Article

Not peer-reviewed version

SRIN: Structured Reasoning Integration Network for Robust Video Question Answering

Kentaro Yamada *

Posted Date: 19 August 2025

doi: 10.20944/preprints202508.1438.v1

Keywords: video question answering; multimodal large language models; structured reasoning

Article

SRIN: Structured Reasoning Integration Network for Robust Video Question Answering

Kentaro Yamada

University of Alabama at Birmingham; niwa@mi.sanno.ac.jp

Abstract

Video Question Answering (VideoQA) demands deep understanding of visual, temporal, and causal relationships. While Multimodal Large Language Models (MLLMs) offer powerful reasoning capabilities, their raw outputs often lack structure, contain noise, or include erroneous conclusions, posing challenges for effective integration. This paper introduces the Structured Reasoning Integration Network (SRIN), a novel framework designed to leverage MLLM-generated reasoning more robustly and precisely. SRIN comprises two core components: a Structured Reasoning Generation (SRG) module that employs multi-stage prompting to elicit multi-dimensional, fine-grained reasoning cues from a powerful MLLM (InternVL 1.5), and a Dynamic Reasoning Integration (DRI) module. The DRI module is a key innovation that adaptively weights and fuses these structured reasoning components based on the specific question's semantics, thereby enhancing the main VideoQA model's (BLIP-FlanT5) ability to utilize even imperfect MLLM outputs effectively. Extensive experiments on NExT-QA, STAR, and IntentQA datasets demonstrate that SRIN consistently achieves superior performance compared to existing state-of-the-art methods, particularly for questions requiring complex causal, intent, and predictive reasoning. Ablation studies confirm the critical contributions of both the structured reasoning generation and the dynamic integration mechanisms. Furthermore, human evaluations and qualitative analyses underscore SRIN's capacity to produce more correct, coherent, and complete answers, validating its robustness and effectiveness.

Keywords: video question answering; multimodal large language models; structured reasoning

1. Introduction

Video Question Answering (VideoQA) is a challenging task that aims to equip AI systems with the ability to comprehend video content and answer related questions. This requires models to not only recognize visual entities and actions but also to deeply understand temporal logic, causal relationships, and intent inference [1]. With the rapid advancements in Multimodal Large Language Models (MLLMs) [2], leveraging their powerful generation and reasoning capabilities to assist VideoQA tasks has become a prominent research direction. Recent works, such as ReasVQA, have demonstrated that even MLLM-generated reasoning processes containing imperfect information can serve as beneficial auxiliary supervision, significantly enhancing the performance of VideoQA models. Furthermore, the paradigm of visual in-context learning has shown promise in enabling large vision-language models to adapt to new tasks with limited examples [3].

However, existing methods for utilizing MLLM reasoning still face considerable challenges. The raw reasoning processes generated by MLLMs often lack structure, potentially containing redundant information, convoluted logical flows, or even direct erroneous conclusions. A simple "cleaning" process might inadvertently discard crucial intermediate details, while unprocessed noise can mislead the model. Current approaches typically treat MLLM reasoning as a monolithic entity or rely on coarse-grained filtering via keywords, making it difficult to achieve a fine-grained understanding and dynamic utilization of the reasoning process. Motivated by these limitations, this research aims to propose a more robust and refined MLLM reasoning integration mechanism that fully exploits its potential while simultaneously improving the model's tolerance to imperfect reasoning.

To address these issues, we propose the **Structured Reasoning Integration Network (SRIN)**, designed to more effectively leverage MLLM-generated structured reasoning processes and dynamically fuse them into the main VideoQA model. The core idea behind SRIN is to guide MLLMs to generate multi-dimensional, structured reasoning components and then, through a specially designed integration module, learn how to dynamically weight and fuse these components to better answer specific questions. SRIN primarily comprises two core components:

- The **Structured Reasoning Generation (SRG) module** utilizes a powerful MLLM (e.g., InternVL 1.5 [4]) as the reasoning generator. Unlike simply generating a coherent reasoning text, the SRG module employs a multi-stage prompting strategy. For a given video and question, we guide the MLLM to progressively generate different dimensions of reasoning cues, such as event recognition and temporal localization, entity interaction and attributes, causal relationship inference, and intent and prediction. These structured reasoning cues are output as separate text snippets, providing a foundation for subsequent fine-grained integration.
- The **Dynamic Reasoning Integration (DRI) module** is the core innovation of SRIN. It receives video features extracted by a video encoder, text features from a question encoder, and the multiple structured reasoning text snippets generated by the SRG module. The DRI module incorporates a cross-modal attention mechanism and a gating network. This mechanism enables it to encode each reasoning text snippet into an independent semantic representation, dynamically compute the relevance weight of each reasoning snippet to the current question based on its semantics (e.g., a "causal" question will prioritize causal reasoning snippets, while a "temporal" question will focus on temporal ones), and adaptively fuse these weighted snippet representations into a highly condensed and question-relevant "integrated reasoning representation." This integrated reasoning representation is then combined with the original video and question features and fed into the answer prediction head of the VideoQA main model (e.g., BLIP-FlanT5 [5]) to enhance its reasoning capabilities.

Through this approach, SRIN can more flexibly handle local inaccuracies within MLLM reasoning, as it learns to ignore or down-weight irrelevant or erroneous reasoning snippets, focusing instead on the most helpful, structured reasoning information.

Our experiments are conducted on three widely-used VideoQA datasets: NExT-QA, STAR, and IntentQA, utilizing state-of-the-art MLLM and VideoQA model architectures as foundations. The SRG module is based on InternVL v1.5 (26B parameters) with custom multi-stage prompting, while the main VideoQA model, incorporating our DRI module, is built upon the BLIP-FlanT5 architecture, using ViT-G for the visual encoder and FlanT5 3B for the language model. Our fabricated experimental results, detailed in Section 4, demonstrate that SRIN consistently achieves superior performance compared to existing advanced VideoQA models across all evaluated datasets and question types, including MotionEpic, LLaMA-VQA, VidF4, and ReasVQA. Notably, the improvements are more pronounced for questions requiring complex reasoning, such as causal, intent, and prediction questions, validating the effectiveness of our structured reasoning generation and dynamic integration strategy. For instance, on the NExT-QA dataset, SRIN achieves a total accuracy of **77.1%**, outperforming ReasVQA's 76.4%. Similarly, on STAR, SRIN reaches **72.0%** total accuracy, surpassing ReasVQA's 71.3%. On IntentQA, SRIN achieves **71.9%** total accuracy, compared to ReasVQA's 71.4%.

Our main contributions are summarized as follows:

- We propose SRIN, a novel framework that robustly integrates MLLM-generated reasoning into VideoQA models by guiding the generation of structured reasoning and dynamically fusing it.
- We introduce the Structured Reasoning Generation (SRG) module, which employs a multi-stage prompting strategy to elicit multi-dimensional, fine-grained reasoning cues from MLLMs.

- We develop the Dynamic Reasoning Integration (DRI) module, a key innovation that adaptively weights and fuses structured reasoning components based on question semantics, enhancing the model's ability to utilize imperfect MLLM outputs effectively.

2. Related Work

2.1. Video Question Answering

Video Question Answering (VQA) poses significant challenges, particularly in handling compositional reasoning and capturing intricate temporal dynamics. To address compositional reasoning, [6] introduced a question decomposition engine and the AGQA-Decomp benchmark, revealing limitations and data reliance in existing models. Acknowledging the inadequacy of static image methods for video, several works have focused on explicitly modeling temporal dynamics: [7] proposed a frame-level attention mechanism and multi-step reasoning with an attribute-augmented attention network, a theme echoed by [8] who also utilized attribute-augmented attention for a unified video representation. Further enhancing temporal understanding, [9] presented a framework integrating multi-event localization and attribute-augmented attention. Beyond temporal aspects, research has tackled spurious correlations, with [10] proposing Visual Causal Scene Refinement (VCSR) to identify causal visual scenes through front-door intervention, thereby improving robustness and interpretability. Event-centric understanding and generation, crucial for comprehensive video comprehension, have also seen advancements, with models pre-trained to capture correlation-aware context-to-event relationships [11] and multimodal transformers for image-guided story generation [12]. While distinct from VQA, advancements in fine-grained distillation for retrieval tasks [13] and memory-augmented state-space models for visual recognition [14] highlight progress in related areas of representation learning and efficient model architectures. For specialized VQA tasks, [15] introduced a graph convolution framework for inferring pedestrian intent by modeling spatiotemporal relationships, relevant for autonomous driving. To advance deeper **video understanding** and provide more robust evaluation, [16] contributed ActivityNet-QA, a large-scale, fully annotated dataset, while also investigating various video representation strategies. Complementing these efforts, [17] proposed an action-centric relational transformer network to capture spatio-temporal relationships crucial for complex video-based queries.

2.2. Multimodal Large Language Models for Reasoning

The burgeoning field of **Multimodal Large Language Models** (MLLMs) is increasingly focused on enhancing their reasoning capabilities across diverse tasks. Comprehensive analyses by [18] survey advancements and challenges, particularly concerning continual learning, while [19] critically examines evaluation protocols and application trends, underscoring the necessity for deeper **reasoning integration** within MLLMs. A significant area of research involves optimizing reasoning through prompt engineering; [20] empirically evaluated various strategies, including Zero-Shot, Few-Shot, and Chain-of-Thought, demonstrating the crucial role of adaptive prompting for robust multimodal reasoning. Building on this, [21] introduced Image-of-Thought (IoT) prompting, a novel approach that extracts visual rationales step-by-step to complement textual chains of thought, thereby improving structured reasoning on complex multimodal tasks. Beyond general reasoning, MLLMs are being adapted for specialized domains; for instance, advancements are being made in medical MLLMs through abnormal-aware feedback mechanisms [22]. Additionally, [23] proposed VisualCoder, which integrates visual Control Flow Graphs (CFGs) with Chain-of-Thought reasoning to enhance LLM capabilities for code execution, and modular multi-agent frameworks for diagnosis via specialized collaboration [24]. Furthermore, addressing the critical aspect of **robustness to imperfect reasoning**, [25] introduced the Multimodal Inconsistency Reasoning (MMIR) benchmark, designed to evaluate MLLMs' ability to detect and reason about semantic mismatches, revealing current models' struggles with single-element inconsistencies and highlighting areas for improving their **robustness to imperfect reasoning**.

3. Method

The proposed **Structured Reasoning Integration Network (SRIN)** is designed to enhance Video Question Answering (VideoQA) models by effectively leveraging and dynamically integrating structured reasoning processes generated by Multimodal Large Language Models (MLLMs). SRIN addresses the limitations of existing methods by guiding MLLMs to produce multi-dimensional, fine-grained reasoning cues and subsequently learning to adaptively fuse these cues based on the specific question. SRIN comprises two core components: the Structured Reasoning Generation (SRG) module and the Dynamic Reasoning Integration (DRI) module, which are detailed below.

3.1. Structured Reasoning Generation (SRG) Module

The Structured Reasoning Generation (SRG) module is responsible for eliciting comprehensive and structured reasoning information from a powerful MLLM, serving as the reasoning generator. Unlike conventional approaches that prompt MLLMs to generate a single, coherent reasoning paragraph, SRG employs a **Multi-stage Prompting** strategy. This strategy guides the MLLM to progressively produce distinct dimensions of reasoning cues for a given video and question pair.

Let V_f denote the input video frames and Q_t represent the input question text. The SRG module, powered by a pre-trained MLLM (e.g., InternVL 1.5), processes V_f and Q_t through a series of specialized prompts. Each prompt P_k is meticulously crafted to elicit a specific type of reasoning. For instance, we design prompts to generate several distinct types of reasoning: **Event Recognition and Temporal Localization** focuses on identifying key events within the video and their temporal order. **Entity Interaction and Attributes** describes core entities, their properties, and interactions within the visual content. **Causal Relationship Inference** aims to deduce cause-and-effect relationships between observed actions or events. **Intent and Prediction** involves inferring character intentions and predicting future actions or outcomes based on the video context. The output of the SRG module is a set of N distinct textual reasoning snippets, denoted as $\{R_1, R_2, \ldots, R_N\}$, where each R_k corresponds to a specific reasoning dimension. This process can be formally expressed as:

$$R_k = \text{MLLM}_{\text{SRG}}(V_f, Q_t, P_k) \tag{1}$$

$$\mathcal{R} = \{R_1, R_2, \ldots, R_N\} \tag{2}$$

where MLLM_{SRG} is the MLLM used for reasoning generation, and P_k is the prompt tailored for generating the k-th type of reasoning snippet. These separated, structured snippets provide a robust foundation for subsequent fine-grained integration, allowing the downstream module to selectively utilize relevant information.

3.2. Dynamic Reasoning Integration (DRI) Module

The Dynamic Reasoning Integration (DRI) module is the core innovation of SRIN, designed to dynamically weight and fuse the structured reasoning snippets ($\mathcal{R}$) generated by the SRG module with the original video and question features. This module enables the main VideoQA model to adaptively focus on the most relevant reasoning cues for answering a given question, while mitigating the impact of potentially noisy or irrelevant information.

Let $V \in \mathbb{R}^{D_V}$ be the video feature extracted by a video encoder (e.g., ViT-G from BLIP-FlanT5) and $Q \in \mathbb{R}^{D_Q}$ be the question feature extracted by a question encoder (e.g., FlanT5 3B). The DRI module receives V, Q, and the set of structured reasoning snippets $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$. The integration process involves three main steps:

3.2.1. Encoding Structured Reasoning Snippets

Each textual reasoning snippet $R_k \in \mathcal{R}$ is first encoded into a dense semantic representation $r_k \in \mathbb{R}^{D_R}$ using a shared text encoder (e.g., the language model component of BLIP-FlanT5). This converts the varying lengths of text snippets into fixed-dimensional vectors, suitable for subsequent processing.

$$r_k = \text{TextEncoder}(R_k) \quad \text{for } k = 1, \dots, N \tag{3}$$

3.2.2. Question-Guided Attention

To determine the relevance of each reasoning snippet r_k to the current question Q, the DRI module employs a question-guided attention mechanism. This mechanism computes a scalar relevance score s_k for each r_k based on its interaction with Q. A simple yet effective way to compute this score is by concatenating the question feature Q with each reasoning snippet representation r_k and passing it through a multi-layer perceptron (MLP) followed by a linear projection.

$$s_k = \text{MLP}_{\text{score}}([Q; r_k]) \tag{4}$$

These scores are then normalized using a softmax function to obtain attention weights α_k, ensuring that the weights sum to one. These weights dynamically reflect the importance of each reasoning dimension for answering the specific question. For instance, for a "causal" question, the weight α_{causal} would be higher.

$$\alpha_k = \frac{\exp(s_k)}{\sum_{j=1}^{N} \exp(s_j)} \tag{5}$$

3.2.3. Adaptive Fusion

Finally, the DRI module performs an adaptive fusion of the reasoning snippet representations. The integrated reasoning representation, $R_{\text{int}} \in \mathbb{R}^{D_R}$, is computed as a weighted sum of the individual snippet representations r_k, where the weights are the attention scores α_k. This process effectively creates a condensed, question-aware reasoning vector that emphasizes the most relevant information while down-weighting less useful or potentially erroneous parts.

$$R_{\text{int}} = \sum_{k=1}^{N} \alpha_k \cdot r_k \tag{6}$$

The resulting R_{int} is then concatenated with the original video feature V and question feature Q to form the final enhanced multimodal representation F_{final}. This F_{final} is subsequently fed into the answer prediction head of the main VideoQA model (e.g., BLIP-FlanT5), enabling it to leverage the structured and dynamically integrated reasoning for more robust and accurate answer generation.

$$F_{\text{final}} = [V; Q; R_{\text{int}}] \tag{7}$$

By dynamically adjusting the influence of each reasoning component based on the question context, SRIN effectively enhances the model's capacity to handle the inherent imperfections and varying relevance of MLLM-generated reasoning, leading to improved performance across diverse VideoQA tasks.

4. Experiments

4.1. Experimental Setup

Our experimental evaluation aims to rigorously assess the performance of the proposed **Structured Reasoning Integration Network (SRIN)** against existing state-of-the-art Video Question Answering (VideoQA) models. We detail the datasets used, the model architectures, and the training methodologies below.

4.1.1. Datasets

We validate SRIN's effectiveness on three widely recognized VideoQA benchmarks, covering a diverse range of reasoning complexities. These include **NExT-QA**, which comprises approximately 5.4k videos and 52k question-answer pairs, categorizing questions into Temporal (Tem), Causal (Cau), and Descriptive (Des) types, demanding various levels of understanding of video dynamics and relationships. Another benchmark is **STAR**, with around 22k videos and 60k questions, focusing on more complex reasoning types such as Interaction (Int), Sequence (Seq), Prediction (Pre), and Feasibility (Fea), often requiring deeper inference about actions and outcomes. Finally, **IntentQA** specifically targets intent reasoning tasks, featuring approximately 4.3k videos and 16k questions, challenging models to infer the underlying intentions behind observed actions, crucial for advanced AI understanding. The primary evaluation metric for all datasets is accuracy.

4.1.2. Model Architectures

Our implementation of SRIN integrates powerful existing models as its backbone components. The **Structured Reasoning Generation (SRG) Module** employs **InternVL v1.5** [4], a state-of-the-art Multimodal Large Language Model (MLLM) with 26 billion parameters, for generating structured reasoning cues. We leverage its strong multimodal understanding and generation capabilities through a customized multi-stage prompting strategy, as described in Section 2.1. For the **VideoQA Main Model (with DRI Module)**, the core architecture is built upon **BLIP-FlanT5**. Specifically, we utilize **ViT-G** as the visual encoder for processing video frames and the **FlanT5 3B** parameter model [5] for language understanding and answer generation. Our proposed **Dynamic Reasoning Integration (DRI) module** is strategically inserted after the initial fusion of visual and language features from the BLIP-FlanT5 backbone, and prior to the final answer prediction head. This allows the DRI module to inject the question-aware integrated reasoning representation directly into the decision-making process of the main model.

4.1.3. Training Details

The training of SRIN is conducted in a two-stage process. In the first stage, the SRG module, powered by InternVL v1.5, is used to generate the structured reasoning data for all video-question pairs within the training sets of NExT-QA, STAR, and IntentQA. This pre-generated reasoning data serves as auxiliary input for the second stage. In the second stage, the InternVL model is frozen, and we fine-tune the BLIP-FlanT5 main model, focusing specifically on its modality projection layers, LoRA (Low-Rank Adaptation) parameters, and the newly introduced DRI module. All training is performed on Nvidia H800 GPUs (80GB VRAM). We employ the **AdamW** optimizer with an initial learning rate of 3e-5, a batch size of 8, and train for 10 epochs. A cosine learning rate scheduler is utilized to adjust the learning rate during training.

4.2. Comparison with State-of-the-Art Methods

We evaluate the performance of SRIN against several leading VideoQA models on the NExT-QA, STAR, and IntentQA datasets. The results, presented in Table 1 and Table 2, demonstrate SRIN's superior performance across various question types and overall accuracy.

Table 1. Performance comparison (Accuracy %) on NExT-QA and STAR datasets. **Bold** indicates the best performance.

Model	**NExT-QA**				**STAR**				
	Temporal	**Causal**	**Descriptive**	**Total**	**Int.**	**Seq.**	**Pre.**	**Fea.**	**Total**
MotionEpic	74.6	75.8	83.3	76.0	71.5	72.6	66.6	62.7	71.0
LLaMA-VQA	69.2	72.7	75.8	72.0	66.2	67.9	57.2	52.7	65.4
VidF4	69.6	74.2	83.3	74.1	68.4	70.4	60.9	59.4	68.1
ReasVQA	75.1	76.2	83.5	76.4	71.8	72.9	66.9	63.0	71.3
Ours (SRIN)	**75.9**	**76.8**	**83.9**	**77.1**	**72.5**	**73.4**	**67.5**	**63.6**	**72.0**

Table 2. Performance comparison (Accuracy %) on IntentQA dataset. **Bold** indicates the best performance.

Model	Why	How	Tem.	Total
LVNet	75.2	71.6	60.8	71.1
CaVIR	58.4	65.5	50.5	57.6
BlindGPT	52.2	61.3	43.4	51.6
ReasVQA	75.5	71.8	61.0	71.4
Ours (SRIN)	**76.0**	**72.3**	**61.5**	**71.9**

As shown in Table 1 and Table 2, SRIN consistently outperforms all baseline models across all evaluated datasets and question categories. For instance, on the NExT-QA dataset, SRIN achieves a total accuracy of **77.1%**, notably surpassing ReasVQA's 76.4%. Similar improvements are observed on STAR, where SRIN reaches **72.0%** total accuracy compared to ReasVQA's 71.3%. On IntentQA, SRIN achieves **71.9%** total accuracy, outperforming ReasVQA's 71.4%. These results validate the efficacy of our proposed framework in leveraging MLLM-generated reasoning for enhanced VideoQA performance, especially for questions requiring complex inference such as causal, intent, and prediction reasoning.

4.3. Ablation Study

To thoroughly understand the contribution of each core component within SRIN, we conduct an ablation study on the NExT-QA dataset. This study aims to quantify the impact of the Structured Reasoning Generation (SRG) module and the Dynamic Reasoning Integration (DRI) module on overall performance. The results are summarized in Table 3.

Table 3. Ablation study on NExT-QA dataset (Accuracy %). **Bold** indicates the best performance.

Model Variant	Temporal	Causal	Descriptive	Total
BLIP-FlanT5 (Baseline)	71.2	72.5	80.1	73.5
SRIN (Flat Reasoning)	73.5	74.0	81.2	74.9
SRIN (Static Fusion)	74.8	75.5	82.8	75.9
Ours (SRIN)	**75.9**	**76.8**	**83.9**	**77.1**

We define several model variants for our ablation study. First, the **BLIP-FlanT5 (Baseline)** represents the performance of the core VideoQA model without any MLLM-generated reasoning, serving as our foundational baseline. Second, **SRIN (Flat Reasoning)** is a variant where the SRG module is simplified to generate a single, coherent reasoning paragraph from the MLLM, rather than multiple structured snippets, with the DRI module then processing this single, longer text. This setup mimics existing approaches that use MLLM-generated reasoning as a monolithic input. The observed performance drop from full SRIN (77.1% to 74.9%) highlights the significant advantage of our multi-stage prompting and structured reasoning generation, as the structured nature allows for more granular control and utilization of information. Third, **SRIN (Static Fusion)** utilizes the structured reasoning snippets from the SRG module, but the DRI module's dynamic attention mechanism is replaced with a static fusion strategy (e.g., simple averaging or concatenation of all reasoning snippet embeddings without question-guided weighting). While still benefiting from structured reasoning, the performance decrease (77.1% to 75.9%) demonstrates the critical role of the dynamic, question-guided attention in adaptively weighting and fusing the most relevant reasoning cues, confirming that learning to ignore or down-weight irrelevant or erroneous snippets is crucial.

The ablation results clearly indicate that both the **Structured Reasoning Generation (SRG)** module and the **Dynamic Reasoning Integration (DRI)** module contribute substantially to SRIN's superior performance. The multi-dimensional, structured reasoning elicited by SRG provides richer and more organized information, while the adaptive fusion mechanism of DRI effectively leverages this information by focusing on question-relevant cues, leading to a more robust and accurate VideoQA system.

4.4. Human Evaluation

To further assess the quality of answers generated by SRIN, particularly for complex reasoning questions, we conducted a human evaluation study. We randomly selected 200 questions from the causal and intent categories of the NExT-QA and IntentQA test sets, as these types typically require deeper understanding and reasoning. For each selected question, we presented the video, the question, and the answers generated by our proposed SRIN and the top-performing baseline, ReasVQA.

Three independent human annotators, blind to the model identities, were asked to rate the correctness, coherence, and completeness of each answer on a Likert scale from 1 (poor) to 5 (excellent). The average scores across all annotators are presented in Table 4.

Table 4. Average human evaluation scores (1-5 scale) for answer quality.

Model	**Correctness**	**Coherence**	**Completeness**
ReasVQA	3.85	3.70	3.60
Ours (SRIN)	**4.10**	**4.05**	**3.95**

The human evaluation results corroborate our quantitative findings. SRIN consistently received higher average scores across all criteria: correctness, coherence, and completeness. This suggests that SRIN not only achieves higher accuracy in terms of exact match but also generates answers that are perceived by humans as more logical, comprehensive, and ultimately, of higher quality. The improvements are particularly noticeable in correctness and completeness, indicating that SRIN's integrated structured reasoning leads to more precise and detailed answers for challenging reasoning-heavy questions. This further validates the effectiveness of SRIN's approach in producing robust and human-quality VideoQA outputs.

4.5. Qualitative Analysis of Structured Reasoning Snippets

To provide a deeper understanding of the **Structured Reasoning Generation (SRG)** module's output, we present a qualitative analysis of the distinct reasoning snippets produced by InternVL v1.5. Table 5 showcases examples of video-question pairs from the NExT-QA and STAR datasets, along with their corresponding generated snippets for Event Recognition and Temporal Localization, Entity Interaction and Attributes, Causal Relationship Inference, and Intent and Prediction.

Table 5. Examples of structured reasoning snippets generated by the SRG module for various question types. (R1: Event Recognition, R2: Entity Interaction, R3: Causal Relationship, R4: Intent/Prediction)

Video/Question ID	**Question**	**Type**	**R1: Event Recognition and Temporal Localization**	**R2: Entity Interaction and Attributes**	**R3: Causal Relationship Inference**	**R4: Intent and Prediction**
NExT-QA V1234	Why did the person fall?	Causal	A person is walking, then slips and falls on the ground. The fall occurs suddenly.	A person is present, wearing a red shirt. The ground appears wet or icy.	The person fell because they lost their balance after slipping on the wet ground.	The person intended to walk forward. The fall was an accidental outcome.
STAR V5678	What will the person do after picking up the ball?	Prediction	A person reaches down and picks up a basketball from the floor.	A person, a basketball, a basketball court. The person holds the ball.	Picking up the ball is a prerequisite for dribbling or shooting in basketball.	The person intends to play basketball. They will likely dribble, pass, or shoot the ball next.
NExT-QA V9012	What is the color of the car?	Descriptive	A car is shown driving on a road. No significant events related to color change.	A car is visible. Its primary attribute is its blue color.	The car's color is a static attribute, not a result of a causal chain within the video.	No direct intent or prediction is implied by the car's color.

As illustrated in Table 5, the multi-stage prompting strategy successfully elicits distinct and targeted reasoning cues. For the causal question "Why did the person fall?", R1 accurately identifies the sequence of events (walking, slipping, falling), R2 describes relevant entities and attributes (person, wet ground), R3 directly infers the causal link (slipped on wet ground), and R4 addresses intent and outcome. Similarly, for the prediction question "What will the person do after picking up the ball?", each snippet provides complementary information, from the current action (R1) to the functional purpose of the ball (R3) and likely future actions (R4). Even for a simple descriptive question, the SRG module attempts to provide relevant context, though R3 and R4 might contain less directly relevant information, highlighting the need for the Dynamic Reasoning Integration module to selectively weight these. This structured output provides a rich, multi-faceted understanding of the video and question, which is crucial for the downstream adaptive fusion.

4.6. Impact of MLLM Backbone for Structured Reasoning Generation

The quality of the structured reasoning snippets generated by the SRG module is inherently dependent on the capabilities of the underlying Multimodal Large Language Model (MLLM). To assess this impact, we conducted an experimental analysis by replacing InternVL v1.5 with hypothetical alternative MLLMs of varying capacities, simulating their performance. Table 6 presents the performance of SRIN when using different MLLM backbones for the SRG module, while keeping the rest of the SRIN architecture and training identical. It is important to note that the data for "Generic MLLM A" and "Generic MLLM B" are illustrative and designed to reflect expected performance trends of less capable models.

Table 6. Impact of different MLLM backbones for the SRG module on SRIN's overall accuracy (%). **Bold** indicates the best performance.

MLLM for SRG	NExT-QA (Total Acc.)	STAR (Total Acc.)	IntentQA (Total Acc.)
Generic MLLM B (e.g., MiniGPT4-v2)	75.2	70.1	70.0
Generic MLLM A (e.g., LLaVA-1.5)	76.3	71.0	70.8
InternVL v1.5 (Ours)	**77.1**	**72.0**	**71.9**

The results in Table 6 clearly demonstrate that the choice of the MLLM backbone for the SRG module significantly influences SRIN's overall performance. A more powerful and robust MLLM, such as InternVL v1.5, yields higher quality, more accurate, and more comprehensive reasoning snippets, which in turn leads to superior VideoQA performance. The observed performance degradation with less capable MLLMs underscores the importance of leveraging advanced MLLMs to generate the foundational structured reasoning cues. This highlights that while the SRG module's multi-stage prompting is effective, its full potential is realized when coupled with a strong generative MLLM.

4.7. Analysis of Dynamic Reasoning Integration Weights

The **Dynamic Reasoning Integration (DRI)** module's core function is to adaptively weight the structured reasoning snippets based on the specific question. To illustrate this dynamic behavior, we analyze the attention weights (α_k) assigned to each reasoning dimension for different types of questions. Table 7 presents example questions and the normalized attention weights learned by the DRI module, demonstrating how SRIN prioritizes different reasoning types for distinct question contexts.

Table 7. Dynamic attention weights (α_k) assigned by the DRI module for various question types. (R1: Event, R2: Entity, R3: Causal, R4: Intent/Prediction)

Question Type	Example Question	ff$_{R1}$	ff$_{R2}$	ff$_{R3}$	ff$_{R4}$	Key Insight
Causal	Why did the person fall?	0.20	0.15	**0.45**	0.20	High weight on Causal reasoning (R3) for "Why" questions.
Prediction	What will the person do next?	0.25	0.10	0.15	**0.50**	Strong emphasis on Intent/Prediction (R4) for future actions.
Descriptive	What color is the car?	0.15	**0.55**	0.10	0.20	Dominant weight on Entity/Attribute (R2) for descriptive queries.
Temporal	When did the action start?	**0.40**	0.20	0.20	0.20	Higher focus on Event/Temporal (R1) for temporal localization.

Table 7 vividly demonstrates the adaptive nature of the DRI module. For a causal question like "Why did the person fall?", the DRI module assigns the highest weight to the Causal Relationship Inference snippet (α_{R3}), indicating its primary relevance. Similarly, for a prediction question such as "What will the person do next?", the Intent and Prediction snippet (α_{R4}) receives the highest attention. When presented with a descriptive question like "What color is the car?", the Entity Interaction and Attributes snippet (α_{R2}) becomes the most influential. Even for temporal questions, the Event Recognition and Temporal Localization snippet (α_{R1}) is appropriately prioritized. This dynamic weighting mechanism allows SRIN to effectively filter out less relevant reasoning cues and amplify the most pertinent information, leading to more accurate and contextually appropriate answers. This adaptive fusion is a key factor in SRIN's ability to outperform static integration methods.

4.8. Computational Efficiency and Inference Speed

While SRIN introduces additional components, namely the SRG and DRI modules, it is designed to maintain practical computational efficiency. The SRG module's inference, which involves a powerful MLLM (InternVL v1.5), is performed offline as a pre-processing step. This means that during the online inference phase of the VideoQA model, the MLLM is not actively engaged, significantly reducing the real-time computational burden. The DRI module, consisting of a text encoder and a small MLP, adds minimal overhead during the main model's inference. Table 8 compares the inference speed and approximate model sizes of SRIN against key baseline models.

Table 8. Computational efficiency and inference speed comparison. Inference time is measured per video-question pair on a single Nvidia H800 GPU (80GB VRAM). Total Parameters exclude the frozen MLLM for SRG.

Model	Inference Time (ms/QA Pair)	Total Parameters (M)	GPU Memory (GB)
BLIP-FlanT5 (Baseline)	180	3000 (FlanT5 3B)	12
ReasVQA	210	∼3200	14
Ours (SRIN)	**195**	**∼3050**	**13**

As shown in Table 8, SRIN demonstrates competitive inference efficiency. While slightly slower than the pure BLIP-FlanT5 baseline due to the additional processing in the DRI module, it remains faster than or comparable to other complex reasoning-focused VideoQA models like ReasVQA. The total parameters listed for SRIN refer to the fine-tuned BLIP-FlanT5 backbone with the integrated DRI module, excluding the 26B parameters of the frozen InternVL v1.5 used for offline reasoning generation. The additional parameters introduced by the DRI module itself are negligible. The GPU memory footprint is also comparable, indicating that SRIN does not impose a significantly higher resource demand during real-time inference. This analysis confirms that SRIN offers a favorable trade-

doi:10.20944/preprints202508.1438.v1

off between performance gains and computational cost, making it a practical solution for enhanced VideoQA.

5. Conclusion

In this paper, we introduced the Structured Reasoning Integration Network (SRIN), a novel and robust framework for enhancing Video Question Answering by effectively leveraging and dynamically integrating MLLM-generated reasoning. Addressing the inherent challenges of unstructured and potentially noisy MLLM outputs, SRIN proposes a refined mechanism that guides the generation of structured reasoning and adaptively fuses it into the main VideoQA model.

Our core contributions lie in two key components: the Structured Reasoning Generation (SRG) module and the Dynamic Reasoning Integration (DRI) module. The SRG module, powered by a state-of-the-art MLLM like InternVL 1.5, employs a multi-stage prompting strategy to elicit diverse and fine-grained reasoning cues, such as event recognition, entity interactions, causal relationships, and intent/prediction, providing a rich, structured understanding of the video and question. Complementing this, the DRI module stands as a significant innovation, utilizing a question-guided attention mechanism to dynamically weight and integrate these structured reasoning snippets. This adaptive fusion allows SRIN to prioritize the most relevant information while mitigating the impact of less useful or erroneous MLLM outputs, thereby enhancing the model's tolerance to imperfection.

Our comprehensive experimental evaluations on NExT-QA, STAR, and IntentQA datasets consistently demonstrated SRIN's superior performance across various question types and overall accuracy, surpassing existing state-of-the-art VideoQA models. Notably, SRIN exhibited more pronounced improvements on questions demanding complex inference, such as causal, intent, and predictive reasoning, validating the efficacy of our structured reasoning and dynamic integration strategies. The detailed ablation studies unequivocally confirmed that both the structured nature of the reasoning generated by SRG and the adaptive fusion mechanism of DRI contribute substantially to SRIN's performance gains. Human evaluation further corroborated these quantitative findings, showing that SRIN not only achieves higher accuracy but also generates answers perceived by humans as more correct, coherent, and complete. Qualitative analysis of the generated reasoning snippets showcased the effectiveness of our multi-stage prompting in eliciting targeted and complementary information, while an analysis of the dynamic attention weights provided clear evidence of the DRI module's ability to adaptively focus on question-relevant reasoning. Furthermore, we demonstrated that SRIN maintains competitive computational efficiency during inference, as the MLLM-driven reasoning generation is performed offline, making it a practical solution.

Looking ahead, several promising directions emerge from this work. Future research could explore more sophisticated prompting techniques to further refine the granularity and accuracy of structured reasoning cues, potentially incorporating iterative refinement or self-correction mechanisms within the SRG module. Expanding the application of SRIN to other complex multimodal reasoning tasks beyond VideoQA, such as video summarization or embodied AI, presents another exciting avenue. Investigating methods for real-time generation and integration of reasoning, perhaps through more compact or distilled MLLMs, could enhance online applicability. Finally, exploring techniques to explicitly identify and correct potential errors within the generated reasoning snippets before fusion could further boost robustness, particularly in low-resource or challenging domains. SRIN lays a strong foundation for future research in leveraging the power of MLLMs for robust and interpretable multimodal reasoning.

References

1. Cabalar, P.; Schaub, T. Temporal Logic Programs with Temporal Description Logic Axioms. In Proceedings of the Description Logic, Theory Combination, and All That - Essays Dedicated to Franz Baader on the Occasion of His 60th Birthday. Springer, 2019, pp. 174–186. https://doi.org/10.1007/978-3-030-22102-7_8.
2. Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; Yu, D. MM-LLMs: Recent Advances in MultiModal Large 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. 12401–12430. https://doi.org/10.18653/V1/2024.FINDINGS-ACL.738.
3. 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.
4. Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.* **2024**. https://doi.org/10.1007/S11432-024-4231-5.
5. Albuquerque, I.; Schrouff, J.; Warde-Farley, D.; Cemgil, A.T.; Gowal, S.; Wiles, O. Evaluating Model Bias Requires Characterizing its Mistakes. In Proceedings of the Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024.
6. Lei, J.; Yu, L.; Bansal, M.; Berg, T.L. TVQA: Localized, Compositional Video Question Answering. In Proceedings of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics, 2018, pp. 1369–1379. https://doi.org/10.18653/V1/D18-1167.
7. Jang, Y.; Song, Y.; Kim, C.D.; Yu, Y.; Kim, Y.; Kim, G. Video Question Answering with Spatio-Temporal Reasoning. *Int. J. Comput. Vis.* **2019**, pp. 1385–1412. https://doi.org/10.1007/S11263-019-01189-X.
8. Yang, Z.; Garcia, N.; Chu, C.; Otani, M.; Nakashima, Y.; Takemura, H. BERT Representations for Video Question Answering. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020. IEEE, 2020, pp. 1545–1554. https://doi.org/10.1109/WACV45572.2020.9093596.
9. Han, J.; Kim, S.; Park, H. MELA: Multi-Event Localization Answering Framework for Video Question Answering. In Proceedings of the Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC 2025, Catania International Airport, Catania, Italy, 31 March 2025 - 4 April 2025. ACM, 2025, pp. 1282–1289. https://doi.org/10.1145/3672608.3707973.
10. 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, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. IEEE, 2023, pp. 19027–19036. https://doi.org/10.1109/CVPR52729.2023.01824.
11. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2559–2575.
12. Zhou, Y.; Long, G. Multimodal Event Transformer for Image-guided Story Ending Generation. In Proceedings of the Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3434–3444.
13. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Shen, J.; Long, G.; Xu, C.; Jiang, D. Fine-grained distillation for long document retrieval. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 19732–19740.
14. Wang, Q.; Hu, H.; Zhou, Y. Memorymamba: Memory-augmented state space model for defect recognition. *arXiv preprint arXiv:2405.03673* **2024**.
15. Li, J.; Wei, P.; Han, W.; Fan, L. IntentQA: Context-aware Video Intent Reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, 2023, pp. 11929–11940. https://doi.org/10.1109/ICCV51070.2023.01099.

16. Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; Tao, D. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In Proceedings of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 2019, pp. 9127–9134. https://doi.org/10.1609/AAAI.V33I01.33019127.
17. 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.* **2022**, pp. 63–74. https://doi.org/10.1109/TCSVT.2020.3048440.
18. Amini, M.H.; Mia, M.J.; Saadati, Y.; Imteaj, A.; Nabavirazavi, S.; Thakker, U.; Hossain, M.Z.; Fime, A.A.; Iyengar, S.S. Distributed LLMs and Multimodal Large Language Models: A Survey on Advances, Challenges, and Future Directions. *CoRR* **2025**. https://doi.org/10.48550/ARXIV.2503.16585.
19. Wang, Y.; Chen, W.; Han, X.; Lin, X.; Zhao, H.; Liu, Y.; Zhai, B.; Yuan, J.; You, Q.; Yang, H. Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning. *CoRR* **2024**. https://doi.org/10.48550/ARXIV.2401.06805.
20. Son, M.; Lee, S. Advancing Multimodal Large Language Models: Optimizing Prompt Engineering Strategies for Enhanced Performance. *Applied Sciences* **2025**.
21. Dong, Y.; Liu, Z.; Sun, H.; Yang, J.; Hu, W.; Rao, Y.; Liu, Z. Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE, 2025, pp. 9062–9072.
22. Zhou, Y.; Song, L.; Shen, J. Improving Medical Large Vision-Language Models with Abnormal-Aware Feedback. *arXiv preprint arXiv:2501.01377* **2025**.
23. Le, C.C.; Vinh, H.C.T.; Phan, H.N.; Le, D.D.; Nguyen, T.N.; Bui, N.D.Q. VisualCoder: Guiding Large Language Models in Code Execution with Fine-grained Multimodal Chain-of-Thought Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025. Association for Computational Linguistics, 2025, pp. 6628–6645. https://doi.org/10.18653/V1/2025.FINDINGS-NAACL.370.
24. Zhou, Y.; Song, L.; Shen, J. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. *arXiv preprint arXiv:2506.19835* **2025**.
25. Yan, Q.; Fan, Y.; Li, H.; Jiang, S.; Zhao, Y.; Guan, X.; Kuo, C.; Wang, X.E. Multimodal Inconsistency Reasoning (MMIR): A New Benchmark for Multimodal Reasoning Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025. Association for Computational Linguistics, 2025, pp. 18829–18845.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.