

# Align and Aggregate: Compositional Reasoning with Video Alignment and Answer Aggregation for Video Question-Answering

Zhaohe Liao<sup>1\*</sup>

Jiangtong Li<sup>2\*</sup>

Li Niu<sup>1†</sup>

Liqing Zhang<sup>1†</sup>

<sup>1</sup> Shanghai Jiao Tong University

{zhaoheliao, ustcnewly, zhang-lq}@sjtu.edu.cn

<sup>2</sup> Tongji University

jiangtongli@tongji.edu.cn

## Abstract

*Despite the recent progress made in Video Question-Answering (VideoQA), these methods typically function as black-boxes, making it difficult to understand their reasoning processes and perform consistent compositional reasoning. To address these challenges, we propose a model-agnostic Video Alignment and Answer Aggregation (VA<sup>3</sup>) framework, which is capable of enhancing both compositional consistency and accuracy of existing VidQA methods by integrating video aligner and answer aggregator modules. The video aligner hierarchically selects the relevant video clips based on the question, while the answer aggregator deduces the answer to the question based on its sub-questions, with compositional consistency ensured by the information flow along question decomposition graph and the contrastive learning strategy. We evaluate our framework on three settings of the AGQA-Decomp dataset with three baseline methods, and propose new metrics to measure the compositional consistency of VidQA methods more comprehensively. Moreover, we propose a large language model (LLM) based automatic question decomposition pipeline to apply our framework to any VidQA dataset. We extend MSVD and NExT-QA datasets with it to evaluate our VA<sup>3</sup> framework on broader scenarios. Extensive experiments show that our framework improves both compositional consistency and accuracy of existing methods, leading to more interpretable real-world VidQA models.*

## 1. Introduction

Video Question-Answering (VidQA) has emerged as a popular research topic in recent years, with potential applications in interactive artificial intelligence and recognition science.

With the development of representing video, question and their alignment, numerous works [6, 14, 21, 23, 34, 35, 49] have achieved considerable success in both open-ended VidQA [22, 53] and multi-choice VidQA [22, 30, 48].

However, existing VidQA methods often function as black-box models, making it difficult to understand the reasoning process behind their predictions and leading to inconsistent compositional reasoning. For example, in Figure 1, HQGA [49] can answer the question “*Is a phone the first object that the person is touching after taking a picture?*” as “Yes”. However, HQGA can neither clearly identify the video clips that contain “*touch a phone*” or “*take a picture*” nor predict all the sub-questions correctly. Therefore, the lack of reasoning transparency can lead to poor compositional consistency, which reveals limited compositional reasoning ability, and further limits the accuracy of VidQA models, particularly on questions that involve temporal relations and multiple visual clues [12].

To tackle this issue, we introduce the Video Alignment and Answer Aggregation (VA<sup>3</sup>) framework, which addresses these challenges by improving their compositional consistency and accuracy. This framework is *model-agnostic* and can be applied to various VidQA methods, such as memory-based [9, 14], graph-based [6, 17, 24, 37, 38, 44, 46], and hierarchy-based [7, 18, 29, 39, 40, 49, 50] methods. In detail, our VA<sup>3</sup> framework includes two additional modules, the video aligner and answer aggregator. The video aligner hierarchically aligns the question with the video clips from the object-level, appearance-level to motion-level. The answer aggregator takes the questions from the same Question Decomposition Graph (QDG) as input and deduces their answers based on their video-question joint representation. To enhance compositional consistency, we further explore a contrastive learning strategy on the edge type of QDG. Overall, the VA<sup>3</sup> framework improves both compositional consistency and accuracy of existing VidQA methods, provides a more transparent compositional reasoning process, and further leads to more in-

\*These two authors contributed equally to this work.

†The corresponding authors.

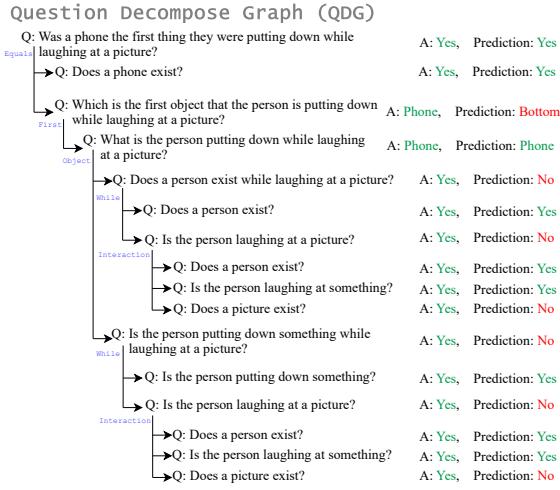


Figure 1. The Question Decomposition Graph (QDG) of a question from AGQA-Decomp [12]. The predicted answer to each question is from HQGA [49]. Green (*resp.*, Red) represents the predicted answer is right (*resp.*, wrong)

terpretable VidQA models in real-world applications.

As for the evaluation metrics, AGQA-Decomp [12] proposes the compositional accuracy (CA), right for the wrong reasons (RWR), and delta (CA – RWR) system to evaluate the compositional consistency of VidQA methods. 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. In detail, 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 VidQA methods from a balanced viewpoint. More details are in Section 4.

We conduct the experiments on the AGQA-Decomp dataset [12] to verify the effectiveness of our framework. This dataset decomposes the questions into the sub-questions and the directed acyclic graphs, *i.e.* the QDGs, making it applicable to evaluate the compositional consistency for VidQA methods. To validate the effectiveness and compositional consistency of our VA<sup>3</sup> framework, we conduct comprehensive experiments with three baseline methods: HME [9], HGA [24], and HQGA [49], which are the representations of the memory-based, graph-based and hierarchy-based methods, in three different settings: balanced, novel compositions, and more compositional steps. Moreover, we propose an automatic question decomposition pipeline for VidQA datasets with the help of large

language models (LLMs) to generalize our framework to datasets that do not have QDGs (*e.g.*, MSVD [52] and NExT-QA [48]) to verify the applicability of our framework. Additionally, we visualize the aligned video clips and the variation of predicted answers while equipping video aligner and answer aggregator successively to the backbone model on QDG to verify the interpretability of our framework. Our contribution can be summarized as:

- **Dataset:** We propose an automated question decomposition pipeline for any VidQA dataset to generate the QDGs and the sub-questions with the help of LLMs and further extend MSVD and NExT-QA dataset with it.
- **Framework:** We propose a *model-agnostic* VA<sup>3</sup> framework, which provides a more transparent compositional reasoning process and increases both the interpretability and the accuracy of existing VidQA models.
- **Metric:** We extend the compositional consistency metrics as consistency precision (cP), consistency recall (cR) and consistency  $F_1$  (c- $F_1$ ) along with their negative versions for a more balanced and comprehensive evaluation.
- **Experiments:** Comprehensive experiments with three baselines on five benchmark settings of three datasets reveal that our framework significantly boosts these baselines in compositional consistency and accuracy.

## 2. Related Work

### 2.1. Video Question-Answering

While the architecture of VidQA methods has undergone significant changes over the years, the essential components of these methods remain the same: video representation, question representation, and video-question aligned representation. For the video representation, appearance features [20] and motion features [51] were commonly used, then the object-level representation was introduced [25]. For the question representation, most existing works relied on word embeddings [41] with RNNs, while BERT [8] features became widely used in more recent works [34, 49]. In the early research, the video-question alignment was implemented using cross-modal attention [15, 32] or memory networks [9, 14], then graph reasoning [6, 17, 24, 37, 38] became popular. Recently, the natural hierarchy in video representation [18, 29, 40, 49, 50] received more attention.

Despite these advancements, existing methods still face challenges in achieving satisfactory levels of compositional consistency [12]. To address these challenges, in this paper, we propose a *model-agnostic* framework for compositional reasoning by combining the visual alignment and the answer aggregation to improve current VidQA methods.

### 2.2. Compositional Reasoning in VidQA

The practice of decomposing a complex question into simpler questions has been observed in various tasks [4, 54].

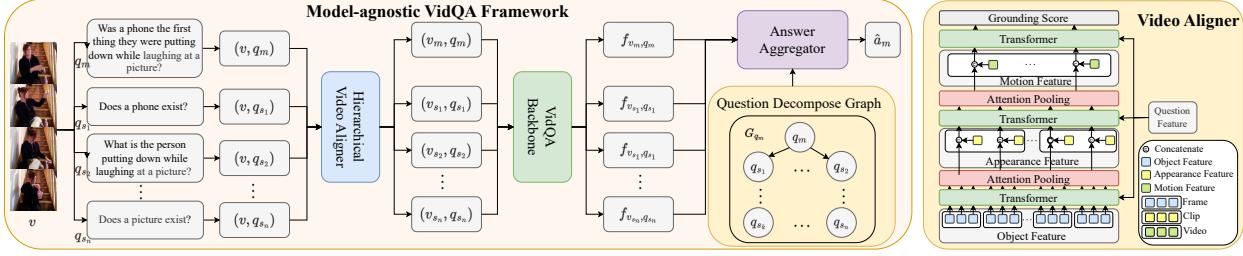


Figure 2. Our *model-agnostic* Video Alignment and Answer Aggregation (VA<sup>3</sup>) framework.  $(v, q)$  is a video-question pair, where  $q_m$  denotes main-question and  $q_{s_1}, \dots, q_{s_n}$  denote the  $n$  sub-questions derived from  $q_m$ .  $v_m$  and  $\{v_{s_1}, \dots, v_{s_n}\}$  denote the aligned videos according to corresponding questions.  $f_{v_m, q_m}$  and  $\{f_{v_{s_1}, q_{s_1}}, \dots, f_{v_{s_n}, q_{s_n}}\}$  denote the video-question joint features.  $G_{q_m}$  is the question decomposition graph (QDG) associated with  $q_m$ , which is a direct acyclic graph describing the compositional relationship among questions. Moreover,  $G_{q_m}$  stores in which manner the questions are decomposed (*i.e.*, the operators in the decomposition program) as the attribute of edges.  $\hat{a}_m$  denotes the predicted answer for  $q_m$ .

In VidQA, most of earlier efforts [16] broke down questions into modular programs that were defined in a neural modular network [42] to answer the question. AGQA [16] explored spatio-temporal scene graphs to represent the programs for VidQA. However, such a reasoning program cannot be directly used by existing VidQA methods. To address this issue, AGQA-Decomp [12] transferred each reasoning program into several sub-questions and QDG to evaluate the compositional consistency of existing VidQA methods.

The Neural Modular Network (NMN)-based methods (*e.g.*, DSTN [42]) modularized the VidQA task into multiple modules (*e.g.*, *FindObj*, *TemporalFilter*, *etc.*) and generated the reasoning program through a modular policy. Although the NMN-based approaches provide perfect interpretability, there still exist three main challenges: 1) the basic modulars and logic rules have to be pre-defined, making any novel modulars and logic rules incompatible; 2) compared to conventional neural networks, training NMNs can be more challenging because the learning process optimizes the composition strategy other than the individual modules; 3) as the number of modules increases, the search space for optimal modules grows exponentially, which hinders its scalability to more complex tasks or larger datasets.

### 2.3. Video Grounding

Video grounding [2, 13] seeks to identify the most relevant moment in a video based on language queries [31, 36, 45, 55], and has received growing attention from downstream video-language tasks [1, 33–35]. Previous works such as IGV [35] and EIGV [34] have focused on differentiating between causal and environment clips in VidQA through a simple grounding indicator and encouraging sensitivity to semantic changes in the causal scene, respectively. In contrast to these approaches, our framework hierarchically aligns the question with video clips from object-level, appearance-level, to motion-level to provide a more refined video context along with an answer aggregator. This

approach enhances both the generalization and compositional reasoning abilities of existing VidQA methods, leading to more effective and accurate models.

## 3. Our Method

As described in Section 1, current VidQA models suffer from insufficient compositional reasoning ability. Therefore, we propose our VA<sup>3</sup> framework, consisting a hierarchical video aligner and a QDG-based answer aggregator.

### 3.1. Model-agnostic VidQA Framework

Our VidQA framework is illustrated in Figure 2. For each original question in dataset, it is decomposed (either oracularly or with our question decomposition pipeline) into several sub-questions. Formally, the original question (regarded as main question)  $q_m$  and its decomposed sub-questions  $q_{s_i}^m$  form a question cluster on corresponding QDG  $G_{q_m}$ . First, we introduce a hierarchical video aligner, which selects the question-related video clips by hierarchically interacting the video clips with the question among the object-level, appearance-level and motion-level. After that, the questions in the cluster and their corresponding video clips are send into a VidQA model, regarded as  $\mathcal{F} : (v, q) \rightarrow f_{v, q} \in \mathbb{R}^h$ , where  $h$  is the hidden dimension for joint feature space, to generate the joint feature  $f_{v, q}$ . The answer aggregator takes all the joint features from the questions cluster and associates each joint feature with regard to the question node in the QDG. By aggregating the information in joint features of each node through QDG, we adjust the question representations and predict the answers.

### 3.2. Video Alignment

The structure of video aligner is shown in Figure 2. The video aligner takes video-question pair  $(v, q)$  as input, and gives the video clips that are most relevant to the question. Former research on localizing video clips with ques-

tions [34, 35] failed to use the natural hierarchy of video features, thus limited the representation ability of the aligner.

In our video aligner, a video  $v$  is represented within three-level features: object feature  $\mathbf{F}_o \in \mathbb{R}^{n_c \times n_f \times n_o \times h_v}$ , appearance feature  $\mathbf{F}_a \in \mathbb{R}^{n_c \times n_f \times h_v}$  and motion feature  $\mathbf{F}_m \in \mathbb{R}^{n_c \times h_v}$ , where  $n_c$ ,  $n_f$  and  $n_o$  represents the number of clips per video, frames per clip and objects per frame respectively, and  $h_v$  is the hidden dimension for each feature vector. As these three-level features naturally follow a hierarchical relationship, we designed a hierarchical video aligner to capture them for a better alignment.

Our hierarchical video aligner follows a bottom-up video-question interaction scheme. Starting from  $\mathbf{F}_o$ , we aggregate its information among objects with the condition of question, concatenate it with corresponding  $\mathbf{F}_a$ , and further aggregate it with other frames in control of question. Then, such cross-frame representation is concatenated with  $\mathbf{F}_m$ , and interacted with question feature to generate the grounding score. During each aggregation, we fuse the video feature and question feature through a transformer layer, where the video feature is regarded as query while question feature serves as the key and value. Formally, the aggregation from object feature to appearance feature is

$$\begin{aligned}\mathbf{F}_o^j &= \text{TF}(\mathbf{F}_o, \mathbf{F}_q, \mathbf{F}_q); \\ \mathbf{F}_o^a &= \sum_{n_o} \sigma_{n_o} (\mathbf{W}_o \mathbf{F}_o^j + \mathbf{b}_o) \mathbf{F}_o^j; \mathbf{F}_a^c = [\mathbf{F}_o^a || \mathbf{F}_a];\end{aligned}\quad (1)$$

where  $\sigma_{n_i}$  is the softmax function along  $n_i$  dimension, TF is the transformer encoder layer,  $\mathbf{W}_o$  and  $\mathbf{b}_o$  are trainable parameters, and  $[\cdot || \cdot]$  denotes the concatenation operator. The updated appearance feature  $\mathbf{F}_a^c$  is used to produce aggregated motion feature  $\mathbf{F}_m^c$  in a similar manner. Further, we use the  $\mathbf{F}_m^c$  to generate the binary indicator of relevant clip for the video, which can be formulated as

$$\begin{aligned}\mathbf{F}_m^j &= \text{TF}(\mathbf{F}_m^c, \mathbf{F}_q, \mathbf{F}_q); \mathbf{s}_{rel} = \text{MLP}_1(\mathbf{F}_m^j); \\ \mathbf{s}_{irr} &= \text{MLP}_2(\mathbf{F}_m^j); I = \text{Gumble-Softmax}([\mathbf{s}_{rel} || \mathbf{s}_{irr}]),\end{aligned}\quad (2)$$

where MLP is the multi-layer linear projection.

Since the ground-truth of aligned video is not applicable in VidQA dataset, we exploit the contrastive learning [34] to guide this module. Formally, given a video-question pair  $(v, q)$  in training data, the indicator specifies the relevant video clips  $\hat{v}_r$  and irrelevant video clips  $\hat{v}_c$  within  $v$ . Thus, the anchor  $\mathbf{f}_{\hat{v}_r, q}$ , positive sample  $\mathbf{f}_{\hat{v}', q}$ , and negative sample  $\mathbf{f}_{\hat{v}_c, q}$  of the contrastive loss is presented as

$$\mathbf{f}_{\hat{v}_r, q} = \mathcal{F}(\hat{v}_r, q); \mathbf{f}_{\hat{v}', q} = \mathcal{F}(v', q); \mathbf{f}_{\hat{v}_c, q} = \mathcal{F}(\hat{v}_c, q), \quad (3)$$

where  $\hat{v}'$  is acquired by replacing  $\hat{v}_c$  in  $v$  with random sampled clips. Therefore, the contrastive loss is defined as

$$\mathcal{L}_{al}^c = -\log \frac{\exp(\mathbf{f}_{\hat{v}_r, q}^T \mathbf{f}_{\hat{v}', q})}{\exp(\mathbf{f}_{\hat{v}_r, q}^T \mathbf{f}_{\hat{v}', q}) + \exp(\mathbf{f}_{\hat{v}_r, q}^T \mathbf{f}_{\hat{v}_c, q})}. \quad (4)$$

Moreover, the answer prediction can be formulated as

$$P(\hat{a} | \hat{v}_r, q) = \sigma(\mathbf{W}_{o_1} \mathbf{f}_{\hat{v}_r, q} + \mathbf{b}_{o_1}), \quad (5)$$

where  $\sigma$  is softmax function and  $\mathbf{W}_{o_1}$  and  $\mathbf{b}_{o_1}$  are the trainable parameters. Therefore, the total loss of each video-question pair  $(v, q)$  for video aligner can be formulated as

$$\mathcal{L}_{al} = \text{CE}(P(\hat{a} | \hat{v}_r, q), a) + \mathcal{L}_{al}^c. \quad (6)$$

where CE refers to cross entropy and  $a$  represents the ground-truth answer for the video-question pair  $(v, q)$ .

### 3.3. Answer Aggregation

Existing VidQA methods predict the answers of different questions independently, which ignores the correlations among questions from the same cluster, leading to insufficient compositional consistency. Therefore, we introduce an answer aggregator to supplement the main-question with sub-questions and enhance the compositional consistency. Specifically, assume  $\{\mathbf{f}_{q_i, v_i} | i \in \{s_1, \dots, s_n, m\}\}$  is the video-question joint feature extracted by backbone VidQA model for all questions in QDG  $G_{q_m} = (V_{q_m}, E_{q_m})$ . We explore the graph attention network (GAT) to aggregate the joint feature along the given QDG. Formally, given the  $k$ -th layer of the GAT, the main question  $q_m$  and its sub-questions  $\{q_{s_1}, \dots, q_{s_n}\}$ , the information aggregation for node associated with  $q_i$  is formulated as

$$\begin{aligned}s(\mathbf{f}_{q_i, v_i}^k, \mathbf{f}_{q_j, v_j}^k) &= \mathbf{a}_k^T \text{LeakyReLU}(\mathbf{W}_s^k [\mathbf{f}_{q_i, v_i}^k || \mathbf{f}_{q_j, v_j}^k]); \\ \alpha_{i,j} &= \sigma_j(s(\mathbf{f}_{q_i, v_i}^k, \mathbf{f}_{q_j, v_j}^k)); \\ \mathbf{f}_{q_i, v_i}^{k+1} &= \text{ReLU} \left( \sum_{q_j \in \{(q_i, q_j) \in E_{q_m}\}} \alpha_{i,j} \mathbf{W}_g^k \mathbf{f}_{q_j, v_j}^k \right),\end{aligned}\quad (7)$$

where  $i, j \in \{s_1, \dots, s_n, m\}$ , and  $\mathbf{a}_k$ ,  $\mathbf{W}_s^k$  and  $\mathbf{W}_g^k$  are the trainable parameters for the  $k$ -th layer of GAT. Finally, the outputs of all layers is concatenated together and projected to predict the answer, which is formulated as

$$\begin{aligned}\mathbf{f}_{q_i, v_i}^a &= \mathbf{W}_{o_2} [\mathbf{f}_{q_i, v_i}^1 || \dots || \mathbf{f}_{q_i, v_i}^K] + \mathbf{b}_{o_2}; \\ P_{ag}(a_i | v_i, q_i) &= \sigma(\mathbf{f}_{q_i, v_i}^a),\end{aligned}\quad (8)$$

where  $\mathbf{W}_{o_2}$  and  $\mathbf{b}_{o_2}$  are trainable parameters.

Moreover, to enhance the compositional consistency, we introduce an additional contrastive training scheme. As the type of relation (*i.e.*, edge) between the questions provides crucial clues in question decomposing and compositional reasoning, we introduce a heuristic prior, where edges with the same type shall have similar representations, and the distance between different types of edges shall be relatively large. By leveraging this prior, we can raise the level of abstraction for more accurate and consistent answer reasoning. Formally,  $\{\mathbf{f}^e = \mathbf{W}_e [\mathbf{f}_{q_i, v_i}^a || \mathbf{f}_{q_j, v_j}^a] + \mathbf{b}_e | e \in E_{q_m}\}$

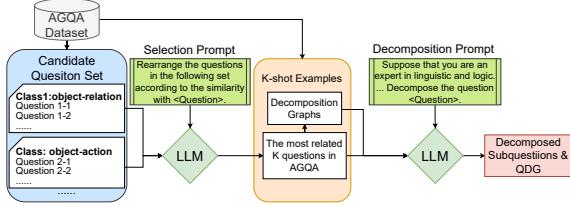


Figure 3. The automatic question decomposition pipeline. The question to be decomposed is denoted as  $\langle \text{Question} \rangle$ .

denotes the set of node relation representation for edges in graph  $G_{q_m}$ , where  $\mathbf{W}_e$  and  $\mathbf{b}_e$  are the trainable parameters. Moreover, we use  $t_e \in T$  to denote the class of edge  $e$ , where  $T$  is the set of edge types. For  $t \in T$ , we use  $t^c = T/\{t\}$  to represent the complementary set of  $t$ , and use  $e_t$  to denote a random edge sampled from all edges with type  $t$ . Thus, the triplet loss for main question  $q_m$  is

$$\mathcal{L}_c^{q_m} = \mathbb{E}_{e \in E_{q_m}} \max(d(\mathbf{f}^e, \mathbf{f}^{e_{t_e}}) - d(\mathbf{f}^e, \mathbf{f}^{e_{t_e^c}}) + m, 0), \quad (9)$$

where  $d(\cdot, \cdot)$  is Eular distance and  $m$  is margin. Thus, the answer aggregation loss for question cluster  $Q = \{q_m, q_{s_1}, \dots, q_{s_n}\}$  is formulated as

$$\mathcal{L}_{ag} = \mathcal{L}_c^{q_m} + \mathbb{E}_{(v_i, q_i) \in \mathcal{D}} \text{CE}(P_{ag}(a_i | v_i, q_i), a_i), \quad (10)$$

where  $\mathcal{D} = \{(v, q) | q \in Q\}$ . Both  $\mathcal{L}_{ag}$  and  $\mathcal{L}_{al}$  are consequently applied during training.

### 3.4. Automatic Question Decomposition Pipeline

For some VidQA dataset (*e.g.*, MSVD and NExT-QA), the QDGs are not applicable as they only provide the main questions. To address this issue, we explore an automatic question decomposition pipeline for VidQA data using the knowledge in LLMs. Since directly asking the LLM to decompose questions may result in poor results, and random examples could cause unstable quality as the chosen examples may have different compositional structure with the queried question, we proposed the decomposition pipeline as shown in Figure 3. Firstly, we construct a candidate example set based on the AGQA-Decomp dataset manually, in which each subset consists of a few main questions chosen with a main question type in AGQA-Decomp dataset to cover as many types of main questions as we can. Then, we construct a selection prompt which ask the LLM to select the most similar K question, and they form K-shot examples with their QDGs. Finally, these K-shot examples are provided to LLM with a decomposition prompt asking the LLM to decompose the target question, resulting in the decomposed sub-questions and corresponding QDGs with better quality. More details and explanations of our pipeline and prompts are in the supplementary material.

## 4. Metrics

Compositional consistency measures whether a method can provide the correct answer for the right reason. AGQA-Decomp [12] propose compositional accuracy (CA), right for the wrong reasons (RWR), and Delta (CA-RWA), which offer some insight on it. However, as we are to illustrate in Section 4.1, they cannot fully reveal the reasoning capabilities, leading to asymmetric and unstable problem. To address this issue, we extend them with compositional precision (cP), recall (cR), and  $F_1$  (c- $F_1$ ), along with their negative versions, providing a more comprehensive assessment of reasoning consistency. We name  $q_j$  as parent question, and  $q_i, q_k$  as children question of  $q_j$  for QDG edges like  $q_i \leftarrow q_j \rightarrow q_k$ . Note that we only consider the **1-st** order parent-children relation, and the intermediate sub-questions can be both parent or children regarding the viewpoint.

### 4.1. Another View of Existing Metrics

CA and RWR are designed to evaluate the accuracy of the parent questions, conditional on whether all their children questions are correct. Let  $N_-^+$  denote the number of correctly answered main-question with any sub-question incorrect, while  $N_+^-$  denotes the number of falsely answered main question with all sub-questions correct.  $N_-^-$  and  $N_+^+$  is similarly defined. Thus, CA and RWR is formulated as

$$\text{CA} = N_+^+ / (N_+^+ + N_-^+); \text{RWR} = N_-^- / (N_-^- + N_+^-), \quad (11)$$

and  $\text{Delta} = \text{RWR} - \text{CA}$ . Therefore, both CA and 1 – RWR can be viewed as “**precisions**”, as the conditions only inspect the correctness of the children questions and missed out the correctness of the parent questions, leading to asymmetric and unstable problems illustrated in Table 1.

For row 1 to row 4, the corresponding model shall have the same compositional consistency, because in all 210 parent questions, 110 of them can be answered for the right reasoning and the opposite. However, the CA, RWR and Delta varies significantly among these models, which indicate that the CA-RWR-Delta metrics cannot treat  $N_+^+$ ,  $N_-^-$ ,  $N_-^+$  and  $N_+^-$  asymmetrically and cannot correctly identify the compositional consistency. Moreover, such failure also lead to instability when facing imbalanced child question accuracy distribution, as shown in row 5 to row 7. These models have similar compositional consistency, but the CA, RWR and Delta may vary to the extreme opposite value.

### 4.2. Our Metrics

As described in Section 4.1, CA and 1 – RWR can both be viewed as “**precisions**”, thus we denote them as **consistency precision (cP)** and **negative consistency precision (NcP)** respectively to simplify the following formulation. For a more comprehensive view on compositional consistency, we are to introduce the corresponding “recalls”.

	Data Count				Existing Metrics			Our Metrics		
	$N_+^+$	$N_+^-$	$N_-^+$	$N_-^-$	CA	RWR	Delta	Acc.	c-F <sub>1</sub>	Nc-F <sub>1</sub>
1	100	100	0	10	50.00	0.00	-50.00	47.61	66.67	16.67
2	10	100	0	100	9.09	0.00	-9.09	4.76	16.67	66.67
3	100	0	100	10	100.00	90.91	-9.09	95.23	66.67	16.67
4	10	0	100	100	100.00	50.00	-50.00	52.38	16.67	66.67
5	99	100	1	0	49.75	100.00	-50.25	50.00	66.22	0.00
6	100	99	0	1	50.25	0.00	50.25	50.00	66.88	0.99
7	99	99	1	1	50.00	50.00	0.00	50.00	66.44	0.98

Table 1. The conterexamples. Acc. is parent question accuracy.

**Definition 1 (Consistency Recalls)** *Given a VidQA model  $M$ , the consistency recall ( $cR$ ) and negative consistency recall ( $NcR$ ) is defined as*

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

These metrics condition on the correctness of main questions. Clearly, neither cP and cR can represent the compositional consistency solely, since they only take part of the condition into consideration. To combine both perspectives, we introduce their weighted harmonic mean, *i.e.*, consistency F-scores for a robust and symmetric representation.

**Definition 2 (Consistency F-Score)** *Given a VidQA model  $M$ , the consistency F-score ( $cF_\beta$ ) and negative consistency F-score ( $NcF_\beta$ ) is defined as*

$$cF_\beta = \frac{(1+\beta^2)cP \cdot cR}{\beta^2 cP + cR}; \quad NcF_\beta = \frac{(1+\beta^2)NcP \cdot NcR}{\beta^2 NcP + NcR}. \quad (13)$$

In such definition,  $\beta$  is a hyper-parameter to balance cP and cR. To measure the two kinds of error in a equal weight, we set  $\beta = 1$ , and thus use c-F<sub>1</sub> to measure the compositional consistency of models. Such metric considers the compositional consistency in a symmetric manner, raising a comprehensive evaluation on model’s ability. As shown in Table 1, our c-F<sub>1</sub> and Nc-F<sub>1</sub> metrics raises more reasonable evaluations (row 1 to row 4), and is also more stable facing extreme cases (row 5 to row 7). Note that although c-F<sub>1</sub> and Nc-F<sub>1</sub> provides a balanced view of compositional consistency, we still need to the accuracy, cP, and cR for a detailed analysis regarding prediction ability and compositional bias. **More analysis and comparison between our metrics and original ones are in the supplementary.**

## 5. Experiments

### 5.1. Experiment Setting

**Dataset** As described in Section 1, we conduct our experiment on AGQA-Decomp benchmark, which extends AGQA 2.0 by decomposing each question into several sub-questions with a QDG, and evaluates the compositional

reasoning ability by supplying extensive challenging complex questions with their decomposed sub-questions and answers. Moreover, we test the improvement on MSVD and NExT-QA dataset to verify the applicability of our VA<sup>3</sup> framework and question decomposition pipeline.

**Baselines and Metrics** We conduct experiment on all three categories of VidQA mothods, *i.e.*, memory-based, graph-based and hierarchy-based methods. Specifically, we choose HME [9], HGA [24] and HQGA [49] as the baseline methods from each category respectively. For the evaluation metrics, we measure the open-ended, binary, and overall accuracy for main questions and sub-questions. Moreover, to illustrate the improvement in terms of compositional consistency, we evaluate the cR, cP, c-F<sub>1</sub>, NcR, NcP and Nc-F<sub>1</sub>. More detailed settings are in supplementary material.

### 5.2. Main Results

The result of our framework with various baselines is shown in Table 2. Compared the 1-st to 3-rd row with the 4-th to 6-th row correspondingly, our framework outperforms all baseline models, including memory-, graph- and hierarchy-based models, significantly in terms of both accuracy and compositional consistency. For accuracy, the overall main question accuracy improves 1.23% to 2.71%, while the sub-question accuracy raises 3.16% to 3.29%. The accuracy improvement on sub-questions are usually more than that on main questions. For video aligner, it is more hard to align the corresponding video clips for main questions, besides, for answer aggregator, it is also more challenge to aggregate all the sub-questions to deduce the main question, leading to such improvement gap. Moreover, the accuracy improvements on binary and open-ended questions are not equal. The reasons include the the following two aspects: 1) the video aligner helps open-ended questions more since they are more sensitive to irrelevant clips as they have to choose answer from a much larger candidate set than binary questions, and may be misled by the actions in irrelevant clips more easily; 2) the open-ended questions provide and receive more severe information in answer aggregation, making the answer aggregator contribute more on them.

Moreover, the compositional consistency also raises significantly compared to the baseline models. Specifically, the c-F<sub>1</sub> significantly improves 2.97% to 3.54%, while the Nc-F<sub>1</sub> raises 0.11% to 0.64%. The c-F<sub>1</sub> indicates how much the model answers **correctly** with **correct** inference, therefore, the c-F<sub>1</sub> is the most important overall measurement for the reasoning ability of models. The significant improvement in terms of c-F<sub>1</sub> further indicates that our framework does help the VidQA models reasoning correctly and consistently. And the Nc-F<sub>1</sub>, which measures if the model is still consistent even when the answer is incorrect, shows that our framework also slightly helps improve the overall consistency on the incorrect main questions. Furthermore,

	Main Accuracy			Sub Accuracy			Compositional Consistency					
	Open	Binary	All	Open	Binary	All	cP	cR	c-F <sub>1</sub>	NcP	NcR	Nc-F <sub>1</sub>
HME	36.29	51.41	<b>41.59</b>	29.68	71.73	<b>56.59</b>	53.06	46.02	<b>49.29</b>	56.57	63.33	<b>59.76</b>
HGA	41.18	56.62	<b>46.61</b>	36.03	73.11	<b>59.75</b>	64.02	57.85	<b>60.78</b>	57.35	63.55	<b>60.29</b>
HQGA	41.05	50.40	<b>44.34</b>	33.22	70.03	<b>56.77</b>	54.49	38.55	<b>45.16</b>	53.89	69.04	<b>60.53</b>
VA <sup>3</sup> (HME)	39.91 <sup>+3.62</sup>	52.26 <sup>+0.85</sup>	<b>44.30</b> <sup>+2.71</sup>	35.23 <sup>+5.55</sup>	73.56 <sup>+1.83</sup>	<b>59.75</b> <sup>+3.16</sup>	57.78 <sup>+4.72</sup>	48.66 <sup>+2.64</sup>	<b>52.83</b> <sup>+3.54</sup>	55.81 <sup>-0.76</sup>	64.58 <sup>+1.25</sup>	<b>59.87</b> <sup>+0.11</sup>
VA <sup>3</sup> (HGA)	43.04 <sup>+1.86</sup>	56.82 <sup>+0.20</sup>	<b>47.88</b> <sup>+1.27</sup>	42.18 <sup>+6.15</sup>	74.69 <sup>+1.58</sup>	<b>62.98</b> <sup>+3.23</sup>	67.52 <sup>+3.50</sup>	60.38 <sup>+2.53</sup>	<b>63.75</b> <sup>+2.97</sup>	57.47 <sup>+0.12</sup>	64.83 <sup>+1.28</sup>	<b>60.93</b> <sup>+0.64</sup>
VA <sup>3</sup> (HQGA)	42.35 <sup>+1.30</sup>	51.53 <sup>+1.13</sup>	<b>45.57</b> <sup>+1.23</sup>	35.28 <sup>+2.06</sup>	74.01 <sup>+3.98</sup>	<b>60.06</b> <sup>+3.29</sup>	56.02 <sup>+1.53</sup>	42.50 <sup>+3.95</sup>	<b>48.33</b> <sup>+3.17</sup>	55.26 <sup>+1.37</sup>	68.04 <sup>-1.00</sup>	<b>60.99</b> <sup>+0.46</sup>

Table 2. The comparison with baseline methods on AGQA-Decomp [12]. The overall measurements are highlighted in bold, and the improvements of our framework are highlighted as superscript.

	Novel Comp. Setting			More Comp. Step Setting		
	Accuracy	c-F <sub>1</sub>	Nc-F <sub>1</sub>	Accuracy	c-F <sub>1</sub>	Nc-F <sub>1</sub>
HME	31.54	35.88	68.94	44.28	48.32	63.71
HGA	33.40	35.44	65.18	47.26	48.42	63.45
HQGA	34.21	38.91	67.96	46.64	50.65	62.97
VA <sup>3</sup> (HME)	33.45 <sup>+1.91</sup>	38.26 <sup>+2.38</sup>	69.08 <sup>+0.14</sup>	46.07 <sup>+1.79</sup>	49.87 <sup>+1.55</sup>	64.01 <sup>+0.30</sup>
VA <sup>3</sup> (HGA)	35.27 <sup>+1.87</sup>	40.47 <sup>+5.03</sup>	65.33 <sup>+0.15</sup>	48.38 <sup>+1.12</sup>	51.08 <sup>+2.66</sup>	63.58 <sup>+0.13</sup>
VA <sup>3</sup> (HQGA)	36.33 <sup>+2.12</sup>	40.76 <sup>+1.85</sup>	68.20 <sup>+0.24</sup>	47.91 <sup>+1.27</sup>	51.75 <sup>+1.10</sup>	63.26 <sup>+0.29</sup>

Table 3. The comparison with baseline methods on the AGQA-Decomp [12] *novel composition* setting and *more composition step* setting [16]. Comp. is the abbreviation for compositional. The improvements of our framework are highlighted as superscript.

we can also find that the improvement on c-F<sub>1</sub> are naturally more significant than on Nc-F<sub>1</sub> as our constraint on answer aggregation mainly focus on correctly deducing the main question based on the correct sub-questions.

### 5.3. Generalization Ability

We further test the improvement of our framework when generalizing to new situations on two extra settings, *i.e.* *novel composition* and *more composition step* [16]. The *Novel composition* setting tests if models can generalize to unseen composition types, and the *more composition step* setting tests the generalization ability when facing more complex questions than training. The results are in Table 3.

When facing novel composition setting, the clear accuracy drop compared to Table 2 indicates generalizing to novel composition setting is much harder for VidQA models than the standard setting. However, our framework is still capable of significantly boosting baseline methods under this challenging setting. The main question accuracy raises 1.87% to 2.12%, while the c-F<sub>1</sub> improves 1.85% to 5.03% and the Nc-F<sub>1</sub> improves 0.14% to 0.24%, on different baselines, implying that our framework provides better generalization ability on unseen composition types, in terms of both accuracy and compositional consistency.

For the more composition step setting, our framework still significantly improves the baseline methods on both the accuracy and compositional consistency, indicating our framework is effective when generalizing to more complex questions. In detail, there is a 1.12% to 1.76% accuracy

	Main Accuracy			Sub Accuracy			Consistency	
	Open	Binary	All	Open	Binary	All	c-F <sub>1</sub>	Nc-F <sub>1</sub>
HME	36.29	51.41	41.59	29.68	71.73	56.59	49.29	59.76
EIGV [34]	38.84	51.48	43.29	31.77	72.42	57.77	48.05	59.31
Our Aligner	39.49	51.50	43.72	32.27	72.81	58.20	49.21	59.71
AA. +AA.	38.21	51.35	42.83	33.78	72.33	58.44	52.19	59.73
AA. +AA. + $\mathcal{L}_c^{q_m}$	38.81	52.05	43.46	34.12	72.91	58.93	53.49	60.05
VA <sup>3</sup> (HME)	39.91	52.26	44.30	35.23	73.56	59.75	52.83	59.87

Table 4. The ablation study on our Video Aligner and Answer Aggregator. VA. is video aligner and AA. is answer aggregator.

improvement for main questions, while the c-F<sub>1</sub> improves 1.10% to 2.66% while the Nc-F<sub>1</sub> raises 0.13% to 0.30%.

### 5.4. Ablation Study

To measure the effectiveness and necessity of our modules, we conduct ablation studies in this section. To clearly reveal the contribution of each module and loss, we choose HME [9] as the baseline method in this section, since the improvement over HME is the largest among all three baselines. For the video aligner, we test its improvement over the EIGV [34] aligner brought by the hierarchy structure. For the answer aggregator, we test how much it boosts the original model even without our contrastive loss  $\mathcal{L}_c^{q_m}$ , then measure the improvement of applying  $\mathcal{L}_c^{q_m}$  on answer aggregator. The results are summarized in Table 4.

By comparing row 3 with rows 1 and 2, we can conclude that the video aligner substantially improves the accuracy of both main questions and sub-questions, but helps little on compositional consistency. This is reasonable since video aligners do not exploit the relations between main questions and sub-questions, thus cannot improve the compositional consistency among them. Moreover, the comparison between row 3 and row 2 shows that our aligner outperforms the video grounding module in EIGV due to its hierarchical structure, which could effectively extract information from different level of video feature and align them with the question. As for the answer aggregation module, by comparing row 4 with row 1, we can infer that answer aggregator, even without the contrastive loss  $\mathcal{L}_c^{q_m}$ , can signifi-

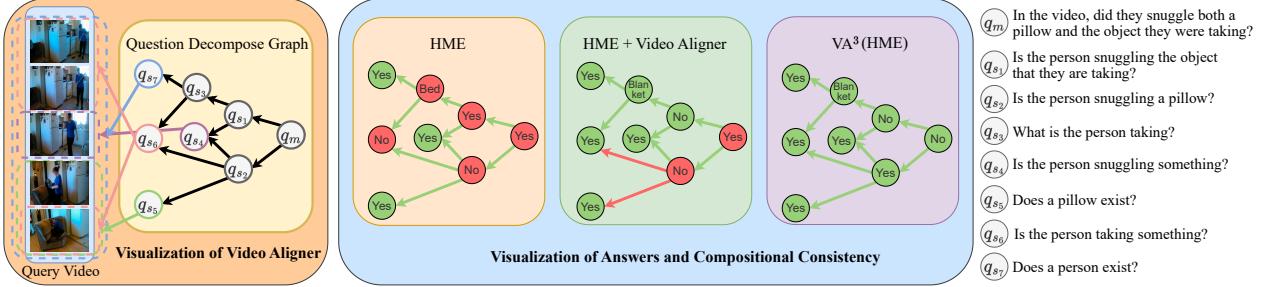


Figure 4. Quantitive results of Video Aligner and the visualization of improvements on accuracy and compositional consistency brought by our modules. Best viewed in color and zoom in. **More visualizations and explanations are in the supplementary material.**

cantly improve the accuracy and compositional consistency, as the information exchange along main-sub questions relation helps both correct answering and ensuring their consistency. Further, as the comparison between row 4 and row 5 implies, the contrastive loss  $\mathcal{L}_c^{q_m}$  further improves the accuracy and compositional consistency. Finally, the improvement of row 6 over row 3 and 5 indicates that our combination of video aligner and answer aggregator is mutually beneficial on both accuracy and compositional consistency, proving the effectiveness of our framework.

## 5.5. Qualititative Study

In Figure 4, we firstly visualize the result of video aligner. As the left part of Figure 4 shows, our video aligner successfully aligns the related video clips (denoted by the dotted boxes) with the corresponding questions. Therefore, the VidQA backbones can use more accurate information to improve the accuracy of both main questions and sub-questions. Moreover, we also visualize the predicted answers of HME, HME with video aligner, and VA<sup>3</sup>(HME) respectively. The original model failed to answer the main question correctly due to several factual errors in sub-questions and the potential reasoning failure. With the help of video aligner, we may reduce the irrelevant noise by only processing the most correlated clips, thus eliminating several factual errors (*i.e.*,  $q_{s1}, q_{s3}$  and  $q_{s6}$ ), but may not help on the main question since the whole video is fatal in generating its answer. Moreover, the video aligner may not correct the compositional reasoning failure as it does not use inter-question relation information. However, with the help of the answer aggregator, we can correct the reasoning failure by aggregating the information from sub-questions (*e.g.*, correct the answer of  $q_{s2}$  by aggregating video-question joint features of  $q_{s4}, q_{s5}$  and  $q_{s6}$ ), and deduce the correct answer of main questions with higher compositional consistency.

## 5.6. Applicability

To verify that our framework is generally applicable, we further apply our framework on MSVD [52] and NExT-QA [48] datasets extended by our automatic question de-

	HME	HGA	HQGA	VA <sup>3</sup> (HME)	VA <sup>3</sup> (HGA)	VA <sup>3</sup> (HQGA)
MSVD	33.75	36.71	41.23	<b>38.51<sup>+4.76</sup></b>	<b>41.24<sup>+4.53</sup></b>	<b>44.46<sup>+3.23</sup></b>
NExT-QA	48.72	50.04	51.65	<b>53.23<sup>+4.51</sup></b>	<b>54.11<sup>+4.07</sup></b>	<b>55.23<sup>+3.58</sup></b>

Table 5. The comparison with baseline methods on MSVD and NExT-QA datasets. Improvements are highlighted as superscripts.

composition pipeline. The results are summarized in Table 5. By comparing the 1st to 3rd columns with 4th to 6th columns in the table, we could find that our framework, with our automatic decomposition pipeline, significantly boosts the performance of baselines on both MSVD and NExT-QA dataset. Specifically, the overall accuracy improves 3.23% to 4.76% on MSVD, while improves 3.58% to 4.51% on NExT-QA, which indicates that with the help of our question decomposition pipeline, our framework can still raise the performance significantly even on datasets which originally do not have QDGs, further implying the applicability of our framework in real world scenarios.

## 6. Conclusion

In this work, we have focused on the VidQA from interpretability and proposed a model-agnostic align-and-aggregate framework for VidQA. It firstly aligns the video representation towards both main question and sub-questions, then aggregates the video-question joint representation through the QDG. Further, we have revisited the compositional consistency metrics and have proposed more comprehensive c-F scores. Extensive experiments on various VidQA models have revealed that our framework improves both compositional consistency and accuracy significantly, leading to more interpretable VidQA models.

## Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant No. 62076162), the Shanghai Municipal Science and Technology Major/Key Project, China (Grant No. 2021SHZDZX0102).

## References

- [1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. [3](#)
- [2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. [3](#)
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.
- [4] Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, and Liang Lin. Visual question reasoning on general dependency tree. In *CVPR*, pages 7249–7257, 2018. [2](#)
- [5] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011.
- [6] Anoop Cherian, Chiori Hori, Tim K. Marks, and Jonathan Le Roux. (2.5+1)D spatio-temporal scene graphs for video question answering. In *AAAI*, pages 444–453, 2022. [1, 2](#)
- [7] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In *IJCAI*, pages 636–642, 2021. [1](#)
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. [2](#)
- [9] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007, 2019. [1, 2, 6, 7](#)
- [10] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling. In *arXiv:2111.1268*, 2021.
- [11] Tsu-Jui Fu\*, Linjie Li\*, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling. In *CVPR*, 2023.
- [12] Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *CVPR*, pages 5046–5055, 2022. [1, 2, 3, 5, 7](#)
- [13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. [3](#)
- [14] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585, 2018. [1, 2](#)
- [15] Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. Structured two-stream attention network for video question answering. In *AAAI*, pages 6391–6398, 2019. [2](#)
- [16] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *CVPR*, pages 11287–11297, 2021. [3, 7](#)
- [17] Mao Gu, Zhou Zhao, Weike Jin, Richang Hong, and Fei Wu. Graph-based multi-interaction network for video question answering. *IEEE Trans. Image Process.*, 30:2758–2770, 2021. [1, 2](#)
- [18] Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. Multi-scale progressive attention network for video question answering. In *ACL*, pages 973–978, 2021. [1, 2](#)
- [19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [2](#)
- [21] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028, 2020. [1](#)
- [22] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 1359–1367, 2017. [1](#)
- [23] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and Conquer: Question-guided spatio-temporal contextual attention for video question answering. In *AAAI*, pages 11101–11108, 2020. [1](#)
- [24] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116, 2020. [1, 2, 6](#)
- [25] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yuet-ting Zhuang. Multi-interaction network with object relation for video question answering. In *ACM MM*, pages 1193–1201, 2019. [2](#)
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [27] Dohwan Ko, Ji Soo Lee, Miso Choi, Jaewon Chu, Jihwan Park, and Hyunwoo J Kim. Open-vocabulary video question answering: A new benchmark for evaluating the generalizability of video question answering models. In *ICCV*, 2023.
- [28] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2023.
- [29] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, pages 9969–9978, 2020. [1, 2](#)

- [30] Jiangtong Li, Li Niu, and Liqing Zhang. From Representation to Reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR*, pages 21241–21250, 2022. 1
- [31] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueling Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, pages 3032–3041, 2022. 3
- [32] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond RNNs: Positional self-attention with co-attention for video question answering. In *AAAI*, pages 8658–8665, 2019. 2
- [33] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. Interventional video relation detection. In *ACM MM*, 2021. 3
- [34] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. Equivariant and invariant grounding for video question answering. In *ACM MM*, pages 4714–4722, 2022. 1, 2, 3, 4, 7
- [35] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2918–2927, 2022. 1, 3, 4
- [36] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, pages 11235–11244, 2021. 3
- [37] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. HAIR: hierarchical visual-semantic relational reasoning for video question answering. In *ICCV*, pages 1678–1687, 2021. 1, 2
- [38] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge To Answer: Structure-aware graph interaction network for video question answering. In *CVPR*, pages 15526–15535, 2021. 1, 2
- [39] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *ACM MM*, pages 2871–2879, 2021. 1
- [40] Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. Multilevel hierarchical network with multiscale sampling for video question answering. In *IJCAI*, pages 1276–1282, 2022. 1, 2
- [41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 2
- [42] Zi Qian, Xin Wang, Xuguang Duan, Hong Chen, and Wenwu Zhu. Dynamic spatio-temporal modular network for video question answering. In *ACM MM*, pages 4466–4477, 2022. 3
- [43] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.
- [44] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video question answering. In *ACL*, pages 6167–6177, 2021. 1
- [45] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, pages 7026–7035, 2021. 3
- [46] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Trans. Multim.*, 24:3369–3380, 2022. 1
- [47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.
- [48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 1, 2, 8
- [49] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*, pages 2804–2812, 2022. 1, 2, 6
- [50] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, 2022. 1, 2
- [51] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 2
- [52] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueling Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, pages 1645–1653, 2017. 2, 8
- [53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1
- [54] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380, 2018. 2
- [55] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xi- uqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106, 2020. 3