partially supported by NSF 2421782 and MPS-AI-00010515.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. 2

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[4] Jianfeng Dong, Xiaoman Peng, Daizong Liu, Xiaoye Qu, Xun Yang, Cuizhu Bao, and Meng Wang. Temporal sentence grounding with relevance feedback in videos. *Advances in Neural Information Processing Systems*, 37:43107–43132, 2024. 2

[5] Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, and Liqiang Nie. Object-shot enhanced grounding network for egocentric video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24190–24200, 2025. 1, 2, 6, 3

[6] Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 907–916, 2026. 5

[7] 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, 6, 3

[8] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018. 2

[9] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 806–822, 2020. 2

[10] Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. Rgnet: A unified clip retrieval and grounding network for long videos. In *European Conference on Computer Vision*, pages 352–369. Springer, 2024. 1, 2

[11] 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, 2, 6, 3

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3

[13] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 2

[14] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 2

[15] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36: 61437–61449, 2023. 2

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[17] 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. 2

[18] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[19] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 5, 6

[20] Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2025. 2, 6, 3

[21] 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

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[23] Zijia Lu, ASM Iftekhar, Gaurav Mittal, Tianjian Meng, Xiawei Wang, Cheng Zhao, Rohith Kukkala, Ehsan Elhamifar, and Mei Chen. Decafnet: Delegate and conquer for efficient temporal grounding in long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24066–24076, 2025. 1, 2

[24] Kaijing Ma, Xianghao Zang, Zerun Feng, Han Fang, Chao Ban, Yuhan Wei, Zhongjiang He, Yongxiang Li, and

Hao Sun. Llavilo: Boosting video moment retrieval via adapter-based multimodal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2798–2803, 2023. 2

[25] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023. 2

[26] Sho Maeoki, Kohei Uehara, and Tatsuya Harada. Interactive video retrieval with dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 952–953, 2020. 2

[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. 1, 2

[28] 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. 5, 6, 2

[29] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913, 2025. 5, 1

[30] 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

[31] 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, 2023. 1, 2, 6, 3

[32] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Spotem: Efficient video search for episodic memory. In *International Conference on Machine Learning*, pages 28618–28636. PMLR, 2023. 1, 2

[33] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2, 6, 3

[34] 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: 38863–38886, 2023. 5, 1

[35] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025. 2

[36] Qwen Team. Qwen3 technical report, 2025. 1

[37] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 2

[38] 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. 3

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[40] Qiaoqiao Wei, Hui Zhang, and Jun-Hai Yong. Focused and collaborative feedback integration for interactive image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18643–18652, 2023. 2

[41] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 2

[42] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023. 2

[43] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023. 2

[44] 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

[45] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553, 2023. 2

[46] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12870–12877, 2020. 2

[47] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 2

[48] 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. 3