Parse, Align and Aggregate: Graph-driven Compositional Reasoning for Video Question Answering

Jiangtong Li, Zhaohe Liao, Fengshun Xiao, Tianjiao Li, Qiang Zhang, Haohua Zhao, Li Niu, Guang Chen, Liqing Zhang, Changjun Jiang

***Abstract*—Video Question-Answering (VideoQA) enables machines to interpret and respond to complex video content, advancing human-computer interaction. However, existing multimodal large language models (MLLMs) often provide incomplete or opaque explanations and existing benchmarks mainly focus on the correction of final answers, limiting insight into their reasoning processes and hindering both transparency and verifiability. To address this gap, we propose the Question Parsing, Video Alignment and Answer Aggregation framework (QPVA^3), which leverages a compositional graph to drive visual and logical reasoning in VideoQA. Specifically, QPVA^3 consists of three core components, the planner, executor, and reasoner to generate the compositional graph and conduct graph-driven reasoning. For the original question, the planner parses it into the compositional graph, capturing the underlying reasoning logic and structuring it into a series of interconnected questions. For each question in compositional graph, the executor aligns the video by selecting relevant video clips and generates answers, ensuring accurate, context-specific responses. For each question with its first-order descents, the reasoner aggregates answers by integrating reasoning logic with visual evidence, resolving conflicts to produce a coherent and accurate response. Moreover, to assess the performance of existing MLLMs in the reasoning processes of VideoQA, we introduce novel compositional consistency metrics and construct a VideoQA benchmark (QPVA^{3}Bench) with 3,492 question-video tuples, each annotated with detailed compositional graphs and fine-grained answers. We evaluate the QPVA^3 framework on QPVA^{3}Bench and 5 other VideoQA benchmarks. Experimental results demonstrate that our framework improves both consistency and accuracy compared to baselines, leading to a more transparent and verifiable VideoQA system. This approach has the potential to advance the field, as supported by our comprehensive evaluation and benchmarking efforts. Code and dataset are available at https://github.com/QPVA3/QPVA3-PAMI.**

***Index Terms*—Multi-modal Large Language Model, Multi-modal Reasoning Framework, Multi-modal Benchmark, Compositional Reasoning, Video Question-Answering.**

I. Introduction

VIDEO Question-Answering (VideoQA) has recently emerged as a prominent research area, with significant applications in interactive artificial intelligence and recognition science. The advent of Large Language Models (LLMs) and their multimodal extensions (MLLMs) has spurred the development of numerous models, such as Video-LLaMA [1], [2], and Video-LLaVA [3], which exhibit substantial capabilities in understanding and reasoning over video data. These models have achieved notable success on several challenging VideoQA benchmarks [4]–[10].

However, existing MLLMs often offer limited explanations for their answers to complex questions. This lack of transparency makes it challenging to comprehend the full reasoning process, reducing users' trust in the models and making it difficult to identify and correct issues in their predictions. For instance, in Fig. 1, Video-LLaVA answers the question, "Is the boy skilled at eating?" with "No" and explanation comprising only a few vague clues. This explanation neither identifies the specific video clips that show "how the boy is eating", nor does it outline the reasoning process behind the answer. Therefore, the lack of reasoning transparency makes the question-answering (QA) process unverifiable, as users cannot trace how conclusions are drawn. This not only hinders our understanding of the reasoning logic but also limits our ability to enhance MLLM accuracy, especially when addressing questions involving temporal relationships and multiple visual cues [11].

In VideoQA, perception and cognition are two essential stages involved in understanding the video scene and answering questions. Specifically, low-level perception is required to understand the spatiotemporal characteristics of the video, while high-level cognition aims to comprehend the logic behind the video and question, reasoning along the logical structure. VoT [12] attempts to construct a cognition-level VideoQA reasoning framework by providing detailed analysis of the video at the object and action levels, and then performing reasoning based on the fine-grained video representation. While VoT improves MLLM performance and offers additional reasoning clues, it still falls short in capturing the logical structure of questions and fully explaining the reasoning process. Considering that existing MLLMs have exhibited formidable video understanding abilities and can handle relatively simple questions [2], we propose constructing the cognition system starting from the question itself. By reconceptualizing VideoQA as a compositional reasoning

This work is supported in part by the National Natural Science Foundation of China (NFSC) (No. 62402341, 62471287), and the Shanghai Municipal Science and Technology Major Project, China (Grant No. 2021SHZDZX0102).

Jiangtong Li and Zhaohe Liao contribute equally in this work.

Corresponding Authors: Li Niu and Liqing Zhang

Zhaohe Liao, Haohua Zhao, Li Niu, and Liqing Zhang are with the School of Computer Science and Technology, Shanghai Jiao Tong University, Shanghai, China. E-mail: {zhaoheliao, haoh.zhao, ustcnewly, zhang-lq}@sjtu.edu.cn

Jiangtong Li, Guang Chen and, Changjun Jiang are with the School of Computer Science and Technology, Tongji University, Shanghai, China. E-mail: {jiangtongli, guangchen, cjjiang}@tongji.edu.cn

Fengshun Xiao, Tianjiao Li, and Qiang Zhang are with the BiliBili. E-mail: {xiaofengshun, litianjiao01, zhangqiang}@bilibili.com

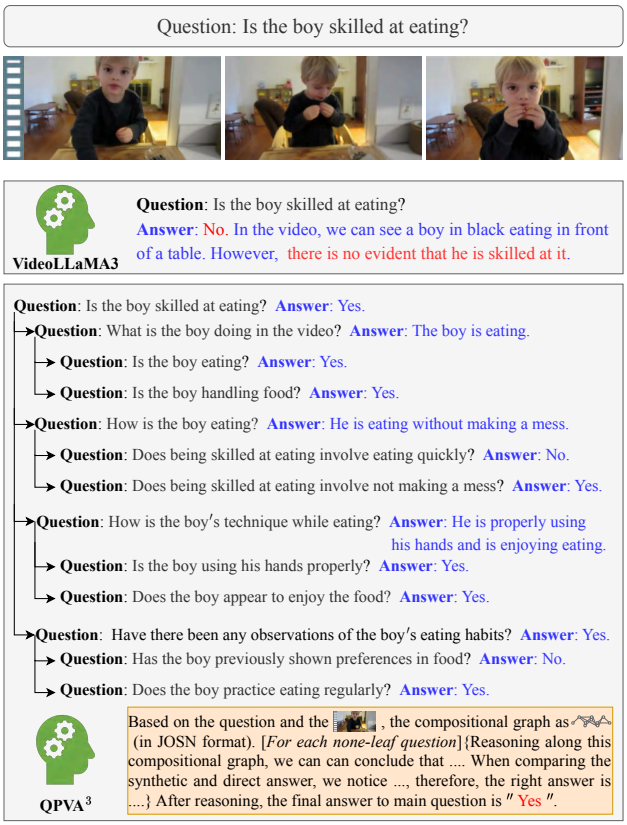


Fig. 1: The reasoning pipeline between VideoLLaMA3 and our QPVA^3. The VideoLLaMA3 can only answers the question with only a few vague clues, however, QPVA^3 can provide the compositional graph and reasoning process.

task, we aim to enhance both the reasoning capabilities and transparency of the system. To improve the cognition within the MLLM framework, we first parse complex questions into compositional graphs, and then align and aggregate relevant information according to the logical structure.

In this work, we introduce the Question Parsing, Video Alignment, and Answer Aggregation framework (QPVA^3), which addresses these challenges by reconceptualizing VideoQA as a compositional reasoning task. By focusing on the logical structure of questions, our framework enhances reasoning transparency and is applicable to various MLLMs [1], [3], [13]. Specifically, our QPVA^3 framework incorporates three core components, the planner, the executor, and the reasoner, that collaboratively represent and process the reasoning of the original question within a compositional graph. The planner utilizes an MLLM to parse the complex question into a compositional graph, effectively capturing its underlying logical structure and segmenting it into manageable sub-questions. The executor employs a video aligner and an MLLM to generate answers for each question in the compositional graph. By selecting relevant video clips for each question through the video aligner, the executor ensures that background scenes do not affect the predictions [14], leading to accurate, context-specific responses. The reasoner leverages both an LLM and an MLLM to aggregate individual answers into a coherent, comprehensive response. The LLM performs compositional reasoning along the compositional graph and check for logical consistency among the answers. Meanwhile, the MLLM arbitrates these conflicts using the video content, ensuring that the response is both logically sound and accurate.

Existing VideoQA benchmarks, such as MVBench [9] and NExT-QA [4], focus on the predicted answer of each video-question pair, ignoring the reasoning consistency during the process. To evaluate the reasoning consistency in VideoQA, we propose a new data annotation pipeline that parse questions into compositional graphs without requiring fine-grained spatiotemporal scene graphs of videos [6], [11]. Specifically, for each video-question pair, we create compositional graphs and corresponding answers with the assistance of ChatGPT-4o, which are then checked and corrected by human experts. The video-question pairs are selected from five existing datasets: MSVD, MSRVTT, TGIF-QA, NExT-QA, and Causal-VidQA, based on question type and the complexity of the compositional graphs. For each combination of question type and complexity level, we include about 250 questions in our benchmark to reduce the bias. After filtering and annotation, we construct QPVA^{3}Bench, consisting of 3,492 question-video pairs with corresponding compositional graphs and fine-grained answers.

As for the evaluation metrics, AGQA-Decomp [11] proposes the compositional accuracy (CA), right for the wrong reasons (RWR), and delta (CA-RWA) to evaluate the compositional consistency. However, these metrics only focus on reasoning failure based on the sub-questions correctness without considering the main question correctness, leading to asymmetric and unstable problems. To address this, we extend it to provide a symmetric and stable measurement for compositional consistency. Our metrics include consistency precision (cP), consistency recall (cR), and consistency F_1 (c-F_1) along with their negative versions. These metrics can evaluate the compositional consistency of MLLMs in a balanced viewpoint.

To evaluate the performance of our QPVA^3 framework, we conduct experiments on QPVA^{3}Bench and other five VideoQA benchmarks, including AGQA-Decomp [11], Causal-VidQA [5], STAR [6], and MVBench [9]. Furthermore, we analyze the effectiveness of individual modules in QPVA^3 framework and present examples illustrating how the reasoning process is enhanced. Our contribution is summarized as:

- **Framework**: We introduce the QPVA^3 framework, which incorporates three core components to enhance both the reasoning transparency and accuracy of existing method.
- **Benchmark**: We construct the **QPVA^{3}Bench**, a new VideoQA benchmark built with a human-in-the-loop pipeline that provides compositional graphs and fine-grained answers for detailed evaluation.
- **Metrics**: We propose **novel metrics** grounded in a theoretical analysis that proves the symmetry and stability of our system over prior methods, enabling a more balanced assessment of compositional reasoning.
- **Experiment**: Extensive experiments demonstrate **QPVA^3** framework improves both compositional consistency and accuracy over baselines on multiple benchmarks, leading to a more transparent and verifiable system.

This paper extends our previous work [15], with several important improvements. **For algorithmic**, we extend the

TABLE I: A contribution comparison of this work to our previous publication [15].

Category	CVPR 2024 Version (VA^3)	This Work (QPVA^3)
Algorithm	**Question Parser:** Utilized an external, text-only LLM pipeline requiring few-shot examples.	**Planner:** Upgraded to a unified, MLLM-based Planner that leverages video context to generate compositional graphs end-to-end.
	Answer Aggregator: Employed a training-dependent Graph Neural Network requiring dataset-specific fine-tuning.	**Reasoner:** Introduced a training-free, MLLM-integrated Reasoner for graph-based logical inference, improving generalization and transferability.
Theory	-	**Formal Metric Analysis:** Provided a theoretical proof demonstrating the asymmetry and instability of the CA-RWR system. Formally proved that our proposed c-F_1 system is symmetric, stable, and continuous.
Benchmark	-	**New Benchmark (QPVA^{3}Bench):** Constructed a new, high-quality benchmark with (1) broader scope, (2) question-specific applicability, and (3) higher accuracy via LLM+human annotation.
Experiment	Experiments were conducted on AGQA-Decomp and its variants.	**Evaluation:** Conducted new experiments on QPVA^{3}Bench and multiple MLLM-specific benchmarks (MVBench, STAR, etc.).
	Visualizations showed the effectiveness of the modules.	**Analysis:** Provided detailed ablation studies, visualization and statistical significance analysis, proving the effectiveness of our framework.

original VA^3 pipeline into the QPVA^3 framework by replacing the external text-only parser and the training-dependent GNN with a unified MLLM-based Planner and a training-free Reasoner, which enhances end-to-end integration, generalization, and transferability. **For theory**, we provide a formal analysis that proves the asymmetry and instability of the previous CA-RWR metrics, while also proving that our proposed c-F_1 system is symmetric, stable, and continuous. **For benchmark**, we introduce a novel compositional reasoning benchmark, QPVA^{3}Bench, which is constructed via a comprehensive LLM-assisted and human-verified pipeline that generates question-specific compositional graphs from a diverse range of reasoning types without relying on scene graphs. **For experiments**, we conduct extensive evaluations on our QPVA^{3}Bench and broader evaluation benchmarks, including MVBench, STAR, and Causal-VidQA, and provide statistical analysis demonstrating the superiority of QPVA^{3}Bench.

II. Related Works

A. VideoQA Dataset and Benchmark

VideoQA extends ImageQA by incorporating the temporal dimension to answer questions about dynamic visual content. Recent VideoQA benchmarks have progressed to complex tasks demanding skills like temporal [16], physical [17], [18], evidence [4], [6], and commonsense reasoning [5], [7], as well as long video understanding [8]. Concurrently, the rise of MLLMs in video understanding has led to test-only benchmarks like MVBench [9] and MMT-Bench [19] for performance evaluation. Despite these advancements, using compositional visual events for answer prediction remains underexplored. For instance, AGQA [20] and AGQA-Decomp [11] use fine-grained spatiotemporal scene graphs from Action Genome [21] to build compositional reasoning datasets. However, annotating these scene graphs is labor-intensive, which limits the creation of scalable compositional reasoning datasets for broader tasks like evidence [4] and commonsense reasoning [5]. Our proposed pipeline addresses this by using ChatGPT-4o to parse VideoQA questions, rather than videos, into compositional graphs with fine-grained answers. Human experts then verify and correct the outputs, ensuring both efficiency and accuracy.

B. VideoQA Methodology

Prior to the proliferation of LLMs, VideoQA methods focused primarily on video-question alignment. These methods implemented this alignment with cross-modal attention [22] or memory networks [23], before graph reasoning became popular [24]. Recently, exploiting the natural hierarchy in video representation has gained traction [25]–[27]. HQGA [26], VGT [27], and CoVGT [28] align the question and video hierarchies from low-level entities to high-level activities.

Video-MLLMs typically combine a pre-trained visual encoder for feature extraction [29], a vision-language aligner, and an instruction-tuned language decoder for generating responses [30]. Much research has focused on designing effective visual alignment approaches. Finsta [31] uses structure-aware spatio-temporal alignment of video and text scene graphs to improve VLM representations. Video-ChatGPT [13] introduces spatial and temporal pooling to caption information independently. Video-LLaMA [1] employs a Video Q-Former to capture temporal changes, while VideoLLaMA2 [2] uses a Spatial-Temporal Convolution Connector to efficiently capture spatio-temporal features. LLaVA-NeXT [32] uses the AnyRes algorithm to balance performance and cost across different resolutions. Video-LLaVA [3] aligns video and image features before LLM projection for a unified, computationally efficient representation.

Despite these advancements, existing MLLMs still provide limited explanations for their answers to complex questions. This opacity obscures their reasoning, which undermines user trust and hampers error correction. To address this, we propose the QPVA^3 framework for compositional reasoning. The planner, executor, and reasoner collaboratively parse, align, and aggregate information to improve reasoning transparency.

C. Compositional Reasoning

Parsing complex questions into simpler sub-questions is a technique used in tasks like ImageQA [33]. Early bench-

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3650864

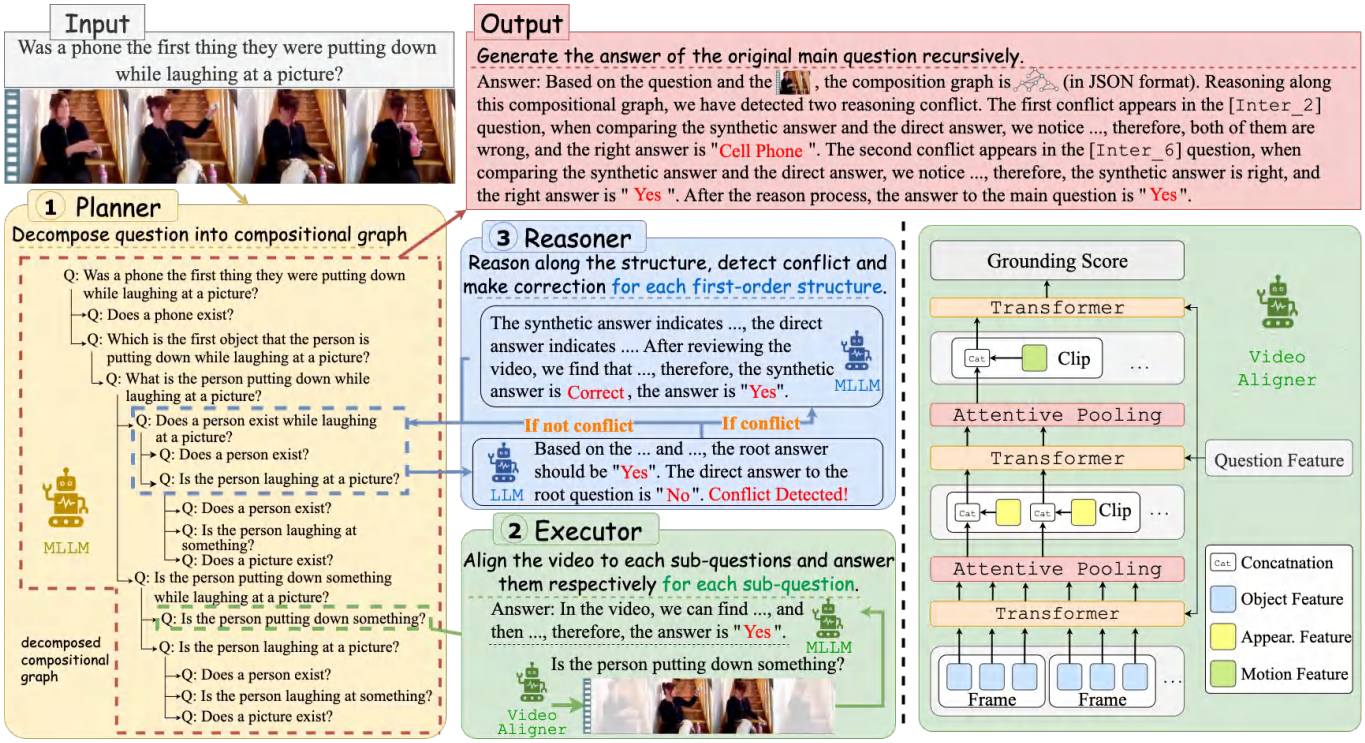


Fig. 2: The overall framework of our QPVA^3. The planner, the executor, and the reasoner work collaboratively to parse questions, align videos, and aggregate answers during the reasoning of the original question.

marks [20], [34] parsed questions into modular programs, often executed by neural modular networks (NMNs) [35], to derive answers. For instance, ACMN [33] uses dependency trees to parse questions. The GQA benchmark [34] uses structured programs to compute answers from scene graphs. Building on GQA [34], AGQA [20] represents reasoning programs for VideoQA using spatio-temporal scene graphs. However, existing VideoQA methods cannot directly use the programs from AGQA. AGQA-Decomp [11] addresses this by transforming programs into sub-questions within compositional graphs to evaluate the consistency of VideoQA methods.

Neural modular networks (NMNs) are widely used for compositional reasoning in ImageQA [36] and VideoQA [35]. They parse questions into sub-programs and assemble the corresponding modules to find the answer. While NMNs are interpretable, their reliance on pre-defined modules that limit their flexibility and scalability. VoT [12] aims for transparency by reasoning over fine-grained object and action-level video representations. However, it struggles to capture the logical structure or adequately explain its reasoning. Therefore, we propose the QPVA^3 framework, which starts by parsing questions to frame VideoQA as a compositional reasoning task. This approach improves both reasoning capability and transparency, moving towards a more cognitive-level process.

D. Causal Reasoning in VidQA

Recent efforts in VideoQA have explored causal inference techniques to mitigate spurious cross-modal biases and enhance reasoning robustness. For example, IGV series models [14], [37], [38] focus on differentiating between causal and environmental clips in VideoQA. CMCIR [39] introduces front- and back-door interventions to disentangle visual-linguistic spurious correlations. KPI [40] addresses dataset bias by using front-door intervention as a knowledge proxy to mitigate its effects. Similarly, VCSR [41] present a Visual Causal Scene Refinement method that performs causal analysis to isolate critical video segments/frames as the "visual causal scene" for each question. It employs question-guided front-door interventions to find causally evidence and separates non-causal scenes via a contrastive learning mechanism. CRA [42] further addresses the video question answering task with a cross-modal causal grounding, which aligns answer prediction with temporal grounding by leveraging bidirectional contrastive learning and multi-modal deconfounding via applying front-door intervention on vision and back-door on language.

These causal reasoning frameworks explicitly construct causal graphs or apply intervention operations to remove confounders and highlight true cause–effect relations in videos. In contrast, our QPVA^3 method takes a graph-driven compositional reasoning approach rather than performing formal interventions, we parse each complex question into a hierarchy of sub-questions and align video evidence to these sub-queries, then aggregate the intermediate answers via a question parse graph. Such structured pipeline inherently guides the model to focus on relevant events and their relationships (*e.g.*, temporal or causal dependencies mentioned in the question) without relying on spurious correlations. Therefore, QPVA^3 achieves robust and interpretable VideoQA by enforcing compositional reasoning constraints, offering an alternative yet complementary route to capturing causal dependencies in video narratives.

III. QPVA^3 Framework

QPVA^3 consists of a planner, an executor, and a reasoner that collaboratively parse, align, and aggregate textual and visual information to answer the question. The planner (Subsec. III-A) uses an MLLM to parse the question into a compositional graph, capturing its logical structure and creating manageable sub-questions. The executor (Subsec. III-B) uses a video aligner and an MLLM to generate an answer for each sub-question in the graph. The reasoner (Subsec. III-C) uses both an LLM and an MLLM to aggregate the sub-answers into a final, coherent response. Finally, we provide optimization objectives and extra discussion (Subsec. III-D).

A. Planner

The planner in our QPVA^3 framework drives compositional reasoning by managing the complex logic across visual and textual data. To overcome the limitations of prior vision-agnostic external parsers [15], the planner is designed to be vision-aware, ensuring the decomposition is grounded in the video content. For a given video and question, the planner begins by parsing the question's underlying structure and intent. It deconstructs the question into a compositional graph that defines the logical relationships among sub-questions. This decomposition is applied recursively until the leaf nodes are perceptual questions that an MLLM can directly answer. This graph structure enables the framework to solve complex questions from low-level perception with high-level cognition.

The parsing process follows specific instructions to ensure a coherent breakdown of the input question. The planner takes the initial question $\mathcal{Q}$ and constructs a compositional graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (sub-questions) and $\mathcal{E}$ is the set of edges indicating the compositional relationship. The compositional graph is outputted in JSON format, facilitating data handling and interoperability. This ensures that the graph $\mathcal{G}$ accurately captures the compositional reasoning structure. Therefore, we prompt the MLLMs with:

> **Instruction**: Suppose that you are an expert in linguistics and logic, familiar with parsing complex questions. We will give the rule of hierarchical parsing, please parse the further provided questions based on the video content.
> **Rule**: Hierarchically parsing the complex question into relatively simple sub-questions, until they are simple atomic questions, which are questions that can be directly perceived from the video. The sub-questions could be possibly overlapped. The parsing result shall be given in JSON format as a compositional graph, and each different sub-question shall have an identical ID. The compositional shall end only when occurring the atomic questions.
> **Question**: {original_question}.
> **Video:** {video_tokens}

Beyond parsing, the planner coordinates the reasoning process by managing the executor and the reasoner. As the central coordinator, it orchestrates the other modules to ensure accurate and transparent reasoning. The planner directs the executor to answer each sub-question in the compositional graph. The executor then uses visual-language models to answer these perceptual questions using the video content. Once the executor finishes, the planner recursively engages the reasoner in a bottom-up manner for transparent, controllable aggregation. The reasoner combines sub-question answers within each logical structure to derive higher-level insights. This collaborative process makes the reasoning transparent, as each logical step can be inspected and validated. By separating concerns and managing module communication, the planner ensures effective reasoning across all levels of abstraction.

To exploit the strong logical planning ability of MLLMs, we fine-tune the MLLM to generate the JSON-formatted compositional graph, where the loss can be formulated as:

$$\mathcal{L}_g = \frac{1}{k} \sum_{i=t}^{i+t} -\log p[q_i], \tag{1}$$

where t and k are the input and output sequence length, and sequence $[q_t, q_{t+1}, \ldots, q_{t+k}]$ is the desired output sequences.

The QPVA^3 framework provides enhanced transparency and user engagement. Separating the executor (answering) from the reasoner (reasoning) allows users to clearly trace the logic throughout the compositional graph. This design enables human-in-the-loop interaction, letting users analyze, intervene, and correct reasoning errors in real-time. This interactive process improves reasoning outcomes and fosters greater trust in the framework's conclusions.

B. Executor

The executor in QPVA^3 answers questions by analyzing the provided video content. It employs a two-stage process to answer each question from the compositional graph. First, a video aligner (Subsec. III-B1) selects the most relevant video clips for the given sub-question. Then, an MLLM generates the answer (Subsec. III-B2) using only these selected clips.

1) Video Alignment: Our video aligner is designed to address the challenge of incomplete alignment between questions and video content. **This problem has two sources: questions often lack specific temporal references, hindering grounding models, and scarce annotation data prevents MLLMs from providing precise supervision.** Our hierarchical model for video-question alignment overcomes these issues. It uses a weakly supervised approach that combines contrastive learning with answer prediction to achieve answer-oriented alignment without extra annotations. The structure of the video aligner is shown in Fig. 2. The video aligner processes a video-question pair (v, q) to identify the most relevant video clips. Prior work on this task [14], [37] did not exploit the natural hierarchy of video features, which limited their representational power.

We propose a lightweight video aligner using three transformer layers to hierarchically fuse object, frame, and clip-level video features with question features for selecting relevant clips. A video v is represented by three feature levels: object ($\boldsymbol{F}_o \in \mathbb{R}^{n_c \times n_f \times n_o \times h_v}$), appearance ($\boldsymbol{F}_a \in \mathbb{R}^{n_c \times n_f \times h_v}$), and motion ($\boldsymbol{F}_m \in \mathbb{R}^{n_c \times h_v}$). Here, n_c, n_f, and n_o are the number of clips, frames per clip, and objects per frame, respectively, while h_v is the feature dimension. The aligner uses a bottom-up approach to integrate video and question

features. First, it aggregates object features $\boldsymbol{F}_o$ based on the question, concatenates the result with appearance features $\boldsymbol{F}_a$, and then aggregates this combined feature across frames. The result is then concatenated with motion features $\boldsymbol{F}_m$ and fused with question features to produce grounding scores. Each aggregation step uses a transformer layer, with video features as queries and question features as keys and values. The aggregation from object to appearance is defined as:

$$\boldsymbol{F}_o^j = \text{TF}(\boldsymbol{F}_o, \boldsymbol{F}_q, \boldsymbol{F}_q); \tag{2}$$

$$\boldsymbol{F}_o^a = \sum_{n_o} \sigma_{n_o}\left(\boldsymbol{W}_o \boldsymbol{F}_o^j + \boldsymbol{b}_o\right) \boldsymbol{F}_o^j; \boldsymbol{F}_a^c = [\boldsymbol{F}_o^a || \boldsymbol{F}_a]; \tag{3}$$

where σ_{n_o} is the softmax function over the object dimension, TF is a transformer encoder, $\boldsymbol{W}_o$ and $\boldsymbol{b}_o$ are trainable, and $[\cdot||\cdot]$ is concatenation. The updated feature $\boldsymbol{F}_a^c$ is then used to produce the motion feature $\boldsymbol{F}_m^c$ similarly. Finally, $\boldsymbol{F}_m^c$ is used to generate a binary indicator for relevant clips:

$$\boldsymbol{F}_m^j = \text{TF}(\boldsymbol{F}_m^c, \boldsymbol{F}_q, \boldsymbol{F}_q); \boldsymbol{s}_{rel} = \text{MLP}_1(\boldsymbol{F}_m^j); \tag{4}$$

$$\boldsymbol{s}_{irr} = \text{MLP}_2(\boldsymbol{F}_m^j); I = \text{Gumbel-Softmax}([\boldsymbol{s}_{rel} || \boldsymbol{s}_{irr}]), \tag{5}$$

where MLP is the multi-layer linear projection.

Since the ground-truth of aligned video clips is not available, we exploit contrastive learning [37] to guide this module. For a given training pair (v, q), the indicator identifies relevant ($\hat{v}_r$) and irrelevant ($\hat{v}_c$) clips within the video v. This allows us to define the anchor ($\boldsymbol{f}_{\hat{v}_r,q}$), positive ($\boldsymbol{f}_{\hat{v}',q}$), and negative ($\boldsymbol{f}_{\hat{v}_c,q}$) features for the contrastive loss:

$$\boldsymbol{f}_{\hat{v}_r,q} = \mathcal{F}(\hat{v}_r, q); \quad \boldsymbol{f}_{\hat{v}',q} = \mathcal{F}(v', q); \quad \boldsymbol{f}_{\hat{v}_c,q} = \mathcal{F}(\hat{v}_c, q), \tag{6}$$

where $\hat{v}'$ is formed by replacing irrelevant clips $\hat{v}_c$ with random ones. The function $\mathcal{F}$ is an MLLM, and we use its average generated token feature as the output $\mathcal{F}(v, q)$. Therefore, the contrastive loss is defined as

$$\mathcal{L}_c = -\log \frac{\exp(\boldsymbol{f}_{\hat{v}_r,q}^T \boldsymbol{f}_{\hat{v}',q})}{\exp(\boldsymbol{f}_{\hat{v}_r,q}^T \boldsymbol{f}_{\hat{v}',q}) + \exp(\boldsymbol{f}_{\hat{v}_r,q}^T \boldsymbol{f}_{\hat{v}_c,q})}. \tag{7}$$

While this contrastive loss helps separate relevant and irrelevant clips, it provides no explicit alignment supervision. To create this supervision, we introduce an answer-oriented loss based on next-token prediction, formulated as:

$$\mathcal{L}_a = \frac{1}{k} \sum_{i=t}^{t+k} -\log p[q_i], \tag{8}$$

where, t and k are the input and output lengths, q_i is the i-th ground-truth token, and $p[q_i]$ is its predicted probability.

2) Answer Prediction: After the video aligner identifies relevant clips, an MLLM uses them and the question to generate an answer. The MLLM processes the visual features from these clips alongside the textual question. This process allows the model to concentrate on the visual cues most relevant to questions. Therefore, we prompt the MLLMs with:

> **Instruction**: Given the question and video, answer the question using several words or phrase.
> **Question**: {original_question}.
> **Video:** {video_tokens}

C. Reasoner

The reasoner in QPVA^3 synthesizes sub-question answers to resolve their parent question within the compositional graph. For any parent question, it first infers an answer using the existing results from its children nodes. It then compares this inferred answer to the one generated directly by the executor. If answers conflict, the reasoner uses visual information to adjudicate the response, ensuring final consistency and accuracy.

Answering a parent question using its sub-answers requires advanced logical reasoning. Introducing visual information at this stage can add noise and mislead the model's reasoning process. Therefore, our initial reasoning stage uses an LLM to focus purely on textual logic, avoiding visual distractions. The LLM aggregates sub-question answers using only textual data, applying its advanced reasoning capabilities. This method requires the LLM to understand semantic dependencies within the compositional graph. The LLM also evaluates consistency by comparing its derived answer with the generated ones by the executor. Therefore, we prompt the LLM with:

> **Instruction**: Assume that you are an expert in logical reasoning. In this task, you will be given a question about a video and the directly predicted answer. The question can be parsed into several simpler leaf questions, which are also given with their predicted answers. This hierarchical structure is indicated by indentation in the input questions. The question IDs are provided before each question, and you may refer to the questions the IDs. Your task is to describe how to derive the answer of the main question from the answers of the sub-questions and judge whether derived answer conflicts with directly predicted answer.
> **Rule**: Your output is a JSON only with the following keys:
> **- derivation_process**: The reasoning process of deriving the answer of the root question.
> **- derived_answer**: The derived answer of the root question from the answers of the sub-questions.
> **- is_conflict**: Whether the derived answer conflicts with the directly predicted answer.
> **Compositional Graph**: {first_order_parsing_structure}.

If the logically inferred and video-derived answers conflict, an MLLM begins a second-stage multimodal adjudication. The MLLM then validates both answers by checking their alignment with the video content. If one answer is visually inconsistent, the MLLM defaults to the other. If both answers are incorrect or visually unsupported, the MLLM uses visual cues and the context from the errors to generate a new, more accurate answer. This process involves re-examining the video content and applying multimodal reasoning strategies for a corrected answer. Therefore, we prompt the LLM with:

> **Instruction**: You are an intelligent chatbot designed for evaluating the correctness of generative outputs. You are given a video and two candidate answers for a question. Your task is to evaluate the correctness of these two candidate answers based on the content of the video. If one of them is correct, you should explain the reason why

it is correct from the visual aspect. Otherwise, you should provide the correct answer and explain the reason why the given candidates are both incorrect and why the generated answer is correct from the visual aspect.
Rule: Your output is a JSON only with the following keys:
- wrong_candidate_answer_0_visual_analysis: The explanation and visual evidence to explain answer_0. Be <Empty> if answer_0 answer is correct.
- wrong_candidate_answer_1_visual_analysis: The explanation and visual evidence to explain answer_1. Be <Empty> if answer_1 answer is correct.
- correct_answer_visual_analysis: The explanation and visual evidence that supports the correctness of the chosen candidate answer or generated answer.
- correct_answer: The correct answer chosen form the candidate or generated if neither candidates are correct.
Question: {original_question}.
Candidate Answer 0: {candidate_0}.
Candidate Answer 1: {candidate_1}.
Video: {video_tokens}

The reasoner is central to the logical reasoning capabilities of our QPVA^3 framework. It ensures final answers are logically consistent and visually grounded by integrating sub-answers and using visual verification to resolve conflicts. This two-stage process improves the framework's accuracy, transparency, and controllability, which fosters user trust. Ultimately, the reasoner enhances performance by producing reliable and transparent results, thereby advancing compositional reasoning in VideoQA.

D. Optimization and Discussion

To enable the QPVA^3 framework, we enhance its parsing and alignment capabilities. For parsing, we introduce a compositional graph generation loss, $\mathcal{L}_g$, based on standard next-token prediction. This loss trains the model to generate a question's compositional graph from video and text, improving its ability to parse visually-grounded questions. We use the LoRA method [43] to preserve the MLLM's generalization ability, training on data from AGQA-Decomp [11]. For alignment, we jointly optimize a contrastive loss and an answer prediction loss, which provides weak supervision for the video aligner. This process trains the aligner to select relevant clips from long videos, thus reducing the MLLM's computational load. During this process, the MLLM's weights are frozen, and only the lightweight video aligner is optimized. We use the Gumbel-Softmax technique to enable backpropagation of gradients from $\mathcal{L}_c$ and $\mathcal{L}_a$ to the aligner, which provides a differentiable approximation for the discrete clip selection and is crucial for gradient flow to ensure the aligner is optimized for answer-oriented alignment.

It worth noting that, although QPVA^3 is not explicitly a causal inference model as [39], [41], [42], its compositional graph design inherently captures causal dependencies. Each question is parsed into a directed acyclic graph (DAG) of sub-questions, with edges denoting temporal, logical, or causal dependencies. For example, a cause-effect query (e.g., *"Did X happen because Y happened?"*) produces a graph linking cause Y to effect X, ensuring the model addresses Y first. Aligning each sub-question with the relevant video segment ensures that evidence for precursor events is gathered before reasoning about dependent events. This graph-driven execution mirrors causal reasoning: the model cannot skip intermediate steps and must resolve each link in sequence. Without explicit do-calculus interventions or a dedicated causal module, QPVA^3 still follows the question's dependency graph, preventing shortcuts and requiring visual evidence for each step. Consequently, answers emerge from dependency-aware inferences, improving interpretability and fidelity to true causal relationships implied in the video.

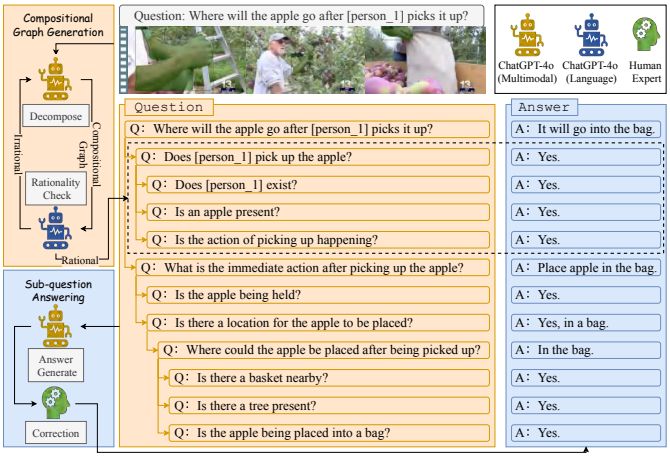


Fig. 3: The data extension pipeline of QPVA^{3}Bench.

IV. QPVA^3 Benchmark

A. Data Selection

To construct the QPVA^{3}Bench, we aim to select representative and diverse video-question (VQ) pairs. These pairs are sourced from five key datasets: MSVD [44], MSRVTT [44], TGIF-QA [45], NExT-QA [4], and Causal-VidQA [5]. The construction process has two main stages: *Representative VQ-Pair Selection* and *Diverse VQ-Pair Sampling*. This method yields a dataset covering a wide range of video content, question types, and complexity. The resulting benchmark is designed to foster models adept at multimodal reasoning.

In the **Representative VQ-Pair Selection** stage, we select the K most representative VQ-pairs from the source datasets to reduce the required number of compositional graph annotations. We define representativeness based on the similarity of both question and video content. Specifically, we represent each VQ-pair by concatenating its video and question embeddings. We generate a 256-dimensional question embedding using the OpenAI `text-embedding-ada-002` model. We compute the video embedding by averaging the embeddings of all its associated questions. We then apply K-means clustering recursively for N iterations, finding $K^{1/N}$ centers at each step, to identify K representative VQ-pairs. We use ChatGPT-4o to annotate the compositional graphs for these selected K VQ-pairs. To improve annotation quality, we provide M few-shot examples in the prompt, selected from AGQA-Decomp based

on VQ-pair similarity. In our experiments, we set $N = 2$, $K = 40,000$, $M = 5$, and obtained 200 cluster centers in each iteration. This process yielded $K_1 = 28,769$ valid VQ-pairs, for which ChatGPT-4o confirmed the logical soundness of their compositional graphs.

In the **Diverse VQ-Pair Sampling** stage, we select VQ-pairs from the representative set to maximize diversity and balance across question types and compositional complexity. We use ChatGPT-4o to classify each question into one of $n_t = 7$ types: **`Counting`**, **`YesOrNo`**, **`When`**, **`What/Who/Which`**, **`Why/How`**, **`Where`**, and **`Other`**. We also measure compositional complexity by the number of sub-questions, creating two categories ($n_c = 2$): **`simple`** (fewer than seven sub-questions) and **`complex`** (seven or more). This process divides the K_1 VQ-pairs into $n_t \times n_c = 14$ distinct groups. From each of these 14 groups, we select the $M = 250$ most diverse VQ-pairs using K-means clustering on their embeddings, yielding a total of 3,500 pairs. We then extend the compositional graph and annotate the sub-question answers for these pairs, followed by manual verification and correction. After this final step, 3,492 questions remained, forming the complete QPVA^{3}Bench.

This systematic process yields a comprehensive dataset for advanced compositional reasoning in multimodal contexts. The representativeness and diversity of QPVA^{3}Bench provide a robust basis for evaluating models on complex reasoning tasks.

B. Data Extension

Constructing QPVA^{3}Bench involves compositional graph generation and sub-question answering to enhance model reasoning on complex videos. The procedure involves generating a compositional graph for a given video and question, followed by answering each derived sub-question. Unlike traditional methods, our approach avoids costly spatial-temporal graph transformations, simplifying the construction of compositional graphs. This independence yields a flexible and efficient method for generating reasoning structures applicable to diverse datasets. The approach streamlines dataset construction and advances multimodal compositional reasoning.

Compositional graph generation begins with a given video and its main question. We then select the five most similar questions from AGQA-Decomp [11] to provide their compositional graphs as few-shot examples. These examples guide ChatGPT-4o in generating a compositional graph for the target question. Next, we use ChatGPT-4o to perform a video-agnostic rationality check, ensuring the graph's logic is sound and its leaf nodes are simple perceptual queries. This check assesses the coherence and validity of the generated graph. If a graph fails this check, we regenerate it, allowing up to three attempts per question. A question is discarded if its graph fails all three attempts. This iterative process ensures all compositional graphs in our dataset are of high quality.

In the **sub-question answering** phase, we answer each sub-question using its corresponding video content. We provide ChatGPT-4o with each sub-question and sampled video frames to generate a detailed answer. We find that answering sub-questions simultaneously resulted in lower accuracy and frequent omissions. To improve performance, we switch to a sequential method that processes each sub-question individually. Next, two graduate students review and correct the responses from ChatGPT-4o for accuracy. This human verification step ensures the quality of the sub-answers and the final dataset.

This methodology improves our dataset by adding complex reasoning structures and detailed annotations. By avoiding spatial-temporal graph transformations, we lower the cost of creating compositional graphs, making the approach more accessible. This improves the support of QPVA^3 for advanced reasoning and offers a scalable, efficient augmentation method for the multimodal field.

C. Dataset Statistics

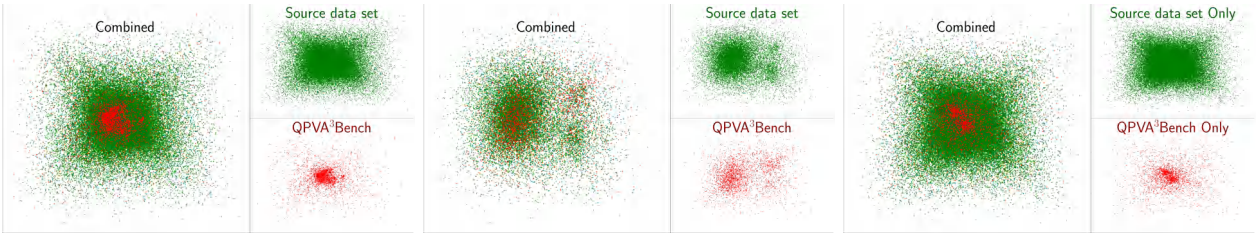
The QPVA^{3}Bench contains 3,492 entries, each with a video, a main question, and a compositional graph. Main questions average 8.13±3.84 words, with answers averaging 4.31±3.00 words. The dataset also includes 9,509± intermediate questions (avg. length: 7.69±2.09) with answers averaging 3.43±3.59 words. It further contains 20,231 leaf questions, which average 6.92±2.17 words, with answers averaging 1.88±2.33 words. Compositional graphs average 9.03±2.29 nodes with a mean hierarchical depth of 2.99±0.42 levels. The associated videos average 16.10±16.64 seconds and 411.5±511.0 frames. These statistics confirm the diversity and complexity of QPVA^3, positioning it as a robust resource for advancing compositional and multimodal reasoning.

D. Bias Analysis

1) Analysis of Sampling Strategy: Our two-stage sampling strategy is designed to construct a representative and diverse benchmark by systematically mitigating semantic and distributional biases from the source datasets. As illustrated by the t-SNE visualizations of video, question, and video-question features in Fig. 4, our QPVA^{3}Bench effectively covers the semantic space of the union of the five source datasets. Furthermore, the distribution of our sampled data is visibly more uniform, avoiding the significant central clustering present in the original data pool. A quantitative analysis of the question type distribution, shown in Fig. 6, reveals that the source datasets are heavily skewed towards **`What/Which`** questions, with less focus on other categories. In contrast, our stratified sampling approach ensures a more balanced distribution across all seven categories, including under-represented types like **`Why/How`** and **`Counting`**. From the analysis, our sampling strategy alleviates the **semantic coverage bias** and **question type distribution bias** inherent in the source datasets.

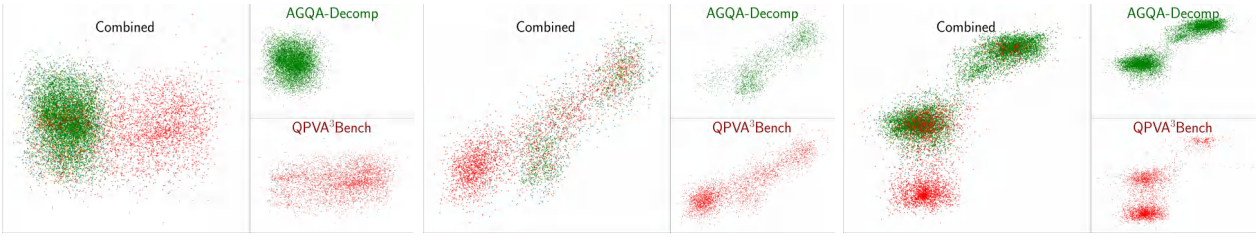
2) Comparison with AGQA-Decomp: As QPVA^{3}Bench is an evaluation-only benchmark, we compare it directly with the test set of AGQA-Decomp. The AGQA-Decomp test set is derived from only 1,814 unique videos and 168k questions, with multiple questions generated for each video via templates based on automatic scene graphs and question programs. This results in high correlation and poor independence among questions for a single video. In contrast, every compositional graph in QPVA^{3}Bench is generated for a unique VQ-pair and verified through a human-in-the-loop process, ensuring higher accuracy and diversity. This template-based generation in AGQA-Decomp also leads to a high degree of repetition among sub-questions and a highly skewed question distribution, where

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3650864



(a) Question Embedding (b) Video Embedding (c) Question-Video Embedding

Fig. 4: t-SNE visualizations of features between QPVA^3 and the union of five source datasets. Video, question, and video-question features are extracted by Qwen-2.5-VL [46]. **Best viewed when zoomed in.**



(a) Question Embedding (b) Video Embedding (c) Question-Video Embedding

Fig. 5: t-SNE visualizations of features between QPVA^3 and AGQA-Decomp. Video, question, and video-question features are extracted by Qwen-2.5-VL [46]. **Best viewed when zoomed in.**

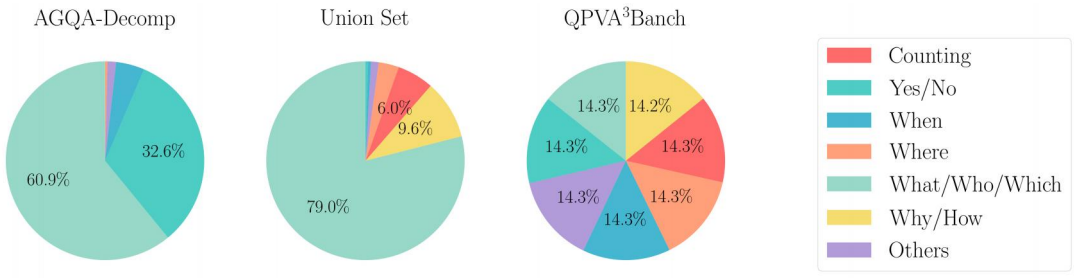


Fig. 6: Question type distribution among the union of five source datasets, QPVA^{3}Bench and AGQA-Decomp.

`What/Which` and **`Is/Are`** types dominate, and all other categories collectively account for less than 5% of the data, as depicted in Fig. 6. The t-SNE visualizations in Fig. 5 further highlight this difference, showing that QPVA^{3}Bench's data points are more dispersed while still covering the semantic range of AGQA-Decomp. To investigate potential data redundancy in AGQA-Decomp, we randomly sampled 10% and 5% of its 168k test questions. We evaluated VideoLLaMA3 and our enhanced VideoLLaMA3+QPVA^3 on these subsets. On the 10% subset, the models achieved Accuracy and cF1 scores of 40.9% vs. 46.7% and 46.8% vs. 52.5%, respectively. On the 1% subset, the scores were 41.6% vs. 47.5% and 47.1% vs. 52.9%. The minimal change in both the absolute scores and the performance gap suggests that a much smaller subset of AGQA-Decomp is sufficient for reliable evaluation, highlighting significant data redundancy.

V. Metrics

A. Limitations of Existing Metrics

Compositional consistency measures whether a method can provide the correct answer for the right reason. To formalize this, we first define the four possible outcomes for a question and its sub-questions. Let M be a model's prediction on a given sample. The status must be one of the following:

- **Positive Consistent** (N_+^+): The model answers both the question and all its sub-questions correctly.
- **Negative Inconsistent** (N_+^-): The model answers all sub-questions correctly, but fails on the question.
- **Positive Inconsistent** (N_-^+): The model answers the question correctly, but fails on at least one sub-question.
- **Negative Consistent** (N_-^-): The model answers the question incorrectly, and fails on at least one sub-question.

Therefore, existing metrics like Compositional Accuracy (CA) and Right-for-Wrong-Reason (RWR) are formulated as:

$$\text{CA} = \frac{N_+^+}{N_+^+ + N_+^-}; \quad \text{RWR} = \frac{N_-^+}{N_-^+ + N_-^-}, \tag{9}$$

and their difference, Delta $=$ CA $-$ RWR. However, these metrics suffer from significant theoretical and practical issues, leading to asymmetric and unstable evaluations.

The practical issues are illustrated in Tab. II. For rows 1 to 4, the underlying compositional consistency of the models

TABLE II: The counterexamples. Acc. is parent question accuracy.

	Data Count				Existing Metrics				Our Metrics		
	N_+^+	N_+^-	N_-^+	N_-^-	CA	RWR	Delta	Acc.	cP	cR	c-F_1
1	100	100	0	10	50.00	0.00	50.00	47.61	50.00	100.00	66.67
2	10	100	0	100	9.09	0.00	9.09	4.76	9.09	100.00	16.67
3	100	0	100	10	100.00	90.91	9.09	95.23	100.00	50.00	66.67
4	10	0	100	100	100.00	50.00	50.00	52.38	100.00	9.09	16.67
5	99	100	1	0	49.75	100.00	-50.25	50.00	49.75	99.00	66.22
6	100	99	0	1	50.25	0.00	50.25	50.00	50.25	100.00	66.88
7	99	99	1	1	50.00	50.00	0.00	50.00	50.00	99.00	66.44

should be similar, as 110 out of 210 questions are answered with correct reasoning. However, CA, RWR, and Delta vary significantly without clear semantic meaning. This suggests an asymmetric treatment of inconsistent cases (N_+^- and N_-^+). Furthermore, the metrics exhibit instability when facing imbalanced data distributions (rows 5 to 7), where models with similar consistency profiles receive wildly different scores. These failures stem from theoretical flaws.

The first flaw is asymmetry. The two types of inconsistency, N_+^- (falsely answered main question with all sub-questions correct) and N_-^+ (correctly answered main-question with any sub-question incorrect), are not treated interchangeably, leading to arbitrary evaluation changes.

Theorem V.1 (Asymmetry of the CA-RWR-Delta System)**.** *Swapping the values of N_+^- and N_-^+ changes the values of CA and RWR, unless $N_+^- = N_-^+$. Consequently, Delta also changes unless $N_+^- = N_-^+$ or $N_+^+ = N_-^-$.*

Proof. Let the metrics be $\text{CA}_1 = \frac{N_+^+}{N_+^+ + N_+^-}$ and $\text{RWR}_1 = \frac{N_-^+}{N_-^+ + N_-^-}$. After swapping N_+^- and N_-^+, the new metrics are $\text{CA}_2 = \frac{N_+^+}{N_+^+ + N_-^+}$ and $\text{RWR}_2 = \frac{N_+^-}{N_+^- + N_-^-}$. The changes are:

$$\Delta\text{CA} = \text{CA}_1 - \text{CA}_2 = \frac{N_+^+(N_-^+ - N_+^-)}{(N_+^+ + N_+^-)(N_+^+ + N_-^+)}, \quad (10)$$

$$\Delta\text{RWR} = \text{RWR}_1 - \text{RWR}_2 = \frac{N_-^-(N_-^+ - N_+^-)}{(N_-^+ + N_-^-)(N_+^- + N_-^-)}. \quad (11)$$

Both ΔCA and ΔRWR are non-zero unless $N_+^- = N_-^+$. Besides, $\Delta\text{Delta} = \Delta\text{CA} - \Delta\text{RWR}$ is also non-zero under the same condition, indicating the system is asymmetric. □

The second flaw is instability, where the metrics can change arbitrarily under minor perturbations, especially in imbalanced scenarios. CA and RWR are relatively independent, meaning one can be changed drastically without affecting the other.

Theorem V.2 (Instability of the CA-RWR-Delta System)**.** *Given a fixed total count $N = N_+^+ + N_+^- + N_-^+ + N_-^-$, RWR can change arbitrarily within [0, 1], while CA is held constant, and vice versa. This instability is amplified when the denominator of the changing metric is small.*

Proof. Let's fix the values of N_+^+ and N_+^-, thereby fixing CA. The remaining sum is $S = N_-^+ + N_-^-$. The RWR is given by $\text{RWR} = \frac{N_-^+}{S}$. By reallocating a small count δ between N_-^+ and N_-^- (*e.g.*, $N_-^{+\prime} = N_-^+ + \delta$, $N_-^{-\prime} = N_-^- - \delta$), the new RWR becomes $\text{RWR}' = \frac{N_-^+ + \delta}{S}$. The change is $\Delta\text{RWR} = \frac{\delta}{S}$.

If S is small, even a tiny perturbation δ (*e.g.*, $\delta = 1$) can cause a large swing in RWR. As $S \to 0$, $|\Delta\text{RWR}| \to \infty$. **This means a model's RWR score can fluctuate from 0 to 1 based on a single sample's outcome, rendering the metric and consequently Delta highly unstable.** A symmetric argument holds for CA when $N_+^+ + N_+^-$ is small. □

B. Proposed Symmetric and Stable Metrics

To address the issues of asymmetry and instability, we propose a new set of metrics based on the concepts of precision and recall. We redefine the problem as measuring how well a model's correct answers are supported by correct sub-question answers. We introduce **consistency precision (cP)** and **consistency recall (cR)**.

Definition V.3 (Consistency Precision (cP) and Recall (cP))**.**

$$\text{cP} = \frac{N_+^+}{N_+^+ + N_+^-}, \quad \text{cR} = \frac{N_+^+}{N_+^+ + N_-^+}. \quad (12)$$

Here, cP measures the precision of the reasoning chain (what fraction of correct sub-answers lead to a correct answer?), while cR measures the recall of correct reasoning (what fraction of correct answers are derived from correct sub-answers?). To provide a single, balanced score, we use their harmonic mean, the consistency F-score.

Definition V.4 (Consistency F-Score (c-F_β))**.**

$$\text{c-F}_\beta = \frac{(1+\beta^2)\text{cP}\cdot\text{cR}}{\beta^2\text{cP}+\text{cR}}. \quad (13)$$

For balanced evaluation, we set $\beta = 1$, yielding the c-F_1 score.

This new system resolves the asymmetry issue because it treats both inconsistencies (N_+^- and N_-^+) symmetrically.

Proposition V.5 (Symmetry of the cP-cR-c-F_1 System)**.** *The c-F_1 score is invariant to swapping the values of N_+^- and N_-^+.*

Proof. The c-F_1 score is the harmonic mean of cP and cR:

$$\begin{aligned}\text{c-F}_1 &= \frac{2\cdot\text{cP}\cdot\text{cR}}{\text{cP}+\text{cR}} \quad (14)\\ &= \frac{2\cdot\frac{N_+^+}{N_+^+ + N_+^-}\cdot\frac{N_+^+}{N_+^+ + N_-^+}}{\frac{N_+^+}{N_+^+ + N_+^-} + \frac{N_+^+}{N_+^+ + N_-^+}} = \frac{\frac{2(N_+^+)^2}{(N_+^+ + N_+^-)(N_+^+ + N_-^+)}}{\frac{N_+^+(N_+^+ + N_-^+) + N_+^+(N_+^+ + N_+^-)}{(N_+^+ + N_+^-)(N_+^+ + N_-^+)}}\\ &= \frac{2(N_+^+)^2}{N_+^+(N_+^+ + N_-^+ + N_+^+ + N_+^-)} = \frac{2N_+^+}{2N_+^+ + N_+^- + N_-^+}.\end{aligned}$$

The final expression depends on the sum $N_+^- + N_-^+$. Swapping the values of N_+^- and N_-^+ does not change this sum, and therefore does not change the value of c-F_1. □

Furthermore, the c-F_1 metric is robust against the instability issues that plague the CA-RWR-Delta system.

Proposition V.6 (Stability of the c-F_1 System)**.** *Given a fixed total N, small perturbations in the distribution of samples lead to small, bounded changes in c-F_1.*

TABLE III: Performance comparison on QPVA^{3}Bench regarding the accuracy, score, and compositional consistency.

Method	#Param	#Frames	Accuracy		Score		Compositional Consistency		
			main	sub	main	sub	cP	cR	cF$_1$
Video-LLaVA [1]	7B	8	37.8	63.1	2.3	3.3	56.6	31.6	40.5
LLaMA-VID [47]	7B	16	41.7	68.0	2.6	3.5	58.7	38.9	46.8
Chat-UniVi [48]	7B	16	44.7	68.8	2.7	3.6	61.7	39.7	48.3
VideoChat2 [9]	7B	16	45.0	65.2	2.7	3.4	62.7	37.1	46.6
VideoLLaMA2 [2]	7B	16	40.8	66.7	2.4	3.4	59.4	36.5	45.2
VideoLLaMA3 [49]	7B	16	41.2	68.0	2.4	3.5	60.7	37.3	46.2
+ VA^3	-	16	43.0	72.1	2.5	2.6	64.0	39.1	48.5
+ QPVA^3	-	16	47.0	76.1	2.9	3.8	67.2	42.5	52.1
LLaVA-OneVision [50]	7B	16	45.8	70.3	2.7	3.7	63.1	41.6	50.1
+ VA^3	-	16	47.1	73.5	2.8	3.8	66.3	44.0	52.9
+ QPVA^3	-	16	**49.5**	**77.3**	**3.0**	3.9	**69.6**	**46.4**	**55.7**
Qwen2-VL [46]	7B	16	41.9	70.2	2.6	3.7	59.2	40.2	47.9
+ VA^3	-	16	43.8	74.0	2.8	3.9	62.2	43.7	51.3
+ QPVA^3	-	16	46.5	76.7	2.9	**4.0**	64.5	46.0	53.7

Proof. From the previous proposition, we have c-F$_1$ $= \frac{2N_+^+}{2N_+^+ + N_+^- + N_-^+}$. Let $D = 2N_+^+ + N_+^- + N_-^+$ be the denominator. Consider a perturbation δ that transfers mass from an inconsistency count (*e.g.*, N_+^-) to the positive consistent count N_+^+. The new values are $N_+^{+\prime} = N_+^+ + \delta$ and $N_+^{-\prime} = N_+^- - \delta$. The new score is:

$$\text{c-F}_1{}' = \frac{2(N_+^+ + \delta)}{2(N_+^+ + \delta) + (N_+^- - \delta) + N_-^+} = \frac{2(N_+^+ + \delta)}{D + \delta}. \tag{15}$$

The change in the score is:

$$\begin{aligned} \Delta\text{c-F}_1 &= \frac{2(N_+^+ + \delta)}{D + \delta} - \frac{2N_+^+}{D} \\ &= \frac{2D(N_+^+ + \delta) - 2N_+^+(D + \delta)}{D(D + \delta)} \\ &= \frac{2\delta(D - N_+^+)}{D(D + \delta)} = \frac{2\delta(N_+^+ + N_+^- + N_-^+)}{D(D + \delta)}. \end{aligned} \tag{16}$$

Since $D \geq 0$ and we assume not all counts are zero, the denominator $D(D + \delta)$ is always positive and grows with the total number of samples involved. Unlike the CA-RWR system, the denominator here is not susceptible to becoming arbitrarily small in a way that would amplify the perturbation δ. The change $|\Delta\text{c-F}_1|$ is bounded and diminishes as the total count increases, ensuring stability. □

VI. Experiments

A. Evaluation Protocols

We evaluate the performance of our QPVA^3 framework on 6 VideoQA benchmarks, including our QPVA^{3}Bench, AGQA-Decomp [11], ActivityNet-QA [51], Causal-VidQA [5], STAR [6], and MVBench [9]. Among these benchmarks, Causal-VidQA [5] and MVBench [9] are evaluated in multiple-choice format by the prediction accuracy. Besides, QPVA^{3}Bench, AGQA-Decomp [11], ActivityNet-QA [51], and STAR [6] are evaluated in open-ended format by the accuracy (the percentage of correctly answered questions) and the average score (where ChatGPT rates each response on a scale of 1 to 5 and calculates the mean of these scores). To ensure consistent comparisons, all open-ended evaluations utilize the `GPT-3.5-turbo`. Moreover, to illustrate the improvement in compositional consistency, we also evaluate the cR, cP, and c-F$_1$ in our QPVA^{3}Bench and AGQA-Decomp.

B. Implementation Details

In our QPVA^3 framework, the LLM and the MLLM are used to handle logical and visual content, respectively. Therefore, we employ LLaMA3.2-8B [30] as the LLM, while VideoLLaMA3 [49], LLaVA-OneVision [50], and Qwen2-VL [46] as alternative choices for the MLLM. We sampled 61-K compositional graphs from the AGQA-Decomp dataset to train the planner. We employ $\mathcal{L}_p$ (Subsec. III-D) to optimize the MLLM for these samples. For executor, we train the video aligner on both main questions and sub-questions in the aforementioned compositional graphs with L_c and L_a, which consists 581-K samples. The aforementioned data splits are mixed, learning rate is set to 2×10^{-5} and trained for two epochs. The video resolution is set as 336 $\times$ 336, and we uniformly sample 16 frames for each video. The maximum token length is set to 4096, and we use AdamW [57] as the optimizer for all the training. Moreover, the training is conducted on 8 NVIDIA-A800 (80G) GPUs. The training takes approximately 14 hours.

C. Results on Compositional Graph Constrained Benchmark

We evaluate compositional consistency and accuracy using only our executor and reasoner modules, guided by the compositional graphs from QPVA^{3}Bench and AGQA-Decomp. Further comparisons and analyses of the compositional graph generated are presented in Subsec. VI-F.

1) Comparison on QPVA^{3}Bench: Tab. III compares our framework against 8 baseline methods on QPVA^{3}Bench. Results show our framework outperforms SOTA baselines on accuracy and score for both main and sub-questions. Moreover, sub-questions show a notably larger accuracy gain of 4.3–5.4% across different MLLMs. Our collaborative reasoning strategy also boosts compositional consistency, improving the cF$_1$ score by over 5.6% compared to baselines.

2) Comparison on AGQA-Decomp: Tab. IV shows performance comparisons against 9 baselines on the AGQA-Decomp test set. As with QPVA^{3}Bench, our framework significantly improves accuracy, score, and compositional consistency over SOTA baselines. The improvement on sub-questions is again more pronounced than on main questions. The large gain in

TABLE IV: Performance comparison on AGQA-Decomp regarding the accuracy, score, and compositional consistency. Results from baseline MLLMs are reproduced by their released version.

Method	#Param	#Frames	Accuracy		Score		Compositional Consistency		
			main	sub	main	sub	cP	cR	cF_1
LLaMA-VID [47]	7B	16	58.2	63.2	2.8	3.2	56.1	58.8	57.4
Chat-UniVi [48]	7B	16	60.3	68.1	3.0	3.4	63.6	69.9	66.6
MiniGPT4-Video [52]	7B	90	54.5	61.8	2.4	2.1	55.6	52.8	54.1
LLaVA-NeXT-Video [32]	7B	16	60.9	68.6	3.0	3.4	64.9	65.9	65.4
VideoChat2 [9]	7B	16	60.5	69.2	3.1	3.5	67.5	66.5	67.0
VideoLLaMA2 [2]	7B	16	60.3	64.0	3.0	3.3	65.6	58.1	61.6
VideoLLaMA3 [49]	7B	16	69.2	76.7	3.7	4.1	72.2	69.7	70.9
+ VA^3	-	16	72.0	80.3	3.9	4.3	75.8	76.1	75.9
+ $QPVA^3$	-	16	**74.6**	**84.1**	**4.1**	**4.5**	**81.0**	**82.6**	**81.8**
LLaVA-OneVision [50]	7B	16	64.3	73.2	3.5	3.9	69.1	66.9	68.0
+ VA^3	-	16	66.1	78.7	3.7	4.2	72.5	77.1	74.7
+ $QPVA^3$	-	16	68.6	81.3	3.9	4.4	78.5	81.3	79.9
Qwen2-VL [46]	7B	16	65.4	73.8	3.5	4.0	68.2	66.2	67.2
+ VA^3	-	16	67.8	79.1	3.6	4.3	73.2	75.0	74.1
+ $QPVA^3$	-	16	70.1	82.1	3.8	**4.5**	79.1	82.0	80.5

TABLE V: Performance comparison on STAR, ActivityNet-QA and Causal-VidQA.

Method	#Param	#Frames	ActivityNet-QA		STAR		Causal-VidQA				
			Accuracy	Score	Accuracy	Score	Acc@D	Acc@E	Acc@P	Acc@C	Acc@All
LLaMA-VID [47]	7B	16	47.5	3.3	30.9	2.5	-	-	-	-	-
LLaMA-Adapter [53]	7B	16	34.2	2.7	-	-	-	-	-	-	-
Chat-UniVi [48]	7B	16	45.8	3.2	30.6	2.5	-	-	-	-	-
Video-LaVIT [54]	7B	24	50.1	3.3	-	-	69.2	71.0	44.4	45.0	57.4
MiniGPT4-Video [52]	7B	90	45.9	3.4	-	-	-	-	-	-	-
VideoChat2 [9]	7B	16	57.2	3.5	38.0	2.7	66.8	75.1	45.8	38.6	56.6
STEP [55]	7B	16	56.0	3.5	39.8	2.8	-	-	-	-	-
VoT [12]	7B	16	-	-	-	-	81.2	83.0	54.7	**58.6**	69.4
LLaVA-NeXT-Video [32]	7B	16	35.4	2.8	-	-	68.9	70.4	38.7	38.7	54.2
VideoLLaMA2 [2]	7B	16	50.4	3.3	35.1	2.5	-	-	-	-	-
VideoLLaMA3 [49]	7B	16	60.6	3.9	38.8	2.6	79.1	79.9	55.9	43.8	64.6
+ VA^3	-	16	62.4	4.0	40.6	2.7	79.6	80.4	57.6	45.2	65.4
+ $QPVA^3$	-	16	**64.3**	**4.1**	43.4	**3.0**	80.4	81.1	59.1	48.2	67.2
LLaVA-OneVision [50]	7B	16	56.0	3.5	36.3	2.6	78.6	78.0	53.1	44.2	63.5
+ VA^3	-	16	57.7	3.6	38.4	2.7	78.9	79.1	55.0	45.8	64.9
+ $QPVA^3$	-	16	59.2	3.7	40.7	2.8	79.6	80.5	57.7	48.3	66.5
Qwen2-VL [46]	7B	16	56.6	3.6	38.4	2.7	80.3	81.5	59.8	50.3	68.0
+ VA^3	-	16	58.1	3.7	40.4	2.9	80.9	82.1	62.1	52.9	68.8
+ $QPVA^3$	-	16	60.0	3.9	**43.5**	**3.0**	**81.8**	**83.0**	**64.9**	55.0	**71.2**

compositional consistency highlights the effectiveness of our reasoner at improving logical coherence.

3) Performance Analysis: Our framework treats VideoQA as a compositional task, using video-aligned execution and language-guided reasoning to improve MLLM consistency and accuracy. This improved compositional reasoning boosts main question accuracy by 4–5% across baselines and benchmarks. Sub-questions show an even greater improvement. This is because sub-questions are simpler and require less specific video information to answer. Thus, sub-questions benefit more from the precise information of video aligner, while main questions gain less from its filtering of irrelevant content. The difficulty of synthesizing main answers from sub-answers also contributes to this performance gap. Moreover, the compositional consistency on the $QPVA^3$Bench and AGQA-Decomp also improves across all SOTA baseline models. The significant improvements in cR, cP, and cF_1 across all baselines stem from our reasoner, which creates an information flow from lower-level sub-answers to resolve higher-level questions. This information flow enhances the coherency between main questions and sub-questions during reasoning along the compositional graph, thereby improving its compositional consistency.

D. Results on Compositional Graph Free Benchmarks

In Tabs. V and VI, we present performance comparisons among 10 to 16 baseline methods on various multiple-choice and open-ended benchmarks, including ActivityNet-QA, STAR, MVBench, and Causal-VidQA.

1) Comparison on Multi-Choice Benchmark: In Tabs. V and VI, we report results on the multiple-choice benchmarks, specifically Causal-VidQA and MVBench. Causal-VidQA focuses on commonsense reasoning and causal relations in video, while MVBench is a comprehensive benchmark designed to evaluate the temporal reasoning capabilities of MLLMs.

On Causal-VidQA, our $QPVA^3$ framework improves performance across all question types: descriptive, explanatory, predictive, and counterfactual. Notably, the performance gain on causal-related questions (*i.e.*, explanatory, predictive, and counterfactual) is substantially higher than on descriptive ones. This discrepancy stems from our framework's design, which is tailored to enhance the multimodal reasoning capabilities of VideoLLMs. These results confirm our approach is particularly effective at modeling causal dependencies in video-based question answering.

On MVBench, our $QPVA^3$ framework improves the accu-

TABLE VI: Experimental results on MVBench. The results in the white area are copied from the corresponding works or MVBench [9], and the results in the blue area are reproduced by us using their published model weights and instructions.

Method	Avg.	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI
LLaMA-VID [47]	41.3	45.5	40.5	58.0	39.5	55.0	53.5	40.0	35.5	18.5	27.5	87.0	41.5	23.0	45.5	41.0	27.0	40.0	34.5	41.5	31.5
LLaMA-Adapter [53]	31.7	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	25.0	31.5	22.5	28.0	32.0
VideoChat [56]	35.5	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0
VideoLLaMA [1]	34.1	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0
VideoChat2 [9]	60.3	66.5	74.0	85.5	51.0	61.5	85.5	68.0	43.5	48.5	35.5	83.5	38.5	66.5	88.0	50.5	**63.5**	46.5	36.0	42.5	70.5
VideoLLaMA3 [49]	67.1	70.5	72.5	**91.5**	43.5	**85.5**	92.5	74.5	42.0	51.5	44.5	92.5	53.0	75.0	92.0	59.0	61.5	76.5	33.5	54.0	75.5
+ QPVA^3	**69.3**	73.5	77.0	**91.5**	46.0	84.5	92.0	**80.0**	**44.5**	**53.0**	44.5	94.0	**54.5**	78.0	91.5	**63.0**	61.0	**78.0**	35.5	**61.5**	**82.0**
LLaVA-OneVision [50]	57.3	73.0	69.5	79.0	46.0	79.5	63.5	76.5	37.5	21.5	40.0	92.0	47.0	46.0	68.5	52.0	55.0	63.5	34.5	51.5	50.0
+ QPVA^3	60.4	79.0	76.0	80.5	46.0	81.0	62.0	82.5	39.0	24.5	42.0	91.0	48.0	48.5	70.5	59.0	56.5	64.5	**37.5**	61.0	59.0
Qwen2-VL [46]	65.7	77.0	80.5	81.5	49.5	75.0	**93.5**	72.5	39.0	47.0	**47.5**	93.5	46.0	81.5	94.5	46.5	58.5	68.0	41.5	55.0	66.5
+ QPVA^3	68.1	**81.5**	**87.0**	82.0	**50.0**	75.0	93.0	78.5	41.0	46.0	**47.5**	**95.0**	48.0	**84.0**	**95.0**	51.0	59.5	68.5	44.0	61.0	74.5

racy of SOTA MLLMs by 2.2% to 3.1%. Accuracy gains are more pronounced on questions requiring complex reasoning compared to basic perception tasks. For instance, the improvement on Counterfactual Inference (CI) is significantly larger than on Fine-grained Pose (FP) questions. This result indicates that our answer aggregation mechanism is key to improving complex reasoning capabilities. Moreover, gains on MVBench and Causal-VidQA show our planner constructs high-quality compositional graphs that improve the reasoning process.

2) Comparison on Open-ended Benchmark: In Tab. V, we present comprehensive comparisons on two additional open-ended datasets, offering further insights into the versatility of our framework. Following the protocol in [55], and owing to the unavailability of the STAR test set answers, we convert the STAR validation set into an open-ended format. This transformation not only broadens our evaluation to more diverse compositional reasoning tasks but also serves as a rigorous test bed for assessing the adaptability of our framework to less constrained answer formats.

On open-ended benchmarks, our QPVA^3 framework significantly improves both accuracy and score. Performance gains are again most pronounced on tasks requiring complex reasoning, where our approach excels at aggregating evidence from sub-questions to capture intricate semantic and temporal relations. This is indicative of the framework's capacity to construct high-quality compositional graphs that enhance the overall reasoning process during answer aggregation.

Consistent improvements across both open-ended and multiple-choice settings demonstrate the robustness and generalizability of our QPVA^3 framework. These findings confirm our framework improves the video reasoning of MLLMs, highlighting its potential for a broad range of VideoQA tasks. Overall, results on various benchmarks confirm our QPVA^3 framework outperforms SOTA baselines in diverse scenarios, creating a more transparent and verifiable VideoQA system.

E. Comparison between VA^3 and QPVA^3

To demonstrate the advancements of QPVA^3 framework over the VA^3 method, we incorporate a direct comparison between them in the experiments, with results detailed in Tabs. III to V. To facilitate a fair comparison across diverse benchmarks, we adapt VA^3 by training its GCN-based answer aggregator on the AGQA-Decomp and the union of five source datasets. However, the limited scope of the training data prevents us from comparing VA^3 against QPVA^3 on evaluation-only open-ended datasets.

TABLE VII: Ablation study on planner. PCG is the planner-generated compositional graph, and DCG is the dataset provided compositional graph.

Method	QPVA^{3}Bench		AGQA	
	Accuracy	Score	Accuracy	Score
VideoLLaMA3 [49]	41.2	2.4	69.2	3.7
+ QPVA^3 (+ DCG)	47.0	2.9	74.6	4.1
+ QPVA^3 (+ PCG)	46.6	2.7	74.4	4.1
LLaVA-OneVision [50]	45.8	2.7	64.3	3.5
+ QPVA^3 (+ DCG)	49.5	3.0	68.6	3.9
+ QPVA^3 (+ PCG)	49.3	3.0	68.1	3.8
Qwen2-VL [46]	41.9	2.6	65.4	3.5
+ QPVA^3 (+ DCG)	46.5	2.9	70.1	3.8
+ QPVA^3 (+ PCG)	46.1	2.9	69.8	3.8

TABLE VIII: Ablation study on the planner, comparing against the external parser (E.P.) from VA^3 [15]. Results are presented as mean±std. The improvements of QPVA^3 compared to the external parser are significant as $p < 0.001$.

Method	QPVA^{3}Bench		CausualQA
	Accuracy	Score	Acc@All
VideoLLaMA3 [49]	41.2 ± 1.2	2.4 ± 0.1	64.6 ± 1.6
+ QPVA^3	47.0 ± 1.5	2.7 ± 0.1	67.2 ± 1.8
+ QPVA^3 (E.P.)	44.5 ± 1.3	2.5 ± 0.1	65.8 ± 1.4
LLaVA-OneVision [50]	45.8 ± 1.2	2.6 ± 0.1	63.5 ± 1.1
+ QPVA^3	49.5 ± 1.6	3.0 ± 0.2	66.5 ± 1.5
+ QPVA^3 (E.P.)	47.2 ± 1.1	2.8 ± 0.2	65.1 ± 1.2
Qwen2-VL [46]	41.9 ± 1.3	2.6 ± 0.1	68.0 ± 1.5
+ QPVA^3	46.5 ± 1.8	2.9 ± 0.2	71.2 ± 1.9
+ QPVA^3 (E.P.)	44.9 ± 1.5	2.7 ± 0.1	69.6 ± 1.4

Our results show that QPVA^3 significantly outperforms VA^3. On the QPVA^{3}Bench, our method achieves an accuracy improvement between 2.4% to 4.0% over VA^3. Similarly, on AGQA-Decomp and Causal-VidQA, QPVA^3 surpasses VA^3 with improvement from 2.4% to 2.6% and from 1.6% to 2.4%, respectively. This substantial improvement is driven by two key innovations. First, we replace the external parser with a finetuned MLLM-based planner, which enhances multi-modal perception and generation quality. Second, we substitute the GCN-based aggregator with a more powerful two-step MLLM-based reasoner, leading to superior reasoning capabilities and generalization performance.

F. Ablation Study

In this section, we study the effect of the planner, the executor, the reasoner, and the number of frames using the main question of QPVA^{3}Bench and Causal-Video.

1) Effect of the Planner: The planner aims to parse the complex question into a compositional graph, thereby capturing its underlying logical structure and segmenting it into manageable sub-questions. To evaluate the effect of the planner, we conduct experiments on QPVA^{3}Bench and AGQA using both the planner-generated compositional graph (PCG) and the dataset-provided compositional graph (DCG), and present the experimental results in Tab. VII. This experiment indicates that substituting the dataset-provided compositional graph with the planner-generated one on QPVA^{3}Bench and AGQA results in similar performance in terms of both accuracy and score (i.e., differences of $\leq$0.4% in accuracy and $\leq$0.1% in score). This indicates that our planner can generate compositional graphs that are comparable to the human-annotated ones.

To address the errors introduced by the external parser (E.P.), we conduct a systematic analysis. The E.P. suffers from two primary issues. First, it generates compositional graphs based on question similarity alone, which may be irrelevant to the specific video content. Second, since the E.P. is not fine-tuned on compositional graph generation, it often produces overly complex leaf-node sub-questions. To quantify these issues, we task three graduate students with annotating all 3,492 questions in the QPVA^{3}Bench. Our analysis reveals that the E.P. generates 472 video-irrelevant graphs (13.5%) and 262 overly complex sub-questions graph (7.5%). In stark contrast, our MLLM-planner significantly reduces these errors, producing only 27 (0.8%) and 32 (0.9%) of each error type, respectively. This highlights the clear advantage of our planner. Furthermore, as shown in Tab. VIII, we compare the performance of the E.P. and our planner within the QPVA^3 framework, with each result averaged over ten runs.

2) Effect of the Executor: The executor aims to select relevant video clips and generates answers for each question in the compositional graph. To evaluate the effect of the executor, we conduct experiments on two settings: 1) removing the video aligner from the executor of QPVA^3 framework; and 2) replacing the video aligner with a pretrained video grounding method, *i.e.*, CG-STVG [58]. In Tab. IX, we present the experimental results for these two settings, as shown in the "- Video Aligner" and "+ QPVA^3 (R.A.)" rows. For the first setting, we observe that a significant performance drop occurs when removing the video aligner from our framework. This is because the video aligner can significantly reduce noisy information in simple perceptual questions, which serves as a solid basis for further accurate reasoning and thereby improves the reasoning performance for complex questions. For the second setting, we notice that by replacing our video aligner with the pretrained CG-STVG [58], the performance drops sharply, indicating that the video grounding method pretrained on descriptive text cannot generalize to questions where the information provided is not comprehensive. Moreover, this large performance drop (even worse than removing the video aligner) also reflects that the pretrained CG-STVG [58] tends to filter out useful frames, further reducing the performance.

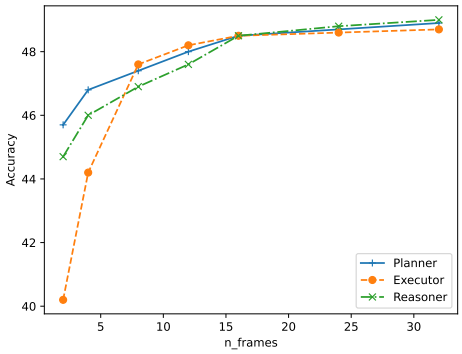


Fig. 7: Accuracy of LLaVA-OneVision on QPVA^{3}Bench at different number of frames for planner, executor and reasoner.

TABLE IX: Ablation study on executor and reasoner. R.A. means replace the video aligner with CG-STVG [58].

Method	QPVA^{3}Bench		CausualQA
	Accuracy	Score	Acc@All
VideoLLaMA3 [49]	41.2	2.4	64.6
+ QPVA^3	47.0	2.7	67.2
+ QPVA^3 (R.A.)	42.1	2.5	64.8
- Video Aligner	45.7	2.5	66.0
- Reasoner	42.5	2.7	65.5
LLaVA-OneVision [50]	45.8	2.6	63.5
+ QPVA^3	49.5	3.0	66.5
+ QPVA^3 (R.A.)	46.5	2.6	63.8
- Video Aligner	48.0	2.8	65.7
- Reasoner	45.9	2.7	64.8
Qwen2-VL [46]	41.9	2.6	68.0
+ QPVA^3	46.5	2.9	71.2
+ QPVA^3 (R.A.)	43.8	2.5	69.2
- Video Aligner	45.2	2.8	70.1
- Reasoner	43.7	2.5	69.4

3) Effect of the Reasoner: The reasoner aims to aggregate individual answers into a coherent, comprehensive response. To evaluate the effect of the reasoner, we conduct experiments without using the reasoner, and provide the experimental results in in Tab. IX "-Reasoner" rows. From the results, we can find that the reasoner achieves the largest improvements in our QPVA^3 framework. This indicates that the reasoner plays a vital part in improving the video understanding capability, as it is the main module that provides intra-question information aggregation and performs context-based logical reasoning.

4) Effect of the Number of Frames: To further investigate how the number of sampled frames influences each module in our QPVA^3 framework, we conduct experiments on QPVA^{3}Bench by fixing the frame count for two modules at 16 while varying it for the remaining module from 2 to 32. The results are presented in Fig. 7.

For the **planner**, we observe that the planner is relatively insensitive to the number of frames. When the number of frames increases from 2 to 16, the accuracy only improves by about 2%. This result suggests that the planner relies on static information from the video to form the compositional graph, and requires fewer frames for effective question parsing.

For the **executor**, the number of frames has a significant impact, particularly at lower frame counts. Specifically, when

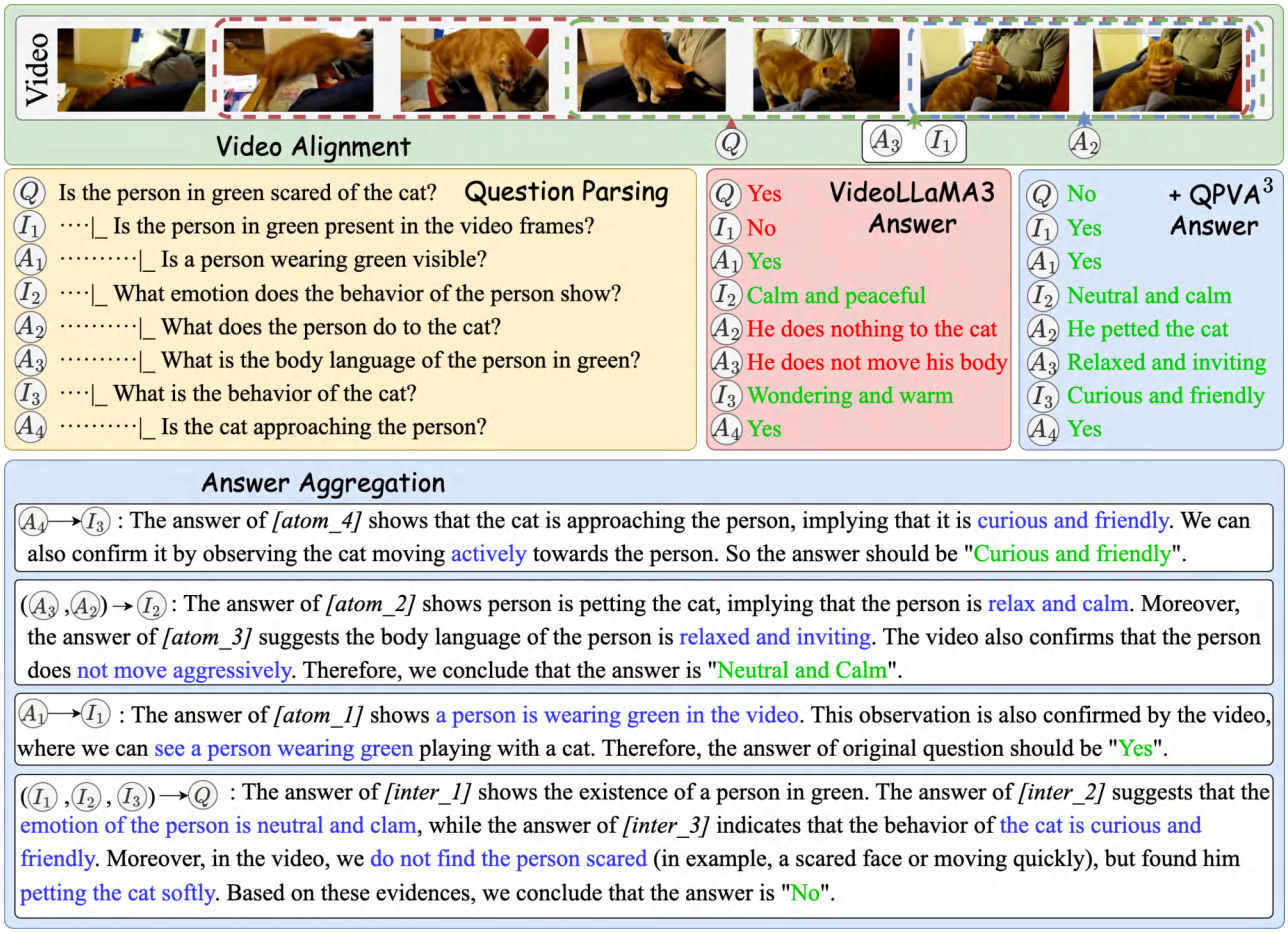


Fig. 8: Qualitative example showcasing how QPVA^3 framework achieves successful reasoning. Best viewed when zoomed in.

the frame count increases from 2 to 8, the accuracy rises by roughly 8%. However, from 8 to 20 frames, the improvement is limited to within 1%. These findings indicate that incorporating dynamic information is crucial for the performance of the executor, but once the model receives enough frames to capture the essential motion cues (around 8–16 frames), further increases in the frame count yield diminishing returns.

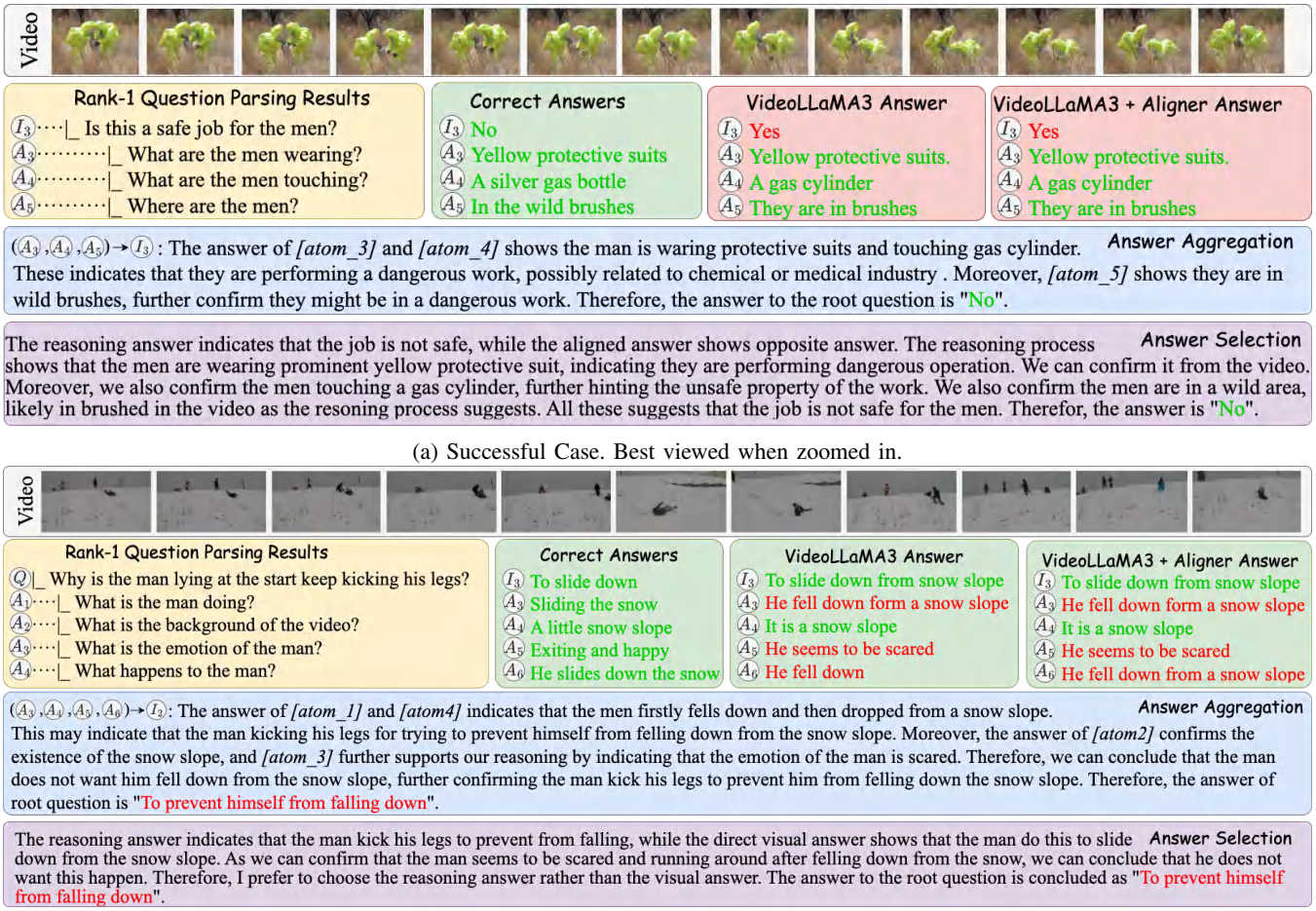
For the **reasoner**, we observe a steady performance gain as the number of frames increases. From 2 to 16 frames, the accuracy improves by about 4%. Since the reasoner is designed to aggregate the outputs of the executor and handle logical conflicts in compositional graph, having more frames provides additional temporal cues for accurate reasoning.

G. Qualitative Results

In Fig. 8, we illustrate the executing and reasoning process of our QPVA^3 framework via a question drawn from QPVA^{3}Bench, where our framework successfully corrects errors in both the main and sub-questions. As illustrated, VideoLLaMA3 fails to recognize that the individual in green is sitting and misidentifies his interaction with the cat (*i.e.*, petting the cat). However, our executor selects the most relevant video segments to address these issues; for instance, sub-question A_2 is associated with the end of the video, where the interaction occurs—thus eliminating extraneous content and enabling accurate identification of the petting action. Moreover, sub-question I_1 is inferred logically: once a sub-question confirms the presence of the person in green, the reasoner deduces the correct answer "Yes" by verifying that the person is sitting. Although the baseline model erroneously inferred that the person was frightened of the cat in the main question, the alignment module offered limited assistance here because a large portion of the video was deemed relevant. Nonetheless, by integrating information from I_1, I_2, and I_3, the reasoning process reveals that the person is seated in a neutral and calm manner while the cat appears friendly, thereby strongly indicating that the person is not scared of the cat. This conclusion is further corroborated by visual evidence.

In Fig. 9, we present two cases showing how the reasoner handles conflict situations. In the successful case shown in Fig. 9a, the executor incorrectly answers the root question (I_3) "Is this a safe job for the men?" with "Yes". However, our reasoner aggregates the answers from the atomic questions, such as A_1 ("Yellow protective suits") and A_2 ("A gas cylinder"). This creates a conflict between the direct, erroneous answer to the root question and the logical inference drawn from the sub-questions, which strongly suggest a hazardous environment. Our reasoner correctly resolves this conflict by prioritizing the evidence-based deduction from the atomic facts, thereby correcting the final answer to "No". This demon-

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3650864



(a) Successful Case. Best viewed when zoomed in.

(b) Failure Case. Best viewed when zoomed in.

Fig. 9: Qualitative example showcasing how the reasoner aggregates conflict information.

strates the reasoner's effectiveness in leveraging detailed, factual sub-answers to override an incorrect holistic judgment. Conversely, the failure case in Fig. 9b highlights a limitation in this two-stage process when faced with incorrect sub-question answers. The reasoner synthesizes information from sub-questions, including the incorrect emotional assessment A_3 ("He seems to be scared"), to infer that the man is kicking his legs "to prevent himself from falling down". The reasoner does identify a conflict between this inferred intent and the direct visual evidence, which shows the man immediately sliding down the slope as a result of the action. However, it inadequately resolves the conflict by favoring the flawed logical chain, which was contaminated by the incorrect sub-answer A_3, over the more direct interpretation of the visual outcome. This failure demonstrates that the reasoner's conflict resolution is somehow dependent on the accuracy of the preceding sub-question answering stage. When multiple sub-answers are inaccurate, they can lead to a logically plausible but factually incorrect narrative that the reasoner may fail to override with visual evidence alone.

VII. Conclusion

In this paper, we addressed the inherent lack of transparency and verifiability in current VideoLLMs during the question answering. To overcome these limitations, we proposed a model-agnostic framework (QPVA^3) that integrates question parsing, video alignment, and answer aggregation. Our approach begins by parsing complex queries into compositional graphs, which then guide a transparent, bottom-up recursive reasoning process through video-aligned question answering.

Furthermore, to rigorously evaluate the reasoning capabilities of VideoLLMs, we introduced a GPT-assisted dataset construction pipeline, which extends existing data by incorporating compositional graphs and corresponding sub-question answers, culminating in the QPVA^{3}Bench. With 3,492 video-question pairs, this benchmark facilitates a detailed assessment of both answer accuracy and the consistency of the reasoning process. In addition, we revisited and enhanced existing compositional consistency metrics by proposing new evaluation measures, compositional precision (cP), recall (cR), and F_1 scores (c-F_1). These metrics comprehensively capture the interplay between answer accuracy and the underlying reasoning process, effectively bridging gaps in current benchmarks.

Extensive experiments demonstrate that our framework significantly improves both compositional consistency and accuracy, leading to more transparent and verifiable VideoQA system. Future work will explore refinements and broader applications of QPVA^3 in complex video reasoning tasks.

REFERENCES

[1] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," in *EMNLP*, 2023, pp. 543–553. 1, 2, 3, 11, 13

[2] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao *et al.*, "VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs," *arXiv preprint arXiv:2406.07476*, 2024. 1, 3, 11, 12

[3] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning United Visual Representation by Alignment before Projection," in *EMNLP*, 2024. 1, 2, 3

[4] J. Xiao, X. Shang, A. Yao, and T. Chua, "NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions," in *CVPR*, 2021, pp. 9777–9786. 1, 2, 3, 7

[5] J. Li, L. Niu, and L. Zhang, "From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering," in *CVPR*, 2022, pp. 21 241–21 250. 1, 2, 3, 7, 11

[6] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "STAR: A Benchmark for Situated Reasoning in Real-World Videos," in *NeurIPS*, 2023. 1, 2, 3, 11

[7] J. Li, P. Wei, W. Han, and L. Fan, "IntentQA: Context-aware Video Intent Reasoning," in *ICCV*, 2023, pp. 11 963–11 974. 1, 3

[8] K. Mangalam, R. Akshulakov, and J. Malik, "EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding," in *NeurIPS*, 2023, pp. 46 212–46 244. 1, 3

[9] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "MVBench: A Comprehensive Multi-modal Video Understanding Benchmark," in *CVPR*, 2024, pp. 22 195–22 206. 1, 2, 3, 11, 12, 13

[10] J. Chen, Z. Luo, Z. Liu, W. Jiang, N. Li, and Y. Fang, "Weak-shot keypoint estimation via keyness and correspondence transfer," in *NeuroIPS*, 2025. 1

[11] M. Gandhi, M. O. Gul, E. Prakash, M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Measuring Compositional Consistency for Video Question Answering," in *CVPR*, 2022, pp. 5046–5055. 1, 2, 3, 4, 7, 8, 11

[12] H. Fei, S. Wu, W. Ji, H. Zhang, M. Zhang, M.-L. Lee, and W. Hsu, "Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition," in *ICML*, 2024. 1, 4, 12

[13] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models," *arXiv preprint arXiv:2306.05424*, 2023. 2, 3

[14] Y. Li, X. Wang, J. Xiao, W. Ji, and T. Chua, "Invariant Grounding for Video Question Answering," in *CVPR*, 2022, pp. 2918–2927. 2, 4, 5

[15] Z. Liao, J. Li, L. Niu, and L. Zhang, "Align and Aggregate: Compositional Reasoning with Video Alignment and Answer Aggregation for Video Question-Answering," in *CVPR*, 2024, pp. 13 395–13 404. 2, 3, 5, 13

[16] S. Choi, K. On, Y. Heo, A. Seo, Y. Jang, M. S. Lee, and B. Zhang, "DramaQA: Character-Centered Video Story Understanding with Hierarchical QA," in *AAAI*, 2021, pp. 1166–1174. 3

[17] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "CLEVRER: Collision Events for Video Representation and Reasoning," in *ICLR*, 2020. 3

[18] J. Chen, J. Yan, Y. Fang, and L. Niu, "Meta-point learning and refining for category-agnostic pose estimation," in *CVPR*, 2024. 3

[19] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu *et al.*, "MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI," in *ICML*, 2024. 3

[20] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning," in *CVPR*, 2021, pp. 11 287–11 297. 3, 4

[21] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs," in *CVPR*, 2020, pp. 10 236–10 247. 3

[22] L. Gao, P. Zeng, J. Song, Y. Li, W. Liu, T. Mei, and H. T. Shen, "Structured Two-Stream Attention Network for Video Question Answering," in *AAAI*, 2019, pp. 6391–6398. 3

[23] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering," in *CVPR*, 2019, pp. 1999–2007. 3

[24] P. Jiang and Y. Han, "Reasoning with Heterogeneous Graph Alignment for Video Question Answering," in *AAAI*, 2020, pp. 11 109–11 116. 3

[25] Z. Guo, J. Zhao, L. Jiao, X. Liu, and L. Li, "Multi-Scale Progressive Attention Network for Video Question Answering," in *ACL*, 2021, pp. 973–978. 3

[26] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T. Chua, "Video as Conditional Graph Hierarchy for Multi-Granular Question Answering," in *AAAI*, 2022, pp. 2804–2812. 3

[27] J. Xiao, P. Zhou, T. Chua, and S. Yan, "Video Graph Transformer for Video Question Answering," in *ECCV*, 2022. 3

[28] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, and T.-S. Chua, "Contrastive Video Question Answering via Video Graph Transformer," *TPAMI*, vol. 45, no. 11, pp. 13 265–13 280, 2023. 3

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *ICML*, 2021, pp. 8748–8763. 3

[30] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The LLaMA 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024. 3, 11

[31] H. Fei, S. Wu, M. Zhang, M. Zhang, T.-S. Chua, and S. Yan, "Enhancing Video-Language Representations with Structural Spatio-Temporal Alignment," *TPAMI*, 2024. 3

[32] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models," *arXiv preprint arXiv:2407.07895*, 2024. 3, 12

[33] Q. Cao, X. Liang, B. Li, G. Li, and L. Lin, "Visual Question Reasoning on General Dependency Tree," in *CVPR*, 2018, pp. 7249–7257. 3, 4

[34] D. A. Hudson and C. D. Manning, "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering," in *CVPR*, 2019, pp. 6700–6709. 4

[35] Z. Qian, X. Wang, X. Duan, H. Chen, and W. Zhu, "Dynamic Spatio-Temporal Modular Network for Video Question Answering," in *ACM MM*, 2022, pp. 4466–4477. 4

[36] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural Module Networks," in *CVPR*, 2016, pp. 39–48. 4

[37] Y. Li, X. Wang, J. Xiao, and T. Chua, "Equivariant and Invariant Grounding for Video Question Answering," in *ACM MM*, 2022, pp. 4714–4722. 4, 5, 6

[38] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Transformer-Empowered Invariant Grounding for Video Question Answering," *TPAMI*, 2024. 4

[39] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE TPAMI*, vol. 45, no. 10, pp. 11 624–11 641, 2023. 4, 7

[40] J. Li, L. Niu, and L. Zhang, "Knowledge Proxy Intervention for Deconfounded Video Question Answering," in *ICCV*, 2023, pp. 2782–2793. 4

[41] Y. Wei, Y. Liu, H. Yan, G. Li, and L. Lin, "Visual causal scene refinement for video question answering," in *ACM MM*, 2023. 4, 7

[42] W. Chen, Y. Liu, B. Chen, J. Su, Y. Zheng, and L. Lin, "Cross-modal causal relation alignment for video question grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 4, 7

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *ICLR*, 2022. 7

[44] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *CVPR*, 2016, pp. 5288–5296. 7

[45] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering," in *CVPR*, 2017, pp. 1359–1367. 7

[46] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv preprint arXiv:2409.12191*, 2024. 9, 11, 12, 13, 14

[47] Y. Li, C. Wang, and J. Jia, "LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models," in *ECCV*, 2024, pp. 323–340. 11, 12, 13

[48] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding," in *CVPR*, 2024, pp. 13 700–13 710. 11, 12

[49] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao, "VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding," *arXiv preprint arXiv:2501.13106*, 2025. 11, 12, 13, 14

[50] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li, "LLaVA-OneVision: Easy Visual Task Transfer," 2024. 11, 12, 13, 14

[51] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering," in *AAAI*, 2019, pp. 9127–9134. 11

[52] K. Ataallah, X. Shen, E. Abdelrahman, E. Sleiman, D. Zhu, J. Ding, and M. Elhoseiny, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens," *arXiv preprint arXiv:2404.03413*, 2024. 12

[53] R. Zhang, J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, and P. Gao, "LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-Initialized Attention," in *ICLR*, 2024. 12, 13

[54] Y. Jin, Z. Sun, K. Xu, L. Chen, H. Jiang, Q. Huang, C. Song, Y. Liu, D. Zhang, Y. Song *et al.*, "Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization," *arXiv preprint arXiv:2402.03161*, 2024. 12

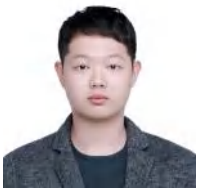
[55] H. Qiu, M. Gao, L. Qian, K. Pan, Q. Yu, J. Li, W. Wang, S. Tang, Y. Zhuang, and T.-S. Chua, "STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training," 2024. 12, 13

[56] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "VideoChat: Chat-centric Video Understanding," *arXiv preprint arXiv:2305.06355*, 2023. 13

[57] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *ICLR*, 2019. 11

[58] X. Gu, H. Fan, Y. Huang, T. Luo, and L. Zhang, "Context-Guided Spatio-Temporal Video Grounding," in *CVPR*, 2024, pp. 18 330–18 339. 14

Jiangtong Li is currently a postdoctoral associate in the School of Computer Science and Technology in Tongji University, Shanghai, China. Before that, he received his B.E. degree and Ph.D. degree from Shanghai Jiao Tong University in 2019, and 2024. His current research interests cover multi-modal modeling, large language model, graph learning and big data in finance.



Zhaohe Liao is currently doctorial student in the Computer Science and Engineering Department at Shanghai Jiao Tong University, Shanghai, China. Before that, he received his B.E. degree from Beijing Institute of Technology in 2021. His current research interests cover multi-modal modeling, large language model, causal inference and compositional reasoning.

Fengshun Xiao is currently a researcher in the AI Platform at Bilibili Inc., Shanghai, China. Before that, he obtained his B.E. degree from Nanjing University, Nanjing, China in 2018 and M.S. degree from Shanghai Jiao Tong University, Shanghai, China in 2021. His current research interests include multi-modal modeling ,large language model and information systems.

Tianjiao Li completed the BSc degree in Computer Science from the University of Nottingham in 2014. He subsequently pursued doctoral studies in functional programming at the University of Leeds from 2014 to 2015. Currently serving as an Expert Engineer at Bilibili, he leads the development of infrastructure for foundation models.

Qiang Zhang is current the Executive Director of AI Platform Department at Bilibili, Shanghai, China. Before joining Bilibili, He worked on large-scale distributed search engineering at Baidu from 2011 to 2017, Beijing, China. He received his Bachelor's and Master's degrees in Computer Science and Technology from Xi'an Jiaotong University in 2008 and 2011, respectively. His current research interests include AI application, machine learning, and distributed high-performance computing.

Haohua Zhao received his Ph.D. degree from Shanghai Jiao Tong University, China, in 2021. Since 2022, he has been an Assistant Researcher with the Department of Computer Science and Engineering at Shanghai Jiao Tong University. His current research interests include machine learning, computer vision, and brain-inspired computing.

Li Niu is currently an associate professor in Computer Science and Engineering Department at Shanghai Jiao Tong University, Shanghai, China. Before joining Shanghai Jiao Tong University, he was a postdoctoral associate at Rice University in Houston, TX, USA. Prior to that, he obtained his B.E. degree from the University of Science and Technology of China, Hefei, China in 2011 and Ph.D. degree from Nanyang Technological University, Singapore in 2017. His current research interests include machine learning, deep learning, and computer vision.

Guang Chen received the B.S. and M.Eng. degrees in mechanical engineering from Hunan University, China, and the Ph.D. degree from the Faculty of Informatics, Technical University of Munich, Germany, in 2016. He is a fully Professor with Tongji University and a Senior Research Associate (guest) with the Technical University of Munich. He is leading the Generalist Embodied AI Laboratory, at Tongji University. His research interests include 3-D vision, embodied artificial intelligence, intelligent robotics, and autonomous driving.

Liqing Zhang received the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1988. He was promoted to full professor position in 1995 at South China University of Technology. He worked as a research scientist in RIKEN Brain Science Institute, Japan from 1997 to 2002. Since September, 2002, he has been a Professor with Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests cover computational theory for cortical networks, visual cognitive representation and inference, statistical learning. He has published more than 250 papers in journals and international conferences.

Changjun Jiang received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1995. He is currently a Professor with the School of Computer Science and Technology, Tongji University, Shanghai, China. His current research interests include concurrency theory, formal verification of software, service-oriented computing, big data in finance, intelligent systems, financial risk management and big data computing.