mathematics

MDPI

Article

ETR: Event-Centric Temporal Reasoning for Question-Conditioned Video Question Answering

Lingmin Pan [1], Ziyi Gao [1], Yueming Zhu [1], Fuchen Chen [1], Chengyuan Zhang [2], Dan Yin [2], Yong Cai [2], Siqiao Tan [1,*] and Lei Zhu [1,*]

[1] College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China; lingminpan@stu.hunau.edu.cn (L.P.); ziyigao@stu.hunau.edu.cn (Z.G.); zhuyueming@stu.hunau.edu.cn (Y.Z.); fuchen1217@hunau.edu.cn (F.C.)

[2] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; cyzhangcse@hnu.edu.cn (C.Z.); yindan@hnu.edu.cn (D.Y.); caiyong911@hnu.edu.cn (Y.C.)

* Correspondence: tsq@hunau.edu.cn (S.T.); leizhu@hunau.edu.cn (L.Z.)

Abstract

Video Question Answering (VideoQA) requires a deep understanding of dynamic video content, integrating spatial reasoning, temporal dependencies, and language comprehension. Existing methods often struggle with long or semantically complex videos due to the lack of question-guided keyframe weight adjustment and the absence of question-aligned cross-modal description generation. To address these challenges, we propose ETR (Event-centric Temporal Reasoning), an adaptive framework for VideoQA. ETR introduces three key mechanisms: (i) a hierarchical weight adjustment selector to identify questions requiring event-centric temporal reasoning; (ii) a T-Route that segments videos into semantically coherent events and dynamically adjusts keyframe weights with question intent; and (iii) a question-conditioned prompting strategy that focuses on key objects to generate textual prompts aligned with a question's semantics. This hierarchical and adaptive design effectively balances visual and textual information, enhances temporal reasoning, and improves object-centric alignment. Experiments on two datasets demonstrate that ETR achieves competitive performance in fine question-aware VideoQA.

Keywords: videoquestion answering; event-centric temporal reasoning; visual-semantic alignment

MSC: 68T45

Academic Editor: Florin Leon

Received: 28 January 2026
Revised: 27 February 2026
Accepted: 5 March 2026
Published: 7 March 2026

1. Introduction

Video Question Answering (VideoQA) [1–3] is a fundamental multi-modal task in artificial intelligence [4–6] that integrates computer vision [7–10], natural language understanding, and temporal reasoning to generate accurate textual answers from dynamic video content. Unlike image-based Visual Question Answering (VQA) [11,12], which operates on static visual inputs, VideoQA requires the joint modeling of spatial semantics (e.g., objects, scenes, and interactions) and temporal dynamics (e.g., action evolution, event progression, and causal relations). The necessity to capture long-form temporal dependencies [13] and continuously evolving visual patterns makes VideoQA particularly challenging for understanding complex real-world scenarios [14–16], while also closely aligning it with practical applications that demand fine spatio-temporal reasoning.

Rather than solely relying on increasingly powerful model architectures, a growing body of research [17–19] has focused on enhancing temporal reasoning capabilities and improving cross-modal fusion mechanisms to advance VideoQA. Despite these efforts, existing methods continue to face critical limitations: they struggle to accurately capture subtle and continuous temporal dynamics, lack effective mechanisms to suppress redundant or irrelevant visual information, and exhibit limited generalization across diverse video types and question contexts. These challenges become particularly pronounced in long-form VideoQA, where complex spatio-temporal structures significantly increase reasoning difficulty and introduce severe information redundancy.

Such redundancy mainly manifests in two forms: visual overload [20], caused by densely sampled frame sequences that are weakly related to the question semantics, and textual overload, arising from generated descriptions containing substantial redundant or marginally relevant information. To mitigate visual redundancy, prior work explores keyframe selection using pre-trained large models [21] or event-level modeling strategies [4,22] that segment videos into semantically coherent units [23,24]. However, whether at the frame or event level, most existing approaches assign similar importance to all selected frames, making it difficult to reflect the true contributions during reasoning. As illustrated in Figure 1, even when key frames are correctly identified and semantically relevant descriptions are generated, nearly uniform weighting prevents the suppression of redundant or weakly relevant information, ultimately impairing the final prediction accuracy. Furthermore, language-based augmentation using large language models (LLMs) [25–27] introduces richer semantics but also exacerbates textual overload. Although recent studies [28–30] attempt to alleviate this issue through question-guided captioning or multi-granularity retrieval, effectively balancing visual and textual information under redundancy constraints remains an open and fundamental challenge for long-form VideoQA.

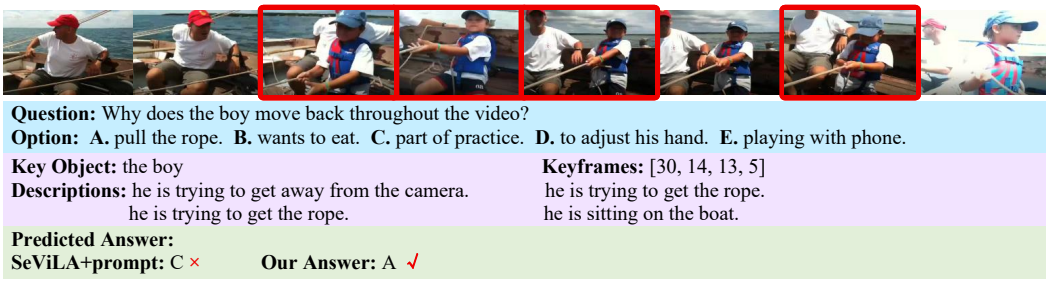

Figure 1. An example illustrating different weighting schemes. We just show parts of frames of the video, and the frames are not strictly consecutive to show a rough video development. The red boxes highlight frames that the models focus on.

Motivation. The unresolved challenges discussed above can be distilled into two closely related issues, which jointly motivate this study. First, existing VideoQA methods [21] typically perform visual localization over entire videos and assign nearly uniform importance to all selected keyframes without accounting for the specific intent of the question. Such question-agnostic weighting inevitably introduces substantial irrelevant visual information, thereby degrading both reasoning efficiency and prediction accuracy. In practice, however, the relevance and importance of visual frames vary significantly depending on the question, highlighting the necessity of adaptive mechanisms that can selectively adjust visual importance in a question-driven manner. This observation further implies the need to identify questions that require event-centric temporal reasoning and to perform adaptive reasoning over events along their temporal progression. Although prior studies [4,22,31] have explored event segmentation to locate question-relevant moments, they largely remain at the level of event localization and fail to explicitly model the internal

temporal evolution and relative importance of events during reasoning. Second, to compensate for limited visual perception, many recent frameworks incorporate auxiliary textual descriptions generated by language models; however, the effectiveness of such descriptions depends not only on semantic relevance but also on the granularity of the conveyed information. While question-guided or program-driven captioning approaches [28] introduce partial flexibility, most existing methods still adopt a one-size-fits-all generation strategy, which prevents fine questions targeting specific objects (e.g., "What is the boy wearing yellow doing?") from facilitating alignment with truly critical visual entities. This mismatch weakens the alignment between textual cues and visual semantics, ultimately limiting reasoning performance. Taken together, these issues reveal a fundamental deficiency in current VideoQA frameworks: the lack of selective, question-conditioned modeling that jointly governs visual importance estimation and textual granularity control across modalities. Motivated by this observation, we seek to answer the following research questions:

- ***RQ1:*** *How does question intent underlie the need for fine event-level reasoning?*
- ***RQ2:*** *How can the relative importance of keyframes be effectively modeled by reasoning over the temporal evolution of events?*
- ***RQ3:*** *How can the generation of object-centric textual descriptions enable precise alignment between visual content and semantic cues?*

Our Method. To address the above research questions, we propose ETR—an Event-centric Temporal Reasoning framework designed to enable precise temporal reasoning and object-centric visual-semantic alignment for VideoQA. At a high level, ETR adopts a hierarchical and question-conditioned design to selectively regulate information flow across temporal and semantic dimensions. To address ***RQ1***, ETR introduces a hierarchical weight-adjustment selector that characterizes question intent using a composite set of indicators, such as question content, the presence of key objects, and frame-question semantic relevance. Based on this characterization, each question is adaptively routed to one of several reasoning routes, ensuring that questions requiring event-centric temporal reasoning are accurately identified. For ***RQ2***, ETR incorporates a novel question-guided fine temporal reasoning route including a two-stage clustering strategy (T-Route) through which videos are adaptively segmented into events and aligned with question semantics to identify the most relevant key events. These events are further decomposed along their temporal progression, and the importance of keyframes is dynamically adjusted in a question-aware manner while accounting for video dynamics, thereby enabling higher-order temporal reasoning. To address ***RQ3***, ETR proposes a question-conditioned prompting strategy that explicitly focuses on key objects. By incorporating object cues extracted from the question into the prompt construction process, ETR guides large language models to generate object-centric descriptions that are more precisely aligned with the underlying question semantics. Through this unified and adaptive design, ETR enables event-centric temporal reasoning and object-centric semantic grounding, providing an efficient and semantically robust solution for VideoQA.

Contributions. Our contributions are summarized below:

- **General Advancements:** We introduce ETR, a novel VideoQA approach that hierarchically and adaptively improves vision–language alignment.
- **Innovative Methodologies:** We propose a hierarchical weight adjustment module alongside a question-guided fine event understanding route, enabling selective weighting of keyframes with question intention. Furthermore, we introduce a novel object-centric description prompting mechanism that generates descriptions centered on key objects of the questions. Together, these components allow the VideoQA model to adaptively leverage the most relevant visual and textual representations for each given question.

- **Comprehensive Experiments:** We conduct extensive performance evaluation and analysis on two commonly used datasets, which demonstrate that our ETR method achieves competitive performance.

Organization. The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the problem formulation and the details of the ETR method. Experimental results are reported in Section 4, limitations are discussed in Section 5, and conclusions are drawn in Section 6.

2. Related Work

Research in VideoQA can be broadly categorized into two settings: normal VideoQA, which focuses on core vision–language reasoning over relatively short video clips, and long-form VideoQA, which addresses the challenges introduced by substantial redundancy and temporally dispersed evidence in extended video sequences. In the following, we review representative works in both directions, with an emphasis on their limitations and how they motivate the design of our framework.

2.1. Normal VideoQA

Early VideoQA methods primarily relied on spatial and temporal attention mechanisms [32,33], often combined with hierarchical architectures [34], to capture localized visual information. However, these approaches were constrained by the limited capacity of RNN-based models to capture long-range temporal dependencies. To alleviate this issue, memory-augmented networks [35,36] were introduced to retain sequential cues and integrate motion and appearance features [37,38], leading to notable improvements on movie-level QA tasks. The advent of Transformer architectures further advanced VideoQA by enabling large-scale video–text pretraining [39–41] and establishing new performance benchmarks. Nevertheless, these models still exhibit limitations in compositional and multi-hop reasoning scenarios [42]. To address these challenges, subsequent studies have explored a variety of enhanced reasoning mechanisms, including graph neural networks for relational and causal reasoning [43–46], modular networks for compositional reasoning [47,48], and neural-symbolic frameworks for structured reasoning [49,50]. Additional efforts incorporate causal inference [51], data-efficient pretraining [52], and bitstream-level modeling [53] to improve robustness and generalization. Despite these advances, existing methods continue to face challenges in fine semantic alignment, interpretable reasoning, and robust generalization across diverse question types, highlighting the need for adaptive architectures that can dynamically adjust reasoning strategies according to question intent.

2.2. Long-Form VideoQA

Long-form VideoQA introduces substantially greater challenges due to extensive visual redundancy, sparsely distributed key evidence, and highly dispersed semantic cues. A prominent line of research therefore focuses on keyframe selection to suppress irrelevant content [21,42,54,55]. Existing approaches include supervised keyframe selection [42], modality-aware filtering [54], diversity-driven selection strategies [55], and more recently, the use of large vision–language models (LVLMs) as adaptive localizers [21]. Beyond frame-level selection, several studies further shift toward event-level modeling by introducing temporal segmentation [4,22,56,57] or exploiting semantic-aware event boundaries [31,58] to construct more structured temporal representations. However, most existing methods primarily emphasize improving localization accuracy while overlooking the fact that even selected keyframes or events may contribute unequally to reasoning under different questions. This observation suggests that effective long-form VideoQA requires not only

accurate localization but also question-conditioned importance modeling over temporally evolving events.

In addition to visual modeling, textual descriptions play a crucial role in bridging semantic gaps between vision and language. Segment-level description generation methods [28,59–61], particularly those incorporating question-aware generation strategies [62,63], have demonstrated promising improvements in semantic alignment and reasoning consistency. Nevertheless, effectively controlling the granularity and focus of generated descriptions remains an open challenge, especially for fine questions targeting specific objects. These limitations collectively motivate approaches that jointly consider temporal importance estimation and object-centric semantic alignment in long-form VideoQA.

3. Methodology

To address these challenges, we introduce an effective and efficient VideoQA framework. This section begins with a description of the notations and problem formulation in Section 3.1, followed by an overview of the complete architecture in Section 3.2. Sections 3.3–3.5 subsequently provide detailed explanations of the key technical components.

3.1. Preliminary

Notations. We denote sets using calligraphic uppercase letters (e.g., $\mathscr{O}$), scalars using Roman letters (e.g., N, n), vectors using bold lowercase letters (e.g., $\boldsymbol{x}$), and matrices using bold uppercase letters (e.g., $\boldsymbol{X}$). The transpose of a matrix is denoted by $\boldsymbol{X}^{\mathrm{T}}$, and the inner product between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is written as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. The Euclidean norm is represented by $|\cdot|$. Models are denoted using typewriter-style uppercase letters, such as $\mathcal{M}$.

Definition 1. *Given a video $\boldsymbol{V}$ consisting of N frames $\{\boldsymbol{f}_i\}_{i=1}^{N}$ and a natural-language question $\boldsymbol{Q}$ composed of W tokens $\{\boldsymbol{\omega}_i\}_{i=1}^{W}$, the objective of VideoQA is to predict the most probable answer $\hat{\boldsymbol{a}}$ from a candidate answer set $\mathscr{A}$. Formally, a VideoQA model $\mathcal{M}$ estimates the conditional probability distribution $p(\boldsymbol{a} \mid \boldsymbol{V}, \boldsymbol{Q})$, and the final prediction is obtained as:*

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a} \in \mathscr{A}} p(\boldsymbol{a} \mid \boldsymbol{V}, \boldsymbol{Q}). \tag{1}$$

To facilitate precise vision–language alignment under complex temporal dynamics, we formalize the reasoning process of ETR as a hierarchical optimization problem that instantiates the conditional probability $p(\boldsymbol{a} \mid \boldsymbol{V}, \boldsymbol{Q})$ in Equation (1). Given the input video $\boldsymbol{V} = \{\boldsymbol{f}_i\}_{i=1}^{N}$ and question $\boldsymbol{Q}$, the model first selects K keyframes based on their relevance to the question:

$$\{\boldsymbol{f}_i^*\}_{i=1}^{K} = \textit{Top-K}\Big(\{\boldsymbol{f}_i\}_{i=1}^{N},\ \mathcal{S}(\boldsymbol{f}_i, \boldsymbol{Q})\Big), \tag{2}$$

where $\mathcal{S}(\boldsymbol{f}_i, \boldsymbol{Q})$ is the frame-question relevance scoring function defined in Section 3.3.2. The hierarchical selector $\mathcal{H}$ then routes each question to an appropriate reasoning path based on the question type $\boldsymbol{T}$, the presence of key objects $\{\boldsymbol{O}_i\}_{i=1}^{O}$, and the relevance scores $\{\mathcal{S}_i\}_{i=1}^{K}$:

$$\mathcal{R} = \mathcal{H}(\boldsymbol{Q}, \{\boldsymbol{O}_i\}, \{\mathcal{S}_i\}) \in \{\textit{T-Route}, \textit{O-Route}, \textit{N-Route}\}, \tag{3}$$

where $\mathcal{R}$ denotes the selected reasoning route. For questions routed to the T-Route, the video is further segmented into events via DBSCAN clustering:

$$\{\boldsymbol{E}_l\}_{l=1}^{L} = \mathcal{C}_{DBSCAN}\Big(\{\boldsymbol{f}_i\}_{i=1}^{N}\Big), \tag{4}$$

and the most question-relevant event $\boldsymbol{E}_j^*$ *is selected by maximizing semantic alignment:*

$$\boldsymbol{E}_j^* = \arg\max_{\boldsymbol{E}_l} Sim(\mathcal{D}_l, \boldsymbol{Q}), \tag{5}$$

where $\mathcal{D}_l$ *are the textual descriptions of event* $\boldsymbol{E}_l$ *generated by Equation (13) in Section 3.4.2. Fine-grained temporal clustering is then applied to obtain refined frame representations:*

$$\{\boldsymbol{f}_i^+\}_{i=1}^K = \mathcal{C}_{k\text{-}means}\left(\boldsymbol{E}_j^*, k = 3\right). \tag{6}$$

For all routes, question-guided object-centric descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ *are generated following the prompt strategies detailed in Section 3.4.2. Finally, the multi-modal representations are fused and fed into the answer predictor* $\mathcal{C}$ *to produce the final answer:*

$$\hat{\boldsymbol{a}} = \mathcal{C}\left(\{\boldsymbol{f}_i^+\}_{i=1}^K \bowtie \{\boldsymbol{D}_i\}_{i=1}^K, \boldsymbol{Q}, \mathscr{A}\right), \tag{7}$$

where $\bowtie$ *denotes feature concatenation. This hierarchical formulation provides a concrete instantiation of Equation (1) by decomposing the conditional probability* $p(\boldsymbol{a} \mid \boldsymbol{V}, \boldsymbol{Q})$ *into a series of interpretable sub-processes: keyframe selection (Equation (2)), adaptive routing (Equation (3)), event segmentation (Equations (4) and (5)), fine-grained clustering (Equation (6)), and multi-modal fusion (Equation (7)). Each sub-process contributes to capturing both coarse-grained event semantics and fine-grained temporal dynamics in a question-adaptive manner.*

3.2. Overview of ETR

As illustrated in Figure 2, the proposed ETR is a unified framework designed for adaptive temporal reasoning and object-centric vision–language alignment in VideoQA. It consists of three sequential modules: the initial processing module Section 3.3, the hierarchical weight adjustment module Section 3.4, and the answer module Section 3.5. Each module is explicitly designed to address a distinct stage of the VideoQA reasoning pipeline, while collectively enabling event-centric temporal reasoning conditioned on question intent.

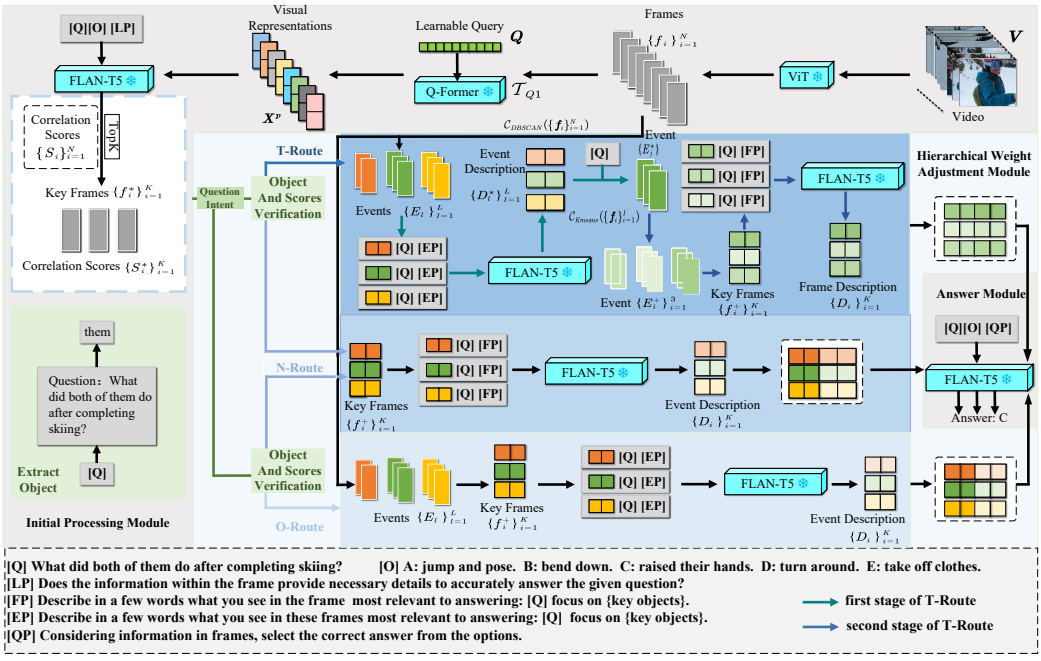

Figure 2. Overview of the ETR framework.

Initial Processing Module. Given an input video $\boldsymbol{V}$, this module first identifies the top-K keyframes $\{\boldsymbol{f}_i^*\}_{i=1}^K$ and estimates their frame-question relevance scores $\{\boldsymbol{S}_i\}_{i=1}^K$. Keyframe identification is performed via a cross-modal matching mechanism that jointly employs a Q-Former and a pre-trained LLM to evaluate the semantic alignment between each frame and the question $\boldsymbol{Q}$. Meanwhile, a key object extractor identifies the key objects from the question, providing object-centric semantic cues for subsequent adaptive reasoning.

Hierarchical Weight Adjustment Module. Based on the inferred question intent, this module selectively performs event-centric temporal reasoning. At its core lies a hierarchical keyframe router $\mathcal{H}$, which assigns each question to one of three reasoning routes according to a composite set of semantic indicators, including question type, the presence of key objects $\{\boldsymbol{O}_i\}_{i=1}^O$, and frame-question relevance scores $\{\boldsymbol{S}_i\}_{i=1}^K$. This routing mechanism enables ETR to dynamically allocate computational and semantic alignment to different levels of temporal granularity. The routing process is defined as follows:

- For questions that require fine event understanding, detailed temporal reasoning is necessary. When no key objects are detected and the frame-question relevance scores are low, indicating insufficient capture of question-specific details, the input is routed to the T-Route. In this route, key events $\{\boldsymbol{E}_j^*\}$ are further analyzed according to their temporal progression to extract fine event-level information. Otherwise, the input is routed to the N-Route, where the keyframes $\{\boldsymbol{f}_i^+\}_{i=1}^K$ and their corresponding descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ are directly utilized for answering.
- For other questions, broader contextual reasoning may be required. When no key objects are detected and the frame-question relevance scores remain low, the input is routed to the O-Route to incorporate additional contextual information. Otherwise, it is processed by the N-Route, which directly employs the selected keyframes $\{\boldsymbol{f}_i^+\}_{i=1}^K$ and descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ for answer prediction.

Answer Module. In the final stage, the selected keyframes $\{\boldsymbol{f}_i^+\}_{i=1}^K$ and the question-guided object-centric descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ are fed into an LLM-based answer predictor $\mathcal{C}$ to generate the final answer $\hat{\boldsymbol{a}}$.

3.3. Initial Processing Module

The initial processing stage identifies preliminary keyframes based on the question $\boldsymbol{Q}$, effectively mitigating the risk of visual information overload. This stage is composed of three modules: a multi-modal encoder, a keyframe selector, and a key object extractor.

3.3.1. Multi-Modal Encoder

Serving as the initial module of the first processing stage, the multi-modal encoder is designed to extract semantically rich representations from both visual and textual modalities. It consists of two parallel branches: a visual encoder that generates the visual embedding for each frame $\boldsymbol{f}_i$, and a textual encoder that produces the corresponding token embeddings $\boldsymbol{\omega}_i$. In our framework, the visual encoder is instantiated using a pre-trained ViT model to capture spatial visual semantics, while the textual encoder is implemented with a pre-trained Flan-T5-XL model to encode linguistic context.

3.3.2. Keyframe Selector

To capture frame-level semantics in a question-aware manner, we introduce a keyframe selector that identifies a set of keyframes $\{\boldsymbol{f}_i^*\}_{i=1}^K$ based on their semantic relevance to the input question $\boldsymbol{Q}$. Given the input frames $\{\boldsymbol{f}_i\}_{i=1}^N$, the selector aims to suppress redundant or weakly related visual content while preserving frames that are most informative for subsequent reasoning. The selection process is performed in two stages: (i) Relevance Scoring, which quantifies the semantic alignment between each frame and the

question; and (ii) Top-K Keyframe Selection, which retrieves the K frames with the highest relevance scores.

Relevance Scoring. For the set of input frames $\{\boldsymbol{f}_i\}_{i=1}^{N}$, a Q-Former $\mathcal{T}_{\mathrm{Q}}$ projects a set of learnable query embeddings $\boldsymbol{\Omega}$ into visual representations $\boldsymbol{X}^{\mathrm{ip}}$ (where the superscript "ip" denotes the initial processing module), calculated as:

$$\{\boldsymbol{X}_i^{\mathrm{ip}}\}_{i=1}^{N} = \mathcal{T}_{\mathrm{Q}}\Big(\{\boldsymbol{f}_i\}_{i=1}^{N}, \boldsymbol{\Omega}\Big). \tag{8}$$

These representations $\{\boldsymbol{X}_i^{\mathrm{ip}}\}_{i=1}^{N}$, together with the question $\boldsymbol{Q}$, the candidate answer set $\mathscr{A}$, and the localization prompt [LP], are fed into the pre-trained Flan-T5-XL model $\mathcal{T}_5$ to compute frame-question relevance scores:

$$\{S_i\}_{i=1}^{N} = \mathcal{T}_5\Big(\{\boldsymbol{X}_i^{\mathrm{ip}}\}_{i=1}^{N}, \boldsymbol{Q}, \mathscr{A}, [\mathrm{LP}]\Big). \tag{9}$$

Following [21], the prompt [LP] is defined as: "Does the information within the frame provide necessary details to accurately answer the given question?"

Top-K Keyframe Selection. Using the frame-level relevance scores $\{\boldsymbol{S}_i\}_{i=1}^{N}$, the top-$K$ frames $\{\boldsymbol{f}_i^*\}_{i=1}^{K}$ are selected according to their relevance ranking. The chosen frame-score pairs, denoted as $\{(\boldsymbol{f}_i^*, S_i^*)\}_{i=1}^{K}$, are then forwarded to the hierarchical weight adjustment module for further refinement and generation of corresponding descriptions.

3.3.3. Key Object Extractor

To robustly identify the most salient entities referenced in the question and provide a reliable foundation for subsequent visual attention guidance, we introduce a key object extractor. This module unifies multiple complementary linguistic cues to stably extract question-centric objects across diverse syntactic structures, thereby facilitating interpretable vision–language alignment. The extractor is implemented as a spaCy-based natural language processing procedure that analyzes the original question $\boldsymbol{Q} = \{\omega_i\}_{i=1}^{W}$ and outputs a set of key objects $\{O_i\}_{i=1}^{O}$.

The extraction pipeline begins with question preprocessing, which removes answer option interference (e.g., descriptions for A, B, and C), converts the text to lowercase, and preserves the question mark to retain interrogative intent. The processed question is then analyzed using dependency parsing and part-of-speech tagging. To comprehensively capture potential object mentions, multiple complementary strategies are employed to collect candidate noun phrases, including subject dependencies (nsubj/nsubjpass), attributes following copular verbs (attr), nouns associated with interrogative words (e.g., what, which, who, whose), and nouns preceding auxiliary verbs (e.g., do, does, did). For each candidate head token, complete noun phrases are assembled by incorporating surrounding modifiers, such as adjectives, compound words, determiners, and possessive markers. Subsequently, invalid candidates are filtered out, including phrases containing stop words (e.g., what, which, question, answer) or those consisting solely of single-letter answer options. This filtering process effectively suppresses spurious or non-informative entities that could otherwise mislead downstream visual attention and routing decisions. The final output is a sorted list of key objects concatenated using the conjunction "and" (e.g., "man and car"), or an empty string if no valid objects are detected.

3.4. Hierarchical Weight Adjustment Module

To enable selective and event-centric temporal reasoning, we design a hierarchical weight adjustment module that explicitly adapts temporal and semantic importance modeling to the intent of the question. This module is composed of three key components:

a hierarchical weight adjustment selector, a hierarchical processor, and a multi-modal fusion module.

3.4.1. Hierarchical Weight Adjustment Selector

To selectively enable event-centric temporal reasoning, we design a hierarchical weight adjustment selector that systematically routes questions into different processing pathways and applies corresponding visual localization calibration strategies. By explicitly decoupling high-level question intent recognition from low-level confidence-aware visual assessment, the selector facilitates adaptive and question-driven semantic modeling.

The selector operates in two successive stages. The first stage, referred to as the problem type selector, performs a coarse-grained analysis to distinguish questions that require detailed information and fine event reasoning from those that do not. Conditioned on this categorization, the second stage adopts a composite selector that jointly considers the presence of key objects and the frame-question relevance scores to determine the most appropriate processing route. Specifically, for questions identified as requiring detailed information, the composite selector further separates them into two branches based on confidence and object cues: one branch corresponds to cases where either (i) there exists at least one frame with low relevance and the boolean flag *non_key_object* is True (i.e., the object set from the Key Object Extractor is empty), or (ii) all frames have low relevance even when the boolean flag *non_key_object* is False; the other branch includes the remaining questions. This decision process is formalized as the following routing function:

$$\mathcal{T}(\cdot) = \begin{cases} \text{T-Route}, & (\exists S_k, S_k < t_1 \wedge non\_key\_object) \\ & \quad \vee (\forall S_k, S_k < t_1 \wedge key\_object) \\ \text{N-Route}, & others \end{cases} \tag{10}$$

Similarly, for the remaining question category, the selector routes inputs with globally low frame-question relevance and missing key objects cues to a context-aware pathway, while all other cases are processed directly. The corresponding routing rule is defined as:

$$\mathcal{N}(\cdot) = \begin{cases} \text{O-Route}, & \forall S_k, S_k < t_1 \wedge non\_key\_object \\ \text{N-Route}, & others \end{cases} \tag{11}$$

3.4.2. Hierarchical Processor

Clustering-based methods [20,31,64] have been widely used for event segmentation in videos. However, most of them focus on boundary partitioning and fail to adapt the event representation to the semantic intent of different questions. As a result, irrelevant visual information is often introduced, which hinders effective reasoning in VideoQA. To address this issue, we propose a hierarchical processor that performs question-guided video reasoning through three complementary processing routes. Instead of treating all frames or events uniformly, the proposed processor adaptively selects different reasoning paths according to the semantic relevance between the question and video content. By explicitly coupling temporal structure modeling with question-aware weighting, the processor generates discriminative vision–language representations tailored to diverse VideoQA scenarios.

Specifically, the Hierarchical Processor consists of three parallel routes: T-Route, which includes a two-stage clustering strategy; O-Route, which includes a one-stage clustering strategy; and N-Route, which has no clustering strategy. These routes correspond to different levels of question–video alignment, ranging from fine temporal reasoning to

efficient direct reasoning. All routes share a unified multi-modal fusion module $\mathcal{F}$ for integrating visual and textual representations.

T-Route: obtaining event-centric temporal reasoning. Many VideoQA questions, such as identifying what happens before, after, or at a critical moment within an event, necessitate precise temporal reasoning. However, existing event segmentation methods [20,31,64] typically yield coarse event boundaries, which fall short in capturing subtle temporal phases and fine object dynamics within a single event. To overcome this limitation, we propose the T-Route, which integrates two complementary components: a two-stage clustering strategy and an object-centric description generation strategy. The former progressively refines the temporal structure by decomposing coarse events into fine sub-events, thus facilitating temporal localization aligned with question intent. The latter, through its explicit focus on question-relevant objects and interactions, enables the assessment of semantic relevance at both event and frame levels. Together, these strategies jointly address the core challenge of adaptively adjusting frame importance in a question-guided manner, thereby assigning higher weights to frames that capture temporally and semantically relevant object states. This synergistic design ultimately promotes more accurate event-centric temporal reasoning in VideoQA.

First Stage. To accurately identify the events most relevant to the question, based on the principle that density-based clustering aggregates temporally coherent frames into semantically consistent events and object-centric event descriptions align event semantics with question intent, we perform coarse event localization by clustering frames to form events and selecting the events most relevant to the question using object-centric descriptions, thereby clarifying event-question relevance and laying a solid foundation for subsequent event-centric temporal reasoning.

Specifically, given the visual embedding for input frames $\{\boldsymbol{f}_i\}_{i=1}^{N}$, we first apply DBSCAN clustering to segment the video into coarse-grained events $\{\boldsymbol{E}_l\}_{l=1}^{L}$:

$$\{\boldsymbol{E}_l\}_{l=1}^{L} = \mathcal{C}_{\text{DBSCAN}}\Big(\{\boldsymbol{f}_i\}_{i=1}^{N}\Big). \tag{12}$$

Event-level textual descriptions are then generated under an object-centric prompt by Flan-T5-XL:

$$\{\boldsymbol{D}_l^*\}_{l=1}^{L} = \mathcal{T}_5\Big(\{\boldsymbol{E}_l\}_{l=1}^{L}, \boldsymbol{Q}, [\text{EP}]\Big), \tag{13}$$

where [EP] represents the event-level description prompt: "Describe in a few words what you see in these frames most relevant to answering: [$\boldsymbol{Q}$] focus on {key objects}.". The semantic relevance between each event and the question is evaluated. Only the most relevant event is selected for further processing, which prevents unnecessary fine reasoning over irrelevant segments. To quantify the relevance of each event to the given question, we compute a set of relevance scores $\{R_l^*\}_{l=1}^{L}$ by measuring the semantic alignment between the event-level descriptions and the question $\boldsymbol{Q}$. The top-1 event with the highest scores is selected as the key event, and we denote the selected event as $\boldsymbol{E}$.

Second Stage. To refine keyframe importance and enhance reasoning accuracy, based on the principle that further clustering reveals intra-event temporal phases and question intent can guide the weighting of frames, we apply fine temporal clustering to the selected events from the first stage and selectively weight frames according to question intent, thereby highlighting temporally critical frames and improving the efficiency of visual information utilization as well as reasoning performance.

The selected event $\{\boldsymbol{E}_j^*\}$ is partitioned into three sub-events using k-means clustering ($k = 3$), corresponding to the early, middle and late phases. Guided by the temporal context implied by the question, key frames are selectively sampled from different sub-events, with higher weights assigned to frames from the most relevant phase. Frame-level textual

descriptions are finally generated to provide fine semantic cues for downstream reasoning. Specifically, for questions posed at a specific moment, several frames from the central cluster and one from each of the other two clusters are selected; for questions asked before a given moment, several frames from the first cluster and one from each of the other two clusters are chosen; for questions asked after a given moment, several frames from the final cluster and one from each of the other two clusters are selected. Finally, we obtain $\boldsymbol{k}$ key frames $\{\boldsymbol{f}_i^+\}_{i=1}^K$. Additionally, the several frames from the same cluster are placed at the end and assigned higher weights. In parallel, it also produces frame-level descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ directly from its input frames. These are generated by Flan-T5-XL under a frame-level prompt [FP], formulated as: "Describe in a few words what you see in the frame most relevant to answering: [$\boldsymbol{Q}$] focus on {key objects}.".

Computational Complexity Analysis. We next analyze the computational complexity of the clustering modules in T-Route. In our framework, both DBSCAN and k-means operate on frame-level visual features, and their complexity is dominated by the number of sampled frames N, which is fixed in our approach. DBSCAN with spatial indexing has an average time complexity of O(N log N). For k-means clustering, the time complexity is O(N · k · T), where k denotes the number of clusters and T is the number of iterations until convergence. Since both k and T are small constants in our design, the overall complexity of the two-stage clustering procedure remains linearithmic with respect to the number of sampled frames. As our method uses a fixed frame sampling strategy, the clustering steps are computationally efficient and do not introduce heavy overhead to the overall framework. Empirically, under our fixed-frame sampling setting (N = 32), the two clustering steps together take less than 0.1 s, accounting for only a small fraction of the total inference time.

O-Route: Recovering Missing Context for Weakly Relevant Segments. In long videos, many frames exhibit low direct relevance to the question but still provide essential contextual information for coherent reasoning. Discarding such frames entirely may cause the model to miss important background cues, while treating them equally with highly relevant frames introduces noise. The O-Route is designed to strike a balance by enhancing weakly relevant frames with event-level contextual semantics.

In the O-Route, the entire video $\{\boldsymbol{f}_i\}_{i=1}^N$ is segmented into events $\{\boldsymbol{E}_l\}_{l=1}^L$ using DBSCAN clustering. Instead of performing fine temporal refinement, representative key frames $\{\boldsymbol{f}_i^+\}_{i=1}^K$ are selected from the center of each event cluster. Event-level textual descriptions are then generated following the same strategy as in the first stage of T-Route. By integrating event-level descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ into frame representations, the O-Route recovers missing semantic information while maintaining temporal coherence across the video.

N-Route: Preserving High-Confidence Visual Evidence. Some questions can be accurately answered using a small number of key frames that are already highly aligned with the question semantics. Applying complex temporal modeling to such frames is unnecessary and may even distort their semantic content. The N-Route therefore provides a streamlined reasoning path that preserves high-confidence visual evidence with minimal processing overhead.

Concretely, the N-Route directly operates on question-relevant key frames $\{\boldsymbol{f}_i^+\}_{i=1}^K$ and generates detailed frame-level descriptions $\{\boldsymbol{D}_i\}_{i=1}^K$ as in the second stage of T-Route without additional event modeling. This design ensures semantic fidelity while improving computational efficiency.

Multi-Modal Fusion. Finally, the key frames selected from all routes and their corresponding question-guided textual descriptions are integrated by a unified multi-modal fusion module:

$$\boldsymbol{X}^m = \{\boldsymbol{f}_i^+\}_{i=1}^K \bowtie \{\boldsymbol{D}_i\}_{i=1}^K, \tag{14}$$

where $\bowtie$ denotes concatenation and m denotes multi-modal fusion. The resulting vision–language representation is used for final answer prediction.

3.5. Answer Module

The answer module functions as the final prediction module, designed to generate a reliable answer by leveraging the fused visual representations along with their associated descriptions. For this purpose, we utilize a frozen Flan-T5-XL model as the answer predictor $\mathcal{C}$, which receives the fused representation $\boldsymbol{X}^m$, the question $\boldsymbol{Q}$, the candidate answers $\mathscr{A}$, and a task-specific QA prompt [QP] as inputs. The resulting predicted answer is expressed as:

$$\hat{\boldsymbol{a}} = \mathcal{C}(\boldsymbol{X}^m, \boldsymbol{Q}, \mathscr{A}, [\mathrm{QP}]), \tag{15}$$

where the QA prompt [QP] is defined as: "Considering information in frames, select the correct answer from the options."

4. Experiments

We assess the effectiveness of the proposed framework by comparing it with state-of-the-art VideoQA baselines on two standard benchmark datasets. Section 4.1 details the experimental setup, followed by systematic quantitative and qualitative evaluations in Sections 4.2–4.4.

4.1. Experimental Setup

Datasets. To thoroughly evaluate the effectiveness and generalizability of our approach, we conduct experiments on two widely used VideoQA benchmarks: NExT-QA [65] and STAR [66], which emphasize complementary reasoning capabilities.

- NExT-QA serves as a comprehensive large-scale benchmark aimed at advancing temporal and causal reasoning in VideoQA. However, its data specificity imposes certain limitations on generalization: it contains 5440 short video clips (primarily focusing on daily-life scenarios with fixed scene settings) and 52,044 human-annotated question–answer pairs, which are grouped into three reasoning categories concentrated on temporal and causal logic: causal (48%), temporal (29%), and descriptive (23%). Specifically, causal questions emphasize understanding the intentions behind actions and their resulting outcomes, temporal questions concentrate on reasoning over action sequences and temporal dependencies, while descriptive questions mainly examine perception-oriented abilities, including object recognition and counting. The dataset is partitioned into 3870 training videos, 570 validation videos, and 1000 testing videos. On average, each question contains 11.6 words, whereas each answer consists of 2.6 words. Compared with STAR, NExT-QA's video scenarios are less diverse, and its reasoning types lack the exploration of situational logic, which makes its model performance difficult to generalize to complex realistic scenes beyond daily simple interactions. More detailed statistics are summarized in Table 1.
- STAR is devoted to situated VideoQA and is designed to facilitate complex reasoning in realistic environments by providing structured situational representations together with logic-based annotations. Similarly, STAR also has limitations in data specificity: the dataset comprises 22,000 trimmed video clips, about 60,000 question–answer pairs, and 240,000 candidate answers. Although its video scenarios are more diverse than NExT-QA, most clips are selected from fixed types of real-world scenes (e.g., home, office), lacking coverage of rare or special scenarios, which restricts the generalization of models trained on it to unconventional situational reasoning tasks. Each question is supplemented with a symbolic reasoning program along with a ground-truth answer, which supports compositional reasoning and enhances interpretability. STAR

categorizes reasoning into four types: Interaction (capturing human–object relationships), Sequence (modeling temporal order), Prediction (inferring upcoming actions), and Feasibility (evaluating whether actions are plausible). In addition, the dataset includes 111 action predicates, 28 object categories, and 24 relationship types. The answer consistency scores—82.5%, 85.3%, 80.4%, and 78.5% across the four reasoning tasks—indicate the varying difficulty levels associated with each reasoning type. In contrast to NExT-QA, STAR's strength lies in situated and compositional reasoning, but its limited scenario coverage and fixed predicate/relationship types still constrain its generalization ability in cross-scenario or cross-domain VideoQA tasks.

Table 1. Distribution of questions in NExT-QA dataset.

	Train	Val	Test	Total
Multi-Choice QA	34,132	4996	8564	47,692
Open-Ended QA	37,523	5343	9178	52,044

Implementation Details. The proposed framework is built upon the BLIP-2 [21] vision–language architecture, which comprises 4.1B parameters and is pre-trained on 129M image–text pairs collected from COCO [67], CC12M [68], Visual Genome [69], and SBU [70]. For the visual stream, a frozen ViT-g/14 model with 1.3B parameters is adopted to extract frame-level visual representations. The language backbone is a frozen Flan-T5-XL model (4B parameters), which is consistently utilized for frame caption generation, event abstraction, and final answer prediction. The model used to extract key objects from the question is `'en_core_web_sm'`. To ensure reproducibility, all experiments are performed under a unified default configuration unless otherwise stated. The number of keyframes is fixed to $K = 4$, the threshold t_1 is 0.2, and the predefined temporal routing weights in T-Route are set to [0.8, 1.0, 1.2, 1.5]. All hyperparameters are selected empirically to attain an appropriate balance among accuracy, robustness, and computational efficiency across the evaluated benchmarks.

Experimental Environment. All algorithms are developed using PyTorch 2.4.1 and executed on a computing server configured with 20 Intel(R) Xeon(R) Platinum 8457C vCPUs, 48 GB of system memory, and two NVIDIA L20 GPUs.

Baseline Methods. We compare our proposed ETR framework with 27 representative VideoQA baselines on the NExT-QA benchmark, which are categorized according to their dependence on large-scale language models.

- **Non-LLM-based methods.** We include 9 conventional VideoQA methods—MIST [4], HiTeA [58], CFMMC-Align [71], CBSR [28], T3T [72], ITMP [73], SHG-VQA [74] and MCG [75]. These models primarily depend on independently trained visual and textual encoders—either built from scratch or obtained through moderate-scale pre-training—without incorporating any large language models. They therefore provide strong baseline systems for evaluating the effectiveness of question-guided grounding and temporal reasoning when only limited semantic prior knowledge is available.
- **Large model-based methods.** We evaluate our framework against 18 recent approaches that integrate LLMs or LVLMs for multi-modal reasoning, including SeViLA [21], Q-ViD [62], LangRepo [60], Mistral [76], LLoVi [61], VideoINSTA [31], Vista-LLaMA [77], MVU [78], ViperGPT [79], Video-STaR [80], InterVideo [81], AssistGPT [82], VideoAgent [83], TOPA [59], IG-VLM [84], LLaVA-NeXT-Video+Selector [85], VGentailment [86] and ProViQ [63]. These approaches cover a wide range of modeling paradigms, spanning from fully end-to-end large language models to reasoning agents powered by LVLMs, thus offering a thorough and challenging benchmark for comparative assessments.

Evaluation Protocols. We follow the standard evaluation protocols established in previous VideoQA benchmarks to guarantee fair and consistent comparisons. Specifically, for the NExT-QA dataset, two complementary metrics are applied according to the question type. For multiple-choice questions, we use accuracy, defined as the proportion of correctly predicted answers relative to the total number of questions. Descriptive questions, including those in binary or counting formats, are evaluated using accuracy for consistency. For the STAR dataset, as well as the validation subset of NExT-QA that contains only multiple-choice questions, accuracy is employed uniformly as the evaluation metric. Formally, accuracy is calculated as: $\text{ACC} = N_{\text{correct}} / N_{\text{total}}$, where N_{correct} is the number of correctly answered questions, and N_{total} represents the total number of evaluated questions. This unified evaluation framework ensures comparability across different datasets and reasoning types.

4.2. Performance Comparison

To assess both the effectiveness and generalizability of our approach, we perform extensive comparisons with baseline methods on the NExT-QA and STAR datasets. The results are presented in Tables 2 and 3, demonstrating that our method consistently surpasses state-of-the-art approaches across diverse architectures, varying model scales, and different levels of task complexity.

Table 2. Quantitative comparison on the NExT-QA. ACC@D, ACC@T, and ACC@C denote the accuracy for descriptive, temporal, and causal questions, respectively, reported in percentage (%). The best results are highlighted in **bold**, and the second-best are underlined.

Method	Reference	Large Model	Zero-Shot	ACC@C	ACC@T	ACC@D	ACC@All
		Non-LLM-based methods					
MIST [4]	CVPR'23	–	–	54.6	56.6	66.9	57.2
HiTeA [58]	ICCV'23	–	–	62.4	58.3	75.6	63.1
CBSR [28]	TCSVT'24	–	–	59.3	58.0	67.9	60.3
MCG [75]	TIP'24	–	–	57.3	56.5	68.2	58.8
CFMMC-Align [71]	TCSVT'24	–	–	49.7	48.6	61.0	51.3
T3T [72]	arxiv'25	–	–	59.6	59.2	68.9	61.0
ITMP [73]	arxiv'25	–	–	58.1	58.0	68.9	60.4
		Large-model-based methods					
AssistGPT [82]	arxiv'23	ChatGPT-4	✓	60.0	51.4	67.3	58.4
ViperGPT [79]	ICCV'23	GPT-3	✓	56.4	49.8	–	60.0
SeViLA [21]	NeurIPS'23	Flan-T5-4B	✓	61.3	61.5	**75.6**	63.6
Q-ViD [62]	arxiv'24	Flan-T5-4B	✓	60.6	56.4	58.4	60.2
LangRepo [60]	NeurIPS'24	Mistral-7B	✓	57.8	45.7	61.9	54.6
LangRepo [60]	NeurIPS'24	Mistral-12B	✓	**64.4**	51.4	69.1	60.9
Mistral [76]	arxiv'24	Mistral-7B	✓	51.0	48.1	57.4	51.1
LLoVi [61]	EMNLP'24	Mistral-7B	✓	55.6	47.9	63.2	54.3
LLoVi [61]	EMNLP'24	Mistral-12B	✓	60.2	51.2	66.0	58.2
VideoINSTA [31]	EMNLP'24	Llama3-8B	✓	–	–	–	58.3
Videoagent [83]	ECCV'24	ChatGPT-3.5	✓	–	–	–	48.8
Vista-llama [77]	CVPR'24	Vicuna-7B	✓	–	–	–	60.7
TOPA [59]	NeurIPS'24	Llama2-7B	✓	61.3	53.4	68.3	59.9
TOPA [59]	NeurIPS'24	Llama3-8B	✓	61.9	53.0	64.5	59.5
TOPA [59]	NeurIPS'24	Llama2-13B	✓	63.6	57.2	68.9	62.1
ProViQ [63]	ECCV'24	GPT-3.5-turbo	✓	55.6	60.1	–	63.8
MVU [78]	ICLR'25	LLaVA-v1.5-13B	✓	55.7	48.2	64.2	55.4
DYTO [87]	ICCV'25	Qwen-VL-Chat-9B	✓	–	–	–	53.4
LLaVA-NeXT-Video+Selector [85]	CVPR'25	LLaVA-NeXT-Video+Selector-8.5B	✓	–	–	–	63.4
VGentailment [86]	CVPR'25	VideoLLAVA-7B	✓	–	–	–	60.8
ETR	Ours	Flan-T5-4B	✓	62.9	**61.7**	75.5	**64.4**

Table 3. Quantitative comparison on the STAR dataset. ACC@I, ACC@S, ACC@P and ACC@F denote accuracy for interaction, sequence, prediction and feasibility questions, respectively, reported in percentage (%). The best results are highlighted in **bold**, and the second-best are underlined.

Method	Reference	Large-Model	ACC@I	ACC@S	ACC@P	ACC@F	ACC@Avg
Non-LLM-based methods							
SHG-VQA [74]	CVPR'23	–	48.0	42.0	35.3	32.5	39.5
SHG-VQA* [74]	CVPR'23	–	45.8	42.8	34.6	29.9	38.3
Large-model-based methods							
Video-STaR [80]	arXiv'24	Vicuna-7B-v1.5	–	–	–	–	33.0
SeViLA [21]	NeurIPS'23	Flan-T5-4B	48.3	45.0	44.4	40.8	44.6
Q-ViD [62]	arXiv'24	Flan-T5-4B	47.4	44.8	**44.7**	42.8	44.9
InterVideo [81]	arXiv'22	VIT-L-1.4	43.8	43.2	42.3	37.4	41.6
IG-VLM [84]	ACCESS'24	CogAgent	39.8	**47.4**	40.5	**43.6**	44.4
TOPA [59]	NeurIPS'24	Llama2-7B	36.4	45.6	39.3	36.3	41.3
TOPA [59]	NeurIPS'24	Llama3-8B	40.8	43.1	39.4	34.5	41.4
TOPA [59]	NeurIPS'24	Llama2-13B	41.6	46.2	44.2	36.7	43.0
ETR	Ours	Flan-T5-4B	**49.4**	45.3	43.8	42.4	**45.2**

4.2.1. Quantitative Comparison on the NExT-QA

Table 2 summarizes the overall performance comparison between ETR and 23 representative approaches on the NExT-QA benchmark, including 7 methods without using large models and 16 large-model-based methods. As reported, ETR achieves 64.4% accuracy, ranking first among all evaluated models while maintaining a relatively compact model scale and computational efficiency.

Within the category of non-LLM methods, MIST [4], HiTeA [58], CFMMC-Align [71], CBSR [28], MCG [75], T3T [72], and ITMP [73] report ACC@All scores ranging from 51.3% to 63.1%, with HiTeA [58] being the strongest performer in this group. Although these methods have progressively improved visual–language fusion and reasoning strategies, they are still limited in fully modeling complex temporal dynamics and rich cross-modal semantics in challenging video scenarios. In contrast, ETR reaches 64.4% accuracy, exceeding the best-performing non-LLM competitor by 1.3 percentage points, highlighting its stronger robustness and reasoning capacity.

For large-model-based approaches, performance varies substantially across different model scales and reasoning frameworks. (i) Lightweight models, such as SeViLA [21] and Q-ViD [62] equipped with Flan-T5-4B, obtain 60.2–63.6%. (ii) Medium-scale models, including LangRepo [60], Mistral [76], LLoVi [61], Vista-LLaMA [77], DYTO [87], LLaVA-NeXT-Video+Selector [85], and VGentailment [86], generally achieve 51.1–63.4%, showing competitive yet slightly inferior performance overall. (iii) Large-scale systems, such as AssistGPT [82], ViperGPT [79], VideoAgent [83], and ProViQ [63], despite leveraging GPT-3.5, GPT-4, or LLaVA-13B, yield total accuracy mainly between 48.8% and 63.8% and do not surpass ETR.

Overall, ETR delivers the best total accuracy performance among all compared methods. Its superiority can be attributed to the T-Route module, which effectively filters and enhances question-relevant visual evidence, thereby strengthening visual–semantic alignment and enabling more reliable high-level reasoning while remaining considerably more parameter-efficient than many LLM-based counterparts.

4.2.2. Quantitative Comparison on the STAR Dataset

Table 3 presents the accuracy comparison on the STAR dataset. It is worth noting that the backbone of the conventional method SHG-VQA [74] is SlowR50-K400, while its improved variant SHG-VQA* uses ResNext101-ImageNet1K. As a representative symbolic reasoning-based VideoQA model, SHG-VQA primarily relies on video–text alignment and

manually designed reasoning modules. While effective on simpler datasets, both versions exhibit unstable performance on STAR, which contains complex causal structures and cross-temporal event sequences, achieving only 38.3–39.5% and struggling to maintain coherent reasoning chains.

Recent large-model-based approaches, such as InterVideo [81], Video-STaR [80], SeViLA [21], Q-ViD [62], TOPA [59] and IG-VLM [84], leverage large-scale pre-trained knowledge to mitigate the limitations of conventional methods. These frameworks improve video–language integration and semantic reasoning to some extent, yet they often suffer from inconsistent alignment between visual evidence and textual reasoning when tasks require compositional or cross-temporal inference.

ETR addresses these challenges through an adaptive weight adjustment mechanism combined with a hierarchical alignment framework. This framework dynamically assigns importance weights to key frames based on the objects emphasized in the question and captures multi-level visual–language interactions to facilitate alignment with the central subject. By establishing a robust connection between low-level visual representations and high-level semantic reasoning, this structured alignment not only improves the model's ability to separate causal and situational cues in visually entangled scenarios but also enhances interpretability and reasoning stability. Under this design, ETR achieves 45.2% on the STAR dataset, surpassing all conventional and LLM-based baselines by up to 6.9 percentage points. These results demonstrate that question-guided weight adjustment and hierarchical visual–language alignment are critical for advancing situated video understanding in complex temporal and causal contexts.

4.3. Ablation and Comparative Studies

We first perform ablation studies to assess the contribution of each key module in Table 4. Components are progressively removed or replaced to quantify their impact on addressing the task's core challenges. To systematically validate the effectiveness of ETR in addressing the research questions, we then conduct experiments focusing on three aspects: the effectiveness of the hierarchical selector, the capability of T-Route for event-centric temporal reasoning, and the alignment of semantic descriptions with key question-relevant objects. All experiments are performed on the NExT-QA dataset, and the results are summarized in Tables 5–11.

4.3.1. Combined Effect of Key Components

To investigate the cumulative and interactive effects of ETR's core modules, we conduct a systematic ablation study on the NExT-QA dataset, as summarized in Table 4. The study examines three key modules: hierarchical selector, T-Route, and object-centric prompt.

Table 4. Combined effect of key components.

Hierarchical Selector	T-Route	Object-Centric Prompt	ACC@C	ACC@T	ACC@D	ACC@All
-	-	-	62.40	59.35	76.88	63.63
-	-	✓	61.94	60.72	76.46	63.77
✓	✓	-	62.86	60.86	75.19	64.09
✓	✓	✓	62.93	61.73	75.48	64.45

Starting from the baseline model with only a coarse keyframe selection, the overall accuracy reaches 63.63%. Incorporating the object-centric prompt increases the accuracy slightly to 63.77%, indicating that text descriptions aligned with question-relevant objects provide weakly supervised semantic guidance, improving alignment between textual se-

mantics and query intent while complementing the visual information stream. Further integrating the hierarchical selector and T-Route raises accuracy to 64.45%. The hierarchical selector adaptively identifies questions that require event-centric temporal reasoning, while T-Route dynamically adjusts the attention weights of visual cues based on event-level information, optimizing alignment between the question and video content. This mechanism not only improves keyframe selection precision but also mitigates temporal information loss from video truncation or frame sampling, enabling more accurate modeling of causal and semantic relationships within events. Overall, the complete ETR configuration achieves an absolute improvement of 0.82 percentage points over the baseline (Table 4). Despite its moderate model size, this consistent gain across modules reflects their complementary roles: the hierarchical selector identifies questions requiring detailed event understanding, T-Route refines the attention on critical visual cues while compensating for missing temporal information, and the object-centric prompt maintains query-focused semantic alignment—particularly effective for fine questions. Together, these components work synergistically, and the lightweight N-Route bypass ensures computational efficiency while delivering stable accuracy improvements. These results demonstrate that a coarse-to-fine selection strategy combined with target-oriented semantic guidance forms an effective and mutually reinforcing approach for robust video–language reasoning.

4.3.2. Impact of Keyframe Selection Strategy

Table 5 compares different routing strategies and provides insights into the optimal performance conditions of our hierarchical routing mechanism. The non strategy, which applies a general approach to all questions, achieves 63.77% accuracy. Applying all questions uniformly through non T-Route or T-Route yields lower accuracy (63.67% and 63.11%, respectively), revealing two key insights: (1) Over-processing harms simple questions: The performance drop of all T-Route (63.11%) compared to non (63.77%) indicates that applying fine-grained event reasoning to simple descriptive questions introduces semantic noise rather than benefit. This validates that temporal reasoning should be deployed selectively. (2) Under-processing harms complex questions: The moderate performance of non (63.77%) suggests that while it handles simple questions well, it fails to capture the temporal dynamics needed for complex temporal questions. Building on this observation, our proposed hierarchical selector adaptively identifies questions requiring event-centric temporal reasoning, increasing accuracy to 64.45%. This semantics-driven selection achieves optimal performance by matching each question to its most appropriate processing route—deploying expensive T-Route only when the expected gain outweighs the risk of noise introduction. The 0.68–1.34% improvement over single-route baselines demonstrates that intelligent routing, rather than module complexity alone, is the key to effective VideoQA.

Table 5. Impact of hierarchical selector.

Method	ACC@C	ACC@T	ACC@D	ACC@All
non	61.94	60.72	76.46	63.77
non T-Route	62.05	60.93	75.09	63.67
all T-Route	61.67	59.93	74.83	63.11
Ours	62.93	61.73	75.48	64.45

4.3.3. Impact of T-Route

Using Table 6, we analyze the effect of the attention weight adjustment mechanism in T-Route on overall QA performance. When using uniform weights, the accuracy is only 63.97%. In contrast, our approach performs two rounds of clustering—first based on

question-relevant objects and then on the temporal cues of events—to dynamically adjust the attention weights of key visual cues. This allows the model to more effectively facilitate the alignment of the relevant subjects, actions, and critical events associated with the question. These results indicate that the dynamic weight adjustment mechanism in T-Route enhances the model's sensitivity to both semantic and temporal information, improving event-level reasoning and cross-modal alignment. The benefits are particularly pronounced for fine or temporally complex questions, demonstrating a clear advantage over uniform weighting strategies.

Table 6. Impact of T-Route.

Method	ACC@C	ACC@T	ACC@D	ACC@All
unified weight	62.74	60.73	75.10	63.97
Ours	62.93	61.73	75.48	64.45

4.3.4. Impact of the k Value in k-Means Clustering

Table 7 analyzes the impact of different values of k on overall model performance. Specifically, performance reaches its lowest point at 63.51% when k = 1. This is likely because treating all high-confidence time steps as a single pattern mixes distinct temporal structures, failing to effectively capture the sequential characteristics of events. As k increases to 4, overall performance slightly drops to 64.25% compared with k = 3. Although temporal relationships are captured to some extent, over-segmentation may introduce redundancy and noise, leading to degraded performance. Overall, k = 3 achieves the best balance between expressive power and model stability, enabling more effective modeling of the underlying temporal structure in high-confidence time steps.

Table 7. Impact of the k value in k-means clustering.

Method	ACC@C	ACC@T	ACC@D	ACC@All
1	62.36	60.36	74.19	63.51
4	62.55	61.60	75.74	64.25
Ours(3)	62.93	61.73	75.48	64.45

4.3.5. Impact of the t_1 Value

Table 8 analyzes the impact of different values of t_1 on overall model performance. The t_1 value serves as a threshold to route samples with low semantic similarity between key frames and the question to event-level understanding. When t_1 is too small, many samples that require event-level analysis are instead answered directly using weakly relevant key frames, limiting the model's ability to capture overall event structure and temporal logic, which reduces accuracy. As t_1 increases, performance first improves and then declines. When t_1 becomes too large, too many samples are unnecessarily routed to temporal reasoning; for samples whose key frames already have high semantic relevance, this extra processing introduces redundant temporal noise and interferes with the model's judgment. To empirically validate our threshold selection, we randomly sampled multiple videos from our dataset and analyzed the distribution of frame-question relevance scores. Across this random sample, we found that on average, 18.23% of frames have relevance scores below $t_1 = 0.2$, with a standard deviation of 4.68% across different videos. This consistent proportion, approximately one-fifth of frames falling below the threshold, demonstrates that the threshold is appropriately calibrated across a wide range of video content. The 18.23% figure represents an optimal balance: it is high enough to meaningfully impact performance, as the 0.8–1.3% overall improvement comes precisely from these frames that

benefit from deeper temporal analysis, yet it is low enough that the majority of frames (81.77%) can be processed efficiently through simpler routes, maintaining fast inference. The moderate standard deviation of 4.68% indicates that this balance is consistently achieved across diverse video types, rather than being driven by a small subset of atypical examples. Thus, the best performance is achieved at $t_1 = 0.2$, where the model reaches an optimal balance between direct key-frame question answering and event-level understanding.

Table 8. Impact of the t_1 value.

Method	ACC@C	ACC@T	ACC@D	ACC@All
0.1	62.60	61.60	75.49	64.24
0.2	62.93	61.73	75.48	64.45
0.3	62.60	61.60	75.49	64.24
0.6	62.60	60.81	79.54	64.21
1.0	62.55	60.75	79.53	64.17

4.3.6. Impact of Object-Centric Descriptions

Table 9 investigates the effect of different prompt designs on the quality of generated textual descriptions and their downstream impact on VideoQA performance. The effectiveness of textual descriptions is highly dependent on prompt design: poorly constructed prompts may produce low-quality outputs that not only fail to improve performance but can also introduce semantic noise. Specifically, using a General Question Prompt ('GQP') yields an accuracy of 63.49%, whereas a Generic Declarative Prompt ('GDP'), a declarative Question-Guided Prompt ('QDP'), or an interrogative Question-Guided Prompt ('QQP') achieve accuracies of 63.63%, 64.09%, and 63.55%, respectively. In contrast, prompts that explicitly facilitate alignment with the key objects mentioned in the question generate more targeted and semantically aligned descriptions, resulting in a significant performance improvement, with peak accuracy reaching 64.45%. We hypothesize that generic or non-object-centric prompts often produce ambiguous or irrelevant content, weakening the semantic alignment between visual evidence and question intent. These findings highlight that precise and structured prompt design is critical for generating high-quality textual descriptions and supporting effective video reasoning.

Table 9. Impact of captioning prompt design. "GDP", "GQP", "QDP" and "QQP" denote general declarative prompt, general question prompt, question-guided declarative prompt and question-guided question prompt.

Method	ACC@C	ACC@T	ACC@D	ACC@All
GDP	61.82	60.31	76.86	63.63
GQP	61.71	69.92	77.12	63.49
QDP	62.86	60.86	75.19	64.09
QQP	61.97	60.20	76.09	63.55
Ours	62.93	61.73	75.48	64.45

4.3.7. Performance Analysis

Table 10 presents the performance analysis. We compare our proposed method (Ours) with the method of all T-Route under the same hardware and dataset conditions. The results show that Ours achieves an average per-video inference latency of 1.01 s, which is approximately 20% faster than the baseline's 1.26 s. Both total wall-clock time and total GPU time are reduced by 25%. The primary source of acceleration lies in the optimization of the T5 generation stage, where the average time per call decreases from 1.8–2.4 s in the baseline to 0.92 s in Ours (a reduction of over 50%), while peak and average memory con-

sumption remain nearly identical (24.29 GB/19.2 GB). If a full temporal sequence modeling capability were further introduced into the baseline, it would inevitably incur significant additional computational overhead and increased latency. In contrast, Ours achieves an effective balance between inference speed and temporal understanding capability under its current design.

Table 10. Ours vs. all T-Route performance comparison.

Metric	Ours	All T-Route
Avg. per-video wall-clock latency	1.01 s	1.26 s
Avg. per-video GPU latency	1.01 s	1.26 s
Total wall-clock time	1.41 h	1.87 h
Total GPU time	1.403 GPU-h	1.869 GPU-h
Average peak memory	24.29 GB	24.29 GB
Average memory usage	19.23 GB	19.30 GB
T5 Generate—avg. per call	0.92 s	1.8–2.4 s
Typical 8-video batch wall time	7.5–8.5 s	8–11 s
Overall compute cost (time × resources)	Clearly lower	Higher

To further analyze the computational characteristics of the proposed method, we additionally report the inference cost under different frame sampling settings, as summarized in Table 11. All experiments follow the same hardware configuration and fixed-frame sampling strategy, where a constant number of frames (24/32/48) is uniformly sampled from each video. Under this setting, the computational cost is determined by the number of sampled frames rather than the original video duration. Following common GPU profiling practice, we report results from a steady-state iteration after warm-up to eliminate initialization overhead and ensure consistent measurement of runtime performance. As shown in Table 11, the computational cost scales consistently with the number of sampled frames. The visual encoding FLOPs increase from 37,172.75 GFLOPs at 24 frames to 49,563.66 GFLOPs at 32 frames and 74,345.49 GFLOPs at 48 frames, exhibiting a strictly linear growth pattern. Correspondingly, the total inference time increases moderately from 12.45 s to 13.41 s and 15.39 s, while the peak GPU memory rises from 20.14 GB to 24.19 GB and 32.31 GB, respectively. Notably, the memory growth ratio is lower than the FLOPs scaling ratio, indicating that only frame-dependent feature representations scale with the number of sampled frames, whereas model parameters remain constant. In the feature post-processing stage, k-means and DBSCAN are jointly applied to the frame-level representations, and their computational complexity is determined by the number of sampled frames N. Specifically, k-means has a time complexity of $O(N \cdot k \cdot T)$, while DBSCAN exhibits an average complexity of $O(N \log N)$ with spatial indexing. Since both clustering methods operate on low-dimensional feature representations, their runtime and memory overhead remain significantly smaller than that of the visual encoding stage. Overall, these results demonstrate that under a fixed-frame sampling strategy, inference time, FLOPs, and memory consumption are primarily governed by the number of sampled frames rather than the original video length. This property ensures stable and controllable resource usage for long-video inputs and supports the scalability of the proposed method in practical deployment scenarios.

Table 11. Comparison of inference time, FLOPs, and GPU memory consumption under different frame sampling settings.

Frames	FLOPs (GFLOPs)	Inference Time (s)	Peak Memory (GB)	Avg Memory (GB)
24	37,172.75	12.45	20.14	16.37
32	49,563.66	13.41	24.19	19.16
48	74,345.49	15.39	32.31	24.77

4.4. Qualitative Analysis

To further demonstrate the effectiveness of the proposed method, we present several representative qualitative results on the NExT-QA dataset, as shown in Figure 3. Each case includes the original question–answer pair with candidate options, the detected key objects, the selected key frames (with higher weights assigned to the latter two frames), and the generated descriptions that align with the critical objects. From these visualizations, we can clearly observe that our hierarchical question-guided framework adaptively locates informative frames and salient objects based on the semantic alignment between the video content and the question target. By highlighting the most relevant and discriminative evidence, our method effectively suppresses redundant information and supports more reliable and fine video reasoning.

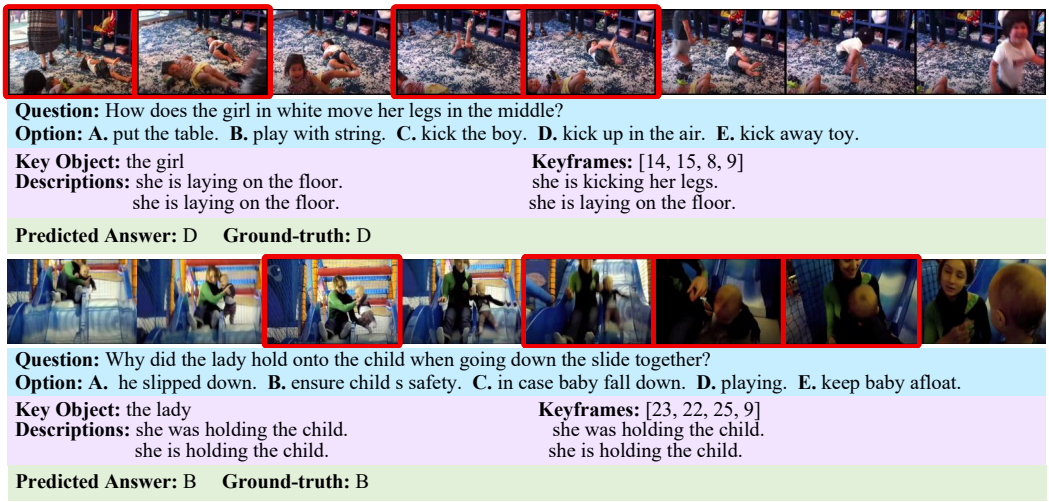

Figure 3. An example illustrating different weighting schemes. We just show parts of frames of the video, and the frames are not strictly consecutive to show a rough video development. The red boxes highlight the frames our model focuses on.

5. Limitations

Despite the overall effectiveness of our approach, several limitations remain evident in our experimental analysis. To better understand the underlying causes of model failures in long-form VideoQA, we examine representative error cases and categorize them into two primary types based on the nature of the observed shortcomings.

- Failure Type I—Insufficient Robustness to Distractions. The model sometimes struggles to filter out visually distracting or temporally irrelevant content when constructing event representations. As illustrated in the first example of Figure 4, although the model captures a general understanding of the video—which depicts a boy unwrapping a gift—it is misled by an intermediate segment where the boy briefly shows the gift to a woman. This distraction causes the model to incorrectly select an answer involving "share with the girl" rather than the correct action "unwrap it." This suggests

that the model's current mechanism for assessing the importance of event segments remains insufficient for suppressing task-irrelevant but salient visual cues.

- Failure Type II—Insufficient Sensitivity to Fine-Grained Visual Details. In other cases, the model demonstrates a coarse understanding of the scene but fails to capture fine-grained visual information essential for accurate reasoning. The second example in Figure 4 illustrates this limitation: given a video of a band performance, the model correctly identifies the general action (e.g., "playing an instrument") but overlooks critical details regarding how the action is performed (e.g., bowing vs. plucking). This indicates a need for enhanced perceptual granularity in frame-level feature extraction and cross-modal alignment.

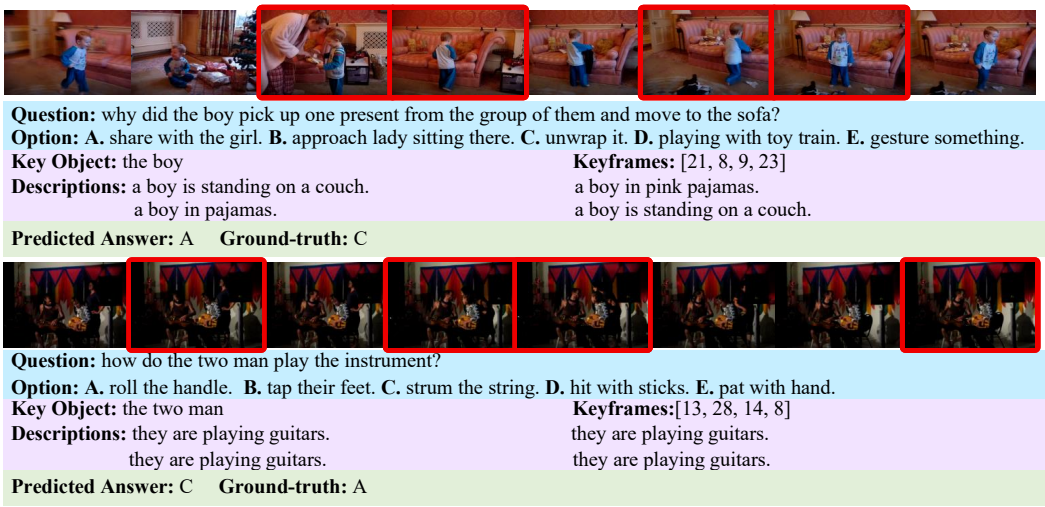

Figure 4. An example illustrating failure cases of weighting schemes. We just show parts of frames of the video, and the frames are not strictly consecutive to show a rough video development. The red boxes highlight frames our model focuses on.

The core design of ETR operates on general semantic representations extracted from video content—including visual features and semantic similarity calculations—rather than relying on the internal parameters or architecture of any specific language model. This means that the temporal enhancement functions of ETR, such as temporal segmentation, key event localization, and frame importance weighting, are performed independently of the language model used for description generation or answer prediction. However, the current implementation and all reported results in this paper are based on Flan-T5-XL. Models with different scales or architectures may vary in their semantic representation quality and temporal reasoning capacity, which could affect the final question-answering performance. To address this limitation and more thoroughly validate the applicability of ETR, we plan to conduct systematic experiments across diverse language models in future work. Specifically, we will: (i) evaluate ETR on language models of different scales, including small models (e.g., T5-base), medium models (e.g., Flan-T5-XL), and larger models (e.g., LLaMA or GPT variants); (ii) quantify the performance gain brought by ETR on each model by comparing the "ETR + LLM" framework against using the base LLM alone; and (iii) extend our experiments to models with different architectures and pre-training objectives, including lightweight models designed for efficiency and open-domain pre-trained models. These experiments will help establish whether ETR consistently provides temporal enhancement across different language models and further strengthen the generalizability of our framework.

6. Conclusions

In this paper, we propose ETR, a framework for VideoQA designed to enable precise temporal reasoning and object-centric visual-semantic alignment. The key innovations of ETR include: (i) Hierarchical weight adjustment module, which dynamically identifies questions requiring event-centric temporal reasoning and leverages detailed event analysis along with question intent to adjust the weights of key event cues, thereby enhancing question-relevant visual information and improving reasoning accuracy; (ii) Question-focused key-object description generation mechanism, which produces object-centric textual descriptions aligned with the question semantics, ensuring that key visual objects are effectively represented and reinforcing visual-semantic alignment for more precise answer prediction. Experimental results on two benchmark datasets demonstrate that ETR efficiently integrates critical visual cues with semantic context, significantly enhancing the comprehension of complex dynamic videos and providing robust, efficient, and high-precision performance for VideoQA.

Author Contributions: Conceptualization, L.P. and Z.G.; methodology, L.P. and L.Z.; software, L.P. and Y.Z.; validation, L.P., Z.G. and F.C.; formal analysis, C.Z. and L.Z.; investigation, L.P. and Z.G.; resources, C.Z. and S.T.; data curation, L.P., D.Y. and Y.C.; writing—original draft preparation, L.P.; writing—review and editing, C.Z. and L.Z.; visualization, L.P. and Z.G.; supervision, L.Z.; project administration, L.Z. and S.T.; funding acquisition, L.Z., C.Z. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China under Grant 62202163, Grant 62472161 and Grant 62372150, Yuelu Mountain Laboratory Seed Industry Special Project Grant on YLS-2025-ZY04016, the Natural Science Foundation of Hunan Province under Grant 2023JJ30169.

Data Availability Statement: The research data are openly available at https://doc-doc.github.io/docs/nextqa.html (NExT-QA), accessed on 4 March 2026, and https://bobbywu.com/STAR/ (STAR), accessed on 4 March 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, J.; Shao, J.; Cao, R.; Gao, L.; Xu, X.; Shen, H.T. Action-centric relation transformer network for video question answering. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *32*, 63–74. [CrossRef]
2. Zang, C.; Wang, H.; Pei, M.; Liang, W. Discovering the real association: Multimodal causal reasoning in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2023; pp. 19027–19036.
3. Chen, G.; Liu, X.; Wang, G.; Zhang, K.; Torr, P.H.; Zhang, X.P.; Tang, Y. Tem-adapter: Adapting image-text pretraining for video question answer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2023; pp. 13945–13955.
4. Gao, D.; Zhou, L.; Ji, L.; Zhu, L.; Yang, Y.; 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*; Computer Vision Foundation: New York, NY, USA, 2023; pp. 14773–14783.
5. Zhu, L.; Wu, R.; Zhu, X.; Zhang, C.; Wu, L.; Zhang, S.; Li, X. Bi-Direction Label-Guided Semantic Enhancement for Cross-Modal Hashing. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 3983–3999. [CrossRef]
6. Zhu, L.; Zhang, C.; Song, J.; Liu, L.; Zhang, S.; Li, Y. Multi-graph based hierarchical semantic fusion for cross-modal representation. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
7. Zhu, L.; Yu, W.; Zhu, X.; Zhang, C.; Li, Y.; Zhang, S. MvHAAN: Multi-view hierarchical attention adversarial network for person re-identification. *World Wide Web* **2024**, *27*, 59. [CrossRef]
8. Tian, C.; Zheng, M.; Li, B.; Zhang, Y.; Zhang, S.; Zhang, D. Perceptive self-supervised learning network for noisy image watermark removal. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7069–7079. [CrossRef]

9. Tian, C.; Song, M.; Fan, X.; Zheng, X.; Zhang, B.; Zhang, D. A Tree-guided CNN for image super-resolution. *IEEE Trans. Consum. Electron.* **2025**, *71*, 3631–3640. [CrossRef]
10. Tian, C.; Zhang, X.; Liang, X.; Li, B.; Sun, Y.; Zhang, S. Knowledge distillation with fast CNN for license plate detection. *IEEE Trans. Intell. Veh.* 2023, *early access*.
11. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2015; pp. 2425–2433.
12. Shih, K.J.; Singh, S.; Hoiem, D. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2016; pp. 4613–4621.
13. Kim, M.; Cho, M.; Lee, H.; Lee, S. Spatio-temporal Feature-level Augmentation Vision Transformer for video-based person re-identification. *Pattern Recognit.* **2025**, *168*, 111813. [CrossRef]
14. Zhu, L.; Wu, R.; Liu, D.; Zhang, C.; Wu, L.; Zhang, Y.; Zhang, S. Textual semantics enhancement adversarial hashing for cross-modal retrieval. *Knowl.-Based Syst.* **2025**, *317*, 113303. [CrossRef]
15. Tian, C.; Zheng, M.; Lin, C.W.; Li, Z.; Zhang, D. Heterogeneous window transformer for image denoising. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 6621–6632. [CrossRef]
16. Tian, C.; Zheng, M.; Jiao, T.; Zuo, W.; Zhang, Y.; Lin, C.W. A Self-Supervised CNN for Image Watermark Removal. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7566–7576. [CrossRef]
17. Zeng, K.H.; Chen, T.H.; Chuang, C.Y.; Liao, Y.H.; Niebles, J.C.; Sun, M. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2017; Volume 31.
18. Yang, Z.; Garcia, N.; Chu, C.; Otani, M.; Nakashima, Y.; Takemura, H. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2020; pp. 1556–1565.
19. Zhu, L.; Xu, Z.; Yang, Y.; Hauptmann, A.G. Uncovering the temporal context for video question answering. *Int. J. Comput. Vis.* **2017**, *124*, 409–421. [CrossRef]
20. Wang, Z.; Yu, S.; Stengel-Eskin, E.; Yoon, J.; Cheng, F.; Bertasius, G.; Bansal, M. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; Computer Vision Foundation: New York, NY, USA, 2025; pp. 3272–3283.
21. Yu, S.; Cho, J.; Yadav, P.; Bansal, M. Self-chained image-language model for video localization and question answering. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 76749–76771.
22. Qian, T.; Cui, R.; Chen, J.; Peng, P.; Guo, X.; Jiang, Y.G. Locate before answering: Answer guided question localization for video question answering. *IEEE Trans. Multimed.* **2023**, *26*, 4554–4563. [CrossRef]
23. Lyu, L.; Liu, D.; Zhang, C.; Zhang, Y.; Ruan, H.; Zhu, L. Temporal Map-Based Boundary Refinement Network for Video Moment Localization. *Electronics* **2025**, *14*, 1657. [CrossRef]
24. Zhu, L.; Zhang, C.; Song, J.; Zhang, S.; Tian, C.; Zhu, X. Deep multigraph hierarchical enhanced semantic representation for cross-modal retrieval. *IEEE MultiMed.* **2022**, *29*, 17–26. [CrossRef]
25. Chen, J.; Zhu, D.; Haydarov, K.; Li, X.; Elhoseiny, M. Video ChatCaptioner: Towards enriched spatiotemporal descriptions. *arXiv* **2023**, arXiv:2304.04227. [CrossRef]
26. Wang, J.; Chen, D.; Luo, C.; Dai, X.; Yuan, L.; Wu, Z.; Jiang, Y.G. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv* **2023**, arXiv:2304.14407.
27. Zhang, C.; Lu, T.; Islam, M.M.; Wang, Z.; Yu, S.; Bansal, M.; Bertasius, G. A simple llm framework for long-range video question-answering. *arXiv* **2023**, arXiv:2312.17235.
28. Wu, X.; Wu, J.; Zhu, L.; Senhadji, L.; Shu, H. Collaborative Aware Bidirectional Semantic Reasoning for Video Question Answering. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *35*, 2074–2086. [CrossRef]
29. Gong, H.; Li, L.; Zhang, J.; Sun, Y.; Gao, Y.; Yan, C. Spatial-Temporal Clue Reasoning Chain for Long Video Question Answering. *IEEE Trans. Circuits Syst. Video Technol.* 2025, *early access*.
30. Kugo, N.; Li, X.; Li, Z.; Gupta, A.; Khatua, A.; Jain, N.; Patel, C.; Kyuragi, Y.; Ishii, Y.; Tanabiki, M.; et al. VideoMultiAgents: A Multi-Agent Framework for Video Question Answering. *arXiv* **2025**, arXiv:2504.20091.
31. Liao, R.; Erler, M.; Wang, H.; Zhai, G.; Zhang, G.; Ma, Y.; Tresp, V. VideoINSTA: Zero-shot Long Video Understanding via Informative Spatial-Temporal Reasoning with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 6577–6602.
32. Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; Kim, G. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2017; pp. 2758–2766.

33. Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1645–1653.
34. Zhao, Z.; Lin, J.; Jiang, X.; Cai, D.; He, X.; Zhuang, Y. Video question answering via hierarchical dual-level attention network learning. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1050–1058.
35. Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; Fidler, S. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2016; pp. 4631–4640.
36. Kim, J.; Ma, M.; Kim, K.; Kim, S.; Yoo, C.D. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2019; pp. 8337–8346.
37. Gao, J.; Ge, R.; Chen, K.; Nevatia, R. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2018; pp. 6576–6585.
38. Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; 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*; Computer Vision Foundation: New York, NY, USA, 2019; pp. 1999–2007.
39. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; 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*; Computer Vision Foundation: New York, NY, USA, 2021; pp. 7331–7341.
40. Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J.S.; Cao, J.; Farhadi, A.; Choi, Y. Merlot: Multimodal neural script knowledge models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23634–23651.
41. Fu, T.J.; Li, L.; Gan, Z.; Lin, K.; Wang, W.Y.; Wang, L.; Liu, Z. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv* **2021**, arXiv:2111.12681.
42. Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; Niebles, J.C. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2022; pp. 2917–2927.
43. Huang, D.; Chen, P.; Zeng, R.; Du, Q.; Tan, M.; Gan, C. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2020; Volume 34, pp. 11021–11028.
44. Liu, F.; Liu, J.; Wang, W.; Lu, H. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2021; pp. 1698–1707.
45. Xiao, J.; Yao, A.; Liu, Z.; Li, Y.; Ji, W.; Chua, T.S. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2022; Volume 36, pp. 2804–2812.
46. Peng, L.; Yang, S.; Bin, Y.; Wang, G. Progressive graph attention network for video question answering. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 2871–2879.
47. Le, T.M.; Le, V.; Venkatesh, S.; Tran, T. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2020; pp. 9972–9981.
48. Dang, L.H.; Le, T.M.; Le, V.; Tran, T. Hierarchical object-oriented spatio-temporal reasoning for video question answering. *arXiv* **2021**, arXiv:2106.13432.
49. Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; Tenenbaum, J.B. CLEVRER: Collision Events for Video Representation and Reasoning. *arXiv* **2019**, arXiv:1910.01442.
50. Chen, Z.; Mao, J.; Wu, J.; Wong, K.; Tenenbaum, J.; Gan, C. Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning. *arXiv* **2021**, arXiv:2103.16564. [CrossRef]
51. Li, Y.; Wang, X.; Xiao, J.; Ji, W.; Chua, T.S. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2022; pp. 2928–2937.
52. Falcon, A.; Lanz, O.; Serra, G. Data augmentation techniques for the video question answering task. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 511–525.
53. Kim, N.; Ha, S.J.; Kang, J.W. Video question answering using language-guided deep compressed-domain video feature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2021; pp. 1708–1717.

54. Lu, H.; Ding, M.; Fei, N.; Huo, Y.; Lu, Z. Lgdn: Language-guided denoising network for video-language modeling. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25198–25211.
55. Han, W.; Chen, H.; Kan, M.Y.; Poria, S. Self-Adaptive Sampling for Efficient Video Question-Answering on Image–Text Models. *arXiv* **2023**, arXiv:2307.04192.
56. Wei, Y.; Liu, Y.; Yan, H.; Li, G.; Lin, L. Visual causal scene refinement for video question answering. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 377–386.
57. Xu, J.; Lan, C.; Xie, W.; Chen, X.; Lu, Y. Retrieval-based video language model for efficient long video question answering. *arXiv* **2023**, arXiv:2312.04931. [CrossRef]
58. Ye, Q.; Xu, G.; Yan, M.; Xu, H.; Qian, Q.; Zhang, J.; Huang, F. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2023; pp. 15405–15416.
59. Li, W.; Fan, H.; Wong, Y.; Kankanhalli, M.S.; Yang, Y. Topa: Extending large language models for video understanding via text-only pre-alignment. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 5697–5738.
60. Kahatapitiya, K.; Ranasinghe, K.; Park, J.; Ryoo, M.S. Language repository for long video understanding. *arXiv* **2024**, arXiv:2403.14622. [CrossRef]
61. Zhang, C.; Lu, T.; Islam, M.M.; Wang, Z.; Yu, S.; Bansal, M.; Bertasius, G. A Simple LLM Framework for Long-Range Video Question-Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 21715–21737.
62. Romero, D.; Solorio, T. Question-instructed visual descriptions for zero-shot video question answering. *arXiv* **2024**, arXiv:2402.10698.
63. Choudhury, R.; Niinuma, K.; Kitani, K.M.; Jeni, L.A. Video Question Answering with Procedural Programs. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 315–332.
64. Park, J.; Ranasinghe, K.; Kahatapitiya, K.; Ryoo, W.; Kim, D.; Ryoo, M.S. Too Many Frames, not all Useful: Efficient Strategies for Long-Form Video QA. *arXiv* **2024**, arXiv:2406.09396. [CrossRef]
65. Xiao, J.; Shang, X.; Yao, A.; Chua, T.S. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2021; pp. 9777–9786.
66. Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J.B.; Gan, C. Star: A benchmark for situated reasoning in real-world videos. *arXiv* **2024**, arXiv:2405.09711. [CrossRef]
67. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755.
68. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2556–2565.
69. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
70. Ordonez, V.; Kulkarni, G.; Berg, T. Im2text: Describing images using 1 million captioned photographs. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1143–1151.
71. Guo, K.; Tian, D.; Hu, Y.; Lin, C.; Qian, Z.; Sun, Y.; Zhou, J.; Duan, X.; Gao, J.; Yin, B. CFMMC-Align: Coarse-Fine Multi-Modal Contrastive Alignment Network for Traffic Event Video Question Answering. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10538–10550. [CrossRef]
72. Song, Z.; Hu, Z.; Ma, Y.; Li, J.; Hong, R. Video Flow as Time Series: Discovering Temporal Consistency and Variability for VideoQA. *arXiv* **2025**, arXiv:2504.05783. [CrossRef]
73. Liang, L.; Sun, G. Leveraging Static Relationships for Intra-Type and Inter-Type Message Passing in Video Question Answering. *arXiv* **2025**, arXiv:2504.02417.
74. Urooj, A.; Kuehne, H.; Wu, B.; Chheu, K.; Bousselham, W.; Gan, C.; Lobo, N.; Shah, M. Learning situation hyper-graphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2023; pp. 14879–14889.
75. Yu, T.; Fu, K.; Zhang, J.; Huang, Q.; Yu, J. Multi-granularity contrastive cross-modal collaborative generation for end-to-end long-term video question answering. *IEEE Trans. Image Process.* **2024**, *33*, 3115–3129. [CrossRef]
76. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Hanna, E.B.; Bressand, F.; et al. Mixtral of experts. *arXiv* **2024**, arXiv:2401.04088. [CrossRef]

77. Ma, F.; Jin, X.; Wang, H.; Xian, Y.; Feng, J.; Yang, Y. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Computer Vision Foundation: New York, NY, USA, 2024; pp. 13151–13160.
78. Ranasinghe, K.; Li, X.; Kahatapitiya, K.; Ryoo, M.S. Understanding Long Videos with Multimodal Language Models. *arXiv* **2024**, arXiv:2403.16998. [CrossRef]
79. Surís, D.; Menon, S.; Vondrick, C. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2023; pp. 11888–11898.
80. Zohar, O.; Wang, X.; Bitton, Y.; Szpektor, I.; Yeung-Levy, S. Video-star: Self-training enables video instruction tuning with any supervision. *arXiv* **2024**, arXiv:2407.06189.
81. Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv* **2022**, arXiv:2212.03191. [CrossRef]
82. Gao, D.; Ji, L.; Zhou, L.; Lin, K.Q.; Chen, J.; Fan, Z.; Shou, M.Z. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv* **2023**, arXiv:2306.08640.
83. Wang, X.; Zhang, Y.; Zohar, O.; Yeung-Levy, S. Videoagent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 58–76.
84. Kim, W.; Choi, C.; Lee, W.; Rhee, W. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access* **2024**, *12*, 193057–193075. [CrossRef]
85. Hu, K.; Gao, F.; Nie, X.; Zhou, P.; Tran, S.; Neiman, T.; Wang, L.; Shah, M.; Hamid, R.; Yin, B.; et al. M-LLM based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; Computer Vision Foundation: New York, NY, USA, 2025; pp. 13702–13712.
86. Liu, H.; Ilievski, F.; Snoek, C.G. Commonsense video question answering through video-grounded entailment tree reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; Computer Vision Foundation: New York, NY, USA, 2025; pp. 3262–3271.
87. Zhang, Y.; Zhao, Z.; Chen, Z.; Ding, Z.; Yang, X.; Sun, Y. Beyond training: Dynamic token merging for zero-shot video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2025; pp. 22046–22055.