Multi-Factor Adaptive Vision Selection for Egocentric Video Question Answering

Haoyu Zhang [1 2] **Meng Liu** [† 3] **Zixin Liu** [2] **Xuemeng Song** [4] **Yaowei Wang** [2 1] **Liqiang Nie** [† 1]

Abstract

The challenge of interpreting the world from a human perspective in Artificial Intelligence (AI) is particularly evident in egocentric video question answering, which grapples with issues like small object recognition, noise suppression, and spatial-temporal reasoning. To address these challenges, we introduce the Multi-Factor Adaptive vision Selection (MFAS) framework. MFAS integrates a patch partition and merging module for enhanced small object recognition, a prior-guided patch selection module for noise suppression and focused analysis, and a hierarchical aggregation network to aggregate visual semantics guided by questions. Extensive experiments on several public egocentric datasets have validated the effectiveness and generalization of our framework. Code and data are available in `https://github.com/Hyu-Zhang/EgoVideoQA`.

1. Introduction

In the contemporary landscape of human-centric applications, the field of egocentric video understanding has emerged as a crucial area of research within Computer Vision (CV) and embodied Artificial Intelligence (AI) (Plizzari et al., 2023; Akiva et al., 2023). Distinct from third-person or exocentric videos, egocentric videos are characterized by their intricate scene composition, constrained informational content, and irregular motion dynamics, posing significant challenges for associated computational tasks (Xu et al., 2023b; Radevski et al., 2023). Egocentric Video Question Answering (VideoQA) stands out as a particularly versatile task in this domain, drawing extensive interest from both the academic and industrial sectors. This line of research holds potential for advancing smart assistive technologies like visual assistants, and in broader applications such as augmented reality and smart glasses (Nagarajan et al., 2023).

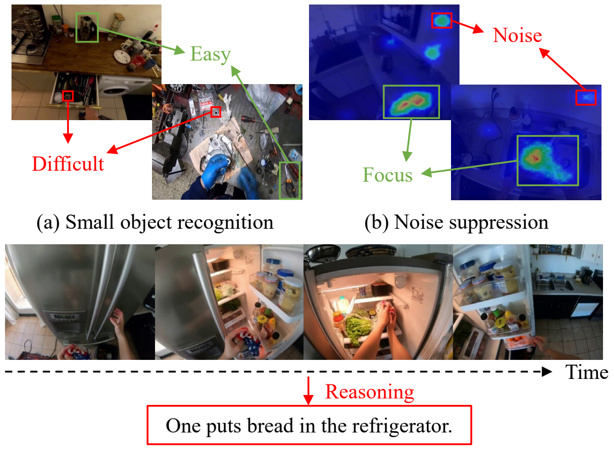
Egocentric VideoQA presents a higher level of complexity compared to exocentric VideoQA, largely due to the divergence in human perception from the representations in many internet datasets. A notable domain gap exists, resulting in a weak inductive bias between the exocentric and egocentric domains and complicating the application of transfer learning techniques (Xu et al., 2023b). The release of specialized egocentric VideoQA datasets, such as EgoVQA (Fan, 2019), EgoTaskQA (Jia et al., 2022), and QAEgo4D (Bärmann & Waibel, 2022), has spurred the development of methods tailored for the egocentric domain (Lin et al., 2022; Pramanick et al., 2023; Shen et al., 2023). For instance, Pramanick et al. (Pramanick et al., 2023) adapted a dual-tower model to the egocentric video-text dataset, followed by fine-tuning on an egocentric VideoQA dataset. **However, existing approaches predominantly employ a general framework, addressing the egocentric VideoQA task through fine-tuning strategies.**

Despite these advancements, several egocentric VideoQA-specific challenges remain unaddressed: 1) **Small Object Recognition**. In cluttered settings such as kitchens and laboratories, small objects in egocentric videos are particularly challenging to detect and recognize due to problems such as incompleteness, deformation, and blurring caused by moving shots (Pramanick et al., 2023), as shown in Figure 1(a). 2) **Noise Suppression**. Egocentric videos often contain focused areas like hand-object interactions (highlighted in Figure 1(b)), where redundant and noisy regions can hinder visual information processing. 3) **Spatial-Temporal Reasoning**. The unique first-person perspective in egocentric videos restricts the ability to observe complete behaviors objectively, as illustrated in Figure 1(c). This limitation increases the difficulty in understanding complete activities based on partial observations.

To address these challenges, we introduce a Multi-Factor Adaptive vision Selection (MFAS) framework for egocen-

[†]Corresponding Author [1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China [3]School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China [4]School of Computer Science and Technology, Shandong University, Qingdao, China. Correspondence to: Meng Liu <mengliu.sdu@gmail.com>, Liqiang Nie <nieliqiang@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Small object recognition

(b) Noise suppression

(c) Spatial-temporal reasoning

Figure 1. Illustration of three challenges. (a) Small object recognition, (b) Noise suppression (the images are the attention distribution of the baseline EgoVLPv2 (Pramanick et al., 2023)), and (c) Spatial-temporal reasoning.

tric VideoQA, comprising three main components: a patch partition and merging module, a prior-guided patch selection module, and a hierarchical aggregation network. The patch partition and merging module targets small object recognition by leveraging multi-scale patch information. The prior-guided patch selection module, informed by egocentric data observations and eye gaze habits, aims to filter out redundancy and noise, emphasizing key semantic regions. Lastly, the hierarchical aggregation network facilitates spatial-temporal reasoning by progressively fusing visual information from local to global granularities, guided by the posed question. Extensive experimental results on multiple public datasets underscore the effectiveness of our proposed framework, marking a significant advancement in the field of egocentric VideoQA.

The contributions of our work are threefold:

- We pioneer in recognizing and addressing the unique challenges of the egocentric VideoQA task, being the first to concurrently tackle the issues of small object recognition, noise suppression, and spatial-temporal reasoning in this domain.
- We introduce a prior-guided patch selection module that integrates prior knowledge with spatial and temporal cues, effectively reducing spatial redundancy and highlighting crucial visual regions.
- We devise a patch partition and merging module to integrate multi-scale visual cues, enhancing small object recognition. Besides, we present a hierarchical aggregation network to dynamically adjust the model's receptive field and improve spatial-temporal reasoning.

2. Related Work

2.1. Video Question Answering

VideoQA is a quintessential task within the visual-language domain, playing a pivotal role in enhancing AI systems' ability to interpret and interact with multimedia content. This advancement is instrumental in fostering AI systems that engage in more sophisticated, human-like interactions with digital content, as highlighted in the recent study (Wu et al., 2021; Zhang et al., 2023c). In VideoQA, the agent is asked to analyze a video clip or sequence and answer questions pertinent to the visual content, thus testing its comprehension and inferential abilities (Liu et al., 2023a; Wang et al., 2023b). Over recent years, VideoQA has garnered substantial interest from the research community, leading to diverse methodological advancements. The evolution of this field can be categorized into several prominent approaches: 1) Early Attention-based Methods: Initiatives such as those by (Zhao et al., 2017; Jang et al., 2017) employ attention networks to learn cross-modal representations, effectively bridging the gap between video content and corresponding questions. 2) Memory Network-based Methods: Techniques developed by (Gao et al., 2018; Fan et al., 2019) utilize memory networks to store sequential inputs, allowing for the strategic utilization of information, even from earlier sequences. 3) Graph Neural Network-based Methods: Approaches by (Jiang & Han, 2020; Xiao et al., 2022; Urooj et al., 2023; Bai et al., 2024) leverage graph structures to enhance inference capabilities in VideoQA, facilitating efficient information communication. 4) Modular Network-based Methods: Proposals by (Le et al., 2020; Grunde-McLaughlin et al., 2021) address the limitations of inflexible hand-crafted architectures in handling varying data modalities, video lengths, and question types. And 5) Transformer-based Methods: Recent developments by (Buch et al., 2022; Gao et al., 2023) position transformers at the forefront of VideoQA research, capitalizing on their proficiency in modeling long-term relationships (Yan et al., 2023; Xue et al., 2023).

Despite these significant strides in VideoQA, a common limitation persists: the focus predominantly remains on third-person videos. This orientation overlooks the nuanced perspective of first-person (egocentric) videos, which embody a more direct, human-centric view of the world. Addressing this gap is crucial for developing AI systems that can fully comprehend and interact with a broader spectrum of visual content.

2.2. Egocentric Video Question Answering

The advent of wearable cameras like Google Glass, GoPro, and Nreal X has spurred significant interest in first-person video analysis. This burgeoning research area is characterized by its focus on understanding the unique perspectives

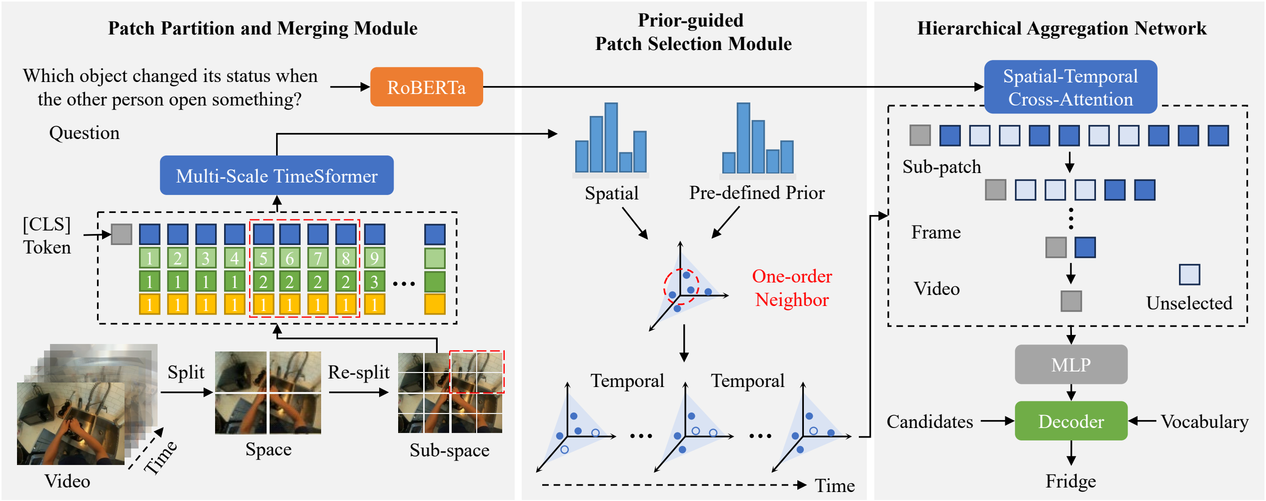


Figure 2. The framework of our proposed MFAS, which consists of the patch partition and merging module, the prior-guided patch selection module, and the hierarchical aggregation network.

captured through egocentric viewpoints (Hazra et al., 2023).

Pioneering contributions in this domain have been significantly bolstered by the introduction of comprehensive egocentric datasets. For instance, the Ego4D dataset (Grauman et al., 2022) has been instrumental in advancing various facets of egocentric video analysis. This includes notable work in egocentric action recognition (Radevski et al., 2023; Wang et al., 2023a), object detection (Akiva et al., 2023; Huang et al., 2023), human (hand)-object interactions (Zhang et al., 2022; Xu et al., 2023b), and visual query localization (Mai et al., 2023; Xu et al., 2023a).

Building upon this foundation, the specific field of egocentric VideoQA has started to gain traction. Early endeavors by (Fan, 2019; Jia et al., 2022; Pramanick et al., 2023) introduced datasets such as EgoVQA and EgoTaskQA, respectively, playing a pivotal role in propelling research in this niche. Lin et al. (Lin et al., 2022) furthered this progress by proposing a dual-encoder framework, adapted for egocentric tasks through pre-training on the Ego4D dataset. Additionally, Pramanick et al. (Pramanick et al., 2023) refined the interplay between visual and textual encoders, tailoring it more closely to the requirements of egocentric VideoQA.

Despite these advancements, there remains a discernible gap in the current body of work. Existing approaches in egocentric VideoQA have not thoroughly addressed the distinct differences between egocentric and exocentric VideoQA. There is a clear need for more nuanced solutions that explicitly cater to the unique characteristics and challenges inherent in first-person video analysis. This gap presents an opportunity for future research to develop methodologies that are more finely attuned to the specificities of the egocentric perspective.

3. Methodology

As depicted in Figure 2, our model is composed of three components: a patch partition and merging module, a prior-guided patch selection module, and a hierarchical aggregation network. Detailed overall process is shown in Algorithm 1 in the Appendix. In the subsequent sections of our paper, we will provide a comprehensive and detailed exploration of each of these components.

3.1. Patch Partition and Merging Module

For the video $\mathcal{V}$, we follow the TimeSformer (Bertasius et al., 2021) model and uniformly sample it to obtain $\mathbf{v} \in \mathbb{R}^{T\times C\times H\times W}$, where T indicates the number of frames and C, H, and W denote channel, height, and width, respectively. To enhance the detection of small-scale objects in the video, a straightforward approach would be to segment the video into smaller patches before processing them with TimeSformer. However, this approach has a critical consideration: detecting objects at different scales is interrelated. Over-emphasizing the recognition of small objects could inadvertently compromise the detection quality of objects at the original scale.

To address this challenge, we propose an extension to the existing TimeSformer model to accommodate multi-scale processing. This extension is based on two key adaptations:

(1) **Patch Partition**. As shown in Figure 2, we first split the video frame into patches via convolution operations. These patches are further subdivided into smaller sub-patches, represented as $\hat{\mathbf{v}} \in \mathbb{R}^{T\times N\times d}$, where N denotes the number of sub-patches and d signifies the feature dimension. In refining the spatial embedding within TimeSformer, we introduce a dual spatial information scheme for each sub-patch.

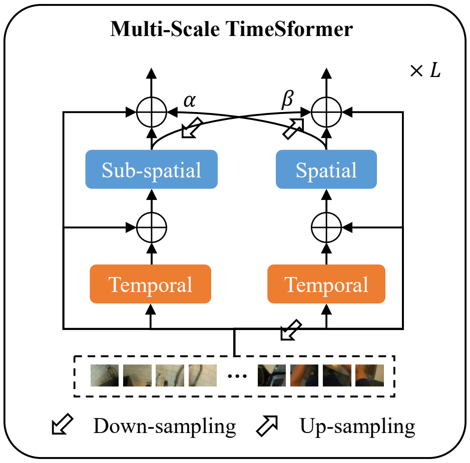


Figure 3. Illustration of the multi-scale TimeSformer (Bertasius et al., 2021). It contains both sub-patch and patch branches.

This includes the sub-patch's positional data in both sub-space and space (see Figure 2), allowing for an innovative update to the original position embedding as follows:

$$\widetilde{\mathbf{v}} = \widehat{\mathbf{v}} + \mathrm{SSE}(\widehat{\mathbf{v}}) + \mathrm{SE}(\widehat{\mathbf{v}}) + \mathrm{TE}(\widehat{\mathbf{v}}), \tag{1}$$

where SSE, SE, and TE denote sub-spatial embedding (sub-patch in sub-space), spatial embedding (sub-patch in space), and temporal embedding (sub-patch in time), respectively. The sequence of sub-spatial embedding is $1, 2, ..., N$, the spatial embedding is $1, 2, ..., N/4$, and the temporal embedding is from 1 to T. The resultant $\widetilde{\mathbf{v}} \in \mathbb{R}^{(T\times N+1)\times d}$ is the final sub-patch representations, including the addition of the [CLS] token at the start of the sequence.

(2) **Patch Merging**. To enhance the TimeSformer's adaptability to multi-scale visual information, we integrate a twin spatial-temporal attention structure, as shown in Figure 3, primarily to encode varied video semantics. This structure allows for the down-sampling of sub-patches into patches, utilizing temporal attention for modeling temporal patch information, followed by spatial attention to facilitate inter-patch interaction within the same frame. In the patch branch, the initial step involves down-sampling $\widetilde{\mathbf{v}}$ to produce a condensed representation $\widetilde{\mathbf{v}}_0^p$. Subsequently, the operation at the l-th layer of this branch is articulated as follows:

$$\mathbf{v}_l^p = \text{S-Att}(\widetilde{\mathbf{v}}_{l-1}^p + \text{T-Att}(\widetilde{\mathbf{v}}_{l-1}^p)), \tag{2}$$

where $\widetilde{\mathbf{v}}_{l-1}^p$ is the video representations at the $(l-1)$-th layer for patch branch. T-Att and S-Att refer to temporal attention and spatial attention, respectively, with $\mathbf{v}_l^p \in \mathbb{R}^{(T\times N/4+1)\times d}$ as the resultant output. In parallel, the sub-patch branch, which bypasses down-sampling, produces $\mathbf{v}_l^{sp} \in \mathbb{R}^{(T\times N+1)\times d}$. A fusion mechanism is employed to facilitate the perception of spatial information at dual scales, expressed as:

$$\begin{cases} \widetilde{\mathbf{v}}_l^p = \mathbf{v}_l^p + \widetilde{\mathbf{v}}_{l-1}^p + \beta \cdot \mathrm{DnSample}(\mathbf{v}_l^{sp}), \\ \widetilde{\mathbf{v}}_l^{sp} = \mathbf{v}_l^{sp} + \widetilde{\mathbf{v}}_{l-1}^{sp} + \alpha \cdot \mathrm{UpSample}(\mathbf{v}_l^p), \end{cases} \tag{3}$$

where UpSample refers to interpolation-based up-sampling, with α and β are learnable parameters. The symbols $\widetilde{\mathbf{v}}_l^p \in \mathbb{R}^{(T\times N/4+1)\times d}$ and $\widetilde{\mathbf{v}}_l^{sp} \in \mathbb{R}^{(T\times N+1)\times d}$ are the output patch and sub-patch representations at the l-th layer. $\widetilde{\mathbf{v}}_{l-1}^p$ and $\widetilde{\mathbf{v}}_{l-1}^{sp}$ ($\widetilde{\mathbf{v}}_0^{sp} = \widetilde{\mathbf{v}}$) are the output at the $(l-1)$-th layer for patch and sub-patch separately. Through the sequential application of L layers, the framework yields refined and effective representations, $\widetilde{\mathbf{v}}_L^p$ and $\widetilde{\mathbf{v}}_L^{sp}$, encapsulating the accumulated spatial-temporal insights at both the patch and sub-patch scales.

For the question q, we use RoBERTa (Liu et al., 2019) as our backbone to extract its representations $\mathbf{q} \in \mathbb{R}^{(|q|+1)\times d}$, where d represents the feature dimension of the embeddings and $|q|+1$ is the number of tokens, including the tokens in the question q as well as the [CLS] token.

3.2. Prior-guided Patch Selection Module

In the realm of egocentric video analysis, it is observed that the user's gaze and the region of hand activity are predominantly central, with peripheral regions often contributing noise rather than informative content. Moreover, there exists a notable redundancy among neighboring sub-patches within a video frame. To address these challenges, we propose a novel prior-guided patch selection module, as depicted in Figure 4 and Algorithm 2. This module synthesizes prior knowledge with spatial and temporal cues to dynamically assess the importance of each sub-patch.

Empirical studies on first-person videos reveal a concentration of user focus on hand-object interaction zones, typically situated in the middle to lower sections of the frame (Xu et al., 2023b; Ohkawa et al., 2023; Zhang et al., 2023a). Leveraging this insight and data from human eye simulations, we initialize a prior matrix $\mathbf{A}_0 \in \mathbb{R}^{N^{1/2}\times N^{1/2}}$, defined as:

$$\begin{cases} \mathbf{A}_0[i,j] = 1, \text{ if } i \geq \frac{1}{3}N^{\frac{1}{2}} \text{ and } \frac{1}{6}N^{\frac{1}{2}} \leq j \leq \frac{5}{6}N^{\frac{1}{2}}, \\ \mathbf{A}_0[i,j] = 0, \text{ otherwise.} \end{cases} \tag{4}$$

This matrix is stored in memory as an initial reference. Furthermore, to incorporate multi-scale spatial relationships, we integrate spatial attention scores $\mathbf{S}^p \in \mathbb{R}^{M\times T\times N/4}$ and sub-spatial attention scores $\mathbf{S}^{sp} \in \mathbb{R}^{M\times T\times N}$ derived from the L-th layer of the modified TimeSformer:

$$\mathbf{S} = \frac{1}{M}\sum_{m=1}^{M}(\mathbf{S}_m^{sp} + \mathrm{UpSample}(\mathbf{S}_m^p)), \tag{5}$$

where M denotes the number of attention heads and $\mathbf{S} \in \mathbb{R}^{T\times N}$ is the merged attention score. The patch selection module operates on the principle that the relevance of each sub-patch in the t-th frame ($t \in [1, T]$) is determined by a combination of current spatial relationships, previous temporal relationships, and data priors. The selection process is formulated as:

$$\mathbf{A}_t = \mathbf{A}_{t-1} \cup (f(\text{Top-}k(\mathbf{S}_t))) \cap \mathcal{N}(\mathbf{A}_{t-1})), \tag{6}$$

where $\mathbf{A}_{t-1}$ represents the mask matrix for the $(t-1)$-th frame, while $\mathbf{S}_t$ signifies the merged attention score for the t-th frame. The operations $\cup$ and $\cap$ are utilized for intersection and concatenation respectively. The Top-k operation is employed to identify the k most important sub-patches within $\mathbf{S}_t$. Furthermore, the function f is designed to generate binary vectors, marking the positions of these top k elements in $\mathbf{S}_t$ as 1, with other positions being assigned a value of 0. The term $\mathcal{N}$ is used to denote the function responsible for generating a first-order neighborhood. The resulting mask for the t-th frame, denoted as $\mathbf{A}_t \in \mathbb{R}^N$, highlights the key sub-patches by labeling their position as 1. It is worth noting that this description abstracts away basic operations like flattening and repetition for clarity.

In summary, the selection process for each sub-patch within a given frame (t-th frame) is determined by the interplay of spatial dynamics present in the current frame, the temporal context from the previous frame, and the influence of pre-established data priors. This approach ensures a coherent and connected representation of sub-patches by restricting expansion to the immediate neighborhood of regions identified in the preceding frame. This approach enables a detailed evaluation of the relevance of each sub-patch, ultimately leading to the creation of a comprehensive mask for the entire video sequence, denoted by $\mathbf{A} \in \mathbb{R}^{T \times N}$, which encapsulates the vital regions across all frames.

3.3. Hierarchical Aggregation Network

To fully understand the visual semantics from different granularities, we design a hierarchical aggregation network using questions as clues. Specifically, we use a multi-layer spatial-temporal cross-attention network as the backbone for interaction between video and question, avoiding high computational costs due to long sub-patch sequences.

For a video input represented as $\tilde{\mathbf{v}}_L^{sp} \in \mathbb{R}^{(T \times N+1) \times d}$ and a question input $\mathbf{q} \in \mathbb{R}^{(|q|+1) \times d}$, the computational process within the r-th layer of spatial-temporal cross-attention is expressed as:

$$\begin{cases} \mathbf{q}_r = \text{Cross-Att}(\mathbf{q}_{r-1}, \bar{\mathbf{e}}, \bar{\mathbf{e}}), \\ \tilde{\mathbf{v}}_{L,r}^{sp} = \text{Cross-Att}(\bar{\mathbf{e}}, \mathbf{q}_{r-1}, \mathbf{q}_{r-1}), \\ \bar{\mathbf{e}} = \text{S-Att}^{'}(\tilde{\mathbf{v}}_{L,r-1}^{sp} + \text{T-Att}^{'}(\tilde{\mathbf{v}}_{L,r-1}^{sp})), \end{cases} \tag{7}$$

where $\mathbf{q}_{r-1}$ and $\tilde{\mathbf{v}}_{L,r-1}^{sp}$ denote the question and video representations at the $(r-1)$-th layer, respectively. The notations S-Att$^{'}$ and T-Att$^{'}$ denote the adapted spatial and temporal attention mechanisms, which are applied across distinct dimensions. These mechanisms are defined by the expression $\text{Softmax}(\text{MaxPool}(\mathbf{e}\mathbf{e}^{\top})/\sqrt{d} + (-\infty \cdot \neg \mathbf{A}_{r-1}))\mathbf{e}$, where $\mathbf{e}$ is the embedding input and $\neg$ signifies the negation operation, converting 0s to 1s and vice versa. The term $\mathbf{A}_{r-1}$ indicates the mask matrix at the $(r-1)$-th layer, with an added mask for the [CLS] token set to 1 for uniformity. This mask remains unchanged as $\mathbf{A}$ through the initial $R-3$ layers, where R stands for the total count of spatial-temporal cross-attention layers. This consistency underscores our belief in the criticality of detailed, fine-grained interactions for achieving robust high-level hierarchical aggregation.

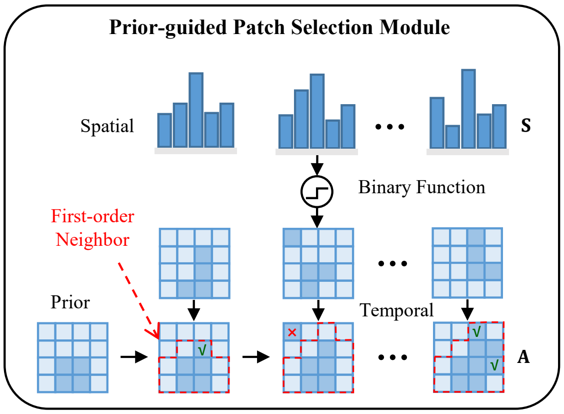


Figure 4. The proposed prior-guided patch selection module. The first-order neighbor is obtained by inflating the one-hop distance (left, right, up, and down) of the core regions (dark marked) at the previous frame. The tick indicates that the patch at current frame lies within the first-order neighbor and satisfies first-order connectivity, while the cross denotes that it does not.

Before inputting the concluding three layers, a systematic down-sampling operation is applied to the video representations, transitioning from sub-patch to patch, then to frame, and ultimately to the video level. This process is defined as:

$$\tilde{\mathbf{v}}_{L,r}^{sp}, \mathbf{A}_r = \text{DnSample}(\tilde{\mathbf{v}}_{L,r}^{sp}, \mathbf{A}_r), \tag{8}$$

where $r \in \{R-2, R-1, R\}$ and DnSample is executed on the spatial dimension (N) at the $R-2$ and $R-1$ layers and on the temporal dimension (T) at the R layer. Correspondingly, the mask $\mathbf{A}_r$ undergoes a similar down-sampling to ensure consistency and retention of relevant information.

Overall, the video representations $\tilde{\mathbf{v}}_{L,r}^{sp}$ reside in the space $\mathbb{R}^{(T \times N+1) \times d}$ for the first $R-3$ layers of the network. However, in the final three layers, there is a dimensional transition, where the representations are first reduced to $\mathbb{R}^{(T \times N/4+1) \times d}$, then to $\mathbb{R}^{(T+1) \times d}$, and ultimately to $\mathbb{R}^{2 \times d}$. Upon navigating through the R spatial-temporal cross-

Table 1. Accuracy comparison with the latest methods on the EgoTaskQA direct/indirect split. The best results are highlighted in bold and the second in underlined. * indicates the reproduced results using the open-source code.

Method	Direct			Indirect		
	Open	Binary	All	Open	Binary	All
Most Likely (Jia et al., 2022)	0.70	50.46	15.40	-	-	-
HGA (Jiang & Han, 2020)	22.75	68.53	36.77	8.66	53.72	28.36
BERT (Kenton & Toutanova, 2019)	-	-	-	11.22	58.24	31.78
PSAC (Li et al., 2019b)	26.97	65.95	38.90	15.31	57.75	32.72
VisualBERT (Li et al., 2019a)	24.62	68.08	37.93	21.05	57.61	37.01
HME (Fan et al., 2019)	27.66	68.6	40.16	18.27	52.55	33.06
ClipBERT (Lei et al., 2021)	27.70	67.52	39.87	11.17	40.71	24.08
HCRN(w/o vision) (Le et al., 2020)	-	-	-	11.38	55.52	30.76
HCRN (Le et al., 2020)	30.23	69.42	42.20	27.82	59.29	41.56
CMCIR* (Liu et al., 2023b)	35.49	65.92	46.04	28.36	58.86	42.00
EgoVLP (Lin et al., 2022)	31.69	71.26	42.51	27.04	55.28	38.69
EgoVLPv2 (Pramanick et al., 2023)	35.56	75.60	46.26	29.14	59.68	42.28
MFAS (Ours)	**38.95**	**75.86**	**48.69**	**32.44**	**63.02**	**45.40**

attention layers[1], the network yields the final question-enhanced video representations, denoted as $\widetilde{\mathbf{v}}^{sp}_{L,R} \in \mathbb{R}^{2\times d}$, and the video-enhanced question representations, represented as $\mathbf{q}_R \in \mathbb{R}^{(|q|+1)\times d}$.

3.4. Decoder and Optimization

3.4.1. Decoder

Our framework consists of a discriminative decoder and a generative decoder, each tailored to distinct tasks.

Discriminative Decoder: The primary objective of this decoder is to accurately select an answer from a set of predefined options. This task is approached by employing a Multi-Layer Perceptron (MLP) as a mapping function:

$$y = \mathrm{MLP}(\widetilde{\mathbf{v}}^{sp}_{L,R}[\mathrm{CLS}]), \tag{9}$$

where y represents the probability distribution over the candidate answers. During the inference phase, the answer corresponding to the highest probability is selected as the final response (Shi et al., 2023).

Generative Decoder: Contrary to the discriminative approach, the target of the generative decoder for the QAEgo4D dataset is to construct an answer, rather than selecting from existing options. To this end, we employ a lightweight Long Short-Term Memory (LSTM) model, coupled with an FC layer, to enable this functionality:

$$y = \mathrm{FC}(\mathrm{LSTM}(\mathbf{q}_R, \widetilde{\mathbf{v}}^{sp}_{L,R}[\mathrm{CLS}])), \tag{10}$$

where FC is the fully-connected layer and y denotes the probability distribution across the vocabulary for each predicted word. Notably, the [CLS] token of $\widetilde{\mathbf{v}}^{sp}_{L,R}$ serves as the initial hidden state for LSTM. During inference, we employ an auto-regressive way to sequentially generate a probability distribution for each word, thereby constructing the complete answer (Lee et al., 2023).

3.4.2. Optimization

The optimization of our model's parameters is primarily guided by the cross-entropy loss (Mao et al., 2023; Zhang et al., 2023b) function, which is mathematically represented as follows:

$$\mathcal{L}_{\mathrm{CLS}} = \frac{1}{Y}\sum_{y}^{Y} -g\log(y), \tag{11}$$

where y represents the predicted probability, g represents the ground truth, and Y is the number of categories.

Additionally, for the QAEgo4D dataset, we enhance our framework with a ranking supervision component, drawing inspiration from prior research (Lei et al., 2020; Bärmann & Waibel, 2022). In the penultimate layer of our hierarchical aggregation network, we implement a unique approach for managing positive and negative samples. For each target frame (positive sample), two frames from different segments of the video are chosen as negative samples. This process is facilitated by the availability of timestamp annotations for target answers within the QAEgo4D dataset. The ranking loss is computed using the Log-Sum-Exp (LSE) loss (Kobayashi, 2023) function, formulated as:

$$\mathcal{L}_{\mathrm{LSE}} = \log(1+\sum_{v\notin\mathcal{G}}\sum_{u\in\mathcal{G}}\exp(f_v(\widetilde{\mathbf{v}}^{sp}_{L,R-1})-f_u(\widetilde{\mathbf{v}}^{sp}_{L,R-1}))), \tag{12}$$

where $f_u(\widetilde{\mathbf{v}}^{sp}_{L,R-1})$ and $f_v(\widetilde{\mathbf{v}}^{sp}_{L,R-1})$ denote the scores of positive and negative samples, respectively, computed using

[1] At layers $R-1$ and R, the network employs self-attention rather than spatial-temporal attention due to the reduced sequence length.

Table 2. Performance comparison with the latest methods on the QAEgo4D dataset and the results are grouped into generative and discriminative settings. The best results are highlighted in bold and the second in underlined. * indicates the reproduced results using the open-source code and † denotes the results utilizing $\mathcal{L}_{\mathrm{LSE}}$ supervision.

Method	Video	Acc	BLEU	METEOR	ROUGE
BlindVQA (Bärmann & Waibel, 2022)	-	9.0	3.6	17.4	25.9
SimpleVQA (Bärmann & Waibel, 2022)	Full	9.3	6.1	17.4	26.1
SimpleVQA† (Bärmann & Waibel, 2022)	Full	9.7	3.6	18.3	27.1
Longformer (Beltagy et al., 2020)	Full	3.0	2.4	15.4	20.9
Longformer† (Beltagy et al., 2020)	Full	6.7	5.4	16.9	24.4
MFAS (Ours)	Sample	9.9	5.4	17.6	26.2
MFAS† (Ours)	Sample	**11.9**	**8.6**	**18.9**	**28.2**
HCRN (Le et al., 2020)	Full	10.3	7.6	17.2	25.7
JustAsk (Yang et al., 2021)	Full	9.6	3.9	17.8	26.7
CMCIR* (Liu et al., 2023b)	Full	9.7	3.1	16.5	24.7
CMCIR* (Liu et al., 2023b)	Sample	9.4	3.9	17.0	25.4
EgoVLP* (Lin et al., 2022)	Sample	10.2	4.6	17.0	25.4
EgoVLPv2* (Pramanick et al., 2023)	Sample	10.3	6.6	17.4	25.8
EgoVLPv2†* (Pramanick et al., 2023)	Sample	11.9	8.6	17.6	26.2
MFAS (Ours)	Sample	10.5	5.8	18.0	26.7
MFAS† (Ours)	Sample	**12.7**	**9.3**	**18.3**	**27.0**

Table 3. Ablation study on the EgoTaskQA indirect split. The best results are highlighted. PPM, PS, and HA denote the patch partition and merging module, prior-guided patch selection module, and hierarchical aggregation network, respectively.

Module			Indirect		
PPM	PS	HA	Open	Binary	All
✓	✓		31.87	62.00	44.52
	✓	✓	32.01	62.22	44.71
✓		✓	31.22	61.28	43.95
✓	✓	✓	**32.44**	**63.02**	**45.40**

a multi-layer perceptron followed by a Softmax function. And the symbol $\mathcal{G}$ denotes the set of frames corresponding to the target moment.

Consequently, the composite loss function for our model is defined as $\mathcal{L} = \mathcal{L}_{\mathrm{CLS}}$ for the EgoTaskQA dataset. In contrast, for the QAEgo4D dataset, the loss function is augmented to $\mathcal{L} = \mathcal{L}_{\mathrm{CLS}} + \lambda\mathcal{L}_{\mathrm{LSE}}$, where λ serves as a balancing coefficient.

4. Experiments

4.1. Datasets

Our method is rigorously evaluated using two public egocentric VideoQA datasets: EgoTaskQA and QAEgo4D.

EgoTaskQA Dataset (Jia et al., 2022): EgoTaskQA stands as a comprehensive benchmark in the realm of egocentric VideoQA, building upon the foundations set by the LEMMA dataset (Jia et al., 2020). This dataset comprises a collection of 2,336 real-world videos, each averaging 36.9 seconds in length, accompanied by a rich set of 40,000 question-answer pairs. EgoTaskQA is distinctive in its bifurcated structure: the "direct" subset, where questions are randomly sampled, and the "indirect" subset, which necessitates multi-step reasoning for accurate answer derivation.

QAEgo4D Dataset (Bärmann & Waibel, 2022): QAEgo4D is an egocentric VideoQA dataset focused on episodic memory (Ramakrishnan et al., 2023; Karuvally et al., 2023) task. It encompasses 1,325 videos, notably longer in duration with an average length of 495.1 seconds, and includes 14,513 question-answer pairs. The extended duration of videos introduces additional complexity in video comprehension.

4.2. Performance Comparison

In the empirical evaluation of our proposed method, we conducted comprehensive comparisons with several state-of-the-art baselines on the EgoTaskQA and QAEgo4D datasets, encompassing multiple settings. The comparative results [2] are systematically presented in Table 1 and 2.

As illustrated in Table 1, our method MFAS demonstrates a significant improvement over all competing methodologies, including the most recent baseline, EgoVLPv2. Notably, under both the direct and indirect settings of EgoTaskQA, MFAS achieves absolute gains of 2.43% and 3.12%, respectively, in the "All" metric. This performance leap underscores the effectiveness of our comprehensive framework in the realm of egocentric VideoQA.

Turning to the results from Table 2, which delineate both

[2] More experimental details can be found in the Appendix.

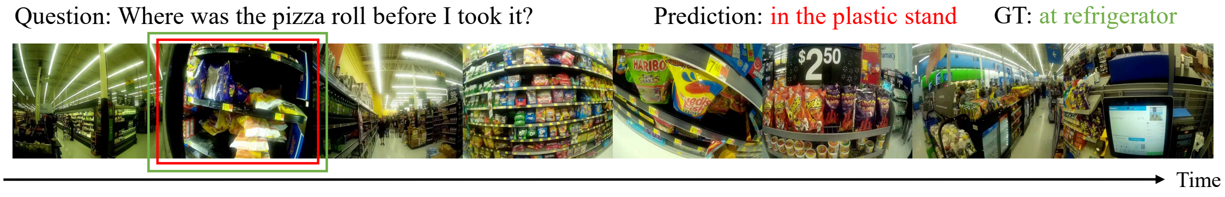

Question: Where was the pizza roll before I took it?
Prediction: in the plastic stand
GT: at refrigerator
Time

Figure 5. The generative case from QAEgo4D dataset.

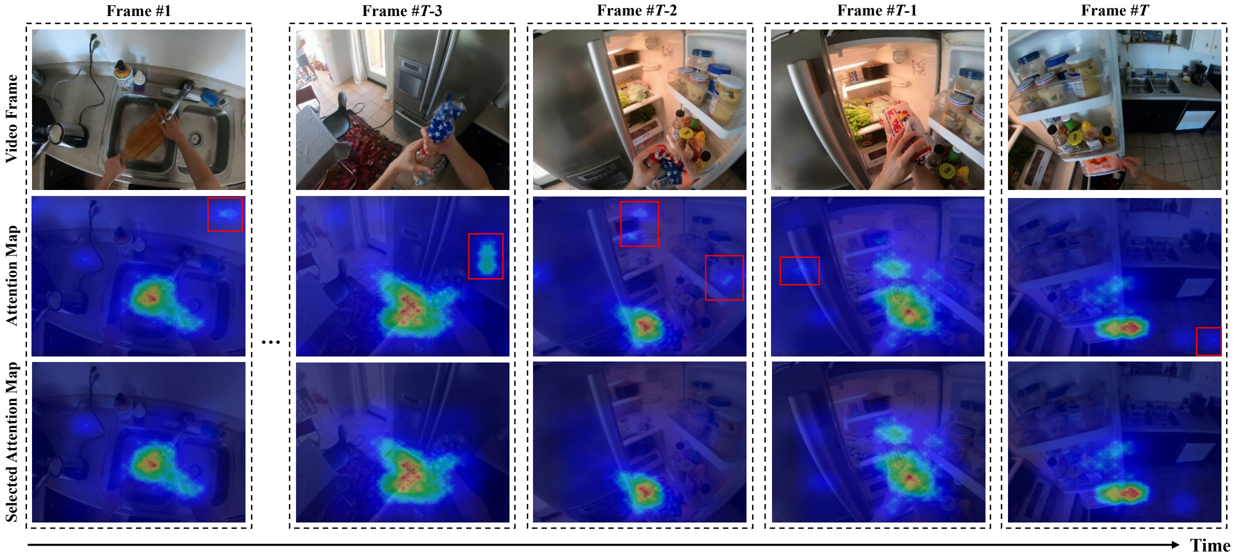


Figure 6. Visualization for prior-guided patch selection. The red boxes are the presentation of the suppressed noise.

generative and discriminative results, our method consistently outperforms the leading baselines across both domains, i.e., CMCIR in the exocentric and EgoVLPv2 in the egocentric. Specifically, MFAS achieves an absolute accuracy improvement of 2.2% in the discriminative setting and 0.8% in the generative setting, compared to these optimal baselines. These results not only validate the superiority of our approach but also demonstrate its robust generalization capabilities across diverse and challenging settings in egocentric long VideoQA.

4.3. Ablation Study

To empirically validate the contribution of the key components in our method, i.e., the patch partition and merging module, the prior-guided patch selection module, and the hierarchical aggregation network, we conducted ablation studies on the EgoTaskQA indirect split. The corresponding modules are individually removed, and the impact on the model's performance is measured. The comparative results of these experiments are detailed in Table 3.

The ablation results provide clear evidence of the importance of each module. Specifically, when the patch partition and merging module is removed and replaced by a combination of RoBERTa and the original TimeSformer, there is a 0.69% decrease in overall accuracy. Similarly, the omission of the prior-guided patch selection module results in a more substantial drop of 1.45% in accuracy. Lastly, replacing the hierarchical aggregation network with a standard cross-attention network leads to a decrease of 0.88% in accuracy. These findings underscore the substantial contribution of each component to the overall effectiveness of our model.

4.4. Case Study

Our investigation into the generative aspect of the QAEgo4D dataset includes a detailed case study, illustrated in Figure 5, with further instances detailed in the Appendix.

The instance highlighted in Figure 5 reveals the capacity of our model to generate contextually coherent answers that may not strictly align with the ground truth. Specifically, the model identifies a "pizza roll" situated both in a refrigerator and on a plastic stand within the video. Despite the deviation from the ground truth, the response of our model, informed by the context of the video, remains logically consistent and pertinent to the narrative depicted. This result illustrates the adeptness of our model at not only comprehending the posed questions and related video content but also its ability to produce viable answers underpinned by a deep understanding of the video context, even in cases where its responses diverge from the predefined ground truth.

4.5. Visualization

To demonstrate the operational efficacy of our prior-guided patch selection module, we visualized the selection process, as exhibited in Figure 6, with additional visualizations provided in the Appendix. These visualizations offer an insightful perspective into how our module processes and refines attention within the video frames.

The comparative analysis of Figure 6 highlights a clear contrast between the initial attention maps (second row) and the refined attention maps (third row). Initially, the standard attention mechanism sometimes highlights non-essential areas, which may detract from an accurate interpretation of the video. However, the refined attention maps, generated by our innovative prior-guided patch selection module, exhibit a concentrated focus on pertinent regions. This approach effectively reduces background noise and emphasizes critical zones, especially those involving interactions between hands and objects. This visual evidence supports the efficacy of our module in optimizing attention allocation, significantly enhancing the ability of our model to discern and interpret relevant video content accurately.

5. Conclusion and Future Work

In this paper, we propose the MFAS framework to address small object recognition, noise suppression, and spatial-temporal reasoning challenges present in egocentric VideoQA. Specifically, we devise a patch partition and merging module to enhance the recognition of small objects and reduce the interference with large targets by considering both coarse and fine-grained spatial semantics. And to reduce redundancy and suppress noise, we design a prior-guided patch selection module to synthesize the prior, spatial, and temporal information, highlighting critical visual regions. Besides, we present a hierarchical aggregation network to incrementally aggregate video semantics at different granularities, improving the spatial-temporal comprehension of egocentric videos. Extensive experiments on two public datasets demonstrate the superiority of the proposed framework. In the future, we plan to mine shot-level semantics for egocentric videos and unify the VideoQA task with different perspectives.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China, No.:62236003, No.:62376137, No.:62376140, and No.:U23A20315; Shenzhen College Stability Support Plan, No.:GXWD20220817144428005; Shandong Provincial Natural Science Foundation, No.:ZR2022YQ59; Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions, No.:2023KJ128; Major Key Project of Peng Cheng Laboratory, No.:PCL2023A08.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, such as it could refine the understanding of the system for egocentric video, thus facilitating visual perception and applications for intelligent robots.

References

Akiva, P., Huang, J., Liang, K. J., Kovvuri, R., Chen, X., Feiszli, M., Dana, K., and Hassner, T. Self-supervised object detection from egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5225–5237, 2023.

Bai, Z., Wang, R., Gao, D., and Chen, X. Event graph guided compositional spatial-temporal reasoning for video question answering. *IEEE Transactions on Image Processing*, 33:1109–1121, 2024.

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*, pp. 65–72, 2005.

Bärmann, L. and Waibel, A. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1560–1568, 2022.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, pp. 813–824, 2021.

Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., and Niebles, J. C. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2022.

Fan, C. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., and Huang, H. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1999–2007, 2019.

Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., and Shou, M. Z. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14773–14783, 2023.

Gao, J., Ge, R., Chen, K., and Nevatia, R. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6576–6585, 2018.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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*, pp. 18995–19012, 2022.

Grunde-McLaughlin, M., Krishna, R., and Agrawala, M. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

Hazra, R., Chen, B., Rai, A., Kamra, N., and Desai, R. Egotv: Egocentric task verification from natural language task descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15417–15429, 2023.

Huang, C., Tian, Y., Kumar, A., and Xu, C. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22910–22921, 2023.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2758–2766, 2017.

Jia, B., Chen, Y., Huang, S., Zhu, Y., and Zhu, S.-C. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *Proceedings of the European Conference on Computer Vision*, pp. 767–786, 2020.

Jia, B., Lei, T., Zhu, S.-C., and Huang, S. Egotaskqa: Understanding human tasks in egocentric videos. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 3343–3360, 2022.

Jiang, P. and Han, Y. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11109–11116, 2020.

Karuvally, A., Sejnowski, T., and Siegelmann, H. T. General sequential episodic memory model. In *Proceedings of the International Conference on Machine Learning*, pp. 15900–15910, 2023.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Kobayashi, T. Two-way multi-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7476–7485, 2023.

Le, T. M., Le, V., Venkatesh, S., and Tran, T. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9972–9981, 2020.

Lee, Y., Lee, K., Park, S., Hwang, D., Kim, J., Lee, H.-i., and Lee, M. Qasa: advanced question answering on scientific articles. In *Proceedings of the International Conference on Machine Learning*, pp. 19036–19052, 2023.

Lei, J., Yu, L., Berg, T., and Bansal, M. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, 2020.

Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., and Liu, J. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7331–7341, 2021.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019a.

Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., and Gan, C. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8658–8665, 2019b.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Lin, K. Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E. Z., Gao, D., Tu, R.-C., Zhao, W., Kong, W., et al. Egocentric video-language pretraining. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 7575–7586, 2022.

Liu, M., Zhang, F., Luo, X., Liu, F., Wei, Y., and Nie, L. Advancing video question answering with a multimodal and multi-layer question enhancement network. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3985–3993, 2023a.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Liu, Y., Li, G., and Lin, L. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mai, J., Hamdi, A., Giancola, S., Zhao, C., and Ghanem, B. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 45–57, 2023.

Mao, A., Mohri, M., and Zhong, Y. Cross-entropy loss functions: theoretical analysis and applications. In *Proceedings of the International Conference on Machine Learning*, pp. 23803–23828, 2023.

Nagarajan, T., Ramakrishnan, S. K., Desai, R., Hillis, J., and Grauman, K. Egoenv: Human-centric environment representations from egocentric video. In *Proceedings of the Conference on Neural Information Processing Systems*, 2023.

Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., and Keskin, C. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12999–13008, 2023.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 8024–8035, 2019.

Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen, D., and Tommasi, T. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023.

Pramanick, S., Song, Y., Nag, S., Lin, K. Q., Shah, H., Shou, M. Z., Chellappa, R., and Zhang, P. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

Radevski, G., Grujicic, D., Blaschko, M., Moens, M.-F., and Tuytelaars, T. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5213–5224, 2023.

Ramakrishnan, S. K., Al-Halah, Z., and Grauman, K. Spotem: efficient video search for episodic memory. In *Proceedings of the International Conference on Machine Learning*, pp. 28618–28636, 2023.

Shen, J., Dudley, J., and Kristensson, P. O. Encode-store-retrieve: Enhancing memory augmentation through language-encoded egocentric perception. *arXiv preprint arXiv:2308.05822*, 2023.

Shi, H., Gu, Y., Zhou, Y., Zhao, B., Gao, S., and Zhao, J. Everyone's preference changes differently: A weighted multi-interest model for retrieval. In *Proceedings of the International Conference on Machine Learning*, pp. 31228–31242, 2023.

Urooj, A., Kuehne, H., Wu, B., Chheu, K., Bousselham, W., Gan, C., Lobo, N., and Shah, M. Learning situation hypergraphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14879–14889, 2023.

Wang, H., Singh, M. K., and Torresani, L. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5250–5261, 2023a.

Wang, Y., Liu, M., Wu, J., and Nie, L. Multi-granularity interaction and integration network for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7684–7695, 2023b.

Wu, B., Yu, S., Chen, Z., Tenenbaum, J. B., and Gan, C. Star: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 1–13, 2021.

Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., and Chua, T.-S. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2804–2812, 2022.

Xu, M., Li, Y., Fu, C.-Y., Ghanem, B., Xiang, T., and Pérez-Rúa, J.-M. Where is my wallet? modeling object proposal sets for egocentric visual query localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2593–2603, 2023a.

Xu, Y., Li, Y.-L., Huang, Z., Liu, M. X., Lu, C., Tai, Y.-W., and Tang, C.-K. Egopca: A new framework for egocentric hand-object interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5273–5284, 2023b.

Xue, F., Chen, J., Sun, A., Ren, X., Zheng, Z., He, X., Chen, Y., Jiang, X., and You, Y. A study on transformer configuration and training objective. In *Proceedings of the International Conference on Machine Learning*, pp. 38913–38925, 2023.

Yan, W., Hafner, D., James, S., and Abbeel, P. Temporally consistent transformers for video generation. In *Proceedings of the International Conference on Machine Learning*, pp. 39062–39098, 2023.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1697, 2021.

Zhang, C., Gupta, A., and Zisserman, A. Helping hands: An object-aware ego-centric video recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13901–13912, 2023a.

Zhang, H., Liu, M., Gao, Z., Lei, X., Wang, Y., and Nie, L. Multimodal dialog system: Relational graph-based context-aware question understanding. In *Proceedings of the ACM International Conference on Multimedia*, pp. 695–703, 2021.

Zhang, H., Liu, M., Li, Y., Yan, M., Gao, Z., Chang, X., and Nie, L. Attribute-guided collaborative learning for partial person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14144–14160, 2023b.

Zhang, H., Liu, M., Wang, Y., Cao, D., Guan, W., and Nie, L. Uncovering hidden connections: Iterative tracking and reasoning for video-grounded dialog. *arXiv preprint arXiv:2310.07259*, 2023c.

Zhang, L., Zhou, S., Stent, S., and Shi, J. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision*, pp. 127–145, 2022.

Zhao, Z., Lin, J., Jiang, X., Cai, D., He, X., and Zhuang, Y. Video question answering via hierarchical dual-level attention network learning. In *Proceedings of the ACM International Conference on Multimedia*, pp. 1050–1058, 2017.

A. Appendix

Algorithm 1 The Pseudo Code of Our MFAS Model

```
Input  : The video $\mathcal{V}$, question $q$, and prior $\mathbf{A}_0$; The patch partition and merging module $\Theta_1$, prior-guided patch selection
         module $\Theta_2$, and hierarchical aggregation network $\Theta_3$.
Output : The answer $a$
Obtain $\mathbf{v}$ through sampling on video $\mathcal{V}$
Output $\widetilde{\mathbf{v}}_L^{sp}$ and attention scores $\mathbf{S}^p$ and $\mathbf{S}^{sp}$ by entering $\mathbf{v}$ into $\Theta_1$
Get $\mathbf{q}$ by feeding $q$ into pre-trained RoBERTa model
Feed $\mathbf{A}_0$, $\mathbf{S}^p$, and $\mathbf{S}^{sp}$ into $\Theta_2$ to get visual mask $\mathbf{A}$
Input $\widetilde{\mathbf{v}}_L^{sp}$, $\mathbf{q}$, and $\mathbf{A}$ into $\Theta_3$ to obtain $\widetilde{\mathbf{v}}_{L,R}^{sp}$ and $\mathbf{q}_R$
Output answer $a$ by decoding $\widetilde{\mathbf{v}}_{L,R}^{sp}$ and $\mathbf{q}_R$
```

Algorithm 2 The Prior-guided Patch Selection Algorithm

```
Input  : The attention scores $\mathbf{S}^p$ and $\mathbf{S}^{sp}$, the selection number $k$
Output : The mask $\mathbf{A} \in \mathbb{R}^{T \times N}$
Initialize the prior $\mathbf{A}_0$ by Eqn. (5) and $\mathbf{A} = [\,]$
Merge attention score $\mathbf{S}$ by Eqn. (6)
for $t \leftarrow 1$ to $T$ do
    $\Omega = \mathrm{index}(\mathrm{Top}\text{-}k(\mathbf{S}[t]))$
    Initialize a zero vector $\mathbf{B}$
    Generate the first-order neighbour $\Gamma$ of $\mathbf{A}_{t-1}$
    for $\tau$ in $\Omega$ do
        if $\tau$ in $\Gamma$ then
            $\mathbf{B}[\tau] = 1$
        end
    end
    $\mathbf{A}_t = \mathbf{A}_{t-1} \cup \mathbf{B}$
    $\mathbf{A} \leftarrow \mathbf{A}_t$
end
Stack and output the mask $\mathbf{A}$
```

A.1. Experimental Settings

A.1.1. Implementation Details

In our experimental setup, following the precedent set (Pramanick et al., 2023), we utilized TimeSformer-B (Bertasius et al., 2021) and RoBERTa-B (Liu et al., 2019) as the foundational backbones for video and question processing, respectively. Video inputs are standardized to a resolution of 224×224. For frame sampling, we adopted different strategies for each dataset: 16 frames (i.e., T=16) for EgoTaskQA and 32 frames (i.e., T=32) for QAEgo4D, to capture the requisite temporal granularity.

In terms of spatial granularity, the videos are partitioned into patches of size 32×32, which are further subdivided into sub-patches of size 16×16, resulting in N=196 sub-patches per frame. The model parameters are meticulously configured, with the selection threshold k set to 3, the number of attention heads M to 12, and the hidden dimension d to 768. The architecture incorporates 6 spatial-temporal attention layers (L) and an equal number of cross-attention layers (R). The balancing coefficient λ in the loss function is fixed at 2. The training regimen extends over 40 epochs with a batch size of 32. Optimization uses the AdamW optimizer (Loshchilov & Hutter, 2017), with a peak learning rate of 2e-4. To ensure computational efficiency and scalability, all experiments were conducted using the PyTorch framework (Paszke et al., 2019) on a cluster of 8 V100 GPUs.

A.1.2. Evaluation Metrics

Following (Jia et al., 2022), we adopted accuracy as the primary evaluation metric for the EgoTaskQA dataset. For the QAEgo4D dataset, we expanded our evaluation criteria to encompass more than just accuracy. Echoing the methodologies established in preceding research (Zhang et al., 2021; Bärmann & Waibel, 2022), we incorporated a comprehensive set of standard metrics typically employed in machine translation evaluation. This suite includes LEU-4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004).

A.2. Comparison under Different Questions

In the EgoTaskQA dataset, questions are categorized into four distinct types to facilitate a comprehensive analysis of models' spatial, temporal, and causal reasoning capabilities. These categories include descriptive (What is the status?), predictive (What will happen next?), explanatory (What caused this?), and counterfactual (What if?) questions. To provide a nuanced evaluation, we have detailed the performance of our MFAS model against various baselines across these question types, with results for both direct and indirect settings presented in Tables 4 and 5.

The comparative analysis reveals that our method consistently surpasses existing methods across most question types in both settings, with particularly notable performance in the "object" category. Here, MFAS demonstrates significant accuracy improvements, notably outperforming the EgoVLPv2 baseline with absolute gains of 6.38% and 3.92% under direct and indirect settings, respectively. This underscores the effectiveness of our patch partition and merging strategy in enhancing object recognition capabilities. However, it is observed that in the direct setting, our model shows a notable performance dip in the "action" category when compared to the CMCIR baseline. This could be attributed to CMCIR's utilization of more comprehensive video information and the straightforward nature of the questions that require minimal inference. Conversely, in the more challenging indirect setting, our model demonstrates superior performance, highlighting the efficacy of our hierarchical aggregation network in facilitating complex visual inference tasks.

Table 4. Comparative accuracy analysis against recent methodologies across all question types within the EgoTaskQA direct split. The highest-performing results are emphasized in bold, and the second-best results are underlined.

	Category	ClipBERT	HCRN	CMCIR	EgoVLP	EgoVLPv2	Ours	Δ
Scope	world	42.15	44.27	46.97	45.35	50.25	**52.96**	2.71↑
	intent	40.94	49.77	44.95	50.41	53.69	**56.00**	2.31↑
	multi-agent	27.63	31.36	30.03	31.90	40.64	**43.45**	2.81↑
Type	descriptive	38.45	43.48	43.79	46.12	52.19	**54.85**	2.66↑
	predictive	31.50	36.56	36.63	38.91	41.41	**44.71**	3.30↑
	counterfactual	46.75	48.00	49.14	44.47	48.16	**51.45**	2.31↑
	explanatory	42.39	40.60	**48.76**	40.22	42.36	44.35	4.41↓
Semantic	action	22.91	14.92	**30.08**	15.96	16.80	17.34	12.74↓
	object	21.80	45.31	41.72	51.47	63.87	**70.25**	6.38↑
	state	54.36	68.28	53.44	64.02	70.90	**76.37**	5.47↑
	change	66.58	67.38	66.08	69.14	72.87	**73.88**	1.01↑
Overall	open	27.70	30.23	35.49	31.69	35.56	**38.95**	3.39↑
	binary	67.52	69.42	65.92	71.26	75.60	**75.86**	0.26↑
	all	39.87	42.20	46.04	42.51	46.26	**48.69**	2.43↑

A.3. Parameter Analysis

We explored the impact of three critical parameters, the selection number k, the balance coefficient λ, and the training epoch number, on the performance of our model. The results are shown in Figure 7.

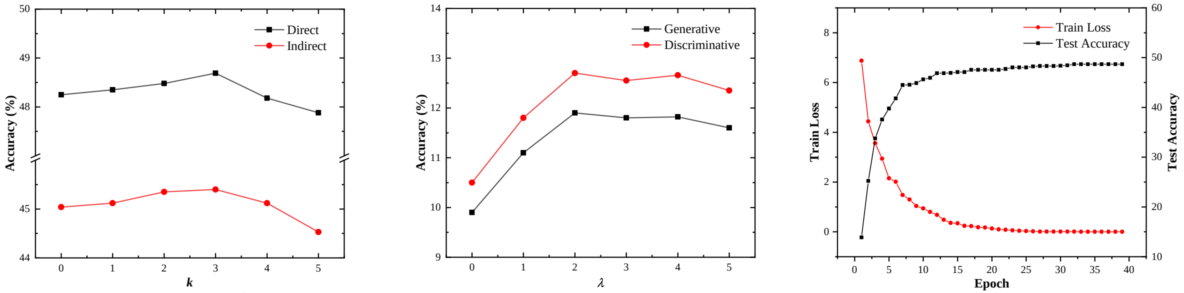
In Figure 7 (a), we present the accuracy trend as the hyperparameter k ranges from 0 to 5 within the EgoTaskQA direct split. This graph reveals an initial increase in accuracy with rising k values, peaking at k=3, before experiencing a subsequent decline. This pattern underscores the importance of an optimal selection number in maximizing model accuracy.

The influence of the balance coefficient λ was similarly assessed, with its values explored between 0 and 5 in the context of the QAEgo4D dataset. As depicted in Figure 7 (b), the accuracy of our model achieves its zenith at λ=2, highlighting the critical role of this parameter in attaining peak performance.

Table 5. Comparison of accuracy for various question types in the EgoTaskQA indirect split against contemporary approaches. The best results are highlighted in bold and the second in underlined.

	Category	ClipBERT	HCRN	CMCIR	EgoVLP	EgoVLPv2	Ours	Δ
Scope	world	26.51	44.04	41.55	41.45	<u>44.90</u>	**48.62**	3.72↑
	intent	14.66	**47.02**	38.13	33.61	40.48	<u>42.55</u>	4.47↓
	multi-agent	20.09	30.11	**40.36**	29.06	<u>32.24</u>	31.21	9.15↓
Type	descriptive	24.35	42.02	41.36	40.30	<u>45.84</u>	**48.79**	2.95↑
	predictive	10.32	<u>46.32</u>	35.38	22.61	43.69	**46.74**	0.42↑
	counterfactual	26.29	<u>43.64</u>	**44.80**	37.70	38.94	41.43	3.37↓
	explanatory	22.46	**39.69**	38.04	35.91	<u>39.10</u>	39.06	0.63↓
Semantic	action	25.25	29.61	38.47	<u>29.71</u>	29.09	**30.13**	0.42↑
	object	10.49	32.20	36.85	32.94	<u>40.19</u>	**44.11**	3.92↑
	state	15.29	<u>41.81</u>	39.79	36.52	41.69	**44.63**	2.82↑
	change	35.26	56.27	48.76	51.84	<u>56.38</u>	**60.78**	4.40↑
Overall	open	11.17	27.82	28.36	27.04	<u>29.14</u>	**32.44**	3.30↑
	binary	40.71	59.29	58.86	55.28	<u>59.68</u>	**63.02**	3.34↑
	all	24.08	41.56	42.00	38.69	<u>42.28</u>	**45.40**	3.12↑

Additionally, Figure 7 (c) showcases the evolution of training loss and test accuracy for 1 to 40 epochs. The depicted trends indicate a consistent reduction in loss and a corresponding enhancement in accuracy, culminating in model convergence. These observations collectively provide valuable insights into the parameter sensitivities and training dynamics of our model.



(a) Parameter k (b) Parameter λ (c) Epoch

Figure 7. Parameter Analysis for (a) Parameter k, (b) Parameter λ, and (c) Epoch.

A.4. More Case Study

We provided two illustrative examples for each scenario across both the QAEgo4D and EgoTaskQA datasets, depicted in Figure 8 to 11. These examples elucidate the following insights:

- Discriminative cases from QAEgo4D. Figure 8 showcases instances where our model adeptly navigates the question and video context to identify suitable answers from available options. The first case illustrates the proficiency of our model in context comprehension, albeit with some limitations due to question ambiguity. The second case highlights the challenges of information loss due to sampling, which occasionally hinders accurate answer retrieval.

- Generative cases from QAEgo4D. In Figure 9, the first example demonstrates the capability of our model to synthesize a correct response by integrating multimodal information. The second example reveals an instance of potential mislabeling in the dataset, where our model, interestingly, generates a coherent and contextually appropriate answer.

- Direct cases from EgoTaskQA. As depicted in Figure 10, the first example underscores the ability of our model to decode complex queries and align its responses with the ground truth. The second example, however, suggests a partial comprehension of the query by our model, reflected in the similarity of its response to the expected answer.

- Indirect cases from EgoTaskQA. In Figure 11, the first scenario indicates a misinterpretation of temporal frame relationships by our model, leading to an incorrect answer. Conversely, the second scenario illustrates the adeptness of our model at handling intricate questions and selecting accurate answers.

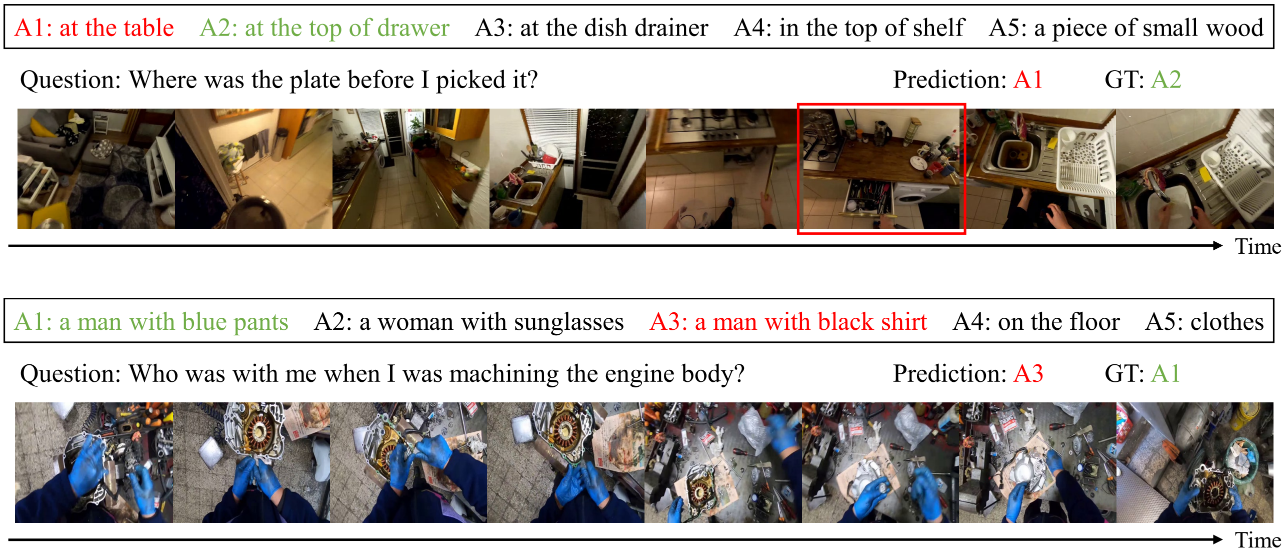
These eight cases collectively offer a detailed perspective on the performance of our model, showcasing its strengths and areas for improvement. Despite the inherent challenges posed by the EgoTaskQA and QAEgo4D datasets, complex questions in the former and extended video lengths in the latter, our model demonstrates commendable efficacy and adaptability.

A.5. More Visualization

Due to spatial constraints in the main manuscript, we provide an expanded visualization in Figure 12, which includes both attention maps and heatmaps to elucidate the operation of our prior-guided patch selection module.

The attention maps (2nd row) and heatmaps (3rd row) in Figure 12 reveal that the standard attention mechanism often erroneously emphasizes irrelevant areas, introducing noise that can obscure the semantic understanding of the video. In contrast, our prior-guided patch selection module adeptly incorporates predefined priors with spatial and temporal data from consecutive frames, refined by local connectivity considerations. This results in more focused attention maps (4th row), directing our model's focus towards pivotal visual elements and thereby bolstering video comprehension.

Additionally, the heatmaps for selected areas (5th row) illustrate a notable observation: in the initial frame (Frame 1), the areas of focus are relatively stable, but as the sequence progresses, these regions adapt dynamically, reflecting the integration of evolving contextual information. This dynamic adjustment underscores the capability of our module to respond to changing visual cues in egocentric videos, showcasing its utility in enhancing the model's interpretative performance.


A1: at the table A2: at the top of drawer A3: at the dish drainer A4: in the top of shelf A5: a piece of small wood

Question: Where was the plate before I picked it? Prediction: A1 GT: A2

A1: a man with blue pants A2: a woman with sunglasses A3: a man with black shirt A4: on the floor A5: clothes

Question: Who was with me when I was machining the engine body? Prediction: A3 GT: A1

Figure 8. The two discriminative cases from QAEgo4D dataset.

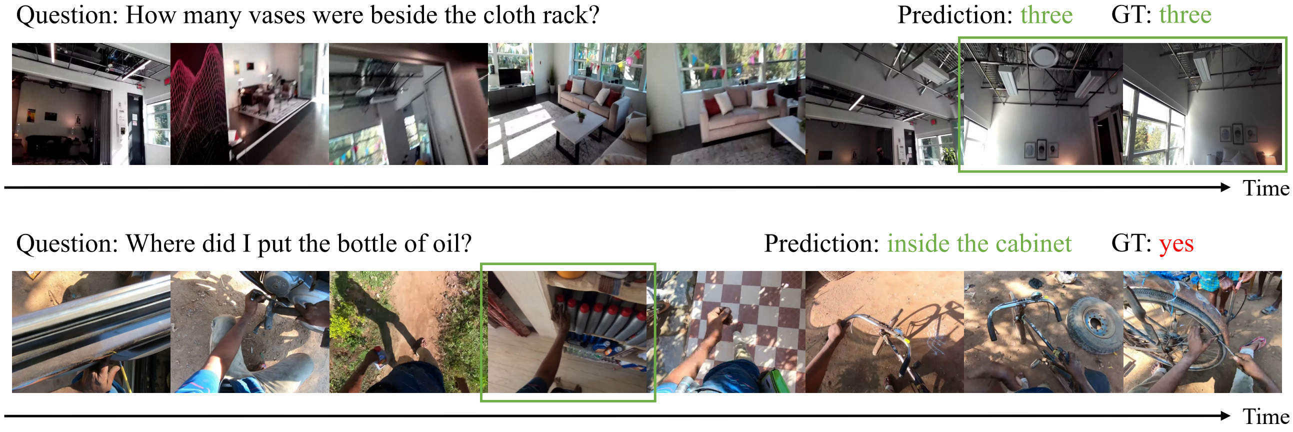

Question: How many vases were beside the cloth rack?
Prediction: three
GT: three
Time
Question: Where did I put the bottle of oil?
Prediction: inside the cabinet
GT: yes
Time

Figure 9. The two generative cases from QAEgo4D dataset.

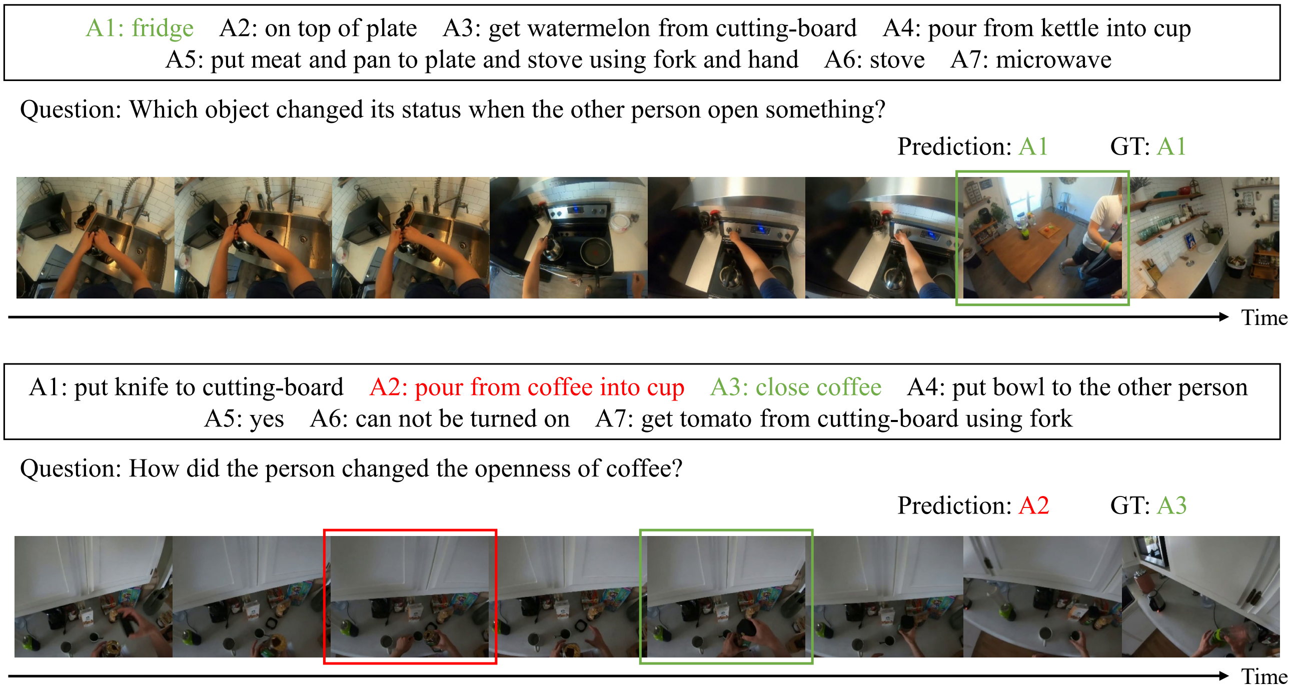

A1: fridge
A2: on top of plate
A3: get watermelon from cutting-board
A4: pour from kettle into cup
A5: put meat and pan to plate and stove using fork and hand
A6: stove
A7: microwave
Question: Which object changed its status when the other person open something?
Prediction: A1
GT: A1
Time
A1: put knife to cutting-board
A2: pour from coffee into cup
A3: close coffee
A4: put bowl to the other person
A5: yes
A6: can not be turned on
A7: get tomato from cutting-board using fork
Question: How did the person changed the openness of coffee?
Prediction: A2
GT: A3
Time

Figure 10. The two discriminative cases from EgoTaskQA direct split.

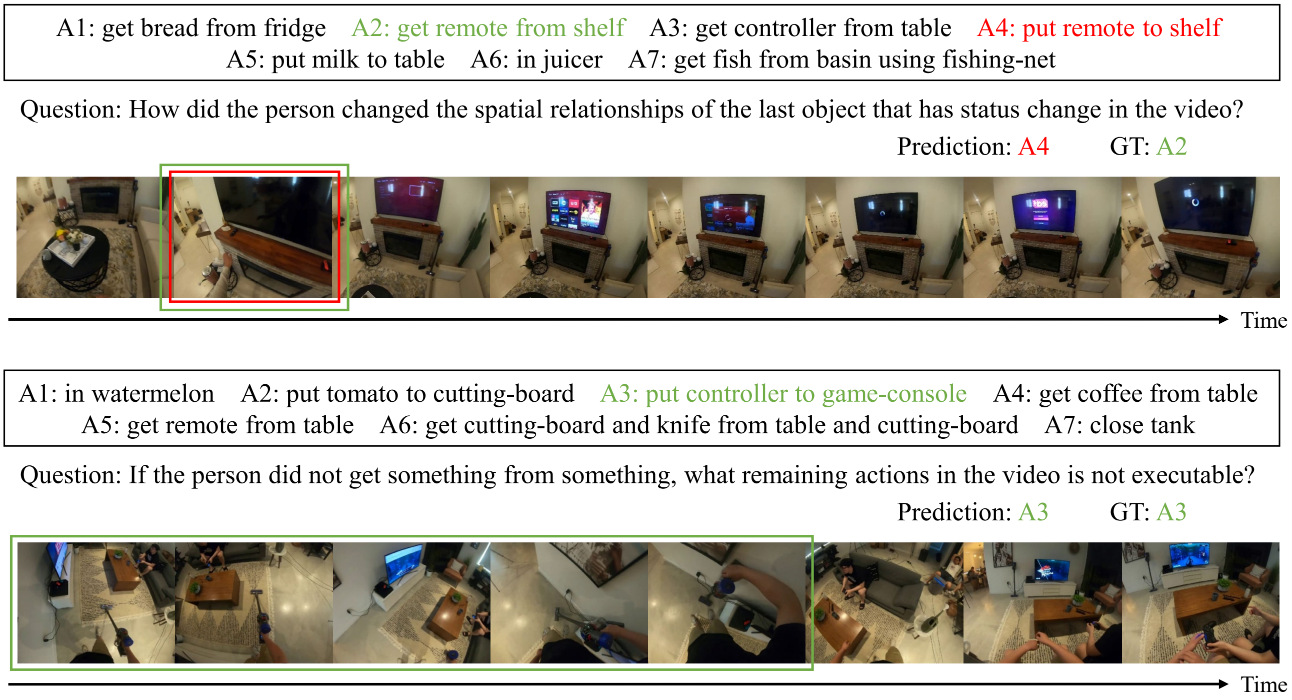

A1: get bread from fridge
A2: get remote from shelf
A3: get controller from table
A4: put remote to shelf
A5: put milk to table
A6: in juicer
A7: get fish from basin using fishing-net
Question: How did the person changed the spatial relationships of the last object that has status change in the video?
Prediction: A4
GT: A2
Time
A1: in watermelon
A2: put tomato to cutting-board
A3: put controller to game-console
A4: get coffee from table
A5: get remote from table
A6: get cutting-board and knife from table and cutting-board
A7: close tank
Question: If the person did not get something from something, what remaining actions in the video is not executable?
Prediction: A3
GT: A3
Time

Figure 11. The two discriminative cases from EgoTaskQA indirect split.

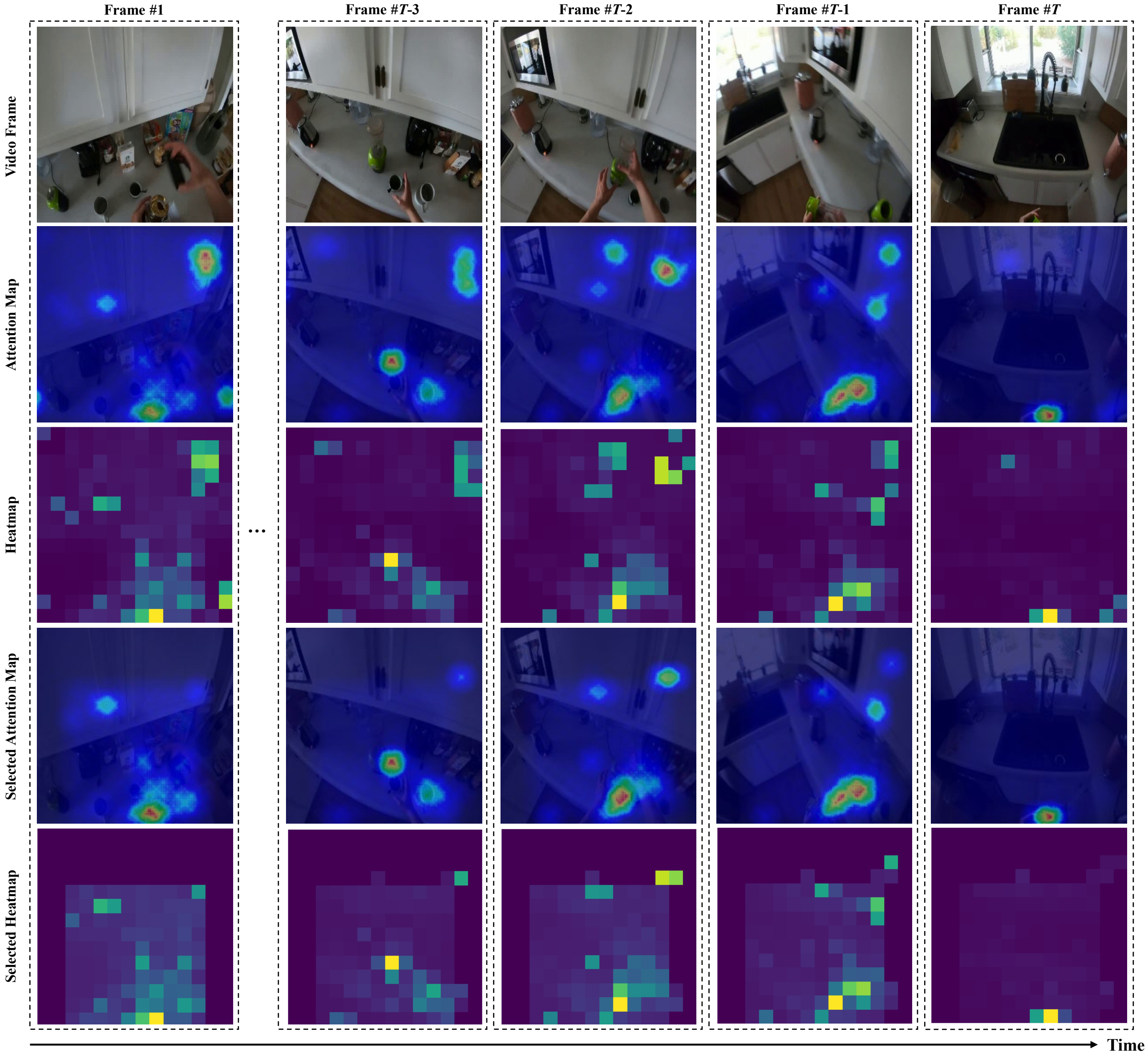


Figure 12. Visualization of our prior-guided patch selection process. “Heatmap” illustrates the initial distribution of attention across sub-patches, and “Selected Heatmap” showcases the refined attention post-prior-guided selection. “Attention Map” and “Selected Attention Map” provide a more intuitive depiction of these distributions.