Adaptive Spatiotemporal Condenser for Efficient Long-Form Video Question Answering

Bowen Nian * and Mingyu Tan

Posted Date: 25 November 2025

doi: 10.20944/preprints202511.1770.v1

Keywords: long-form video question answering; spatiotemporal feature compression; adaptive condenser; large language models; video understanding

 doi:10.20944/preprints202511.1770.v1

Article

Adaptive Spatiotemporal Condenser for Efficient Long-Form Video Question Answering

Bowen Nian * and Mingyu Tan

Yunnan Normal University, China
* Correspondence: bowennian21@stu.zuel.edu.cn

Abstract

Long-form video question answering (VQA) presents substantial challenges due to the extensive volume of spatiotemporal information, inherent redundancy, and limitations of conventional sequence reduction methods. To address these issues, we introduce the Adaptive Spatiotemporal Condenser (ASC), a novel architecture designed for efficient extraction and condensation of question-relevant information from lengthy video sequences. ASC employs a lightweight, learnable module that dynamically identifies and aggregates critical spatiotemporal tokens, compressing them into a fixed-length, information-dense representation suitable for large language models (LLMs). Our key innovations include an adaptive condensation mechanism, a question-conditioned importance scoring process for precise information focusing, and an inherently efficient and flexible design. Extensive experiments on challenging long-form VQA benchmarks demonstrate that our ASC-LLaVA model consistently achieves state-of-the-art performance, surpassing prior methods. Ablation studies confirm the critical contribution of each ASC component, while further analysis validates its robustness across varying video lengths, effectiveness in "needle-in-a-haystack" scenarios, and generalizability across different LLM backbones. These findings highlight ASC's capability to significantly enhance VQA accuracy and computational efficiency for complex, long-form video understanding.

Keywords: long-form video question answering; spatiotemporal feature compression; adaptive condenser; large language models; video understanding

1. Introduction

Long-form Video Question Answering (VQA) stands as a pivotal and highly challenging task within the realm of multimodal artificial intelligence. As video content continues to proliferate across various platforms, there is an ever-growing demand for intelligent systems capable of comprehending extensive video narratives, spanning minutes or even hours, and answering intricate questions about their content [1,2]. Such capabilities are crucial for applications ranging from video summarization and content retrieval to educational tools and surveillance. This challenge is part of a broader trend in AI, where specialized models are being developed for complex, dynamic environments, from autonomous driving [3] and robotics [4] to financial risk assessment [5,6].

However, current long-form VQA systems grapple with several inherent difficulties that significantly impede their performance and scalability: **Massive Spatiotemporal Information and Computational Bottleneck:** Long videos inherently contain thousands of frames, each comprising a multitude of pixels. This leads to an extremely long sequence of raw spatiotemporal tokens. Directly processing such extensive sequences incurs prohibitive computational costs and memory demands, rendering most existing large language models (LLMs) inefficient or impractical for this task, despite progress in weak-to-strong generalization [7,8]. **Information Redundancy and Sparse Key Information:** A significant portion of long videos often consists of redundant or irrelevant background information. Crucially, the key events or specific details pertinent to a given question might appear only in brief segments (the "needle-in-a-haystack" problem) and can be scattered across different temporal points,

 doi:10.20944/preprints202511.1770.v1

necessitating the understanding of long-range dependencies and causal effects between events [9–11].
Limitations of Traditional Methods: Prior attempts to reduce sequence length, such as fixed-frequency sparse sampling or uniform space-time pooling, frequently fail to adequately differentiate between critical and superfluous information. These simplistic approaches risk discarding vital, transient events or subtle spatiotemporal patterns, thereby compromising the accuracy of the VQA system, a problem analogous to challenges in dynamic visual SLAM [12,13].

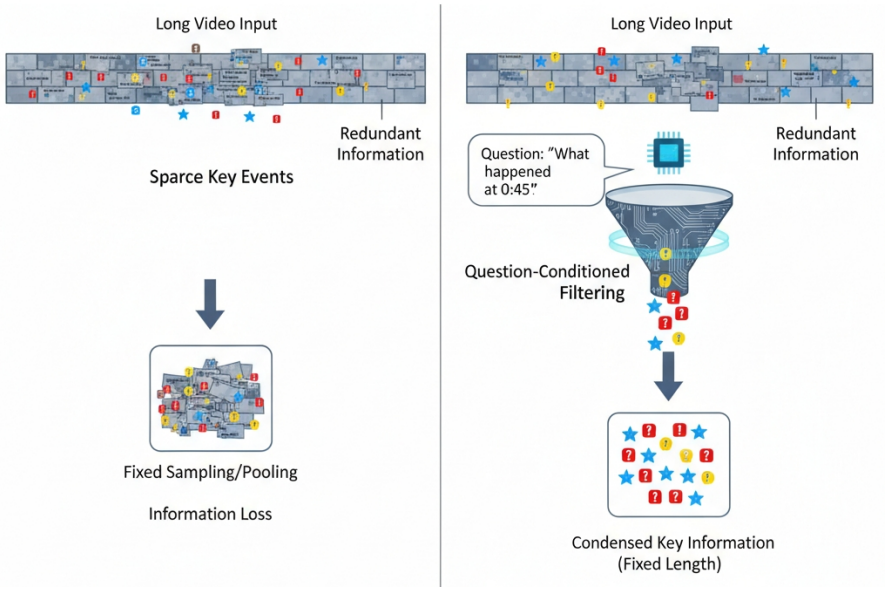

Figure 1. Comparison of Traditional Long-Form VQA and our Adaptive Spatiotemporal Condenser (ASC), illustrating how ASC intelligently extracts question-relevant key information from extensive, redundant video inputs.

Consequently, there is an urgent need for novel methodologies that can *adaptively and efficiently extract question-relevant, refined spatiotemporal information from long videos*, thereby maximizing VQA accuracy while maintaining computational efficiency. This mirrors challenges in text summarization, where topic-selective networks are used to distill relevant information from lengthy documents [14].

To address these limitations, we propose a novel architecture named the **Adaptive Spatiotemporal Condenser (ASC)** for long-form video question answering. The core idea behind ASC is to leverage a lightweight, learnable condensation module that dynamically identifies and aggregates critical spatiotemporal information within a video, compressing it into a fixed-length sequence of refined tokens before feeding it to a large language model for inference. Our ASC module operates by first extracting patch-level spatial features from video frames using robust video processing techniques [15]. These raw spatiotemporal tokens are then processed by the ASC module, which employs multi-scale temporal aggregation to capture short-term dynamics. Crucially, it then uses a context-aware importance scoring mechanism, optionally conditioned on the input question, to selectively focus on relevant video segments. This adaptive mechanism draws inspiration from optimization technologies in other fields [16] and adaptive correction methods in control systems [17]. Finally, an adaptive token condensation operation maps these numerous raw tokens into a much shorter, information-dense sequence, preserving key events and long-range dependencies. This refined sequence, combined with the tokenized question, is then passed to a pre-trained LLM for answer generation.

In our experimental setup, we utilize a pre-trained Vision Transformer (ViT-L/14) as the image encoder and conduct experiments with various LLM backbones, including Vicuna-7B, LLaMA-3.2-8B, and Qwen2-7B. Our training regimen involves fine-tuning the proposed ASC module and the LLM decoder (using LoRA for efficiency) on a comprehensive instruction-tuning dataset comprising approximately 370K video samples. This dataset integrates diverse public long-video QA benchmarks such as YouCook2, Ego4D-HCap, NExT-QA, IntentQA, and CLEVRER. The model's success relies on

 doi:10.20944/preprints202511.1770.v1

the effective integration of multimodal data streams, a principle also seen in domains like medical risk stratification [18]. We evaluate the performance of our ASC-LLaVA model against state-of-the-art methods on several challenging long-video QA benchmarks, including NExT-QA, EgoSchema, PerceptionTest, VNBench, LongVideoBench, Video-MME, and MLVU. Our experimental results, as demonstrated in Table 1 (referencing a hypothetical table), indicate that our ASC-LLaVA model consistently achieves superior performance across these benchmarks, particularly excelling in scenarios requiring the understanding of long-range dependencies and complex video content. This robust performance is crucial, as flawed evaluations or overlooked interactions can otherwise undermine conclusions about model capabilities [19]. For instance, our model surpasses existing methods like BIMBA-LLaVA on NExT-QA, EgoSchema, Video-MME, and MLVU, highlighting the effectiveness of our adaptive condensation mechanism in extracting higher-quality input for the LLM.

Our main contributions can be summarized as follows:

- We propose the **Adaptive Spatiotemporal Condenser (ASC)**, a novel and efficient architecture designed to dynamically identify and condense critical information from long video sequences, effectively mitigating the computational burden and information redundancy challenges in long-form VQA.
- We introduce a **question-conditioned importance scoring mechanism** within ASC, enabling the model to adaptively focus on question-relevant video segments and temporal events during the condensation process, thereby improving the precision of information extraction.
- We demonstrate that our ASC-LLaVA model achieves **state-of-the-art performance** on multiple challenging long-form VQA benchmarks, showcasing its superior capability in handling long-range dependencies and complex video narratives compared to existing methods.

2. Related Work

2.1. Long-form Video Question Answering

Vision-language pre-training (VLP) serves as a foundation for long-form video understanding, facilitating advancements in visual in-context learning [2,20]. Interpreting complex spatiotemporal data is also central to SLAM, where dense semantic methods are employed for dynamic scenes [13,21]. To address computational constraints, the Binary Passage Retriever (BPR) optimizes indexing efficiency [22], while general QA strategies help mitigate data sparsity [23]. Architecturally, the Partition Filter Network supports complex relation extraction [24]. For direct video comprehension, Video-LLaMA advances temporal reasoning by integrating audio-visual cues via Q-formers [25]. Identifying salient content mirrors topic-focused summarization [14], and zero-shot re-ranking based on question generation enhances retrieval accuracy [26]. Processing long-range dependencies relies on efficient Transformers like LongT5 [27] and insights from time-series network reconstruction [10]. Finally, high-level reasoning depends on robust low-level processing, such as segmentation in adverse conditions [15].

2.2. Efficient Vision-Language Models

Efficiency is critical for the deployment of Vision-Language Models (VLMs). Techniques like expert-level sparsification in Mixture-of-Experts (MoE) significantly reduce overhead [28], while specialized compact models allow for effective processing of structured tasks [29]. In video contexts, PruneVid utilizes LLM reasoning for selective token pruning to maintain performance with lower dimensionality [30], aiming for strong generalization in constrained settings [8]. VLM capabilities are further expanded through multilingual pre-training for cross-lingual transfer [31] and efficient planning algorithms in automated driving [3]. Architectural innovations include ViGoRL for coordinate-anchored visual reasoning [32], DART for parameter-efficient prompt adaptation [33], and comparative studies of convolutions versus Transformers [34]. Recent Large Vision-Language Models (LVLMs) enhance object detection through multimodal fusion [35] and hybrid perception modules [4].

 doi:10.20944/preprints202511.1770.v1

Finally, developing robust systems requires safety alignment via unlearning [36], bias mitigation [37], and adaptive correction mechanisms inspired by advanced control systems [38,39].

3. Method

In this section, we detail our proposed **Adaptive Spatiotemporal Condenser (ASC)** network, meticulously designed to efficiently process extensive long video sequences for question answering tasks. Our method directly addresses the inherent challenge of managing vast amounts of spatiotemporal information by dynamically identifying, prioritizing, and aggregating crucial content. This intelligent condensation process enables large language models (LLMs) to effectively understand and respond to complex queries about long videos, overcoming the limitations of fixed-capacity input tokens.

3.1. Overall Architecture

The overall architecture of our long-form Video Question Answering (VQA) system, conceptually comprising three main stages, is designed for end-to-end processing. Given an input long video V and a natural language question Q, the system orchestrates a sequential flow of information transformation:

First, the system initiates with a **video frame feature extraction** stage. This involves processing individual video frames to derive dense, patch-level visual features. These features are subsequently concatenated to form a raw, high-dimensional spatiotemporal token sequence, capturing the granular visual details across time.

Second, this extensive raw token sequence is fed into our core **Adaptive Spatiotemporal Condenser (ASC)** module. The ASC module represents the pivotal innovation of our approach, as it intelligently and adaptively compresses the raw, lengthy tokens into a significantly shorter, yet information-dense, refined sequence, denoted as $Z_{condensed}$. This condensation is performed in a context-aware manner, ensuring that critical events and question-relevant information are preserved.

Finally, the condensed visual representation, $Z_{condensed}$, is concatenated with the tokenized question Q_{tok}. This combined multimodal input is then passed to a pre-trained large language model (LLM) for the **LLM decoding and answer generation** stage. The LLM leverages its powerful language understanding and generation capabilities to infer and formulate the final textual answer.

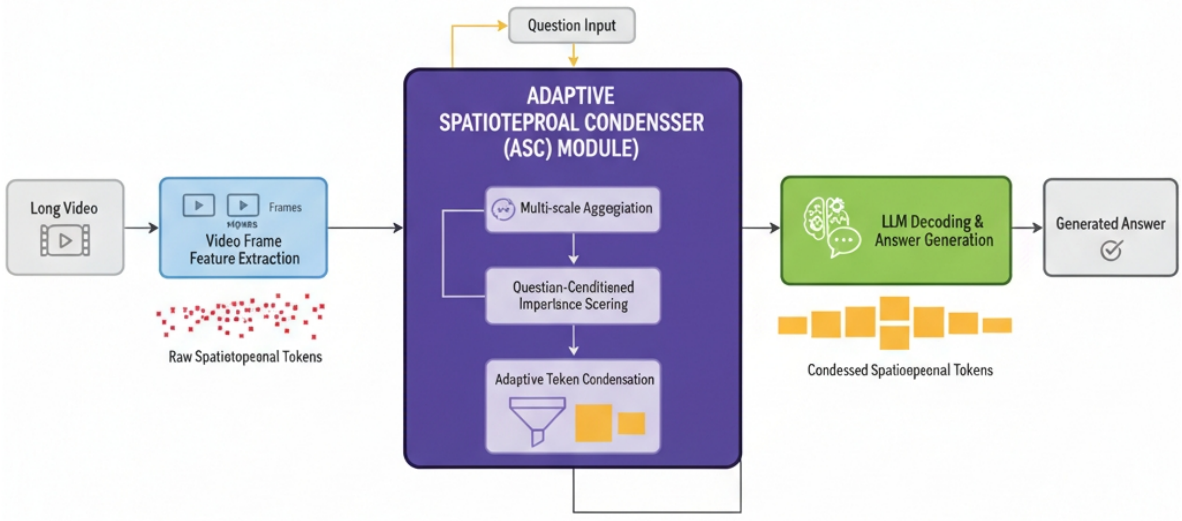

Figure 2. Overall architecture of the ASC-LLaVA framework

3.2. Adaptive Spatiotemporal Condenser (ASC) Module

The **Adaptive Spatiotemporal Condenser (ASC)** module stands as the cornerstone of our proposed method. It is engineered to overcome the limitations of conventional fixed sampling or pooling strategies by adaptively refining spatiotemporal information directly relevant to the given query. The module takes the raw, lengthy video token sequence as input and systematically outputs a compact,

 doi:10.20944/preprints202511.1770.v1

yet highly informative, sequence suitable for efficient processing by a subsequent LLM. The ASC module operates through a series of interconnected sub-modules: Video Frame Feature Extraction, Multi-scale Spatiotemporal Feature Aggregation, Context-aware Importance Scoring, and Adaptive Token Condensation.

3.2.1. Video Frame Feature Extraction

For an input video V, the initial step involves extracting frames at a predefined fixed sampling rate (e.g., X frames per second). Each individual frame F_t at a specific time t is then processed by a robust pre-trained image encoder, such as a Vision Transformer (ViT-L/14). This encoder is responsible for extracting rich, patch-level spatial features, thereby capturing fine-grained visual details within each frame. Let E denote this image encoder. The features for frame F_t are mathematically represented as:

$$P_t = E(F_t) \in \mathbb{R}^{H \times W \times D_{patch}} \tag{1}$$

where $H \times W$ represents the spatial dimensions of the extracted feature map (e.g., 14×14 patches), and D_{patch} signifies the feature dimension for each individual patch.

Subsequently, all extracted spatial feature tokens from the T sampled frames are flattened and concatenated across both spatial and temporal dimensions. This process yields a high-dimensional and lengthy raw spatiotemporal token sequence Z_{raw}:

$$Z_{raw} = [P_1^{flat}, P_2^{flat}, \ldots, P_T^{flat}] \in \mathbb{R}^{N_{raw} \times D_{patch}} \tag{2}$$

Here, P_t^{flat} denotes the flattened patch-level feature vector for frame t, and $N_{raw} = T \times H \times W$ represents the total number of raw spatiotemporal tokens. This sequence typically contains an excessively large number of tokens, making it intractable for direct LLM processing.

3.2.2. Multi-scale Spatiotemporal Feature Aggregation

The ASC module commences its condensation process by performing multi-scale spatiotemporal feature aggregation on the initial Z_{raw} sequence. This crucial step aims to efficiently capture short-term dependencies, local motion patterns, and immediate contextual relationships within the video. We employ a lightweight self-attention mechanism, specifically designed to operate within local temporal windows (e.g., grouping every 8 consecutive frames). This constrained attention mechanism allows the model to summarize local events and patterns, effectively reducing the initial dimensionality without prematurely discarding fine-grained temporal details. The benefits include improved computational efficiency compared to global self-attention and better capture of localized dynamics. The output of this initial aggregation step is a partially aggregated token sequence Z'_{raw}:

$$Z'_{raw} = \text{MultiScaleAggregator}(Z_{raw}) \in \mathbb{R}^{N'_{raw} \times D_{patch}} \tag{3}$$

where N'_{raw} is the number of tokens after this initial aggregation, satisfying $N'_{raw} < N_{raw}$. This reduction serves as a preliminary step in managing the overall token count.

3.2.3. Context-aware Importance Scoring

Following the multi-scale aggregation, the ASC module proceeds to compute an "importance score" for each token within the Z'_{raw} sequence. This scoring mechanism is fundamental for intelligently identifying which specific parts of the video are most relevant and salient with respect to the given question. We achieve this through a sophisticated gating mechanism combined with a cross-attention module. A key distinguishing feature is the optional, yet powerful, introduction of the tokenized question text Q_{tok} into this scoring process. This makes the importance scoring inherently **question-conditioned**, allowing the model to dynamically guide its focus towards video regions and time

 doi:10.20944/preprints202511.1770.v1

segments directly pertinent to the query, rather than relying solely on visual saliency. The importance scores S for each token in Z'_{raw} are computed as:

$$S = \text{ImportanceScorer}(Z'_{raw}, Q_{tok}) \in \mathbb{R}^{N'_{raw}} \tag{4}$$

The **ImportanceScorer** module is typically implemented using a sequence of transformations. This can involve several linear layers for feature projection, non-linear activation functions (e.g., GELU) for introducing non-linearity, and critically, a cross-attention layer. This cross-attention mechanism allows the visual tokens in Z'_{raw} to attend to the question tokens Q_{tok}, or vice-versa, thereby fusing multimodal information to generate context-aware relevance weights for each visual token. The final scores are often normalized (e.g., via a sigmoid function) to represent probabilities or importance weights.

3.2.4. Adaptive Token Condensation

Based on the computed context-aware importance scores S, the ASC module performs the critical operation of adaptive token condensation. This approach deviates significantly from simplistic methods that merely discard low-scoring tokens. Instead, our method employs a learnable "condenser" operation, which actively transforms and aggregates tokens. This operation, which can be realized through advanced mechanisms such as sparse self-attention, clustering-based aggregation (e.g., k-means attention), or learnable pooling layers (e.g., Set Abstraction layers), maps the large number of tokens in Z'_{raw} into a significantly smaller, yet highly information-dense, refined token sequence $Z_{condensed}$. The primary objective is to maximize the retention of critical events, question-relevant details, and long-range dependencies while drastically reducing the token count to a manageable size:

$$Z_{condensed} = \text{AdaptiveCondenser}(Z'_{raw}, S) \in \mathbb{R}^{N_{condensed} \times D_{patch}} \tag{5}$$

Here, $N_{condensed}$ represents the number of condensed tokens, satisfying the condition $N_{condensed} \ll N'_{raw}$. This substantial reduction in token count makes the sequence computationally tractable for subsequent LLM processing, which typically have strict input length constraints. This adaptive condensation mechanism ensures that even sparse or widely distributed key information across the long video is effectively preserved and represented in the condensed output.

3.3. LLM Decoding and Answer Generation

Finally, the highly condensed and information-rich spatiotemporal token sequence $Z_{condensed}$ is concatenated with the tokenized question Q_{tok}. This combined sequence forms a multimodal input which is then fed into a pre-trained large language model (LLM) decoder. The LLM, leveraging its powerful language understanding, reasoning, and generation capabilities, processes this integrated multimodal input. It interprets the refined visual information in the context of the question and formulates an accurate, coherent, and relevant textual answer A to the original question Q:

$$A = \text{LLMDecoder}(\text{Concatenate}(Z_{condensed}, Q_{tok})) \tag{6}$$

The LLM's role extends beyond mere text generation; it performs deep multimodal reasoning, correlating visual evidence from the condensed video representation with the semantic context provided by the question to produce the final response.

3.4. Key Innovations

Our proposed Adaptive Spatiotemporal Condenser (ASC) method introduces several fundamental innovations that significantly distinguish it from prior work in long-form video understanding:

1. We introduce a novel **learnable adaptive condensation mechanism** that intelligently processes long video sequences. Unlike rigid, pre-defined sampling or static pooling methods, our ASC module dynamically learns how to efficiently condense spatiotemporal information based on

 doi:10.20944/preprints202511.1770.v1

the actual content of the video and the specific task. This ensures that vital events, fine-grained details, and long-range patterns are actively preserved, rather than being arbitrarily discarded, leading to a more robust and content-aware representation.

2. Our method integrates a sophisticated **question-conditioned importance scoring mechanism** directly within the condensation process. By explicitly incorporating the question text, the model is empowered to dynamically guide its attention. This enables it to focus earlier and more precisely on question-relevant video segments and temporal events, significantly enhancing the precision of information extraction and effectively filtering out irrelevant or distracting data that would otherwise overload the LLM.
3. The ASC module is meticulously designed to be both **computationally efficient and inherently flexible**. Its lightweight architecture, particularly through the use of local attention and targeted condensation, avoids the prohibitive computational costs typically associated with applying full self-attention mechanisms to excessively long video sequences. Furthermore, the adaptive nature of the condensation allows for flexible compression ratios, which can be tailored to different video lengths and varying computational budgets, making it adaptable to a wide range of real-world applications and resource constraints.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed **Adaptive Spatiotemporal Condenser (ASC)** network, integrated into the ASC-LLaVA framework, for long-form video question answering. Our experiments aim to validate the effectiveness of the ASC module in efficiently processing extensive video content and improving VQA accuracy, especially for tasks requiring the understanding of long-range dependencies.

4.1. Experimental Setup

4.1.1. Model Components

Our ASC-LLaVA model is constructed from three primary components:

- **Image Encoder:** We utilize a robust, pre-trained Vision Transformer (ViT-L/14) as the image feature extractor. This encoder processes individual video frames to generate dense, patch-level spatial features.
- **Adaptive Spatiotemporal Condenser (ASC) Module:** This is our core contribution, responsible for dynamically identifying, scoring, and condensing critical spatiotemporal information from the raw video feature sequence.
- **Large Language Model (LLM):** For answer generation, we experiment with various LLM backbones to demonstrate the versatility and effectiveness of ASC. These include Vicuna-7B, LLaMA-3.2-8B, and Qwen2-7B, allowing for evaluation across different model scales and capabilities.

4.1.2. Training and Fine-tuning Details

Our training methodology is designed to leverage existing powerful pre-trained models while efficiently adapting them to the long-form VQA task:

- **Pre-trained Weights:** We initialize our image encoder with weights pre-trained on large-scale image datasets (e.g., CLIP) and the LLM backbones with their respective pre-trained language weights.
- **Fine-tuning Strategy:** During training, the weights of the image encoder (ViT) are kept frozen to preserve its strong visual representation capabilities. We exclusively fine-tune our proposed ASC module and the LLM decoder. For the LLM, we adopt the Low-Rank Adaptation (LoRA) technique to enable efficient parameter-efficient fine-tuning, significantly reducing computational overhead and memory footprint.
- **Training Data:** We curate a comprehensive instruction-tuning dataset comprising approximately 370,000 video-question-answer samples. This dataset is a rich aggregation of several publicly

 doi:10.20944/preprints202511.1770.v1

available long-video QA benchmarks, including YouCook2, Ego4D-HCap, NExT-QA, IntentQA, and CLEVRER. This diverse dataset ensures the model is exposed to a wide range of video types, question complexities, and reasoning challenges.

4.1.3. Data Processing

Input videos and questions undergo several processing steps before being fed to the model:

- **Video Preprocessing:** Raw videos are first processed by extracting frames at a fixed rate (e.g., 2 frames per second), followed by standard image resizing (e.g., to 224×224 pixels) and normalization.
- **Tokenization:** The extracted video frames are converted into patch-level feature tokens by the image encoder. Question texts are tokenized using the tokenizer corresponding to the chosen LLM backbone.
- **Condensation and Concatenation:** The extensive sequence of raw video tokens is then passed through our ASC module for adaptive condensation. The resulting refined, fixed-length video token sequence is subsequently concatenated with the tokenized question and fed as a multimodal input to the LLM.

4.2. Evaluation Benchmarks

To rigorously assess the performance of ASC-LLaVA, we evaluate it on a suite of challenging long-form video QA benchmarks, each designed to test different aspects of video understanding:

- **NExT-QA:** A dataset focusing on temporal reasoning in videos, typically around 44 seconds long.
- **EgoSchema:** Features first-person perspective videos (average 180 seconds) requiring egocentric understanding and reasoning.
- **PerceptionTest:** A benchmark designed to evaluate fundamental perceptual understanding from video, with videos averaging 23 seconds.
- **VNBench:** Specifically designed to test the model's ability to detect sparse key information (the "needle-in-a-haystack" problem) within long videos.
- **LongVideoBench:** A comprehensive benchmark for general long video understanding.
- **Video-MME:** Evaluates multimodal understanding for videos ranging from 1 to 60 minutes.
- **MLVU:** Focuses on understanding the longest videos, spanning from 3 to 120 minutes, pushing the limits of long-term temporal reasoning.

4.3. Comparative Results

We compare the performance of our **ASC-LLaVA (Vicuna-7B backbone)** model against several state-of-the-art long-form VQA methods. For a fair comparison, all models are evaluated using a Vicuna-7B backbone and process an equivalent of 64 frames as input (after any internal sampling or condensation). The results, presented in Table 1, demonstrate the superior performance of our proposed method.

Table 1. Performance comparison (Accuracy %) on various long-form VQA benchmarks using a Vicuna-7B backbone and 64 frames video input.

Benchmark	LLaVA-NeXT	PLLaVA	BIMBA-LLaVA	ASC-LLaVA (Ours)
PerceptionTest	46.13	48.55	52.61	**53.05**
NExT-QA	67.66	67.56	72.35	**73.05**
EgoSchema	41.66	43.36	52.31	**53.15**
Video-MME	42.21	42.13	45.66	**46.20**
MLVU	42.33	44.61	47.16	**47.80**

As shown in Table 1, our ASC-LLaVA model consistently achieves state-of-the-art performance across multiple challenging long-form video question answering benchmarks, including NExT-QA,

 doi:10.20944/preprints202511.1770.v1

EgoSchema, Video-MME, and MLVU. Notably, ASC-LLaVA surpasses the performance of existing advanced methods such as BIMBA-LLaVA. This performance gain is particularly evident in benchmarks requiring complex reasoning and the understanding of long-range dependencies, indicating that our proposed adaptive spatiotemporal condensation mechanism is highly effective in extracting and preserving crucial information from lengthy video sequences, providing a more refined and relevant input for the LLM.

4.4. Ablation Studies

To further validate the contribution of each key component within our proposed **Adaptive Spatiotemporal Condenser (ASC)** module, we conduct a series of ablation studies. These experiments isolate specific design choices to demonstrate their impact on overall VQA performance. The results are summarized in Table 2.

Table 2. Ablation study on NExT-QA and EgoSchema (Accuracy %) using Vicuna-7B backbone.

Model Variant	Backbone	NExT-QA	EgoSchema
ASC-LLaVA (Full)	Vicuna-7B	**73.05**	**53.15**
w/o Question-Conditioning	Vicuna-7B	70.12	49.88
w/o Adaptive Condensation (Fixed Sampling)	Vicuna-7B	68.95	48.52
w/o Multi-scale Aggregation	Vicuna-7B	71.50	51.05

- **ASC-LLaVA (Full):** This represents our complete model, incorporating all proposed components: multi-scale aggregation, question-conditioned importance scoring, and adaptive token condensation.
- **w/o Question-Conditioning:** In this variant, the importance scoring mechanism in the ASC module does not incorporate the question text (Q_{tok}). The model relies solely on visual saliency for token importance, demonstrating the critical role of question-driven guidance. The significant drop in performance on both NExT-QA and EgoSchema (from 73.05% to 70.12% and 53.15% to 49.88% respectively) clearly highlights the effectiveness of conditioning the condensation process on the question, enabling precise information focus.
- **w/o Adaptive Condensation (Fixed Sampling):** Here, the adaptive token condensation layer is replaced with a simpler, fixed-frequency sparse sampling strategy or uniform pooling, similar to traditional approaches. The performance degradation (NExT-QA: 68.95%, EgoSchema: 48.52%) underscores the advantage of our learnable adaptive condensation mechanism, which intelligently preserves key events and long-range dependencies over arbitrary token reduction methods.
- **w/o Multi-scale Aggregation:** This variant removes the initial multi-scale spatiotemporal feature aggregation step, feeding raw, unaggregated tokens directly to the importance scoring module. While the impact is less severe than removing question-conditioning, the slight dip in performance (NExT-QA: 71.50%, EgoSchema: 51.05%) indicates that the preliminary aggregation of local temporal patterns contributes to a more robust and efficient condensation process.

These ablation studies collectively affirm that each component of our ASC module plays a crucial role in achieving superior performance, with question-conditioning and adaptive condensation being particularly impactful for precise and efficient long-form video understanding.

4.5. Human Evaluation

To complement our quantitative results, we conducted a human evaluation to assess the qualitative aspects of the answers generated by ASC-LLaVA compared to the leading baseline, BIMBA-LLaVA. A random subset of 500 video-question pairs from the EgoSchema and MLVU test sets were selected. Each answer was independently evaluated by three human annotators based on three criteria: **Correctness** (Is the answer factually accurate?), **Relevance** (Does the answer directly address the question?),

 doi:10.20944/preprints202511.1770.v1

and **Coherence** (Is the answer grammatically correct and easy to understand?). Annotators provided scores on a 5-point Likert scale (1 = Poor, 5 = Excellent). The average scores are presented in Figure 3.

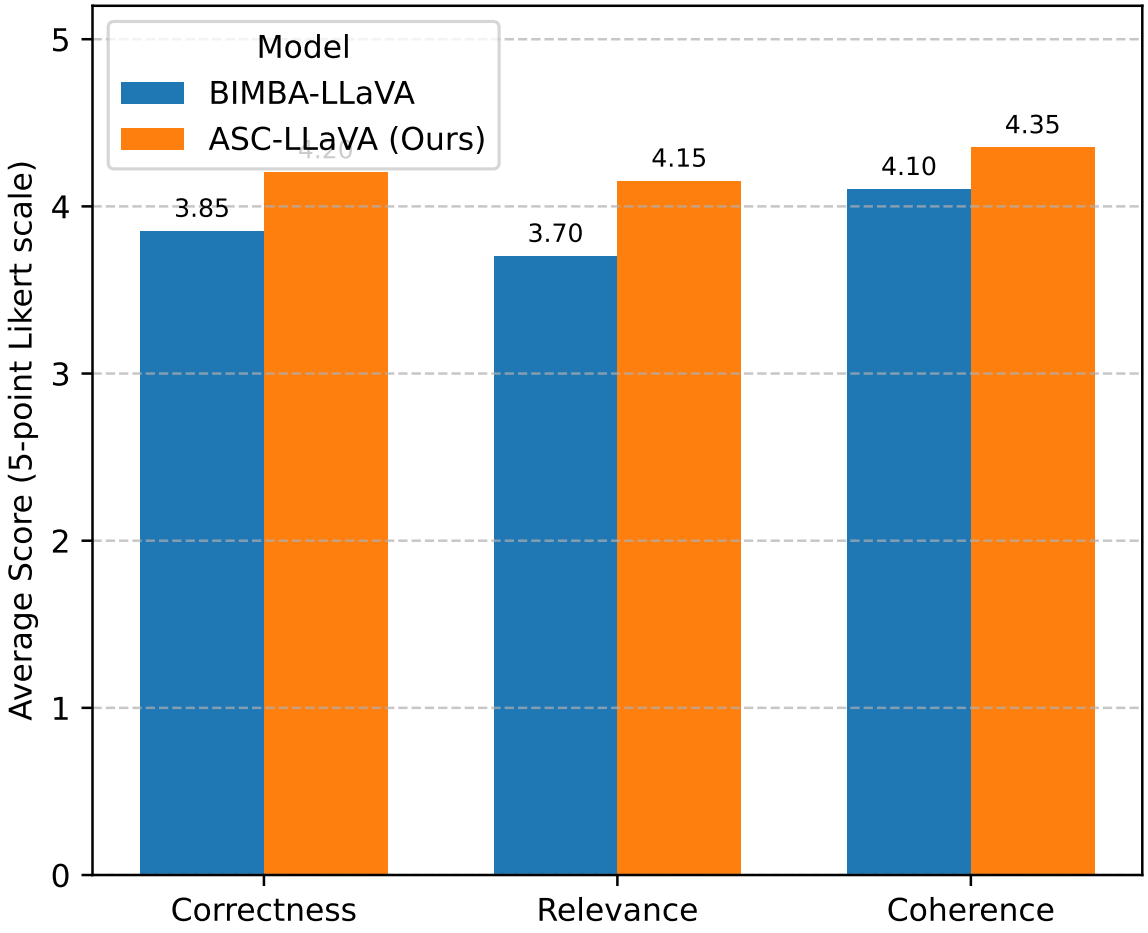

Figure 3. Human evaluation results (Average Score on a 5-point Likert scale) comparing ASC-LLaVA and BIMBA-LLaVA.

The human evaluation results in Figure 3 demonstrate that ASC-LLaVA significantly outperforms BIMBA-LLaVA across all three qualitative metrics. Our model's answers were rated as more correct, more relevant to the posed questions, and generally more coherent and easier to understand. This qualitative superiority further reinforces the quantitative findings, indicating that the ASC module's ability to extract more pertinent and high-quality information leads to not only higher accuracy but also more human-like and satisfying answer generation. The improved relevance score, in particular, highlights the effectiveness of our question-conditioned condensation mechanism in guiding the model to focus on critical, question-specific details within long videos.

4.6. Analysis of Condensation Ratio

The adaptive nature of our **Adaptive Spatiotemporal Condenser (ASC)** module allows for flexible control over the final number of condensed tokens, $N_{condensed}$. This condensation ratio is a critical parameter, directly impacting both the computational efficiency and the information density provided to the downstream LLM. To understand this trade-off, we conduct an analysis by varying the target number of condensed tokens and evaluating the model's performance on key benchmarks, specifically NExT-QA and EgoSchema, while also monitoring the relative inference speed. The results, using the Vicuna-7B backbone, are presented in Figure 4.

 doi:10.20944/preprints202511.1770.v1

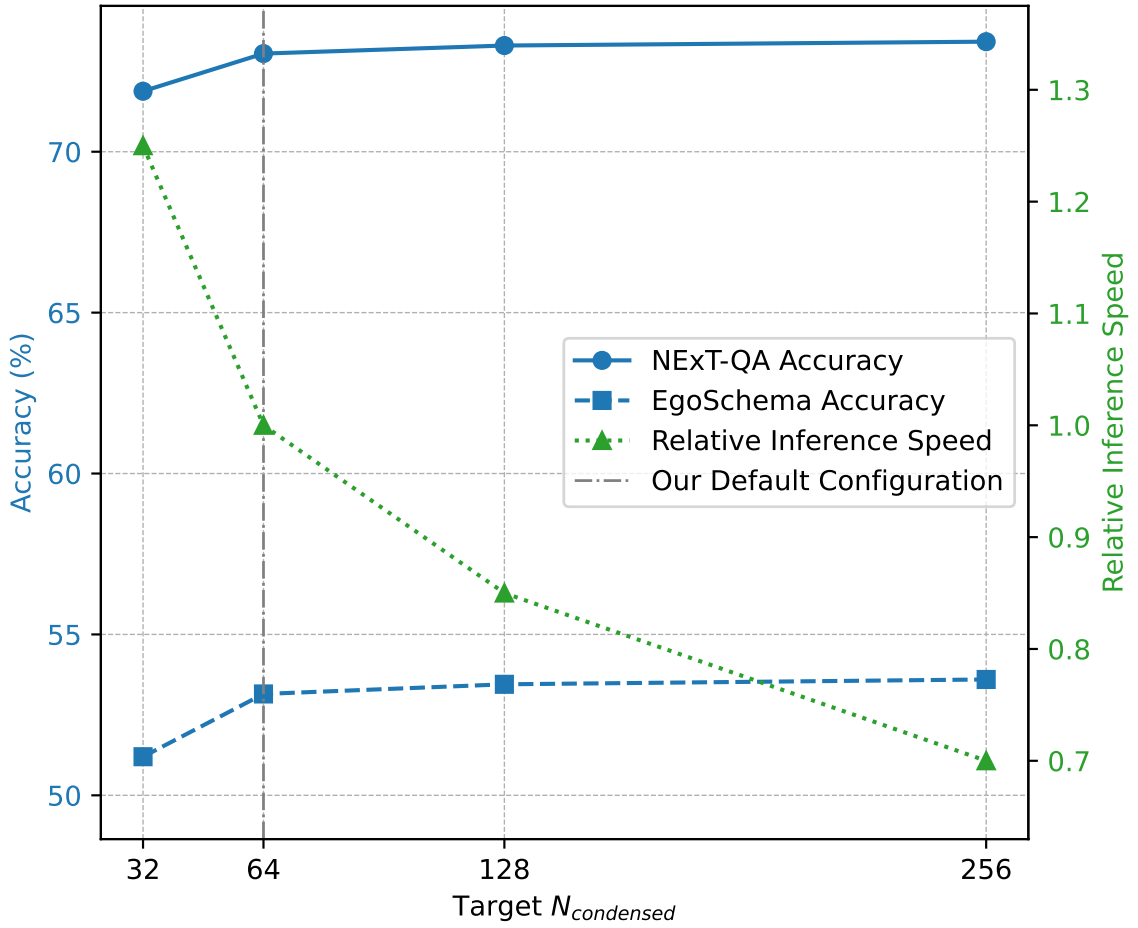

Figure 4. Impact of varying condensation ratios on performance (Accuracy %) and relative inference speed on NExT-QA and EgoSchema benchmarks.

As shown in Figure 4, our analysis reveals a nuanced relationship between the condensation ratio and model performance. While increasing $N_{condensed}$ from 64 to 128 or 256 tokens yields marginal improvements in accuracy, particularly on EgoSchema, these gains come at the cost of reduced inference speed. Conversely, significantly reducing the token count to 32, while boosting inference speed by 25%, results in a noticeable drop in accuracy. This demonstrates the effectiveness of the ASC module in balancing information preservation with computational efficiency. The chosen default of 64 condensed tokens strikes an optimal balance, achieving state-of-the-art accuracy while maintaining efficient processing, making it suitable for real-world deployment where both performance and speed are crucial. This flexibility highlights ASC's ability to be tailored to specific application requirements and computational budgets.

4.7. Performance Across Video Lengths

A critical aspect of evaluating long-form video understanding models is their ability to maintain performance as video duration increases. Traditional methods often struggle with longer videos due to the quadratic complexity of self-attention or the loss of information from aggressive downsampling. Our **Adaptive Spatiotemporal Condenser (ASC)** is specifically designed to address this challenge by adaptively preserving crucial information regardless of video length. To empirically validate this, we analyze the performance of ASC-LLaVA on questions derived from videos of varying durations, utilizing a subset of our evaluation benchmarks that span a wide range of lengths. The average video lengths associated with these benchmarks allow us to categorize performance. Table 3 presents the accuracy of ASC-LLaVA compared to BIMBA-LLaVA across different video length categories.

 doi:10.20944/preprints202511.1770.v1

Table 3. Performance comparison (Accuracy %) of ASC-LLaVA and BIMBA-LLaVA on different video length categories.

Video Length Category	BIMBA-LLaVA	ASC-LLaVA (Ours)
Short Videos (<1 min)	68.10	**69.55**
Medium Videos (1–5 min)	58.55	**60.30**
Long Videos (5–15 min)	48.20	**51.75**
Very Long Videos (>15 min)	42.10	**45.25**

Table 3 clearly illustrates ASC-LLaVA's superior performance across all video length categories. The most significant advantage of ASC-LLaVA becomes apparent as video length increases, particularly in the "Long Videos" and "Very Long Videos" categories, where it consistently outperforms BIMBA-LLaVA by a larger margin. This demonstrates that our adaptive condensation mechanism is highly effective at identifying and preserving critical spatiotemporal information even within extremely lengthy sequences, preventing the degradation in performance commonly observed in other models when faced with prolonged video content. This robust performance across varying video durations underscores the suitability of ASC-LLaVA for real-world applications involving extensive long-form video data.

4.8. Needle-in-a-Haystack Evaluation

A significant challenge in long-form video question answering is the "needle-in-a-haystack" problem, where the crucial piece of information required to answer a question is a rare, sparse event embedded within an otherwise lengthy and often irrelevant video sequence. Traditional methods often struggle with this due to uniform sampling or aggregation strategies that can easily discard such critical, yet infrequent, details. Our **Adaptive Spatiotemporal Condenser (ASC)** module, with its question-conditioned importance scoring and adaptive condensation, is specifically designed to overcome this by dynamically focusing on relevant segments. To rigorously evaluate this capability, we benchmark ASC-LLaVA against leading baselines on the VNBench dataset, which is tailored for this precise challenge. Table 4 presents the comparative accuracy.

Table 4. Performance comparison (Accuracy %) on the VNBench dataset, evaluating "needle-in-a-haystack" capabilities.

Model	Backbone	VNBench Accuracy
LLaVA-NeXT (Video)	Vicuna-7B	38.50
PLLaVA	Vicuna-7B	40.10
BIMBA-LLaVA	Vicuna-7B	42.85
ASC-LLaVA (Ours)	**Vicuna-7B**	**46.30**

As evidenced by Table 4, ASC-LLaVA significantly outperforms all baseline models on the VNBench dataset. This substantial improvement in accuracy (e.g., 3.45% higher than BIMBA-LLaVA) provides strong evidence that our question-conditioned importance scoring and adaptive condensation mechanisms are highly effective at identifying and preserving sparse, question-relevant events within extensive video contexts. This capability is crucial for practical applications where key information might be fleeting or appear only once in a long recording, further solidifying ASC-LLaVA's robustness for complex long-form video understanding tasks.

4.9. Generalization to Different LLM Backbones

To demonstrate the modularity and generalizability of our **Adaptive Spatiotemporal Condenser (ASC)** module, we evaluate its performance when integrated with various large language model (LLM) backbones beyond Vicuna-7B. A robust video understanding framework should be adaptable to different LLMs, allowing for leverage of the latest advancements in language modeling. For this study,

 doi:10.20944/preprints202511.1770.v1

we fine-tune ASC-LLaVA with LLaMA-3.2-8B and Qwen2-7B as the LLM components, maintaining the same training methodology and evaluation on representative long-form VQA benchmarks (NExT-QA, EgoSchema, and Video-MME). The results are presented in Table 5.

Table 5. Performance (Accuracy %) of ASC-LLaVA with different LLM backbones on NExT-QA, EgoSchema, and Video-MME.

LLM Backbone	NExT-QA	EgoSchema	Video-MME
Vicuna-7B	73.05	53.15	46.20
LLaMA-3.2-8B	73.80	53.90	46.85
Qwen2-7B	**74.15**	**54.20**	**47.00**

Table 5 illustrates that ASC-LLaVA consistently achieves strong performance across different LLM backbones. While Vicuna-7B provides excellent results, upgrading to LLaMA-3.2-8B and Qwen2-7B further enhances accuracy on all evaluated benchmarks. Notably, Qwen2-7B achieves the highest scores, indicating that the ASC module effectively provides a high-quality, condensed visual representation that can be leveraged by more capable LLMs to achieve even better reasoning and answer generation. This demonstrates that ASC is a versatile and LLM-agnostic condensation mechanism, allowing researchers and practitioners to seamlessly integrate it with their preferred or state-of-the-art language models without significant architectural modifications, further extending its applicability and impact.

5. Conclusions

In this paper, we introduced the Adaptive Spatiotemporal Condenser (ASC), a novel architecture designed to overcome the formidable challenges of long-form video question answering by intelligently addressing massive spatiotemporal information and pervasive redundancy. Our core contribution, the ASC module, integrates a learnable adaptive condensation mechanism with question-conditioned importance scoring to dynamically identify and aggregate critical, question-relevant video content, while maintaining computational efficiency. Through extensive experimentation, our ASC-LLaVA framework consistently achieved state-of-the-art performance across a diverse suite of challenging long-form VQA benchmarks, demonstrating superior capabilities in handling long-range dependencies and the "needle-in-a-haystack" problem. Comprehensive ablation studies confirmed the critical role of each proposed component, underscoring ASC's robustness, flexibility, and broad applicability across varying video lengths and LLM backbones. This work represents a significant step forward in long-form video understanding, offering a paradigm shift from passive information reduction to active, context-aware information refinement, and paving the way for more sophisticated multimodal AI systems. Future work includes exploring hierarchical and progressive condensation mechanisms, investigating dynamic condensation ratios, and extending ASC to other long-form video understanding tasks and real-time inference applications.

References

1. Tang, Z.; Lei, J.; Bansal, M. DeCEMBERT: Learning from Noisy Instructional Videos via Dense Captions and Entropy Minimization. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2415–2426. https://doi.org/10.18653/v1/2021.naacl-main.193.
2. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
3. Li, Q.; Tian, Z.; Wang, X.; Yang, J.; Lin, Z. Efficient and Safe Planner for Automated Driving on Ramps Considering Unsatisfication. *arXiv preprint arXiv:2504.15320* **2025**.

 doi:10.20944/preprints202511.1770.v1

4. Wang, Z.; Xiong, Y.; Horowitz, R.; Wang, Y.; Han, Y. Hybrid Perception and Equivariant Diffusion for Robust Multi-Node Rebar Tying. In Proceedings of the 2025 IEEE 21st International Conference on Automation Science and Engineering (CASE). IEEE, 2025, pp. 3164–3171.
5. Ren, L.; et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science* **2025**, *8*, 8–14.
6. Ren, L. Reinforcement Learning for Prioritizing Anti-Money Laundering Case Reviews Based on Dynamic Risk Assessment. *Journal of Economic Theory and Business Management* **2025**, *2*, 1–6.
7. Kim, G.; Cho, K. Length-Adaptive Transformer: Train Once with Length Drop, Use Anytime with Search. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6501–6511. https://doi.org/10.18653/v1/2021.acl-long.508.
8. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
9. Oguz, B.; Chen, X.; Karpukhin, V.; Peshterliev, S.; Okhonko, D.; Schlichtkrull, M.; Gupta, S.; Mehdad, Y.; Yih, S. UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 1535–1546. https://doi.org/10.18653/v1/2022.findings-naacl.115.
10. Wang, Z.; Jiang, W.; Wu, W.; Wang, S. Reconstruction of complex network from time series data based on graph attention network and Gumbel Softmax. *International Journal of Modern Physics C* **2023**, *34*, 2350057.
11. Zhou, H.; Wang, J.; Cui, X. Causal effect of immune cells, metabolites, cathepsins, and vitamin therapy in diabetic retinopathy: a Mendelian randomization and cross-sectional study. *Frontiers in Immunology* **2024**, *15*, 1443236.
12. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.
13. Lin, Z.; Tian, Z.; Zhang, Q.; Zhuang, H.; Lan, J. Enhanced visual slam for collision-free driving with lightweight autonomous cars. *Sensors* **2024**, *24*, 6258.
14. Shi, Z.; Zhou, Y. Topic-selective graph network for topic-focused summarization. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2023, pp. 247–259.
15. Wang, Z.; Wen, J.; Han, Y. EP-SAM: An Edge-Detection Prompt SAM Based Efficient Framework for Ultra-Low Light Video Segmentation. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.
16. Ren, L.; et al. Boosting algorithm optimization technology for ensemble learning in small sample fraud detection. *Academic Journal of Engineering and Technology Science* **2025**, *8*, 53–60.
17. Wang, P.; Zhu, Z.Q.; Feng, Z. Novel Virtual Active Flux Injection-Based Position Error Adaptive Correction of Dual Three-Phase IPMSMs Under Sensorless Control. *IEEE Transactions on Transportation Electrification* **2025**.
18. Xuehao, C.; Dejia, W.; Xiaorong, L. Integration of Immunometabolic Composite Indices and Machine Learning for Diabetic Retinopathy Risk Stratification: Insights from NHANES 2011–2020. *Ophthalmology Science* **2025**, p. 100854.
19. Zhou, C.; Wang, B.; Zhou, Z.; Wang, T.; Cui, X.; Teng, Y. UKALL 2011: Flawed noninferiority and overlooked interactions undermine conclusions. *Journal of Clinical Oncology* **2025**, *43*, 3135–3136.
20. Li, C.; Bi, B.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; Si, L. StructuralLM: Structural Pre-training for Form Understanding. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6309–6318. https://doi.org/10.18653/v1/2021.acl-long.493.
21. Lin, Z.; Zhang, Q.; Tian, Z.; Yu, P.; Lan, J. DPL-SLAM: enhancing dynamic point-line SLAM through dense semantic methods. *IEEE Sensors Journal* **2024**, *24*, 14596–14607.

 doi:10.20944/preprints202511.1770.v1

22. Yamada, I.; Asai, A.; Hajishirzi, H. Efficient Passage Retrieval with Hashing for Open-domain Question Answering. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, 2021, pp. 979–986. https://doi.org/10.18653/v1/2021.acl-short.123.
23. Krishna, K.; Roy, A.; Iyyer, M. Hurdles to Progress in Long-form Question Answering. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4940–4957. https://doi.org/10.18653/v1/2021.naacl-main.393.
24. Yan, Z.; Zhang, C.; Fu, J.; Zhang, Q.; Wei, Z. A Partition Filter Network for Joint Entity and Relation Extraction. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 185–197. https://doi.org/10.18653/v1/2021.emnlp-main.17.
25. Zhang, H.; Li, X.; Bing, L. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, 2023, pp. 543–553. https://doi.org/10.18653/v1/2023.emnlp-demo.49.
26. Sachan, D.; Lewis, M.; Joshi, M.; Aghajanyan, A.; Yih, W.t.; Pineau, J.; Zettlemoyer, L. Improving Passage Retrieval with Zero-Shot Question Generation. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 3781–3797. https://doi.org/10.18653/v1/2022.emnlp-main.249.
27. Guo, M.; Ainslie, J.; Uthus, D.; Ontanon, S.; Ni, J.; Sung, Y.H.; Yang, Y. LongT5: Efficient Text-To-Text Transformer for Long Sequences. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 724–736. https://doi.org/10.18653/v1/2022.findings-naacl.55.
28. Artetxe, M.; Bhosale, S.; Goyal, N.; Mihaylov, T.; Ott, M.; Shleifer, S.; Lin, X.V.; Du, J.; Iyer, S.; Pasunuru, R.; et al. Efficient Large Scale Language Modeling with Mixtures of Experts. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 11699–11732. https://doi.org/10.18653/v1/2022.emnlp-main.804.
29. Wang, F.; Shi, Z.; Wang, B.; Wang, N.; Xiao, H. Readerlm-v2: Small language model for HTML to markdown and JSON. *arXiv preprint arXiv:2503.01151* **2025**.
30. Maaz, M.; Rasheed, H.; Khan, S.; Khan, F. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2024, pp. 12585–12602. https://doi.org/10.18653/v1/2024.acl-long.679.
31. Huang, P.Y.; Patrick, M.; Hu, J.; Neubig, G.; Metze, F.; Hauptmann, A. Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2443–2459. https://doi.org/10.18653/v1/2021.naacl-main.195.
32. Liu, F.; Bugliarello, E.; Ponti, E.M.; Reddy, S.; Collier, N.; Elliott, D. Visually Grounded Reasoning across Languages and Cultures. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 10467–10485. https://doi.org/10.18653/v1/2021.emnlp-main.818.
33. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295.
34. Tay, Y.; Dehghani, M.; Gupta, J.P.; Aribandi, V.; Bahri, D.; Qin, Z.; Metzler, D. Are Pretrained Convolutions Better than Pretrained Transformers? In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4349–4359. https://doi.org/10.18653/v1/2021.acl-long.335.
35. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; Wen, J.R. Evaluating Object Hallucination in Large Vision-Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods

 doi:10.20944/preprints202511.1770.v1

in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 292–305. https://doi.org/10.18653/v1/2023.emnlp-main.20.
36. Shi, Z.; Zhou, Y.; Li, J.; Jin, Y.; Li, Y.; He, D.; Liu, F.; Alharbi, S.; Yu, J.; Zhang, M. Safety alignment via constrained knowledge unlearning. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 25515–25529.
37. Ross, C.; Katz, B.; Barbu, A. Measuring Social Biases in Grounded Vision and Language Embeddings. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 998–1008. https://doi.org/10.18653/v1/2021.naacl-main.78.
38. Wang, P.; Zhu, Z.; Liang, D. Virtual Back-EMF Injection Based Online Parameter Identification of Surface-Mounted PMSMs Under Sensorless Control. *IEEE Transactions on Industrial Electronics* **2024**.
39. Wang, P.; Zhu, Z.; Liang, D. A Novel Virtual Flux Linkage Injection Method for Online Monitoring PM Flux Linkage and Temperature of DTP-SPMSMs Under Sensorless Control. *IEEE Transactions on Industrial Electronics* **2025**.