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

Supplementary Material

In this supplementary material, we discuss more details about our new metrics (Section 7). Then, we provide more detail on our automatic question decomposition pipeline (Section 8). Further, we provide more details about our experiments, including the more ablations, improvements of our VA^3 framework conditioned on different composition types and question types, the capability of our framework on more recent VidQA backbones, and more visualized examples and explanations (Section 9). Moreover, we discuss the interpretability in our work (Section 10). Finally, we discuss the potential limitation of our framework along with the future work (Section 11).

7. Further Details on Metrics

7.1. CA-RWR-Delta System

AGQA-Decomp [12] introduced the CA-RWR-Delta system to assess the compositional consistency of VidQA models. For clarity, we refer to Figure 5 to elucidate the concepts and expound on this system. In this context, symbols M^+ and M^-, as well as S_+ and S_-, represent the correctness of the parent questions and all corresponding child questions. Here, the superscript and subscript $+$ denotes '(all) correct', while $-$ signifies '(any) incorrect'. The count of each category among all parent questions is represented by N, accompanied by the relevant superscript and subscript, consistent with the definitions in the main paper. We present the definitions of CA, RWR, and Delta as:

$$\begin{aligned}\mathrm{CA} &= \frac{\mathrm{N}_+^+}{\mathrm{N}_+^+ + N_+^-}, \mathrm{RWR} = \frac{\mathrm{N}_-^+}{\mathrm{N}_-^+ + \mathrm{N}_-^-}, \\ \mathrm{Delta} &= \mathrm{RWR} - \mathrm{CA}.\end{aligned} \tag{14}$$

Comparing Equation (14) with Figure 5, we discern that calculations are conducted horizontally, excluding data conditioned on the correctness of main questions. This horizontal approach can be intuitively perceived as a form of "precision", revealing the potential asymmetry and instability issues associated with using precision as the sole metric, as shown inç. To address this, the main paper recommends incorporating the "recalls" (*i.e.*, vertical computations) into the metric system. Aligned with this notion, we recast CA and $1-$ RWR as cP and NcP, respectively. By expanding them with cR and NcR, which calculate vertically, we subsequently introduce the compositional F-Scores to balance the account for both types of compositional errors for a more comprehensive evaluation.

	M^-	M^+
S_+	N_+^-	N_+^+
S_-	N_-^-	N_-^+

Figure 5. The notation of counters.

7.2. Motivation Behind Our Extension

When introducing new metrics to address limitations in existing ones, it is crucial to determine the necessity of such changes. Empirically, the central query revolves around how much the extreme examples highlighted Table 1 impact the assessments of real-world models. To alleviate such concerns, it is worth noting that the presence of these counterproductive metrics, as observed in our experiments, significantly influenced our decision to perceive existing metrics as precisions and to expand upon them subsequently.

For further insight into the instability issue, kindly refer to Table 7. The row labeled "Longer choose" provides an empirical evidence of the instability in RWR (*e.g.*, 1 - NcP). Upon reverting to the CA-RWR-Delta system, a pronounced RWR discrepancy (approximately 30%) between HME and VA^3(HME) becomes apparent, when the CA (*i.e.*, cP) improvement is much less significant. This might lead one to surmise that the VA^3 framework significantly reduces compositional consistency for this category. However, when observe from the perspective of "recalls", we can find that both model achieves nearly zero recalls. Such circumstance suggests that the RWR value is greatly affected by data imbalances and is therefore unstable, rendering it inaccurate in depicting compositional consistency of models. With the Nc-F_1, we can evaluate and compare the compositional consistency of models more transparently and stably.

Upon deeper analysis, it becomes evident that the instability issue becomes mathematically more pronounced when the proportion of correctly answered sub-questions deviates from 50%—whether it approaches 0% or 100%. Given that the reasoning capabilities of models for various sub-questions and compositional rules might significantly deviate from 50% (as illustrated in Table 6 and Table 7), this instability can undeniably influence the analysis on compositional consistency of models. Furthermore, as VideoQA methodologies advance, this ratio is likely to trend towards 100%. Consequently, the instability and asymmetry are bound to become increasingly prominent for evaluating the VidQA models, regardless of whether conditioned on parent types (and composition types) or not.

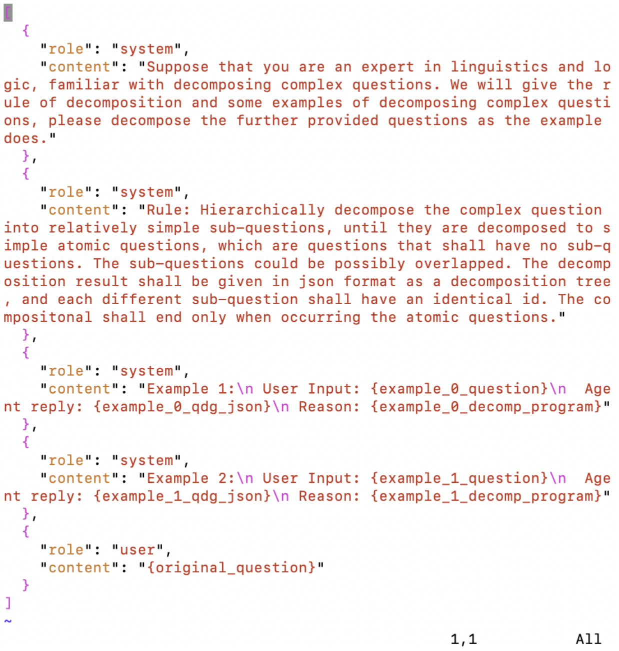
Regarding the issue of asymmetry, it is inherent in the mathematical formulation of any model by the very definition of precision. This is unless the count numbers N_{-}^{-}, N_{-}^{+}, N_{+}^{-}, and N_{+}^{+} happen to be symmetric, an eventuality that is practically improbable. It is essential to understand that this asymmetry naturally arises from the assumption that "the inability (*resp.* ability) of a model to answer a parent question if its corresponding child questions are answered correctly (*resp.* incorrectly)" hurts the compositional consistency of the model equally with "the inability (*resp.* ability) of a model to answer all their sub-questions correct when answering the main question correctly (*resp.* incorrectly)". This is premised on the belief that such discrepancies should be absent if the model is executing proper compositional reasoning. While the original metric offers practical insights, particularly highlighting the failure of models to answer a parent question conditioned on if its child questions are incorrect, it is crucial to recognize that our new metrics do not overlook the utility of the original ones. Indeed, our metrics fully incorporate the insights of the original ones: cP precisely matches compositional accuracy, while NcP equates to $1 - \mathrm{RWR}$. Furthermore, our proposed c-F_β and Nc-F_β metrics allow for a balanced consideration of the two types of compositional errors through parameter β, and these metrics can revert to CA and $1 - \mathrm{RWR}$ as $\beta \to \infty$.

8. Automatic Question Decomposition Pipeline

In this section, we provide more details about our automatic question decomposition pipeline. Our pipeline aims at automatically decompose questions for the VidQA task in high quality with finite financial costs.

LLMs are powerful tools in natural language processing. However, training LLMs to decompose the main questions can be expensive and time-consuming, especially when facing large-scale data ($\sim$ 26M questions and billions of tokens with QDGs and decomposition programs). Instead, large langurage models (LLMs) are shown to perform significantly better with few-shot examples and explanations (*e.g.*, chain-of-think) during prompting [3, 28, 47]. As the number of available main question is large, it would be infeasible to embed all of them into the prompting. Thus, it would be desirable to select few-shot examples in prompting the LLMs to decompose the questions.

Naively, random selection would be a feasible solution. However, given the variety compositional types in the AGQA-Decomp dataset [12] as shown in Table 7, random selection could lead to unstable performance, as the chosen examples may have different composition type with the question to be decomposed (denoted as query question in the following sections), and may have significantly different decomposing program. In that cases, the LLMs may not benefit from the provided examples.



```
[
  {
    "role": "system",
    "content": "Suppose that you are an expert in linguistics and lo
gic, familiar with decomposing complex questions. We will give the r
ule of decomposition and some examples of decomposing complex questi
ons, please decompose the further provided questions as the example
does."
  },
  {
    "role": "system",
    "content": "Rule: Hierarchically decompose the complex question
into relatively simple sub-questions, until they are decomposed to s
imple atomic questions, which are questions that shall have no sub-q
uestions. The sub-questions could be possibly overlapped. The decomp
osition result shall be given in json format as a decomposition tree
, and each different sub-question shall have an identical id. The co
mpositonal shall end only when occurring the atomic questions."
  },
  {
    "role": "system",
    "content": "Example 1:\n User Input: {example_0_question}\n  Age
nt reply: {example_0_qdg_json}\n Reason: {example_0_decomp_program}"
  },
  {
    "role": "system",
    "content": "Example 2:\n User Input: {example_1_question}\n  Age
nt reply: {example_1_qdg_json}\n Reason: {example_1_decomp_program}"
  },
  {
    "role": "user",
    "content": "{original_question}"
  }
]
~
                                                      1,1          All
```

Figure 6. The example of a decomposition prompt template. The {example_*_question}, {example_*_qdg_json}, {example_*_program} and {original_question} shall be replaced with the selected example questions, their QDGs, their programs, and the query question respectively.

To address this issue, we carefully construct a candidate question set to select proper few-shot examples for prompting. Given the fact that questions with similar composition types may have similar decomposing programs, we first cluster the questions based on their composition types. In detail, for each composition type in the $C = 13$ compositional types, we choose $N = 3$ questions with relatively complex, medium, and simple decomposing programs, respectively. These $C \times N$ questions forms the candidate set.

Then, we select the questions which is most similar in terms of compositional structure with the query question from the candidate set. Thus, we ask the LLM to select K questions with the similar compositional type and compositional complexity from the candidate questions. Then, we retrieve the corresponding decomposition graphs of these K questions from the AGQA-decomp dataset in `json` format along with the decompose program as explanations, and send them to the LLM as few-shot examples when asking the LLM to decompose the query question. While larger K gives more flexibility and instruction information, the cost in prompting the LLMs can raise linearly. In balancing the cost and performance, we set $K = 2$ in our experiments. In Fig. 6, we shown an example of a decompose prompt.

Question Type	Question Accuracy		Compositional Consistency											
			cP		cR		c-F_1		NcP		NcR		Nc-F_1	
	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)
Object Exists	49.99	49.94	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
Relation Exists	49.91	50.98	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
Interaction	52.33	57.68	50.07	55.34	92.40	92.90	**64.95**	**69.36**	47.70	48.27	7.00	8.12	**12.21**	**13.90**
Interaction Temporal Loc.	50.73	52.71	61.51	65.41	80.60	73.77	**69.77**	**69.34**	71.27	59.66	48.84	49.87	**57.96**	**54.33**
Exists Temporal Loc.	45.45	47.95	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
First/Last	15.55	16.00	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
Longest/Shortest Action	6.79	5.18	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
Conjunction	50.34	51.82	69.64	67.25	46.27	47.45	**55.60**	**55.64**	59.36	57.51	79.55	75.48	**67.99**	**65.28**
Choose	17.03	22.82	42.70	53.87	20.30	28.56	**27.52**	**37.33**	55.78	56.50	78.68	79.14	**65.28**	**65.93**
Equals	51.60	52.02	55.48	55.23	28.91	29.05	**38.01**	**38.08**	50.35	48.71	75.66	74.10	**60.46**	**58.78**
Overall	16.56	21.56	53.74	58.88	37.40	41.22	**44.10**	**48.49**	56.43	55.55	71.58	71.83	**63.11**	**62.65**

Table 6. The comparison with HME [9] on AGQA-Decomp [12] balanced setting in terms of different question types.

Composition Type	Compositional Consistency											
	cP		cR		c-F_1		NcP		NcR		Nc-F_1	
	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)	HME	VA^3(HME)
Interaction	51.37	56.05	86.85	81.50	**64.55**	**66.42**	54.23	56.22	15.93	27.11	**24.63**	**36.58**
First	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
Last	N/A	N/A	N/A	N/A	**N/A**	**N/A**	N/A	N/A	N/A	N/A	**N/A**	**N/A**
Equals	55.48	55.23	28.91	29.05	**38.01**	**38.08**	50.35	48.71	75.66	74.10	**60.46**	**58.78**
And	85.82	77.76	75.51	72.96	**80.33**	**75.28**	77.67	77.78	87.22	81.94	**82.17**	**79.80**
Xor	37.38	53.51	16.69	30.37	**23.07**	**38.75**	46.26	49.62	71.96	72.22	**56.32**	**58.82**
Choose	46.69	59.65	18.80	29.06	**26.81**	**39.09**	55.77	55.01	82.67	81.52	**66.61**	**65.69**
Longer Choose	26.16	36.79	99.76	99.92	**41.44**	**53.78**	83.33	50.00	0.42	0.05	**0.83**	**0.10**
Shorter Choose	19.99	33.69	99.85	100.00	**33.31**	**50.40**	90.91	100.00	0.38	0.09	**0.76**	**0.18**
After	56.06	54.84	96.12	84.83	**70.82**	**66.62**	86.21	59.62	24.37	24.29	**38.00**	**34.51**
While	53.14	52.79	95.87	85.96	**68.38**	**65.41**	77.61	54.47	14.48	17.94	**24.40**	**26.99**
Before	56.50	57.91	93.72	86.00	**70.50**	**69.22**	77.63	60.58	23.20	25.60	**35.72**	**35.99**
Between	85.08	83.97	87.05	83.48	**86.05**	**83.73**	87.05	86.30	85.07	86.72	**86.05**	**86.51**
Overall	53.07	57.78	46.03	48.67	**49.30**	**52.83**	56.58	55.81	63.34	64.58	**59.77**	**59.88**

Table 7. The comparison with HME [9] on AGQA-Decomp [12] balanced setting in terms of different composition types.

9. Additional Experiments

In this section, we provide more details about our VA^3 framework experiments. The experiment settings are described in Section 9.1. We have also performed extra experiments to test the mprovements by question type and composition type in Section 9.2. We further discussed the capability of our framework on large vision-language pretrained VidQA backbones in Section 9.3. An in-depth ablation study can be found in Section 9.4. Finally, we share more quantitative results in Section 9.5 to show how the video aligner and the answer aggregator enhance the performance and interpretability of VidQA models.

9.1. Experiment Setting

9.1.1 Dataset

AGQA 2.0 [16], which is the backbone of AGQA-Decomp [12] dataset, offers two additional evaluations to assess the generalization capabilities of VidQA models through specific train-test splits. Specifically, the *more composition step* setting includes questions in the training split with fewer compositional steps. It then tests the ability of models to generalize to questions with a higher number of compositional steps. The *novel composition* setting evaluates the capacity of models to generalize to previously unseen composition types. This is done by designating these new types for testing, while excluding them from the training dataset. Consequently, we evaluate our framework under these settings to confirm that it not only enhances performance but also strengthens generalization capabilities.

9.1.2 Training Details

We employ a training batch size of 96 main questions and set the learning rate to 2.5×10^{-4}. The appearance and object features are extracted using a pretrained Resnet-

	Train split		Test split	
	# Videos	# Questions	# Videos	# Questions
MSVD [5]	1,200	31K	770	19,572
NExT-QA [48]	3,870	38K	1,570	14,521
MSVTT [52]	6,513	159K	3,487	85,099
AGQA-Decomp [12]	7,787	26M	1,814	2M
Our Sampled Subset	7,787	480K	1,814	54,125

Table 8. The comparision between some typical VidQA datasets, AGQA-Decomp and our sampleds subset.

	Main Accuracy			Comp. Consistency	
	Open	Binary	**All**	**c-F$_1$**	**Nc-F$_1$**
HME [9]	36.01	51.25	**41.45**	**49.29**	**60.48**
HQGA [49]	41.37	50.73	**44.71**	**44.45**	**59.17**
VIOLETv2 [11]	57.47	56.75	**57.21**	**49.95**	**60.29**
VA^3(HME)	40.01$^{+4.00}$	51.61$^{+0.36}$	**44.15**$^{+2.70}$	**52.85**$^{+3.56}$	**60.85**$^{+0.37}$
VA^3(HQGA)	42.49$^{+1.12}$	51.83$^{+1.10}$	**45.82**$^{+1.11}$	**47.57**$^{+3.12}$	**59.66**$^{+0.49}$
VA^3(VIOLETv2)	58.32$^{+0.85}$	57.34$^{+0.59}$	**57.96**$^{+0.75}$	**53.31**$^{+3.36}$	**60.62**$^{+0.33}$

Table 9. The comparision of accuracy and compositional consistency improvements on our selected subset.

152 [20] with configurations $n_c = 8$, $n_f = 4$, $n_o = 5$, and $h_v = 2048$. Object feature bounding boxes are extracted by a pretrained Faster-RCNN [43]. Motion features are extracted using a pretrained I3D-ResNeXt-101 [19, 51]. During training, we utilize a pretrained BERT [8] to extract question features. The Adam [26] optimizer facilitates model training, with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, when the validation accuracy remain stagnant for 2 epochs, we reduce the learning rate by a factor of 0.5.

9.2. VA^3 Improvement by Question-type and Composition-type

To further reveal how our VA^3 improves the accuracy and consistency, we test the improvements of our framework regarding different question types and different composition types. In Table 6, we compare the improvement of our framework on different question types. Our framework significantly improves the overall accuracy and positive compositional consistencies, while the negative compositional consistency are about the same. Among all the question types, the `Choose` type attains the most improvement on both accuracy and compositional consistency. This could be explained by the natural definition of `Choose` questions, as they are defined as "Compares between two objects, actions, relationships, or time lengths" [12], which naturally requires the combination of information among sub-questions to answer the main question, leading to more ample use of the answer aggregator. Moreover, we noticed that the `Interaction Temporal Localization` questions face a drop on cR, which is defined as $\mathrm{cR} = \frac{\mathrm{N}^{+}_{+}}{\mathrm{N}^{+}_{+}+\mathrm{N}^{+}_{-}}$. Specifically, the N^{+}_{+} and $(\mathrm{N}^{+}_{+} + \mathrm{N}^{+}_{-})$ for HME (*resp.*, VA^3(HME)) are 20692 and 25674 (*resp.*, 21153 and 28674), which indicates the that our VA^3 framework is more effective on main questions than sub-questions of this type. We suspect that the answer aggregator may work better for main questions as they typically requires more information to answer and may benefit more from the intra-QDG information exchange between the questions, especially for the main questions related to multiple vision clues (*i.e.*, `Interaction Temporal Localization`). Therefore, the answer aggregator may cause more correct answers on main question than that on sub-questions, leading to an increased inconsistent between sub-question and main question. This further indicates that the main question may not be able to provide enough guidance for sub-questions, and how to designed a better heuristic for the answer aggregator might be the key for solving this problem.

We further concludes the improvement of our framework regarding different composition types in Table 7. We can also observe that our VA^3 framework significantly increases the positive compositional consistency with a little improvement on negative ones. Though there exists c-F$_1$ decrement on some type of composition types, however, the average decrements (*i.e.*, 3.16%) are significantly smaller than the average increments (*i.e.*, 9.89%). Moreover, these decrements all happens on types with high composition consistency (*i.e.*, the smallest c-F$_1$ that decreases is 68.38%, which is higher than all c-F$_1$s that increase), demonstrating that our models concentrates and improves more on the hard and complex compositional types.

9.3. Vision-Language Pretrained Models

Recently, the vision-language pretrained models [10, 11, 27] have shown impressive performance on VidQA task. However, their compositional consistency remains unexplored. In stressing this issue and verifying the improvement of our VA^3 framework on them, we conduct our experiments on the VIOLETv2 [11]. However, as Table 8 shows, the AGQA-Decomp [12] dataset is over 100 times larger in terms of the number of video-question pair than the typical VidQA datasets which do not contain sub-questions and cannot evaluate the compositional consistency of models. Moreover, these vision-language pretrained models require end-to-end vision-text modeling, which prevents accelerating with precomputed video features. Thus, the original AGQA-Decomp dataset is too large for the vision-language pretrained models to handle. To enable the evaluation of compositional consistency on these vision-language pretrained models, we conduct our experiments on a selected subset of AGQA-Decomp, which has a significant larger scale with MSVD and NExT-QA, while maintaining the distribution AGQA-Decomp dataset. To have the maximal cover on the original AGQA-Decomp dataset and conserve

	Main Accuracy			Comp. Consistency	
	Open	Binary	**All**	**c-F$_1$**	**Nc-F$_1$**
HME	36.29	51.41	**41.59**	**49.29**	**59.76**
VA^3(HME)-NoSubQ	38.92$^{+2.63}$	51.88$^{+0.47}$	**43.48**$^{+1.89}$	**51.36**$^{+2.07}$	**59.80**$^{+0.04}$
VA^3(HME)-Full	39.91$^{+3.62}$	52.26$^{+0.85}$	**44.30**$^{+2.71}$	**52.83**$^{+3.54}$	**59.87**$^{+0.11}$

Table 10. The comparison with baseline methods and fully supervised VA^3 framework when masking out the supervision of all sub questions in AGQA-Decomp.

the original data distribution, we preserve all videos on it and sample the questions for each video while preserving the relative ratio between each composition type. The statistics of sampled result is shown in Table 8, and the result of conducting experiments with existing VidQA models on this sampled subset is shown in Table 9.

For accuracy, compared to the none-pretrained models, there exists a significant improvement, especially for the open questions, as the large scale vision-language pretraining has significantly enlarged the training data of the VidQA model, and may introduce more representation ability, especially for complex objects and actions. Even though, our framework can still improve its accuracy, demonstrating the considerable capability of our framework.

However, in terms of compositional consistency, the vision-language pretrained model does not perform significantly better then traditional ones. This suggests that, even if the vision-language pretraining may introducing more representation ability for the VidQA models, it can hardly increase the compositional reasoning ability of models. This could be reasonable as the extra pretraining data typically does not cover such compositional reasoning tasks, thus may have limited impact on improving the compositional consistency of VidQA models. Despite this, our VA^3 framework can still significantly boost their compositional consistency, as there exists 3.36% c-F$_1$ improvement while the Nc-F$_1$ raises for 0.33%, demonstrating the efficacy and model robustness of our framework in boosting the compositional consistency for various kinds of VidQA models.

9.4. Additional Ablation Study

So far, we have rigorously evaluated the efficacy of our framework in enhancing accuracy, compositional reasoning ability, generalization capability, and interpretability using the extensive AGQA-Decomp dataset. Nevertheless, the influence of sub-question answers during the training of the VA^3 framework remains unexplored. Therefore, we investigated the performance of our framework when all sub-question answers are masked, and related supervisions are omitted. The results are detailed in Table 10.

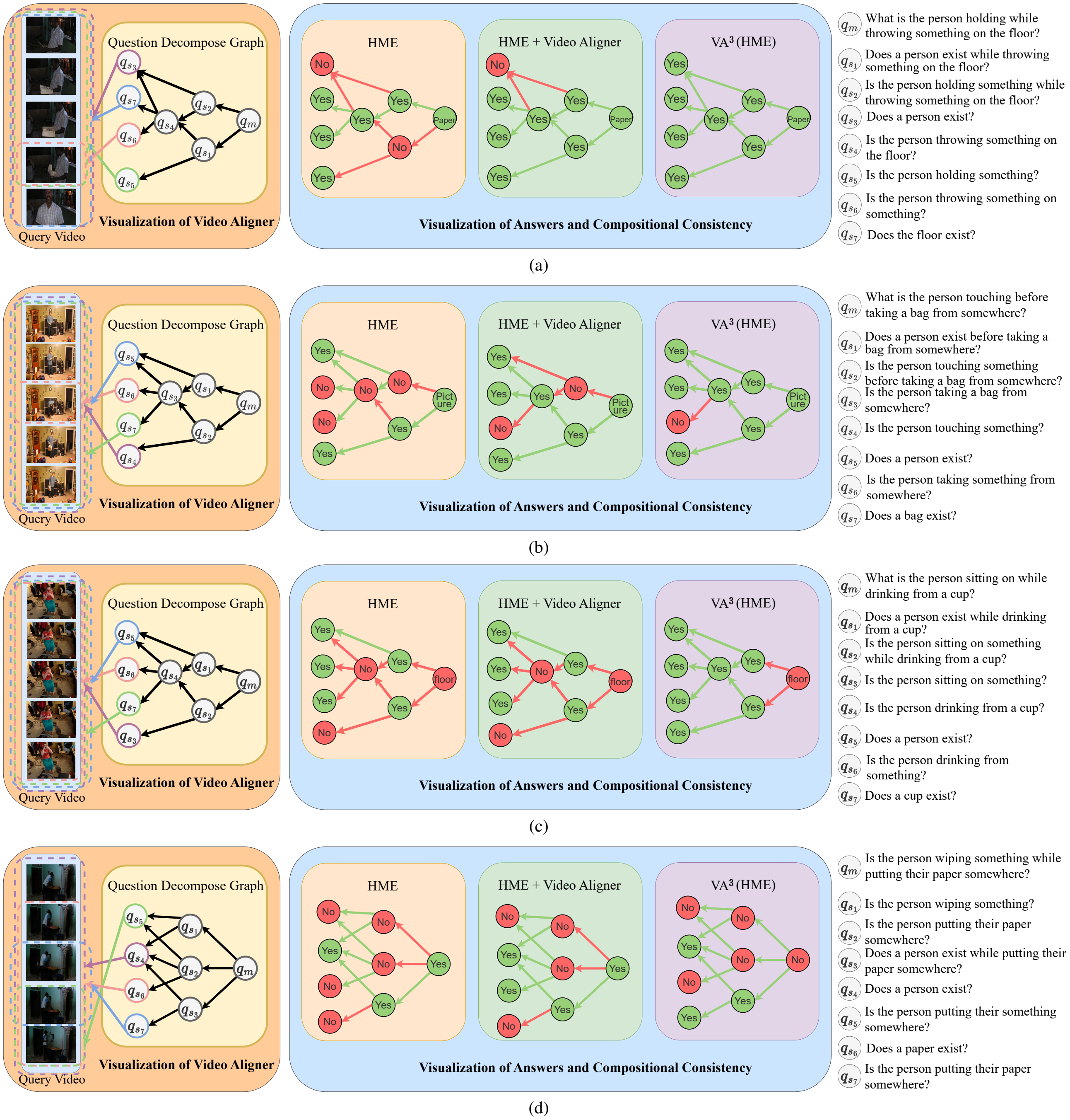
From the table, it is evident that while the gains are not as substantial as with full supervision (row 2 *v.s.* 3), the VA^3 framework still delivers noteworthy improvements in terms of both accuracy and compositional consistency even in the absence of sub-question supervision. This suggests that the inherent structure of our framework, rather than the sub-question supervision, is the primary driver of the observed enhancements. More pointedly, this confirms that the video aligner effectively highlights the most pertinent clips, and the answer aggregator can still consolidates information via the question decomposition graph (QDG), even without guidance from sub-questions.

9.5. Additional Qualitative Results

We give more visualization on the result of video aligner, and the predicted answers of our baseline and framework in Figure 7. Concretely, Figure 7a and Figure 7b quantitatively show how our video aligner and answer aggregator help to improve the accuracy and consistency, while Figure 7c and Figure 7d demonstrate some failure cases on the video aligner and answer aggregator. The result of video aligner is shown as the dotted box

Specifically, in Figure 7a, we can observe that the baseline model fails to answer q_{s_3} and q_{s_1} correctly, leading to inconsistent internal reasoning, as shown in red edges. Our video aligner, as illustrated in the left part of Figure 7a, can help little when the q_{s_3} is incorrectly answered, since the whole video is related to q_{s_3}. However, the sub-questions that are constrained on a certain action (*e.g.*, "throw" in q_{s_1}), can benefit more from the video aligner, since the video aligner can locate related video clips with the action, and then send the video clips to the VidQA model (*i.e.*, HME) to predict the answer, leading to more accurate answer predictions. Moreover, since the information flow within our answer aggregator is bi-directional (*i.e.*, both from the main question to sub-question and from the sub-question to the main question), we are able to correct such isolated errors as q_{s_3} by aggregating the information in video-question joint representations associated with q_{s_2} and q_{s_4} as Figure 7a shows. In Figure 7b, there are a group of incorrect sub-questions (*i.e.*, the related sub-questions q_{s_1}, q_{s_3}, q_{s_6} and q_{s_7}), on where directly apply answer aggregator could face the possibility of mistakenly changing the answer of q_{s_5} from "Yes" to "No". However, with the help of video aligner, we are able to correct some sub-questions (*i.e.*, q_{s_6} and q_{s_3}) by answering questions with more accurate video clips, thus decreases the potential side effects of our answer aggregator, and further leads to the correction of q_{s_1} in answer aggregator. However, such correction failed on q_{s_7}, possibly due to the evidence provided by q_{s_3} is not strong enough to effect such answer.

For the failure cases shown in Figure 7c, the video contains a simple scene with a woman drinking something from a cup repeatedly. For this scene and the listed sub-questions, the whole video is related to all sub-questions, thus the video aligner almost degenerate to a identity mapping, thus cannot fix the incorrect answers. However, with the help of



Query Video
Question Decompose Graph
Visualization of Video Aligner
HME
HME + Video Aligner
VA³ (HME)
Visualization of Answers and Compositional Consistency
Yes
No
Paper
Pict ure
floor
q_m What is the person holding while throwing something on the floor?
q_{s_1} Does a person exist while throwing something on the floor?
q_{s_2} Is the person holding something while throwing something on the floor?
q_{s_3} Does a person exist?
q_{s_4} Is the person throwing something on the floor?
q_{s_5} Is the person holding something?
q_{s_6} Is the person throwing something on something?
q_{s_7} Does the floor exist?
(a)
q_m What is the person touching before taking a bag from somewhere?
q_{s_1} Does a person exist before taking a bag from somewhere?
q_{s_2} Is the person touching something before taking a bag from somewhere?
q_{s_3} Is the person taking a bag from somewhere?
q_{s_4} Is the person touching something?
q_{s_5} Does a person exist?
q_{s_6} Is the person taking something from somewhere?
q_{s_7} Does a bag exist?
(b)
q_m What is the person sitting on while drinking from a cup?
q_{s_1} Does a person exist while drinking from a cup?
q_{s_2} Is the person sitting on something while drinking from a cup?
q_{s_3} Is the person sitting on something?
q_{s_4} Is the person drinking from a cup?
q_{s_5} Does a person exist?
q_{s_6} Is the person drinking from something?
q_{s_7} Does a cup exist?
(c)
q_m Is the person wiping something while putting their paper somewhere?
q_{s_1} Is the person wiping something?
q_{s_2} Is the person putting their paper somewhere?
q_{s_3} Does a person exist while putting their paper somewhere?
q_{s_4} Does a person exist?
q_{s_5} Is the person putting their something somewhere?
q_{s_6} Does a paper exist?
q_{s_7} Is the person putting their paper somewhere?
(d)

Figure 7. Qualitative results of Video Aligner and the visualization of improvements on accuracy and compositional consistency brought by our modules. The green and red nodes represent the correct and wrong answers correspondingly, while green and red edges represent whether the compositional reasoning along these edges is consistent. **Best viewed in color and zoom in.**

answer aggregator, we may correct some failures on sub-questions and increase the compositional consistency. But as for the main question, since the sub-questions can only provide little clues for “which object is the person sitting on”, it also cannot correct the failure on the main question.

As for the failure cases shown in Figure 7d, the video aligner corrects the answer of q_{s_6} but helps little on other sub-questions since their corresponding aligned clips are not that accurate. Moreover, the answer aggregator cannot help much to improve the question accuracy since the interference of false answers are so strong and even mistakenly change the correct answer of main question. However, our answer answer aggregator still improves the compositional consistency even for these incorrect answers.

10. Discussion on Interpretability

In this work, the scope of the interpretability is defined upon "How do the sub-questions assist in answering the main question" in compositional VideoQA. This aspect is pivotal for compositional reasoning, yet it remains overlooked in the majority of existing VidQA models. By addressing this query, our framework enhances the performance and interpretability of existing models in terms of accuracy and compositional consistency.

To shed more light, prior to our research, the prevalent approach among VidQA models was to simply feed the sub-questions alongside the main questions into the training pipeline in parallel. In this configuration, the implicit assumption was that the network would autonomously discern the relationship between sub-questions and the corresponding main question. However, this methodology raises eyebrows, as there is a conspicuous absence of design elements in the network architecture that would promote or elucidate the utilization of such interconnected information. Furthermore, there is a lack of empirical evidence supporting the idea that when sub-questions are processed in tandem with main questions during training, the additional information they provide is effectively harnessed. These observations naturally lead us to pose the ensuing questions:

- Do the sub-questions genuinely augment the main question during the training phase?
- How do the sub-questions facilitate the main question during the inference of answers?

Our VA^3 framework offers qualitative insights, affirming that, within a well-structured framework, sub-questions (potentially with their answers) can enhance the accuracy, consistency, and generalization capability of responses to the main question. Furthermore, our quantitative findings underscore the superior interpretability of our method compared to existing ones, particularly in addressing the second question. The thorough breakdown of these quantitative results, presented in both the main paper and supplementary materials, facilitates a detailed understanding of this query. To achieve this, we deploy a video aligner, elucidating the specific video content that reinforces the relationship between a sub-question and the main question. Moreover, an accompanying answer aggregator interprets which sub-questions are paramount for formulating a response to the primary question. Through these mechanisms, we not only deduce how sub-questions influence the derivation on answer of the main questions but also gain clarity on the rationale behind unsuccessful outcomes. This empowers us to explore: "Why does the model either excel or falter in generating the correct response?" Examples supporting these assertions are detailed in the quantitative results showcased in the main paper and supplementary content. In summation, compared to the existing VidQA models, our VA^3 framework markedly amplifies the interpretability of prevailing methods, equipping us to delve deeper into the decision-making mechanisms of VidQA models.

11. Limitations and Future Work

A limitation of our current approach is the predefined class of examples in the automatic decomposition pipeline. As shown in Sec. 3.4, our decomposition pipeline relies on the prior examples in the AGQA-Decomp [12] dataset. Although it is shown in Sec. 5.6 that this pipeline is capable for various VidQA datasets, its ability when facing the extreme variety of arbitrary input question from real-world users might be unstable in extreme cases. In the future works, as the development of neural language processing techniques, more advanced question decomposition techniques may be explored. In example, using a detailed and comprehensive description on how to decompose questions instead of specific examples in prompting or tuning the large language model may enhance the generalization ability. Studiyng how to construct such description and interact it with large language models in the VidQA scenario could boarder the application of our framework in real-world scenarios.