Cross-modal Causal Relation Alignment for Video Question Grounding

Supplementary Material

1. More Details of CRA

1.1. Compare with previous work

There are two major differences between our method and these works (i.e., SGAE [6] and NExT-OOD [7]).

1) Problem-solving. They address causality but focus on QA accuracy. SGAE uses scene graphs for bias representation, and NExT-OOD employs contrastive loss without considering grammar. Our method explicitly integrates causal intervention in Grounded VideoQA, aligning multimodal features with the grounded video feature v_t to support task objectives.

2) Causality-reasoning. Existing methods rely on manual annotations or pre-trained models for priors, such as temporal labels. In contrast, our automated causal module uncovers true causal effects from limited data. Conventional methods involve complex training, limiting transferability, while our CRA is fully end-to-end and easily adaptable to different models.

Additionally, current works evaluate whether models can effectively distinguish causality by introducing complex biases. Constructing unbiased datasets is challenging and often unrepresentative of real-world scenarios. Therefore, modeling on biased datasets provide a more practical and effective way to validate debiasing methods.

1.2. Proof of Eq.9

$P(a|do(V), do(L))$ can be presented as the following:

$$\begin{aligned}&P(a|do(V), do(L)) = \\ &P(a|do(V), do(L), M = v_t)P(M = v_t|do(V), do(L))\end{aligned} \tag{1}$$

where M is introduced by V and L, while L is deconfounded and there is no back-door path between V and M. Thus, Eq. 1 can be reformulated as:

$$\sum_{v_t} P(a|do(M = v_t), do(L))P(M = v_t|V, do(L)) \tag{2}$$

which the probability of $P(a|do(M = v_t), do(L))$ can be formulated via back-door intervention applied at $M \leftarrow V \leftarrow Z \rightarrow a$ as:

$$\begin{aligned}&\sum_{\hat{v}} P(a|do(M = v_t), do(L), V = \hat{v}))P(V = \hat{v}|do(M = v_t)) \\ &= \sum_{\hat{v}} P(V = \hat{v})P(V = \hat{v}, M = v_t),\end{aligned} \tag{3}$$

where $\hat{v}$ is the feature selected from V to represent the overall distribution of the dataset. Combining Eq. 2 and Eq. 3, we can further calculate Eq. 1 as:

$$\begin{aligned}&P(a|do(V), do(L) = \\ &\sum_{\hat{v}} P(V = \hat{v})P(a|V = \hat{v}, M = v_t, do(L)) * \\ &\sum_{v_t} P(M = v_t|V = v, do(L))\end{aligned} \tag{4}$$

where $*$ is the dot product, and the $\hat{v}$ can be estimated from the clusters center $\widetilde{V}$ of the frame features, embedded by the CLIP model. Additionally, the LCI module explicitly models the linguistic semantic relations through semantic graphs and employs backdoor intervention to isolate the influence of biased pathways, to focus on causal relations rather than superficial features.

1.3. About GSG module

The GSG module generates the key video segment feature v_t by denoising cross-modal attention with an adaptive Gaussian filter, serving as the mediator for the ECI. In ECI, v_t applies front-door intervention to block confounders between the video and the answer, decomposing the causal effect when computing $P(a|do(V))$. GSG's temporal denoising ensures v_t focuses on causally relevant scenes, enhancing visual-language alignment. ECI leverages v_t to quantify intervention effects, improving answer faithfulness (Acc@GQA) and grounding precision (IoP@0.5), validating cross-modal causal consistency.

Moreover, the work [1] depends on manually annotated temporal labels and fixed Gaussian distributions for weighting, which are not well-suited to this task. In contrast, Gaussian attention for grounding in this task is often generated using NG, whereas our approach utilizes adaptive Gaussian filtering to mitigate cross-modal attention noise. This significantly enhances performance, as demonstrated in Table 3 of the paper.

1.4. About CMA module

While Contrastive Learning (CL) has been widely applied in weakly supervised video moment retrieval, our task is fundamentally different. Specifically, VideoQG task involves retrieving video segments based on questions and effectively answering them, which requires more complex textual indexing and ensures causal consistency between video segments and answers. Our CMA, based on the causal model, ensures the alignment of debiased features

	Vid.	Que.	Seg.	Seg. Dur.(s)	Vid. Dir.(s)	Ratio (S./V.)
Train	3,860	34,132	-	-	44.9	-
Val	567	3,358	3,931	7.3	42.2	0.2
Test	990	5,553	6,600	6.7	39.5	0.2
Total	5,417	43,043	10,531	-	-	-

Table 1. Statistics of NExT-GQA dataset.

	Vid.	Que.	Seg.	Seg. Dur.(s)	Vid. Dir.(s)	Ratio (S./V.)
Train	3,032	45,731	45,731	11.5	30.0	0.39
Val	914	7,098	7,098	11.9	30.0	0.40
Test	955	7,377	7,377	11.6	29.7	0.40
Total	4,901	60,206	60,206	-	-	-

Table 2. Statistics of STAR dataset.

rather than merely applying CL in a straightforward manner. From Table 4 of the paper, CMA delivers the most significant improvement. In comparison, Temp[CLIP], which employs simple CL, achieves only a moderate improvement of 0.8 Acc@GQA in the ablation.

2. Datasets Analysis

2.1. NextGQA

NextGQA [4] is a benchmark for the weakly supervised VideoQG task and extends the NextQA [3]. It includes two types of questions: Causal ("why/how"), Temporal ("before/when/after"), and excludes Descriptive ("what/who/where") mostly pertain to global content (e.g., "what event?") or answers can be found almost throughout the whole video (e.g., "where is?"). The dataset contains annotations for 10,531 valid time segments corresponding to 8,911 QA pairs and 1,557 videos, as shown in Table. 1. Most segments are shorter than 15 seconds, with an average duration of 7 seconds, significantly less than the total video length of approximately 40 seconds. These segments occupy an average of only 20% of the full video, and their distribution across the left, middle, and right positions of the video is even.

2.2. STAR

STAR [2] is a situated video question reasoning dataset built with naturally dynamic, compositional, and logical real-world videos, which consists of 4,901 videos, 60,206 questions, and corresponding time segments, as shown in Table 2. The questions are generated programmatically based on situational hypergraphs. Situated reasoning also requires structured situation comprehension and logical reasoning, which is a challenging benchmark for VideoQG models. It features four types of questions: Interaction, Sequence, Prediction, and Feasibility. Video scenes in the dataset are decomposed into hypergraphs containing atomic entities and relationships, such as actions, objects, and interactions.

2.3. Comparison

NextGQA is developed based on NextQA, with its text content also derived from the latter. NExT-QA focuses primarily on causal and temporal reasoning, posing questions like "why" and "how" to explore the sequence and causes of events. In contrast, STAR emphasizes contextual reasoning, involving logical inference based on the context and relationships within the video. While NExT-QA employs multiple-choice and open-ended questions, STAR offers a broader range of question types that require various forms of logical reasoning, such as predicting future actions or assessing the feasibility of events based on the video context.

Method	Acc@GQA↑	Acc@QA↑	Bias Error↓	Unfaithful↓
PH	15.3	59.0	28.5	41.4
CRA	18.2(+2.9)	61.1(+2.1)	27.4(-1.1)	40.0(-1.4)

Table 3. Quantify the de-bias

Method	Text	Acc@GQA	Acc@QA	TIoP@0.5
NG+	Qwen2.5-1.5B	15.0	65.2	21.5
CRA	Qwen2.5-1.5B	16.5	65.4	23.1

Table 4. Ablation of the LLM on NExT-GQA dataset

Furthermore, STAR's questions and answers are generated through automated scripts following standard templates, whereas NextQA relies on human annotations. This suggests that the automatically generated questions and answers in STAR may introduce more systematic and subtler biases. Consequently, as discussed in the main text, our CRA model shows more significant improvements on the STAR dataset compared to its performance on the NextGQA dataset. Additionally, in NextGQA, there are instances where a single QA pair corresponds to multiple time intervals.

2.4. Quantify the de-bias

We employ metrics similar to the Acc@GQA metric. Samples with an IoP < 0.3 can be noted as follows: incorrect answers are denoted bias errors; correct answers are considered unfaithful answers. We effectively reduce bias-induced errors and decrease the occurrence of unfaithful answers, aligning with the improvements observed in the Acc@GQA metric (Table 3).

2.5. LLM ablation experiment

We have addressed this point in the main text from the paper (**Section 4.3.1**). Overall, LLMs possess extensive prior knowledge, and some methods are even trained with detailed temporal annotations, making a fair comparison challenging. Additionally, we conducted experiments using more advanced LLMs as the text model. As shown in Table 4, although the model was not fully adapted to this task due to time constraints, the results still validate the effectiveness of CRA.

Method	Vision	Text	Acc@GQA	Acc@QA	TIoP@0.5
IGV	ResNet	BT	10.2	50.1	18.9
CRA	ResNet	BT	13.2	51.5	23.9

Table 5. Comparison with IGV on NExT-GQA dataset

2.6. Compare with IGV

Most existing work implicitly performs deconfounding, with the effectiveness of their causal modules primarily evaluated using the Acc@VQA metric. Our CRA integrates causal front-door intervention with the Grounded VideoQA task, enabling the performance of the indirectly trained Temporal Grounding to directly quantify the effectiveness of the causal module. Additionally, while IGV constructs a causal model based on scene invariance, our CRA achieves finer alignment through causal intervention compared to the coarse-grained segmentation and recombination of video clips. For a fair comparison, we employed the same backbone as IGV for experimentation, and our method still demonstrated superior performance, as shown in Table 5.

3. Metrics

3.1. Acc@VQA

The Acc@VQA metric evaluates model performance in the VideoQA task. In our experiments, we assessed VideoQA accuracy exclusively on the NextGQA dataset, rather than the NextVQA dataset, as the former is merely a subset of the latter. Furthermore, since the test set of the STAR dataset is evaluated online, we were unable to compute Acc@GQA based on Acc@VQA. Consequently, our evaluation was limited to the validation set of the STAR dataset.

3.2. Acc@GQA

Typically, accuracy in Visual Question Answering (Acc@VQA) represents only the percentage of correctly answered questions. To assess the use of visual evidence, we employ IoU and IoP to evaluate whether the predicted time window aligns with the ground truth. However, evaluating QA and grounding separately does not reveal whether the model correctly infers the answer based on causally relevant video segments. To comprehensively evaluate both aspects, Xiao et al. [4] proposed Grounded QA accuracy (Acc@GQA), which measures the percentage of correctly answered questions where the temporal grounding has an IoP greater than 0.5.

3.3. IoU and IoP

In the Video Temporal Grounding task, the Intersection over Union (IoU) is a crucial metric for assessing the overlap between the predicted time segment and the ground truth segment. Specifically, IoU is calculated using the following formula:

$$IoU = \frac{t_{pred} \bigcap t_{gt}}{t_{pred} \bigcup t_{gt}} \tag{5}$$

t_{pred} represents the predicted time interval, and the ground truth is denoted as t_{gt}. On the timeline, the predicted and actual annotated time segments often do not align perfectly. A higher IoU value indicates greater overlap between the predicted segment and the actual segment, signifying higher model prediction accuracy. The IoU value ranges from 0 to 1, where 1 denotes complete overlap and 0 indicates no overlap at all. The mean Intersection over Union (mIoU) refers to the average IoU value across multiple videos or samples. IoU@0.3 and IoU@0.5 are specific IoU metrics calculated with thresholds of 0.3 and 0.5, respectively. Generally, IoU@0.3 represents the proportion of samples where the IoU value between the predicted and actual time segments is at least 0.3.

Similarly, the Intersection over Prediction (IoP) is another important evaluation metric used to assess the alignment between the predicted time segment and the actual time segment. The IoP is calculated using the following formula:

$$IoP = \frac{t_{pred} \bigcap t_{gt}}{t_{pred}} \tag{6}$$

Unlike IoU, which focuses on the overall overlap between the predicted and actual time segments, IoP emphasizes accuracy within the predicted time segment—specifically, how much of the predicted segment overlaps with the actual annotated segment. The IoP value also ranges from 0 to 1, with a higher value indicating that a larger proportion of the predicted segment correctly matches the actual segment.

4. More Experiments

4.1. NextGQA

As shown in Table 6, a more detailed analysis was conducted based on different question types in the NextGQA benchmark. This benchmark includes two main categories of questions: causal questions, comprising 3,252 examples (58.6% of the total), as mentioned in the dataset analysis section, and temporal questions, comprising 2,301 examples (41.4% of the total). A comparison of these results with the overall performance indicates that the CRA framework achieves superior performance on causal questions. Notably, while Temp[CLIP] and FrozenBiLM achieve identical Acc@GQA scores, Temp[CLIP] exhibits significantly higher IoP@0.5, whereas FrozenBiLM outperforms Acc@VQA. This suggests that larger models, despite leveraging data priors learned from extensive datasets, also introduce more pronounced biases. Nevertheless, the CRA framework significantly mitigates these biases on the FrozenBiLM [5] model when compared to the NG+ method [4].

Additionally, for the temporal question category, CRA achieves the highest Acc@GQA scores across both models. This indicates that CRA demonstrates a higher degree of causal consistency between the retrieved video segments and the answers in the VideoQG task. Regarding the IoP@0.5 metric, the difference between Temp[CLIP] and FrozenBiLM is minimal, suggesting that temporal tasks are less affected by biases introduced during large-scale model pretraining. Consequently, CRA demonstrates robust improvements across various scenarios.

4.2. STAR

In our analysis of the STAR dataset, we categorize the questions into four types: Interaction (2,398 questions, accounting for 33.8% of the total), Sequence (3,586 questions, 50.5%), Prediction (624 questions, 8.8%), and Feasibility (490 questions, 6.9%), as shown in Table 7. It is evident that Interaction and Sequence questions dominate the dataset, comprising over 85% of the questions and significantly influencing the overall performance.

Firstly, for Interaction-type questions, although our approach demonstrates average performance in the Acc@VQA and Acc@GQA metrics, it achieves the best results in the temporal grounding task. This indicates that our model can more accurately locate relevant information when interpreting interactions between people and objects in the video. However, this strong grounding performance does not translate into causal consistency in the answers.

Sequence-type questions stand out, although the proposed method achieves an Acc@VQA score 1.6% lower than FrozenBiLM (NG+) on the Temp[CLIP] model, it surpasses the latter by 6.7% in IoP@0.5 and 2.7% in Acc@GQA. These results highlight the model's exceptional performance in handling temporal reasoning tasks, demonstrating a superior ability to capture the sequence and logic of events. This leads to a deeper understanding of video content and a high degree of causal consistency.

For predictive questions, overall performance is slightly better than that for sequential questions. As shown in the table, the performance of FrozenBiLM consistently surpasses that of Temp[CLIP], including in the IoP@0.5 metric. This suggests that larger-scale models exhibit stronger reasoning capabilities for predicting future events, a benefit derived from the prior knowledge embedded in their training data.

Furthermore, this perspective is further validated in feasibility-related questions. Such questions are characterized by their diversity and complexity, involving not only directly observable information from videos but also implicit conditions and assumptions. These questions typically require a deep understanding of the video context and the ability to infer whether a given scenario is plausible in real life. This often demands sophisticated logical reasoning and consideration of multiple factors. For instance, a question might require the model to determine the feasibility of an action under specific conditions, necessitating not only an understanding of the video content but also reasoning about underlying physical principles and commonsense knowledge. The inherent difficulty of these questions explains why the large-scale FrozenBiLM model performs best in this category. Notably, with the enhancement of the CRA framework, FrozenBiLM achieves an impressive IoP@0.5 score of 41.8%. This finding motivates further development of the CRA framework with even larger models to enhance its capability to handle such complex reasoning tasks.

5. Visualization of CRA on NextGQA dataset

As shown in Figure 1, we present the visualization results of CRA on the NextGQA test set. In Figure 1(a), the question belongs to the Temporal category. After removing Gaussian smoothing, the attention weights exhibit significant oscillations along the temporal axis, preventing the model from effectively estimating the intervals. However, the proposed GSG module successfully mitigates these noise effects, enabling the accurate localization of relevant intervals, thereby improving both IoP@0.5 and IoU@0.5 performance. Nevertheless, the Temp[CLIP] model, despite these enhanced weights, still fails to provide a correct answer to this question. In contrast, the FrozenBiLM model delivers the correct answer by relying solely on the final frame. From the attention weights, it is evident that the model is highly confident, producing a single narrow peak and identifying a short temporal interval. This behavior highlights the inherent bias of large-scale models.

Similarly, as shown in Figure 1(b), our method effectively identifies relevant intervals and correctly answers causal questions. However, it is notable that while FrozenBiLM also answers correctly, it confidently attends to incorrect visual information. This further confirms the more severe spurious correlations introduced by data biases in large models. Additionally, comparing the provided ground truth reveals that our method is not entirely incorrect. The video segment estimated by CRA is sufficient to support the answer, while the ground truth interval appears unnecessarily redundant. This observation underscores the greater importance of IoP@0.5 compared to IoU@0.5, as the task prioritizes the precision of interval estimation.

6. Visualization of CRA on STAR dataset

As illustrated in Figure 2, we present the visualization of CRA on the STAR dataset. As mentioned earlier, Figure 2 (a) depicts a scenario where a man is organizing a wardrobe, an activity that spans the entire video. The ground truth

Que.Type	Method	Model	Acc@GQA	Acc@VQA	IoP@0.5	IoU@0.5
Causal	NG+	Temp[CLIP]	18.2	60.3	29.5	9.8
	NG+	FrozenBiLM	17.6	70.9	23.3	8.1
3,252	CRA	Temp[CLIP]	**20.2**	60.7	**31.9**	**10.8**
58.6%	CRA	FrozenBiLM	**20.2**	**71.3**	27.4	10.0
Temporal	NG+	Temp[CLIP]	12.7	60.5	20.9	8.4
	NG+	FrozenBiLM	14.1	68.6	19.6	8.3
2,301	CRA	Temp[CLIP]	15.4	61.8	23.7	**10.3**
41.4%	CRA	FrozenBiLM	**16.8**	**68.9**	**23.8**	9.0
Total	NG+	Temp[CLIP]	15.9	60.2	25.9	9.2
	NG+	FrozenBiLM	16.1	69.9	21.8	8.2
	CRA	Temp[CLIP]	18.2	61.1	**28.5**	**10.6**
	CRA	FrozenBiLM	**18.8**	**70.3**	25.9	9.6

Table 6. Comparison with state-of-the-art methods on NextGQA test set. We train the Temp[CLIP](NG+) and FrozenBiLM(NG+) models on the NextGQA dataset via the official code.

Que.Type	Method	Model	Acc@GQA	Acc@VQA	IoP@0.5	IoU@0.5
Interaction	NG+	Temp[CLIP]	12.5	52.3	23.6	5.6
2,398	NG+	FrozenBiLM	13.4	54.3	22.8	7.3
33.8%	CRA	Temp[CLIP]	14.1	53.3	**25.5**	**7.5**
	CRA	FrozenBiLM	**14.8**	**54.7**	25.4	5.6
Sequence	NG+	Temp[CLIP]	32.0	59.5	53.2	4.2
3,586	NG+	FrozenBiLM	32.2	62.5	50.6	**8.2**
50.5%	CRA	Temp[CLIP]	**34.9**	60.9	**57.3**	3.8
	CRA	FrozenBiLM	34.0	**62.9**	52.1	4.0
Prediction	NG+	Temp[CLIP]	33.2	61.5	51.9	5.3
624	NG+	FrozenBiLM	37.5	64.8	57.9	**10.4**
8.8%	CRA	Temp[CLIP]	36.5	62.8	56.9	4.6
	CRA	FrozenBiLM	**40.9**	**65.9**	**60.6**	7.2
Feasibility	NG+	Temp[CLIP]	16.1	61.5	28.4	4.2
490	NG+	FrozenBiLM	21.6	**66.1**	33.9	7.1
6.9%	CRA	Temp[CLIP]	17.6	63.3	27.8	**8.6**
	CRA	FrozenBiLM	**25.3**	64.1	**41.8**	7.6
Total	NG+	Temp[CLIP]	24.4	57.3	41.4	4.7
	NG+	FrozenBiLM	25.8	60.1	40.9	**7.8**
	CRA	Temp[CLIP]	26.8	58.6	**44.5**	5.5
	CRA	FrozenBiLM	**27.5**	**60.5**	43.1	5.1

Table 7. Comparison with state-of-the-art methods on STAR val set because the Acc.@GQA metric can not be calculated on the private test set.

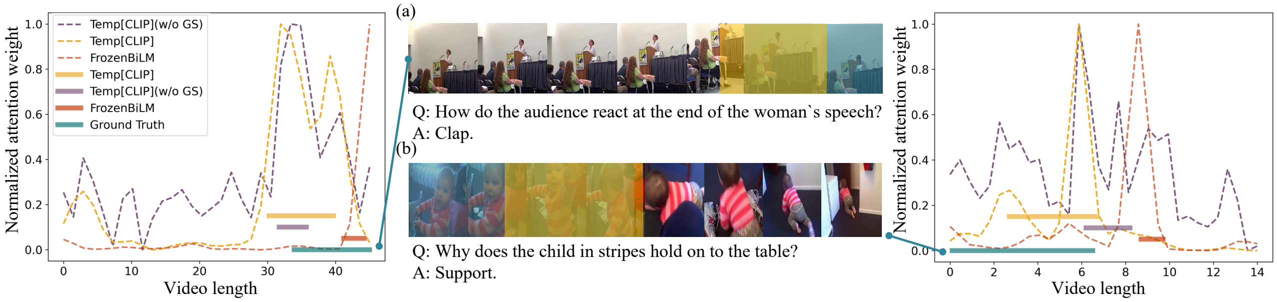


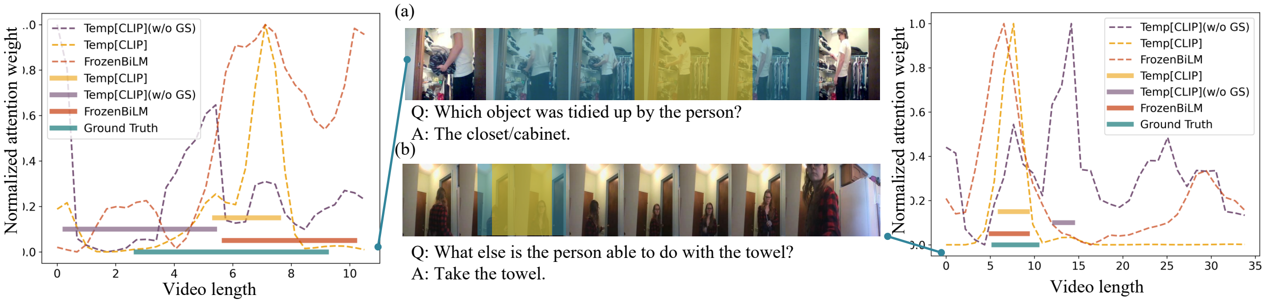
Figure 1. Visualization examples in NextGQA dataset. The numbers ([start time, end time]) indicate the interval range.

segment is centered within the video, which is a reasonable choice. However, the video segments adopted by various methods appear sufficient to serve as the basis for answering the question. This suggests that the dataset may contain some annotation noise and that the evaluation methods could have certain limitations. On the other hand, in the example of a Feasibility-category question as shown in Figure 2 (b), the effectiveness of Gaussian smoothing is reaffirmed. This approach effectively suppresses noise and facilitates better multi-modal alignment.

References

[1] Ran Cui, Tianwen Qian, Pai Peng, Elena Daskalaki, Jingjing Chen, Xiaowei Guo, Huyang Sun, and Yu-Gang Jiang. Video moment retrieval from text queries via single frame annotation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1033–1043, 2022. 1

[2] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum,



Q: Which object was tidied up by the person?
A: The closet/cabinet.

Q: What else is the person able to do with the towel?
A: Take the towel.

Figure 2. Visualization examples in STAR dataset. The numbers ([start time, end time]) indicate the interval range.

and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 2

[3] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2

[4] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 2, 3, 4

[5] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 4

[6] Xu Yang, Hanwang Zhang, and Jianfei Cai. Auto-encoding and distilling scene graphs for image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2313–2327, 2020. 1

[7] Xi Zhang, Feifei Zhang, and Changsheng Xu. Next-ood: Overcoming dual multiple-choice vqa biases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 1913–1931, 2023. 1