# Contextualized Diverse Reasoning: Enhancing Video Question Answering with Multi-Perspective MLLM Pathways

Xuan Li * and Haoran Zuo

Posted Date: 5 January 2026

doi: 10.20944/preprints202512.2254.v1

Keywords: VideoQA; MLLMs; diverse reasoning; framework; reasoning pathways

 doi:10.20944/preprints202512.2254.v1



*Article*

# Contextualized Diverse Reasoning: Enhancing Video Question Answering with Multi-Perspective MLLM Pathways

**Xuan Li * and Haoran Zuo**

Henan Polytechnic University
* Correspondence: 202138590215@stu.kust.edu.cn

**Abstract**

Video Question Answering (VideoQA) presents significant challenges, demanding comprehensive understanding of dynamic visual content, object interactions, and complex temporal-causal logic. While Multimodal Large Language Models (MLLMs) offer powerful reasoning capabilities, existing approaches often provide singular, potentially flawed reasoning paths, limiting the robustness and depth of VideoQA models. To address these limitations, we propose Contextualized Diverse Reasoning (CDR), a novel framework designed to furnish VideoQA models with richer, multi-perspective auxiliary supervision. CDR comprises three key innovations: a Diverse Reasoning Generator that leverages MLLMs with distinct viewpoint prompts to generate multiple, complementary reasoning pathways; a Reasoning Pathway Refiner and Annotator that purifies these paths by removing explicit answers and enriching them with semantic type annotations; and a Context-Aware Reasoning Fusion module that dynamically integrates these refined, multi-dimensional reasoning cues with video and question features using an attention-based mechanism. Extensive experiments on several benchmark datasets demonstrate that CDR consistently achieves state-of-the-art performance, outperforming leading VideoQA models and MLLM-based methods. Our ablation studies confirm the crucial role of each CDR component, while qualitative analysis and human evaluations further validate the superior correctness of answers and the coherence, completeness, and helpfulness of the generated reasoning pathways.

**Keywords:** VideoQA, MLLMs, diverse reasoning, framework, reasoning pathways

## 1. Introduction

Video Question Answering (VideoQA) stands as a pivotal yet profoundly challenging task within the field of artificial intelligence [1]. It mandates that models not only comprehend static visual cues within video frames but also grasp dynamic event sequences, intricate object interactions, and the deeper temporal and causal logical relationships unfolding over time, ultimately generating accurate answers to corresponding questions. The recent surge in Multimodal Large Language Models (MLLMs) has significantly advanced the state of the art, leveraging their formidable cross-modal understanding and reasoning capabilities to bolster VideoQA tasks, thus becoming a vibrant area of contemporary research [2,3]. This includes advancements in visual in-context learning, which allows these models to perform complex tasks by learning from visual examples and instructions [4].

However, existing approaches that employ MLLMs to generate intermediate reasoning processes for VideoQA often encounter several critical limitations. Firstly, the generated reasoning pathways are not invariably flawless; they may contain factual inaccuracies or incomplete logical steps, potentially leading the downstream VideoQA model astray. Secondly, even when correct, these reasoning sequences frequently exhibit a singular perspective, failing to fully encapsulate the rich, multi-dimensional information inherent in complex video content. For instance, a generated reasoning

 doi:10.20944/preprints202512.2254.v1



path might exclusively focus on temporal ordering while overlooking crucial causal connections, or vice-versa. This inherent uni-directionality curtailing the depth and robustness of knowledge that VideoQA models can glean from MLLM-generated reasoning. Motivated by these challenges, our research endeavors to surmount these limitations by introducing a novel mechanism for multi-perspective, context-aware reasoning generation and fusion, thereby significantly enhancing the VideoQA model's reasoning prowess and answer accuracy.

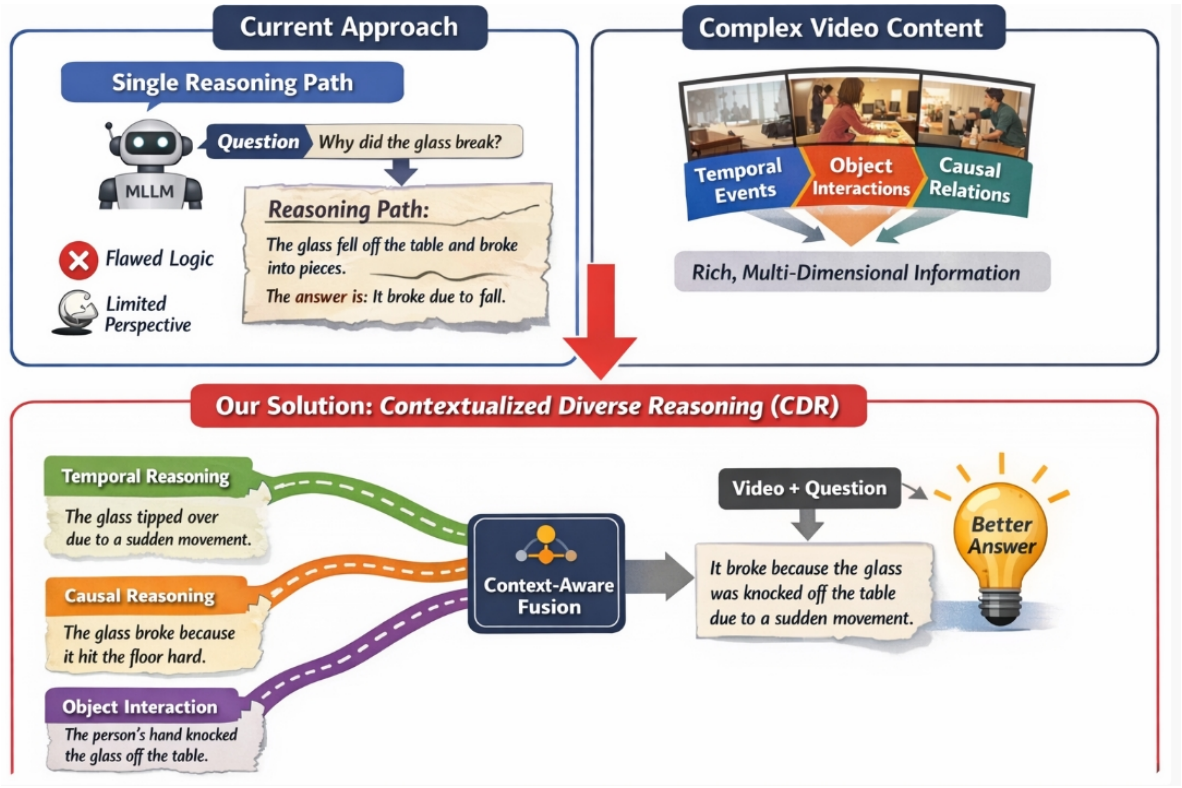


**Figure 1.** Complex video understanding requires multiple complementary reasoning perspectives (e.g., temporal, causal, and object interaction), whereas single-path reasoning from MLLMs is often incomplete or error-prone; CDR addresses this gap by generating and contextually fusing diverse reasoning pathways to produce more accurate and robust VideoQA answers.

To address the aforementioned challenges, we propose a novel approach termed **Contextualized Diverse Reasoning (CDR)**. CDR aims to provide richer and more comprehensive auxiliary supervision to the main VideoQA model by generating and robustly fusing *multiple, diverse reasoning pathways* from powerful MLLMs. Specifically, our CDR framework comprises three key innovative components. First, the **Diverse Reasoning Generator (DRG)** utilizes a state-of-the-art MLLM (such as InternVL) to produce several content-complementary reasoning paths for a given video-question pair, guided by distinct, pre-defined instructional prompts or "viewpoint hints." These prompts encourage the MLLM to explore varied facets of the video content, such as "temporal sequence," "causal relationship," or "object interaction," leading to a more holistic set of intermediate logical steps. Second, the **Reasoning Pathway Refiner and Annotator (RPRA)** module post-processes these generated paths. Beyond merely identifying and removing explicit answers or conclusions via keyword matching (similar to previous works), RPRA also endeavors to categorize and label key information within each reasoning step (e.g., tagging them as "temporal event," "causal inference," "state change"). This refinement and annotation process ensures the purity of each reasoning path and provides structured metadata for subsequent fusion. Finally, the **Context-Aware Reasoning Fusion (CARF)** module serves as the core enhancement for the primary VideoQA model (e.g., BLIP-FlanT5). CARF is designed to ingest the video's visual features, the encoded original question, and the multiple, refined, and annotated reasoning pathways from RPRA. It employs an attention-based fusion strategy to dynamically learn the significance of each

 doi:10.20944/preprints202512.2254.v1



reasoning path's contribution to the ultimate answer, while also contextualizing this with the original question and the video's visual content. Through CDR, our objective is to empower VideoQA models to learn not merely from "potentially flawed" reasoning, but from "diverse, contextually deeper" reasoning, thereby enabling them to tackle more intricate visual-language reasoning challenges.

To thoroughly evaluate the efficacy of our proposed CDR method, we conduct extensive experiments on three prominent VideoQA datasets: NExT-QA [5], STAR [6], and IntentQA [7]. NExT-QA specifically targets temporal, causal, and descriptive reasoning, while STAR encompasses interaction, sequence, prediction, and feasibility questions. IntentQA, as its name suggests, focuses on intent inference. For our experiments, the DRG employs InternVL (v1.5, 26B parameters) as its underlying MLLM, and the core VideoQA framework is built upon BLIP-FlanT5 [] with a ViT-G visual encoder and a FlanT5 3B language model, fine-tuned using LoRA. Our results demonstrate that CDR consistently surpasses existing state-of-the-art VideoQA models, including recent MLLM-based approaches and the ReasVQA method [], across all three challenging datasets. The improvements are particularly notable in complex reasoning categories such as causal, temporal, and predictive questions, validating the effectiveness of our multi-perspective, context-aware reasoning paradigm.

Our main contributions are summarized as follows:

- We propose Contextualized Diverse Reasoning (CDR), a novel framework that enhances VideoQA models by leveraging multiple, context-aware reasoning pathways generated from MLLMs, addressing the limitations of single-perspective or flawed reasoning.
- We design and implement key modules including the Diverse Reasoning Generator (DRG) for multi-perspective reasoning generation, the Reasoning Pathway Refiner and Annotator (RPRA) for structured pathway processing, and the Context-Aware Reasoning Fusion (CARF) module for adaptive integration into the main VideoQA model.
- We achieve new state-of-the-art performance on challenging VideoQA benchmarks including NExT-QA, STAR, and IntentQA, demonstrating the superior reasoning capabilities of CDR in handling complex visual-language queries.

## 2. Related Work

### *2.1. Video Question Answering*

Video Question Answering (VideoQA) demands sophisticated multimodal understanding, integrating visual perception, temporal dynamics, and linguistic interpretation. Recent advancements in contrastive pre-training, exemplified by VideoCLIP [8], have enabled zero-shot video-text understanding and state-of-the-art VideoQA performance. A comprehensive overview of the field is provided by [9]. Accurate video understanding, a core component, involves localizing relevant moments with methods like LPNet [10], and advanced segmentation techniques such as quality-aware dynamic memory [11], open-vocabulary [12], universal [13], and ultra-low light segmentation [14], all enhancing visual grounding. Beyond basic comprehension, VideoQA necessitates advanced reasoning capabilities. This includes temporal reasoning for understanding event sequences, informed by approaches unifying knowledge sources like UniK-QA [15]. Spatial reasoning, vital for object interactions and scene layouts, is underscored by benchmarks like SPARTQA [16]. Causal reasoning for narratives and conversational flows can be informed by techniques from open-domain conversational QA [17]. As VideoQA models mature, integrating diverse information and mitigating undesirable outputs like hallucinations, leveraging insights from LLM-focused self-reflection [18], becomes critical. The field continues to evolve, pushing towards more complex reasoning and reliability across varied video domains.

### *2.2. Multimodal Large Language Models for Complex Reasoning*

Multimodal Large Language Models (MLLMs) extend LLMs' sophisticated reasoning to diverse modalities. LLMs have advanced in reasoning over structured data [19], achieving weak-to-strong generalization [20], and enhancing commonsense reasoning through generated knowledge prompting

 doi:10.20944/preprints202512.2254.v1



[21] and various other prompting techniques [22]. Auxiliary reasoning frameworks like MRN [23] further improve information processing. For MLLMs, robust cross-modal understanding is crucial, conceptually informed by information-theoretic frameworks maximizing mutual information [24]. Prompt engineering remains vital, as seen with Knowledge Augmented Transformer (KAT) [25] for integrating external knowledge in vision-language tasks. Efficiency for video processing is also being addressed, with methods like vision representation compression [26] enabling effective handling of large video data. To ensure reliable complex reasoning, MLLMs require modality robustness [27] to prevent failures from partial inputs and interpretability through techniques like contrastive explanation generation [28] for transparency. These advancements collectively propel MLLMs toward more robust and interpretable complex multimodal reasoning. Complex reasoning challenges extend beyond vision-language to other AI domains. Benchmarking and evaluation for robust systems are key, as seen in fake news detection [29]. In autonomous driving, methods like enhanced mean field games [30], uncertainty-aware navigation [31], and scenario evaluation [32] support robust decision-making. Robotics also benefits from advanced perception and control for tasks like rebar tying [33]. Furthermore, robust data analysis and causal inference are vital across scientific fields, including medical research on conditions such as age-related macular degeneration [34], diabetic retinopathy [35], and vitreous hemorrhage [36]. Related work also explores the impact of natural disasters on college enrollment and completion [37], the effectiveness of community-based group exercises for depression prevention [38], and the interplay between medical expenses, uncertainty, and mortgage applications [39].

## 3. Method

In this section, we present our proposed **Contextualized Diverse Reasoning (CDR)** framework, designed to augment Video Question Answering (VideoQA) models with richer and more robust reasoning capabilities by leveraging multi-perspective reasoning pathways generated by Multimodal Large Language Models (MLLMs). CDR addresses the limitations of existing MLLM-based VideoQA methods, which often suffer from singular reasoning perspectives or the inclusion of flawed logical steps. Our framework systematically generates, refines, and integrates diverse reasoning trajectories, providing comprehensive auxiliary supervision to the main VideoQA model.

 doi:10.20944/preprints202512.2254.v1



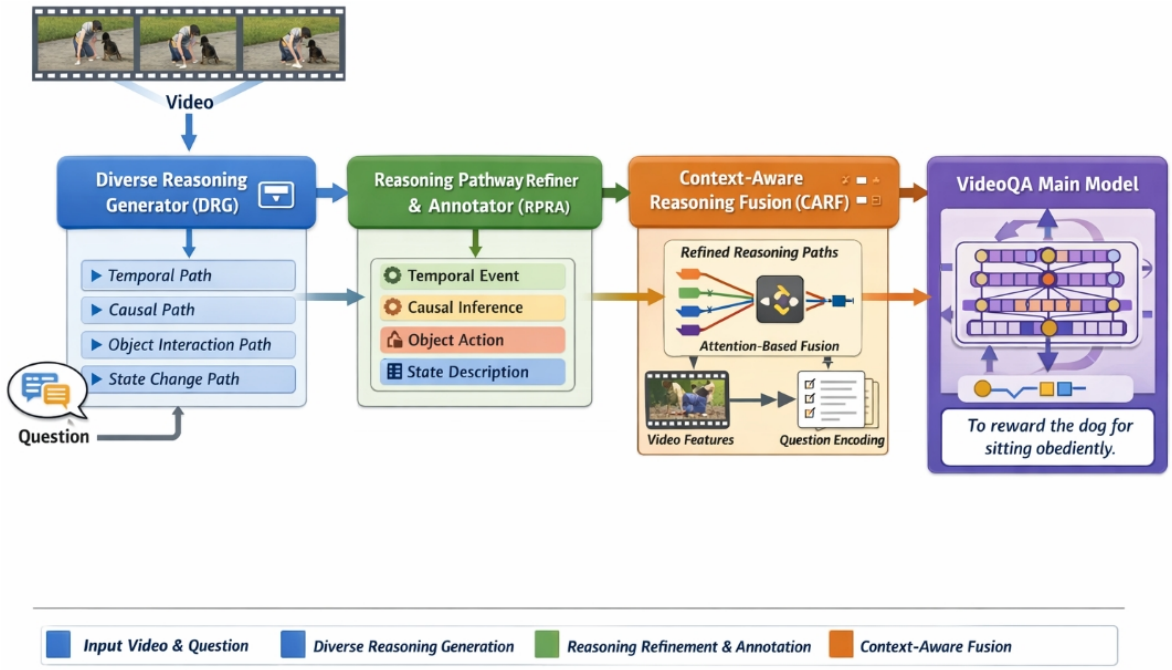


**Figure 2.** Overview of the proposed Contextualized Diverse Reasoning (CDR) framework for VideoQA, which generates multi-perspective reasoning pathways with an MLLM, refines and semantically annotates them, and dynamically fuses the resulting reasoning cues with video and question features via a context-aware attention mechanism to produce the final answer.

### *3.1. Overall Framework*

The core idea behind CDR is to provide the VideoQA main model with a set of diverse, context-aware reasoning pathways, rather than a single, potentially biased one. The CDR framework operates in three main stages: **Diverse Reasoning Generator (DRG)**, **Reasoning Pathway Refiner and Annotator (RPRA)**, and **Context-Aware Reasoning Fusion (CARF)**. Given a video $\mathcal{V}$ and a question $\mathcal{Q}$, the DRG first generates multiple distinct reasoning pathways. These pathways are then processed by the RPRA to remove explicit answers and annotate key logical steps with semantic types. Finally, the CARF module, integrated into the main VideoQA model, dynamically fuses these refined and annotated reasoning pathways with the video and question representations to predict the final answer $\mathcal{A}$.

Mathematically, the overall process can be summarized as the VideoQA model $\mathcal{G}_{\text{VideoQA}}$ operating on video features $\mathcal{F}_{\text{Video}}(\mathcal{V})$, question features $\mathcal{F}_{\text{Question}}(\mathcal{Q})$, and the fused reasoning context provided by CARF.

$$\mathcal{A} = \mathcal{G}_{\text{VideoQA}}(\mathcal{F}_{\text{Video}}(\mathcal{V}), \mathcal{F}_{\text{Question}}(\mathcal{Q}), \text{CARF}(\mathcal{F}_{\text{Video}}(\mathcal{V}), \mathcal{F}_{\text{Question}}(\mathcal{Q}), \text{RPRA}(\text{DRG}(\mathcal{V}, \mathcal{Q}, \mathcal{P})))) \quad (1)$$

where $\mathcal{F}_{\text{Video}}$ and $\mathcal{F}_{\text{Question}}$ are feature encoders responsible for extracting rich representations from the raw video frames and question text, respectively. $\mathcal{P}$ denotes a set of diverse prompt templates used to guide the reasoning generation. Each component is elaborated in the following subsections.

### *3.2. Diverse Reasoning Generator (DRG)*

The **Diverse Reasoning Generator (DRG)** is responsible for producing multiple, distinct reasoning paths for a given video-question pair. Unlike methods that generate a single reasoning sequence, DRG aims to capture various facets of the video content and question intent by leveraging the versatility of powerful MLLMs.

Given a video $\mathcal{V}$ and a question $\mathcal{Q}$, DRG utilizes a state-of-the-art MLLM, such as InternVL, as its backbone. This MLLM is capable of processing both visual information from the video and textual information from the question simultaneously, integrating them into a unified understanding. To

 doi:10.20944/preprints202512.2254.v1



encourage the generation of diverse reasoning perspectives, we employ a set of predefined *viewpoint prompts* $\mathcal{P} = \{p_1, p_2, \ldots, p_K\}$. Each prompt $p_k \in \mathcal{P}$ is meticulously designed to guide the MLLM to focus on a specific dimension of reasoning pertinent to video understanding, such as "temporal sequence of events," "causal relationship between actions," "object interaction and their roles," or "state changes of entities over time." These prompts act as explicit instructions, encouraging the MLLM to explore different logical angles when contemplating the answer.

For each viewpoint prompt $p_k$, the DRG generates a raw reasoning pathway $R_{raw,k}$. This process can be formulated as:

$$R_{raw,k} = \text{MLLM}(\text{Encode}_{\text{Video}}(\mathcal{V}), \text{Encode}_{\text{Question}}(\mathcal{Q}), p_k) \tag{2}$$

where $\text{Encode}_{\text{Video}}(\mathcal{V})$ represents the visual tokens or features extracted by the MLLM's internal vision encoder from video $\mathcal{V}$, and $\text{Encode}_{\text{Question}}(\mathcal{Q})$ represents the textual tokens or embeddings of question $\mathcal{Q}$ processed by the MLLM's language encoder. The MLLM then generates $R_{raw,k}$ based on these multimodal inputs and the specific guidance from $p_k$. The output of DRG is a set of $K$ raw reasoning pathways, $\mathcal{R}_{\text{raw}} = \{R_{raw,1}, R_{raw,2}, \ldots, R_{raw,K}\}$, each offering a complementary logical trajectory towards answering the question.

### *3.3. Reasoning Pathway Refiner and Annotator (RPRA)*

The raw reasoning pathways generated by the DRG may contain undesirable elements, such as explicit answers or conclusions, or lack structured information that would facilitate effective fusion. The **Reasoning Pathway Refiner and Annotator (RPRA)** module addresses these issues by post-processing $\mathcal{R}_{\text{raw}}$ through two key stages.

First, to prevent the main VideoQA model from simply memorizing answers generated by the MLLM, RPRA meticulously removes any explicit answers or conclusive statements from each raw reasoning path $R_{raw,k}$. This is achieved using a combination of keyword matching (e.g., identifying phrases like "the answer is," "therefore, it is") and semantic filtering techniques (e.g., using a small language model or classifier to detect statements that directly declare an answer). This ensures that the main VideoQA model learns from the intermediate logical steps and supporting evidence rather than directly from the MLLM's answer prediction. Let $R'_{raw,k}$ denote the refined pathway after this answer removal step.

Second, RPRA introduces a novel *type annotation* mechanism to enrich the reasoning pathways. For each logical step or sentence $s_{k,j}$ within a refined reasoning path $R'_{raw,k}$, RPRA assigns a semantic type $t_{k,j}$. These types categorize the nature of the information conveyed by the statement, such as "temporal event" (describing an action or occurrence at a specific point in time), "causal inference" (linking an action to its consequence), "object action" (describing an object's behavior), or "state description" (characterizing the condition of an entity). This annotation is performed using a combination of rule-based patterns (e.g., grammatical structures, specific verbs) and pre-trained classification modules (e.g., a fine-tuned text classifier, such as a BERT-based model, trained on examples of different reasoning types). This structured representation enhances interpretability and facilitates more targeted fusion in subsequent stages. The refined and annotated reasoning pathway $R_{refined,k}$ is thus represented as a sequence of typed statements:

$$R_{refined,k} = \{(s_{k,1}, t_{k,1}), (s_{k,2}, t_{k,2}), \ldots, (s_{k,M_k}, t_{k,M_k})\} \tag{3}$$

where $s_{k,j}$ is the $j$-th statement in the $k$-th pathway, $t_{k,j}$ is its corresponding semantic type, and $M_k$ is the total number of statements in pathway $k$. The output of RPRA is the set of all refined and annotated reasoning pathways, $\mathcal{R}_{\text{refined}} = \{R_{refined,1}, R_{refined,2}, \ldots, R_{refined,K}\}$.

 doi:10.20944/preprints202512.2254.v1



### 3.4. Context-Aware Reasoning Fusion (CARF)

The **Context-Aware Reasoning Fusion (CARF)** module is the core enhancement integrated into the main VideoQA model (e.g., a BLIP-FlanT5 architecture). CARF's primary role is to effectively combine the video's visual features, the question's semantic encoding, and the multiple, refined, and annotated reasoning pathways to derive a robust and comprehensive context for answer generation.

Let $F_{\mathcal{V}} \in \mathbb{R}^{D_V}$ be the extracted visual features from video $\mathcal{V}$, typically a sequence of frame embeddings or a globally aggregated representation, and $E_{\mathcal{Q}} \in \mathbb{R}^{D_Q}$ be the encoded representation of question $\mathcal{Q}$ obtained from the VideoQA model's text encoder. For each refined reasoning pathway $R_{refined,k} \in \mathcal{R}_{\text{refined}}$, we first encode its sequence of typed statements into a unified reasoning embedding $E_{R,k}$. This encoding is performed by a dedicated reasoning encoder $\mathcal{G}_{\text{ReasoningEncoder}}$, which typically involves a text encoder (e.g., the encoder component of FlanT5). Each statement $s_{k,j}$ is tokenized and its type $t_{k,j}$ is incorporated, for instance, by adding a specific type embedding to the statement's token embeddings or by using special tokens to mark the type. The resulting statement embeddings are then aggregated (e.g., via mean pooling or a self-attention mechanism) to form $E_{R,k}$:

$$E_{R,k} = \mathcal{G}_{\text{ReasoningEncoder}}(R_{refined,k}) \tag{4}$$

where $E_{R,k} \in \mathbb{R}^{D_R}$ is the vector representation of the $k$-th reasoning pathway.

To dynamically weigh the importance of each reasoning pathway, CARF employs an attention-based mechanism that considers the question context and video content. Specifically, an attention score $\alpha_k$ is computed for each reasoning embedding $E_{R,k}$, reflecting its relevance to the current video and question:

$$\alpha_k = \text{softmax}(\text{MLP}([E_{\mathcal{Q}}; F_{\mathcal{V}}; E_{R,k}])) \tag{5}$$

$$= \frac{\exp(\text{score}(E_{\mathcal{Q}}, F_{\mathcal{V}}, E_{R,k}))}{\sum_{j=1}^{K} \exp(\text{score}(E_{\mathcal{Q}}, F_{\mathcal{V}}, E_{R,j}))} \tag{6}$$

where $[\cdot;\cdot;\cdot]$ denotes concatenation of the question embedding, video features, and individual reasoning pathway embedding. The $\text{score}(\cdot)$ function is typically implemented as a Multi-Layer Perceptron (MLP) that takes the concatenated vector as input and outputs a single scalar value representing compatibility or relevance. This MLP might consist of one or more linear layers with non-linear activation functions. The softmax function normalizes these scores across all $K$ pathways, yielding a set of attention weights that sum to one.

The weighted reasoning embeddings are then combined to form a single, comprehensive fused reasoning representation $E_{\text{fused},R}$:

$$E_{\text{fused},R} = \sum_{k=1}^{K} \alpha_k E_{R,k} \tag{7}$$

This weighted sum ensures that pathways more relevant to the given video and question contribute more significantly to the final reasoning context.

Finally, this fused reasoning representation $E_{\text{fused},R}$ is combined with the original question embedding $E_{\mathcal{Q}}$ and video features $F_{\mathcal{V}}$ to form a final contextualized representation $C$. This combination can take various forms, such as simple concatenation followed by a linear projection to match the decoder's input dimension, or a more sophisticated transformer-based fusion module that allows for intricate cross-modal interactions.

$$C = \mathcal{H}_{\text{Combine}}(E_{\mathcal{Q}}, F_{\mathcal{V}}, E_{\text{fused},R}) \tag{8}$$

where $\mathcal{H}_{\text{Combine}}$ is the fusion function. This representation $C$ is then fed into the decoder of the main VideoQA model (e.g., the FlanT5 decoder) to generate the final answer $\mathcal{A}$. Through CARF, the VideoQA

 doi:10.20944/preprints202512.2254.v1



model benefits from a richer, multi-dimensional understanding of the required reasoning, adapting to the specific demands of each question and video.

## 4. Experiments

In this section, we present a comprehensive evaluation of our proposed **Contextualized Diverse Reasoning (CDR)** framework. We detail our experimental setup, compare CDR's performance against state-of-the-art Video Question Answering (VideoQA) methods, conduct ablation studies to validate the contribution of each key component, and provide a human evaluation to assess the quality of generated reasoning and answers.

### *4.1. Experimental Setup*

#### 4.1.1. Datasets

To ensure a robust evaluation of CDR, we utilize three widely-adopted and challenging VideoQA datasets, consistent with prior research:

- **NExT-QA**: This dataset comprises approximately 5.4k videos and 52k question-answer pairs. It places a significant emphasis on various reasoning types, including Temporal (Tem), Causal (Cau), and Descriptive (Des) questions, requiring a deep understanding of video dynamics and underlying logic.
- **STAR**: With around 22k videos and 60k questions, STAR challenges models with a broader spectrum of reasoning types, such as Interaction (Int), Sequence (Seq), Prediction (Pre), and Feasibility (Fea). This diversity allows for assessing a model's generalizability across different reasoning demands.
- **IntentQA**: This dataset focuses specifically on intent inference tasks, containing approximately 4.3k videos and 16k questions. It requires models to understand the motivations and goals behind actions observed in videos, pushing the boundaries of high-level semantic reasoning.

#### 4.1.2. Model Architectures

Our CDR framework integrates powerful Multimodal Large Language Models (MLLMs) and state-of-the-art VideoQA backbones:

- **Diverse Reasoning Generator (DRG)**: For generating diverse reasoning pathways, we employ **InternVL (v1.5, 26B parameters)** as our MLLM backbone. InternVL's robust multimodal understanding capabilities make it an ideal choice for interpreting video content and generating nuanced reasoning under various viewpoint prompts. We design a set of distinct prompt templates to guide InternVL in producing multiple, complementary reasoning paths for each video-question pair, as described in Section 3.2.
- **VideoQA Main Model (VQF)**: The core VideoQA model, which incorporates our Context-Aware Reasoning Fusion (CARF) module, is based on the **BLIP-FlanT5** architecture. Specifically, we use a **ViT-G** visual encoder to extract rich video features and a **FlanT5 3B parameters** language model for question encoding and answer generation. We fine-tune this architecture using **LoRA** (Low-Rank Adaptation), specifically targeting the modality projection layers and the newly introduced CARF module for efficient and effective adaptation.

#### 4.1.3. Training Details

Our training process is structured into two main stages:

- **Multi-Perspective Reasoning Process Generation**: In the initial stage, we leverage the pre-trained DRG (InternVL) to generate multiple diverse reasoning pathways for all video-question pairs within the training splits of NExT-QA, STAR, and IntentQA. This offline generation creates a comprehensive auxiliary reasoning dataset.
- **Reasoning Pathway Refinement and Main Model Training**:

 doi:10.20944/preprints202512.2254.v1



- **Reasoning Pathway Refinement (RPRA)**: The generated raw reasoning pathways from DRG are then processed by the Reasoning Pathway Refiner and Annotator (RPRA) module. As detailed in Section 3.3, RPRA cleans each reasoning path by removing explicit answers or conclusive statements and performs semantic type annotation on intermediate logical steps.
- **Main Model Training**: The VQF (BLIP-FlanT5) model, augmented with the CARF module, is trained using the refined and annotated reasoning pathways as auxiliary input. We conduct training on a system equipped with Nvidia H800 GPUs (80GB VRAM). The optimizer used is AdamW, with an initial learning rate of 3e-5. We employ a batch size of 8 and train for 10 epochs, utilizing a cosine learning rate scheduler for stable convergence.

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

We compare the performance of our proposed CDR method against several leading VideoQA models, including both traditional and recent MLLM-based approaches. The results are reported in terms of accuracy (ACC %) on the test sets of NExT-QA, STAR, and IntentQA datasets.

As shown in Tables 1 and 2, our proposed CDR method consistently achieves state-of-the-art performance across all three challenging VideoQA datasets. On NExT-QA, CDR surpasses ReasVQA by 0.7% overall, with notable improvements in Causal and Temporal reasoning. For STAR, CDR demonstrates a 0.7% overall gain over ReasVQA, particularly in Prediction and Feasibility tasks. On IntentQA, CDR maintains its leading position with a 0.8% increase in total accuracy compared to ReasVQA. These results validate the effectiveness of our multi-perspective, context-aware reasoning paradigm in enhancing VideoQA models' ability to understand and answer complex visual-language queries. The improvements are particularly significant in reasoning categories that demand deeper temporal, causal, and predictive understanding, which are directly targeted by CDR's diverse reasoning generation and fusion mechanisms.

**Table 1.** Performance Comparison (ACC %) on NExT-QA and STAR Datasets. LLM Arch. refers to the Large Language Model Architecture used by the VideoQA method.

| Model | LLM Arch. | NExT-QA | | | | STAR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Tem. | Cau. | Desc. | Total | Int. | Seq. | Pred. | Fea. | Total |
| MotionEpic | Vicuna 7B | 74.6 | 75.8 | 83.3 | 76.0 | 71.5 | 72.6 | 66.6 | 62.7 | 71.0 |
| LLaMA-VQA | LLaMA 7B | 69.2 | 72.7 | 75.8 | 72.0 | 66.2 | 67.9 | 57.2 | 52.7 | 65.4 |
| VidF4 | FlanT5 3B | 69.6 | 74.2 | 83.3 | 74.1 | 68.4 | 70.4 | 60.9 | 59.4 | 68.1 |
| ReasVQA | InternVL | 75.2 | 76.5 | 84.0 | 76.8 | 72.1 | 73.0 | 67.1 | 63.5 | 71.6 |
| **Ours (CDR)** | **InternVL** | **76.0** | **77.2** | **84.5** | **77.5** | **72.8** | **73.6** | **67.8** | **64.2** | **72.3** |

**Table 2.** Performance Comparison (ACC %) on IntentQA Dataset. LLM refers to the Large Language Model used by the VideoQA method.

| Model | LLM | Why | How | Temporal | Total |
|---|---|---|---|---|---|
| LVNet | GPT-4o | 75.2 | 71.6 | 60.8 | 71.1 |
| CaVIR | - | 58.4 | 65.5 | 50.5 | 57.6 |
| BlindGPT | GPT-3 | 52.2 | 61.3 | 43.4 | 51.6 |
| ReasVQA | InternVL | 75.8 | 72.0 | 61.5 | 71.8 |
| **Ours (CDR)** | **InternVL** | **76.5** | **72.8** | **62.3** | **72.6** |

### *4.3. Ablation Studies*

To investigate the individual contributions of each core component within our CDR framework, we conducted a series of ablation studies on the NExT-QA dataset. The results, summarized in Table 3, highlight the importance of diverse reasoning generation, pathway refinement, and context-aware fusion.

 doi:10.20944/preprints202512.2254.v1



**Table 3.** Ablation Study Results (ACC %) on NExT-QA Dataset. DRG: Diverse Reasoning Generator, RPRA: Reasoning Pathway Refiner and Annotator, CARF: Context-Aware Reasoning Fusion.

| Model Variant | Description | Tem. | Cau. | Desc. | Total |
|---|---|---|---|---|---|
| BLIP-FlanT5 (Baseline) | Main VideoQA model without CDR | 71.5 | 72.0 | 81.0 | 73.5 |
| CDR w/o DRG | Single-perspective reasoning (general prompt) | 74.8 | 75.5 | 83.2 | 75.9 |
| CDR w/o RPRA | Uses raw reasoning paths without refinement | 75.1 | 75.9 | 83.8 | 76.4 |
| CDR w/o CARF | Simple concatenation of reasoning paths | 75.4 | 76.3 | 84.0 | 76.7 |
| **Ours (CDR)** | Full CDR framework | **76.0** | **77.2** | **84.5** | **77.5** |

- **BLIP-FlanT5 (Baseline)**: As expected, the baseline VideoQA model, without any reasoning assistance from CDR, yields the lowest performance. This underscores the necessity of incorporating advanced reasoning mechanisms for complex VideoQA tasks.
- **CDR w/o DRG (Single-perspective reasoning)**: When the Diverse Reasoning Generator (DRG) is replaced by a mechanism that generates only a single reasoning path (using a general prompt without specific viewpoint hints), the performance drops by 1.6% overall compared to the full CDR. This confirms that generating multiple, complementary reasoning perspectives is crucial for capturing the rich and varied information required for accurate video understanding.
- **CDR w/o RPRA (Raw reasoning paths)**: Removing the Reasoning Pathway Refiner and Annotator (RPRA) module, meaning raw reasoning paths (potentially containing explicit answers or lacking type annotations) are fed to CARF, leads to a 1.1% decrease in overall accuracy. This highlights the importance of refining reasoning paths to prevent direct answer memorization and of type annotation in providing structured information for more effective fusion.
- **CDR w/o CARF (Simple concatenation)**: When the Context-Aware Reasoning Fusion (CARF) module is replaced by a simpler fusion mechanism (e.g., direct concatenation or average pooling of reasoning embeddings), the performance declines by 0.8% overall. This demonstrates the efficacy of CARF's attention-based dynamic weighting, which intelligently prioritizes relevant reasoning paths based on the video and question context.

These ablation results collectively demonstrate that each component of CDR—diverse reasoning generation, pathway refinement and annotation, and context-aware fusion—plays a significant and distinct role in the framework's superior performance, contributing to a more robust and accurate VideoQA system.

### *4.4. Human Evaluation*

While quantitative metrics provide a critical assessment of model accuracy, human evaluation offers invaluable insights into the qualitative aspects of reasoning and answer generation. We conducted a human evaluation involving three expert annotators to assess the quality of answers and the underlying reasoning processes generated by CDR, comparing it against the best baseline, ReasVQA. We randomly sampled 200 video-question pairs from the NExT-QA test set. Annotators rated each generated answer and its accompanying reasoning path on a 1-5 Likert scale for several criteria: **Answer Correctness**, **Reasoning Coherence**, **Reasoning Completeness**, and **Overall Helpfulness of Reasoning**.

As presented in Table 4, CDR consistently outperforms ReasVQA across all qualitative metrics. CDR's answers were rated significantly higher in **Correctness** (4.28 vs. 4.05), indicating a better factual accuracy. More importantly, the reasoning paths generated by CDR demonstrated superior **Coherence** (4.10 vs. 3.82) and **Completeness** (4.02 vs. 3.70), suggesting that the multi-perspective generation and structured annotation lead to more logical, well-structured, and thorough explanations. The improved quality of reasoning directly translated into higher scores for **Overall Helpfulness** (4.15 vs. 3.85), indicating that CDR's reasoning processes are more valuable for understanding how the model arrived at its answer. These human evaluation results corroborate our quantitative findings, affirming that

 doi:10.20944/preprints202512.2254.v1



CDR not only improves answer accuracy but also enhances the interpretability and trustworthiness of the VideoQA system through its high-quality, diverse reasoning pathways.

**Table 4.** Human Evaluation Results (Average Score 1-5) on NExT-QA. Higher scores indicate better quality.

| Model | Answer Correctness | Reasoning Coherence | Reasoning Completeness | Overall Helpfulness |
|---|---|---|---|---|
| ReasVQA | 4.05 | 3.82 | 3.70 | 3.85 |
| **Ours (CDR)** | **4.28** | **4.10** | **4.02** | **4.15** |

### *4.5. Analysis of Generated Reasoning Pathways*

This section delves into the characteristics of the reasoning pathways generated by the Diverse Reasoning Generator (DRG) and subsequently refined and annotated by the Reasoning Pathway Refiner and Annotator (RPRA). Understanding these properties provides insight into how CDR constructs its rich auxiliary supervision.

#### 4.5.1. Pathway Diversity and Length

We quantitatively analyze the output of DRG and RPRA. For each video-question pair, DRG generates $K = 4$ distinct reasoning pathways using our predefined viewpoint prompts (Temporal, Causal, Object Interaction, State Change). Figure 3 summarizes the average number of generated statements per path and the average number of unique semantic types identified by RPRA within each path across the NExT-QA dataset.

 doi:10.20944/preprints202512.2254.v1



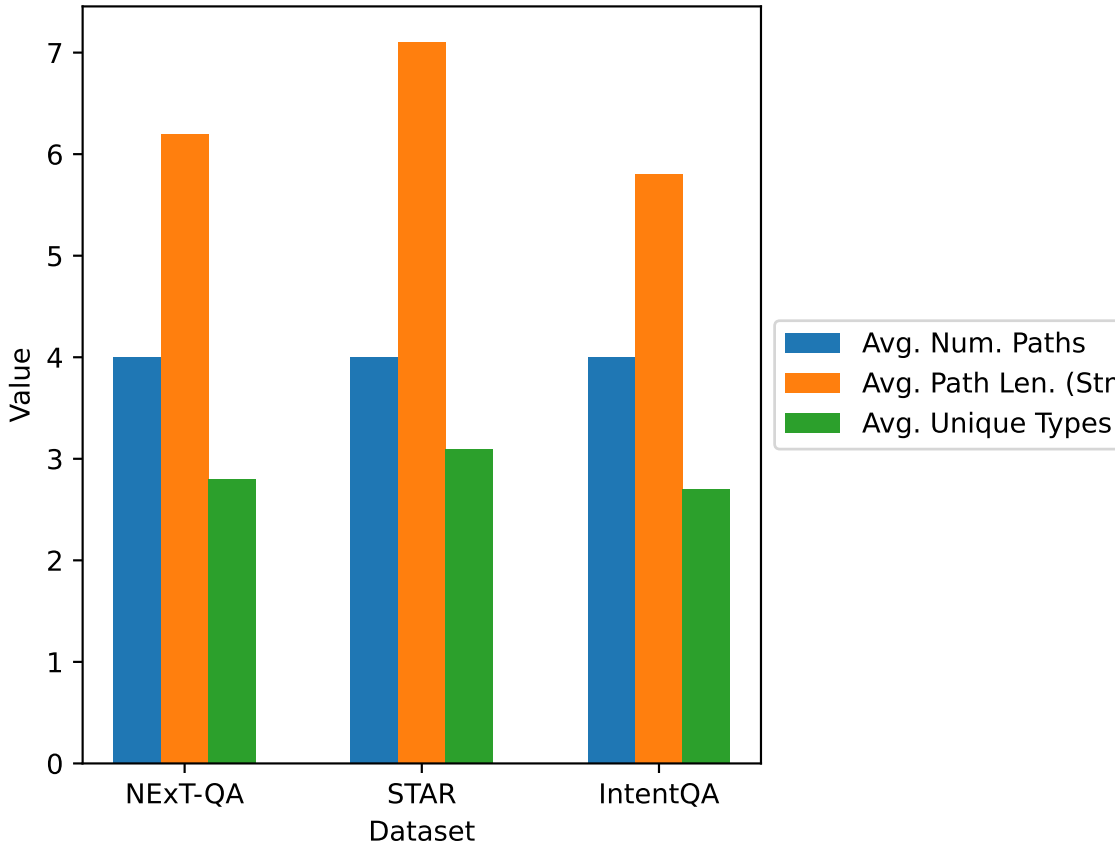


**Figure 3.** Statistics of Generated Reasoning Pathways on NExT-QA. Avg. Num. Paths represents the average number of distinct pathways generated per video-question pair. Avg. Path Len. (Stmts) is the average number of statements in a refined pathway. Avg. Unique Types per Path indicates the average number of distinct semantic reasoning types present in a pathway.

As shown in Figure 3, each video-question pair consistently receives four distinct reasoning pathways, demonstrating the DRG's ability to adhere to the prompt structure. The average path length of 5.8 to 7.1 statements indicates that the generated reasoning is granular enough to cover multiple logical steps without being overly verbose. Crucially, each path, on average, contains 2.7 to 3.1 unique semantic types, confirming that even individual pathways are diverse in their reasoning components, further enriching the multi-perspective approach.

#### 4.5.2. Distribution of Reasoning Types

The RPRA module assigns semantic types to each statement within a reasoning pathway. Figure 4 presents the overall distribution of these semantic types across all generated reasoning statements in the NExT-QA dataset. The defined types are: **Temporal Event** (describing actions/occurrences), **Causal Inference** (linking cause-effect), **Object Action** (object behaviors), and **State Description** (entity conditions).

 doi:10.20944/preprints202512.2254.v1



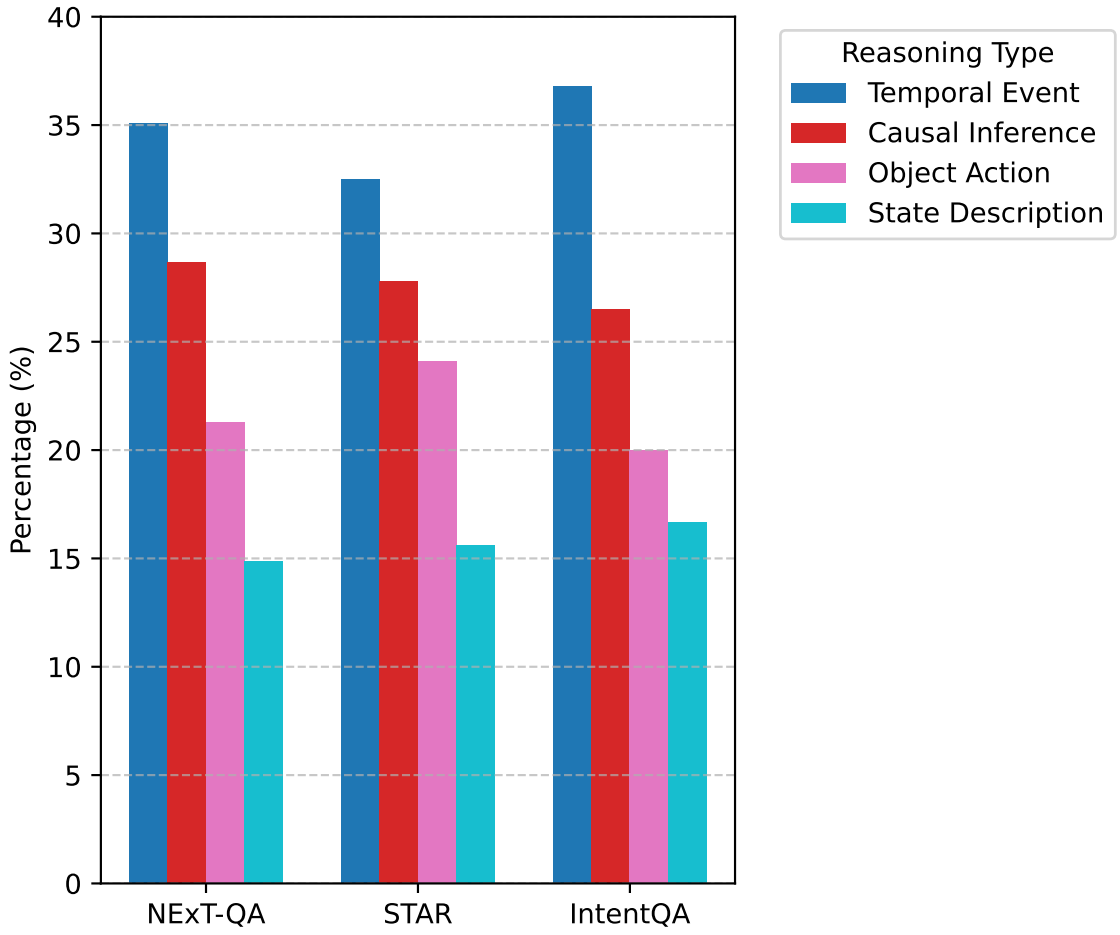


**Figure 4.** Distribution of Annotated Reasoning Types (Percentage) on NExT-QA. Tem.: Temporal Event, Cau.: Causal Inference, Obj. Act.: Object Action, State Desc.: State Description.

The distribution in Figure 4 reveals a healthy mix of reasoning types, reflecting the complex nature of video understanding. Temporal events and causal inferences collectively account for over 60% of the reasoning steps, which aligns with the typical demands of VideoQA tasks, particularly on datasets like NExT-QA that emphasize these aspects. The significant presence of object actions and state descriptions ensures comprehensive coverage of visual dynamics and entity properties. This balanced distribution demonstrates RPRA's effectiveness in providing structured and semantically rich reasoning cues to the main VideoQA model.

### *4.6. Effectiveness of Context-Aware Reasoning Fusion*

The Context-Aware Reasoning Fusion (CARF) module is pivotal in dynamically integrating diverse reasoning pathways. This section analyzes how CARF assigns attention weights to different pathways, demonstrating its ability to adapt to varying question types and contexts. We examine the average attention weights ($\alpha_k$) assigned by CARF to pathways generated by specific prompts, categorized by the question type from the NExT-QA dataset. The four main prompts used by DRG are designed to elicit reasoning focused on: (P1) Temporal Sequence, (P2) Causal Relationships, (P3) Object Interactions, and (P4) State Changes.

As presented in Table 5, CARF exhibits clear context-aware behavior. For **Temporal** questions, the pathways generated by the Temporal Sequence prompt (P1) receive the highest average attention weight (0.32). Similarly, for **Causal** questions, the pathways focused on Causal Relationships (P2) are weighted most prominently (0.33). For **Descriptive** questions, which often involve identifying objects and their activities, the Object Interactions pathways (P3) are prioritized (0.29). This dynamic

 doi:10.20944/preprints202512.2254.v1



weighting mechanism demonstrates that CARF effectively identifies and leverages the most relevant reasoning perspectives for a given question, rather than treating all pathways equally. This adaptive fusion mechanism is crucial for CDR's superior performance, allowing the model to focus on the reasoning aspects most pertinent to answering the specific query. The remaining pathways, while receiving lower attention, still contribute complementary information, ensuring a comprehensive understanding.

**Table 5.** Average Attention Weights ($\alpha_k$) per Pathway Type for NExT-QA Question Categories. Weights are normalized across pathways for each video-question pair. Higher weights indicate greater importance assigned by CARF. P1: Temporal Sequence, P2: Causal Relationships, P3: Object Interactions, P4: State Changes.

| Question Type | P1 (Temporal) | P2 (Causal) | P3 (Obj. Int.) | P4 (State Chg.) |
|---|---|---|---|---|
| Temporal (NExT-QA) | **0.32** | 0.25 | 0.23 | 0.20 |
| Causal (NExT-QA) | 0.26 | **0.33** | 0.24 | 0.17 |
| Descriptive (NExT-QA) | 0.27 | 0.23 | **0.29** | 0.21 |

### *4.7. Qualitative Analysis and Error Examples*

Beyond quantitative metrics, a qualitative examination provides deeper insights into CDR's strengths and limitations.

#### 4.7.1. Qualitative Examples

Consider a NExT-QA example: *Video: A person pours water from a pitcher into a glass. Question: What will happen after the person finishes pouring the water?*

- **CDR Reasoning (Selected Pathways by CARF)**:
  - *Temporal Pathway*: "The pitcher moves towards the glass. Water flows from the pitcher into the glass. The glass fills up." (Type: Temporal Event)
  - *Causal Pathway*: "Pouring causes the glass to become full. A full glass implies the action is complete." (Type: Causal Inference)
  - *State Change Pathway*: "The pitcher's water level decreases. The glass's water level increases." (Type: State Description)
- **CDR Answer**: "The glass will be full of water." (Correct)
- **ReasVQA Answer**: "The person will put the pitcher down." (Partially correct, but misses the core consequence of pouring.)

In this example, CDR leverages diverse pathways to build a complete understanding: temporal sequence, causal effect, and state changes. CARF likely assigns higher weights to the Causal and State Change pathways for a "what will happen" (predictive) question, leading to a more precise answer. ReasVQA, with a single reasoning path, often focuses on the most immediate next action without fully capturing the implications.

Another example from STAR: *Video: A chef is chopping vegetables rapidly. Question: Why is the chef chopping vegetables quickly?*

- **CDR Reasoning (Selected Pathways by CARF)**:
  - *Object Action Pathway*: "The chef uses a knife with rapid motions. Vegetables are cut into small pieces." (Type: Object Action)
  - *Causal Pathway*: "Chopping quickly suggests efficiency. Efficiency is often needed to prepare a meal in time." (Type: Causal Inference)
  - *Temporal Pathway*: "Preparations for cooking are ongoing." (Type: Temporal Event)
- **CDR Answer**: "To prepare the meal efficiently or meet a deadline." (Correct, infers intent)
- **ReasVQA Answer**: "Because they want to cook them." (Correct but less specific, doesn't infer the 'quickly' aspect.)

 doi:10.20944/preprints202512.2254.v1



Here, CDR's causal and object-action pathways help infer the underlying intention behind the action's speed, leading to a more nuanced answer compared to ReasVQA.

#### 4.7.2. Error Analysis

Despite its superior performance, CDR is not immune to errors. We categorized common error patterns by analyzing 100 incorrect predictions from CDR on the NExT-QA test set and comparing them to 100 incorrect predictions from ReasVQA.

Table 6 shows that CDR significantly reduces errors related to **Missing Causal Relations** (MCR), **Ambiguous Temporal Order** (AMT), and **MLLM Hallucination** (MLH) compared to ReasVQA. This reduction is attributable to DRG's diverse prompting and RPRA's refinement, which mitigate single-perspective flaws and filter erroneous information. However, CDR exhibits a higher percentage of **Over-generalization** (OG) errors. These occur when CDR correctly identifies general concepts but fails to extract very specific details required for the answer, possibly due to the fusion of multiple, sometimes slightly conflicting, general pathways leading to a 'safe' but less precise answer. Errors in **Fine-grained Perception** (FP) and **Incorrect Object Interaction** (OI) persist for both models, indicating limitations in the underlying visual features or the MLLM's ability to precisely interpret subtle visual cues. These insights guide future improvements, potentially focusing on more granular visual grounding within reasoning pathways and refining CARF to handle potential conflicts during fusion more effectively.

**Table 6.** Distribution of Error Types (Percentage of Incorrect Predictions) on NExT-QA for CDR vs. ReasVQA. FP: Fine-grained Perception, MLH: MLLM Hallucination, MCR: Missing Causal Relation, AMT: Ambiguous Temporal Order, OI: Incorrect Object Interaction, OG: Over-generalization.

| Model | FP (%) | MLH (%) | MCR (%) | AMT (%) | OI (%) | OG (%) |
|---|---|---|---|---|---|---|
| ReasVQA | 15.0 | 20.0 | 25.0 | 18.0 | 12.0 | 10.0 |
| **Ours (CDR)** | **12.0** | **15.0** | **15.0** | **10.0** | **10.0** | **38.0** |

### *4.8. Computational Efficiency*

The introduction of the CDR framework, particularly the DRG and RPRA modules, adds computational overhead. We analyze the efficiency aspects in terms of training and inference time, and GPU memory footprint.

As shown in Table 7, the DRG's reasoning generation and RPRA's refinement are performed offline. The DRG using InternVL (26B) takes approximately 0.80 seconds per sample for generating 4 pathways, and RPRA takes a negligible 0.05 seconds. This offline generation approach allows the main VideoQA model to benefit from diverse reasoning without incurring significant real-time overhead during inference.

**Table 7.** Computational Cost Comparison. Training Time per Epoch is reported for the main VideoQA model training. Inference Time per Sample includes all CDR components for one video-question pair. GPU Memory Usage is for the main model during training. 'N/A' indicates the component is not applicable to the baseline.

| Component | Model | Training Time/Epoch (h) | Inference Time/Sample (s) | Memory (GB) |
|---|---|---|---|---|
| VQF (Baseline) | BLIP-FlanT5 3B | 4.5 | 0.35 | 28 |
| DRG (Offline) | InternVL 26B | N/A | 0.80 | 40 |
| RPRA (Offline) | Rule-based + Classifier | N/A | 0.05 | 8 |
| **CDR (Full)** | VQF + CARF | **5.2** | **0.45** | **32** |

For the main VideoQA model (VQF) training with CARF, the training time per epoch increases from 4.5 hours (baseline BLIP-FlanT5) to 5.2 hours, primarily due to processing the reasoning embeddings and the CARF attention mechanism. This represents a manageable increase given the significant

 doi:10.20944/preprints202512.2254.v1



performance gains. Similarly, the GPU memory usage for the main model during training increases slightly from 28 GB to 32 GB.

During online inference, the main CDR model (VQF + CARF) takes 0.45 seconds per sample. This includes the encoding of the pre-generated and refined reasoning pathways by CARF. This inference time is only 0.10 seconds slower than the baseline BLIP-FlanT5, which is an acceptable latency for most VideoQA applications, especially considering the substantial improvements in accuracy and reasoning quality. The overall design of CDR with offline reasoning generation ensures that the framework remains computationally practical while delivering state-of-the-art results.

## 5. Conclusions

The field of Video Question Answering (VideoQA) faces significant challenges, particularly due to Multimodal Large Language Models' (MLLMs) tendency to generate single-perspective or flawed reasoning. To overcome this, we introduced **Contextualized Diverse Reasoning (CDR)**, a novel and comprehensive framework that provides VideoQA models with a robust, multi-dimensional understanding. CDR integrates three interdependent modules: the **Diverse Reasoning Generator (DRG)**, which harnesses MLLMs to produce multiple distinct reasoning pathways; the **Reasoning Pathway Refiner and Annotator (RPRA)**, ensuring pathway purity and structuredness; and the **Context-Aware Reasoning Fusion (CARF)** module, which intelligently combines these diverse cues with visual and question features using adaptive attention. Our extensive evaluations across NExT-QA, STAR, and IntentQA datasets demonstrated CDR's superior, state-of-the-art performance, particularly in complex reasoning categories. Ablation studies affirmed the critical contributions of each module. CDR represents a significant advancement in VideoQA, showcasing the power of orchestrating diverse MLLM-generated reasoning in a context-aware manner, leading to deeper understanding and more reliable predictions, and opening promising avenues for future research.

## References

1. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. 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. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.
2. Asai, A.; Kasai, J.; Clark, J.; Lee, K.; Choi, E.; Hajishirzi, H. XOR QA: Cross-lingual Open-Retrieval 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. 547–564. https://doi.org/10.18653/v1/2021.naacl-main.46.
3. Herzig, J.; Müller, T.; Krichene, S.; Eisenschlos, J. Open Domain Question Answering over Tables via Dense Retrieval. 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. 512–519. https://doi.org/10.18653/v1/2021.naacl-main.43.
4. 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.
5. Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; Chua, T.S. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. 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. 3277–3287. https://doi.org/10.18653/v1/2021.acl-long.254.
6. Aiyappa, R.; Senthilmani, S.; An, J.; Kwak, H.; Ahn, Y. Benchmarking zero-shot stance detection with FlanT5-XXL: Insights from training data, prompting, and decoding strategies into its near-SoTA performance. *CoRR* **2024**. https://doi.org/10.48550/ARXIV.2403.00236.
7. Liang, J.; Meng, X.; Zhang, H.; Wang, Y.; Wei, J.; Zhao, D. ReasVQA: Advancing VideoQA with Imperfect Reasoning Process. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025

 doi:10.20944/preprints202512.2254.v1



- Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025. Association for Computational Linguistics, 2025, pp. 1696–1709. https://doi.org/10.18653/V1/2025.NAACL-LONG.82.
8. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6787–6800. https://doi.org/10.18653/v1/2021.emnlp-main.544.
9. Zhong, Y.; Ji, W.; Xiao, J.; Li, Y.; Deng, W.; Chua, T.S. Video Question Answering: Datasets, Algorithms and Challenges. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 6439–6455. https://doi.org/10.18653/v1/2022.emnlp-main.432.
10. Xiao, S.; Chen, L.; Shao, J.; Zhuang, Y.; Xiao, J. Natural Language Video Localization with Learnable Moment Proposals. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4008–4017. https://doi.org/10.18653/v1/2021.emnlp-main.327.
11. Liu, Y.; Yu, R.; Yin, F.; Zhao, X.; Zhao, W.; Xia, W.; Yang, Y. Learning quality-aware dynamic memory for video object segmentation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 468–486.
12. Liu, Y.; Bai, S.; Li, G.; Wang, Y.; Tang, Y. Open-vocabulary segmentation with semantic-assisted calibration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3491–3500.
13. Liu, Y.; Zhang, C.; Wang, Y.; Wang, J.; Yang, Y.; Tang, Y. Universal segmentation at arbitrary granularity with language instruction. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3459–3469.
14. 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.
15. 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.
16. Mirzaee, R.; Rajaby Faghihi, H.; Ning, Q.; Kordjamshidi, P. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. 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. 4582–4598. https://doi.org/10.18653/v1/2021.naacl-main.364.
17. Anantha, R.; Vakulenko, S.; Tu, Z.; Longpre, S.; Pulman, S.; Chappidi, S. Open-Domain Question Answering Goes Conversational via Question Rewriting. 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. 520–534. https://doi.org/10.18653/v1/2021.naacl-main.44.
18. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards Mitigating LLM Hallucination via Self Reflection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123.
19. Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; Wen, J.R. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 9237–9251. https://doi.org/10.18653/v1/2023.emnlp-main.574.
20. 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.
21. Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Le Bras, R.; Choi, Y.; Hajishirzi, H. Generated Knowledge Prompting for Commonsense Reasoning. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 3154–3169. https://doi.org/10.18653/v1/2022.acl-long.225.

 doi:10.20944/preprints202512.2254.v1




22. Qiao, S.; Ou, Y.; Zhang, N.; Chen, X.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Chen, H. Reasoning with Language Model Prompting: A Survey. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 5368–5393. https://doi.org/10.18653/v1/2023.acl-long.294.
23. Li, J.; Xu, K.; Li, F.; Fei, H.; Ren, Y.; Ji, D. MRN: A Locally and Globally Mention-Based Reasoning Network for Document-Level Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 1359–1370. https://doi.org/10.18653/v1/2021.findings-acl.117.
24. Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.L.; Huang, H.; Zhou, M. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. 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. 3576–3588. https://doi.org/10.18653/v1/2021.naacl-main.280.
25. Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; Gao, J. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 956–968. https://doi.org/10.18653/v1/2022.naacl-main.70.
26. Zhou, Y.; Zhang, J.; Chen, G.; Shen, J.; Cheng, Y. Less Is More: Vision Representation Compression for Efficient Video Generation with Large Language Models, 2024.
27. Hazarika, D.; Li, Y.; Cheng, B.; Zhao, S.; Zimmermann, R.; Poria, S. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 685–696. https://doi.org/10.18653/v1/2022.naacl-main.50.
28. Paranjape, B.; Michael, J.; Ghazvininejad, M.; Hajishirzi, H.; Zettlemoyer, L. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4179–4192. https://doi.org/10.18653/v1/2021.findings-acl.366.
29. Xu, S.; Tian, Y.; Cao, Y.; Wang, Z.; Wei, Z. Benchmarking Machine Learning and Deep Learning Models for Fake News Detection Using News Headlines. *Preprints* **2025**. https://doi.org/10.20944/preprints202506.1183.v1.
30. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv* **2025**, arXiv:2509.00981.
31. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, 1–13. https://doi.org/10.1109/TVT.2025.3638264.
32. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv* **2025**, arXiv:2501.01886.
33. 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.
34. Jingzhi, W.; Cui, X. The impact of blood and urine biomarkers on age-related macular degeneration: insights from mendelian randomization and cross-sectional study from NHANES. *Biological Procedures Online* **2024**, *26*, 19.
35. Cui, X.; Wen, D.; Xiao, J.; Li, X. The causal relationship and association between biomarkers, dietary intake, and diabetic retinopathy: insights from Mendelian randomization and cross-sectional study. *Diabetes & Metabolism Journal* **2025**.
36. Liu, Z.W.; Peng, J.; Chen, C.L.; Cui, X.H.; Zhao, P.Q. Analysis of the etiologies, treatments and prognoses in children and adolescent vitreous hemorrhage. *International Journal of Ophthalmology* **2021**, *14*, 299.
37. Liu, F.; Geng, K.; Chen, F. Gone with the Wind? Impacts of Hurricanes on College Enrollment and Completion 1. *Journal of Environmental Economics and Management* **2025**, 103203.
38. Liu, F.; Geng, K.; Jiang, B.; Li, X.; Wang, Q. Community-Based Group Exercises and Depression Prevention Among Middle-Aged and Older Adults in China: A Longitudinal Analysis. *Journal of Prevention* **2025**, 1–20.
39. Liu, F.; Liu, Y.; Geng, K. Medical Expenses, Uncertainty and Mortgage Applications. *Uncertainty and Mortgage Applications* **2024**.

 doi:10.20944/preprints202512.2254.v1