Improving Video Understanding through Reliable Question-Relevant Frame Localization and Spatial Guidance

Anonymous ACL submission

Abstract

Video Question Answering (Video QA) is a challenging task that requires models to accurately identify and contextualize relevant information within abundant video contents. Conventional approaches attempt to emphasize related information in specific frames by considering the visual-question relationship. However, the absence of ground-truth of causal frames makes such a relationship can only be learned *implicitly*, leading to the "misfocus" issue. To address this, we propose a novel training pipeline called "Spatial distillation And Reliable Causal frame localization", which leverages an off-the-shelf image QA model to make the video QA model better grasp relevant information in temporal and spatial dimensions of the video. Specifically, we use the visual-question and answer priors from an image QA model to obtain pseudo ground-truth of causal frames and *explicitly* guide the video QA model in the temporal dimension. Moreover, due to the superior spatial reasoning ability of image models, we transfer such knowledge to video models via knowledge distillation. Our model-agnostic approach outperforms previous methods on various benchmarks. Besides, it consistently improves performance (up to 5%) across several video QA models, including pre-trained and non pre-trained models.

1 Introduction

Video question answering (Video QA) is an important field of research that requires machines to identify occurrences or events such as scenes, objects, temporal relationships, and causality in videos. This task poses a critical challenge as videos often contain a wealth of information that is sparsely distributed, requiring machines to comprehend questions and correctly locate relevant information in order to provide accurate answers. Recently, researchers have developed modules to encourage machines to focus on frames crucial to answer the question, which we term as "*causal frames*" (Li et al., 2022b). Existing strategies typically implicitly acquire this knowledge merely relying on the interaction between video and question without direct training objective, as manually annotating frame-by-frame causal information for each video is costly and impractical. Specifically, they use either soft probability to focus on frames inside the attention layers (Fu et al., 2021; Luo et al., 2020; Piergiovanni et al., 2022; Wang et al., 2022; Zellers et al., 2021; Li et al., 2020; Yang et al., 2021) or hard selection mechanisms to train the video QA model by selected frames (Li et al., 2022b,a; Buch et al., 2022).

Despite the advancements made by these methods, video QA models still face a significant issue we refer to as "*misfocus*" – focusing on irrelevant or useless regions for answering a question – especially when the critical clue to causal frames is absent in the question. This is particularly prevalent in questions that require an understanding of temporal relationships or causality, as shown in Figure 1 (a). In this temporal-related example, the question prompt only refers to the object "ornament" and the action "put ... on the tree". Thus, in Figure 1 (b), previous methods (Li et al., 2022b,a), which rely on attention scores between the question and visual features, often prioritize the first two frames. However, to properly answer this question, the machine needs to focus on the last two frames showing a man patting a baby's back, which is not explicitly mentioned in the question and thus is difficult to be learned implicitly by existing methods. This example highlights that video QA models can barely locate useful spatial and temporal regions in a video without the ground-truth causal frames information. Quantitative evidence illustrating this issue is provided in Section 4.1.1.

To overcome this issue, we propose "Spatial distillation And Reliable Causal frame localization" (SpARC), a novel training strategy designed to encourage video QA models to focus on relevant parts

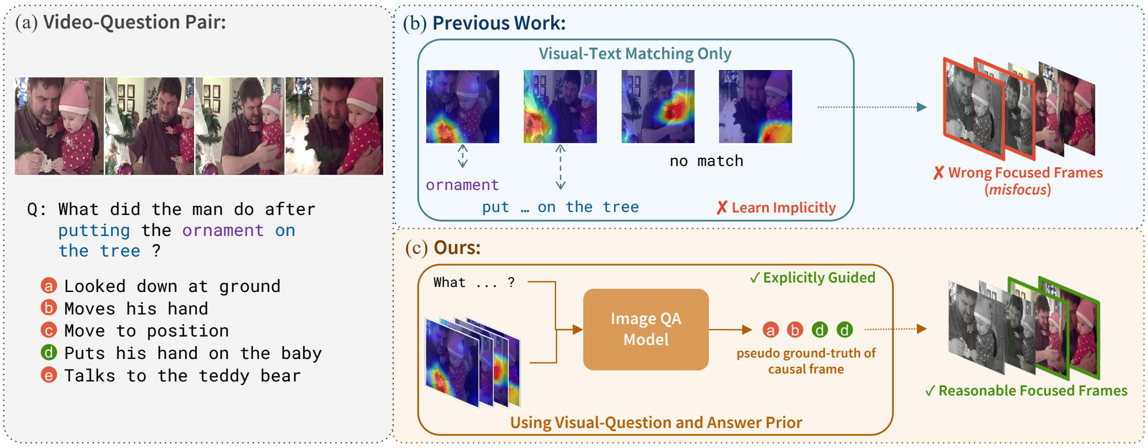

Figure 1: **Comparison of (b) Prior works and (c) Our works.** (a) An example that requires understanding of temporal relationship. (b) Previous work (Li et al., 2022b,a) focused on specific frames by interaction of video and question through implicit learning, as ground-truth causal frames are not available. This approach often led to models focusing on incorrect frames. (c) In contrast, our approach uses the visual-question and answer prior in image QA model to generate reliable pseudo ground-truth of causal frames. This approach directly guides the video QA model to focus on question-relevant frames, avoiding the "misfocus" issue.

of a video. We leverage the knowledge within an off-the-shelf image QA model (Li et al., 2021) to provide pseudo ground-truth of causal frames, which can be used as an explicit signal to direct the video QA model to better locate relevant information in the video during training. As the knowledge of causal frames is related to the temporal dimension of the video, we refer to it as "temporal guidance". In Figure 1 (c), for the last two frames, a well-trained image QA model would predict "d" as it is the only choice related to the input. The frames that lead to the correct answer can be considered crucial to answering the question, therefore, the pseudo ground-truth of causal frames. This explicit information can then be used to provide temporal guidance to the video QA model. Notice that such guidance is only used in training phase; model would process the entire video during inference. For more examples and practical predictions of image QA model (Li et al., 2021), please refer to Figure 3.

In addition to temporal guidance, we also leverage the property that image models have superior spatial comprehension capabilities (compared to video models) (Lin et al., 2022; Kae and Song, 2020; Li et al., 2017; Lee et al., 2022). We treat the predicted probabilities from an image model as its spatial knowledge and then distill it to the video model. Specifically, when the video QA model is fed with a single image, it is expected to make a similar prediction as the image QA model. This approach offers "spatial guidance", making the video QA model better attending to important spatial features. By integrating both spatial and temporal guidance, SpARC enhances the model's ability to comprehend videos and questions. Besides, unlike methods (Arnab et al., 2021; Chen et al., 2022; Ding et al., 2022) using modules or layer interactions to handle spatial-temporal information, we decouple the spatial and temporal approaches, making our method model-agnostic.

We illustrate the effectiveness of SpARC on several video QA benchmarks, including NExT-QA (Xiao et al., 2021), its ATP-hard subset (Buch et al., 2022), and AGQA2.0 (Grunde-McLaughlin et al., 2022), which all require both spatial and temporal understanding to answer questions accurately. We also show the broad applicability of SpARC by presenting consistent improvement (up to 5%) on different video QA architectures, including HGA (Jiang and Han, 2020) and VGT (Xiao et al., 2022b). Furthermore, when integrated into pre-trained video-language models, our method still demonstrates its efficacy, whereas previous model-agnostic work doesn't show such success.

To summarize our contributions: (*i*) We address the "misfocus" issue by leveraging the video-question and answer priors in the image QA model to provide explicit causal frame guidance to the video QA model. (*ii*) Our model-agnostic approach enhances models' spatial-temporal compositional reasoning ability by providing spatial and temporal guidance from an image QA model during training. (*iii*) Our method achieves superior performance on various video QA benchmarks. Additionally, in contrast to previous model-agnostic work that shows inferior performance on pre-trained models, SpARC has broader applicability by demonstrating improvement in both pre-train and non pre-trained video QA models.

2 Related Work

2.1 Image Question Answering

In recent times, visual-language (VL) tasks have received significant attention, with image question answering (image QA) (Antol et al., 2015) being a notable task as it requires reasoning about both textual comprehension and the understanding of relative spatial information. In contrast to video-based tasks, image QA task places a greater emphasis on fine-grained spatial reasoning ability, thereby necessitating stronger spatial understanding ability.

Early work (Anderson et al., 2018; Santoro et al., 2017; Norcliffe-Brown et al., 2018; Cadene et al., 2019; Li et al., 2019a) extracted visual features by object detection backbones such as Faster R-CNN (Ren et al., 2015) and used graph-based or simple cross-attention approaches to model object interactions and improve reasoning capability. Recent work (Li et al., 2019b, 2021; Bao et al., 2022; Kim et al., 2021; Gan et al., 2020; Tan and Bansal, 2019) incorporated transformer (Vaswani et al., 2017) architecture to model the interactions between visual and language information. These models utilized cross-modal pre-training objectives such as image-text matching (Li et al., 2019b, 2021; Bao et al., 2022; Kim et al., 2021; Gan et al., 2020; Tan and Bansal, 2019), word-patch alignment (Kim et al., 2021; Gan et al., 2020), and masked object prediction (Tan and Bansal, 2019) to guarantee that model can handle both semantic understanding of the question and spatial information in the image correctly.

Regardless of the approach used (object-based or transformer-based), the goal of prior research is to ensure the accurate comprehension of relationships between objects relevant to the posed questions. This results in promising spatial understanding abilities among existing image QA models.

2.2 Video Question Answering

Video question answering (video QA) (Zhong et al., 2022; Xiao et al., 2021; Grunde-McLaughlin et al., 2022) is also a highly challenging task among all vision-language tasks. This is because video QA necessitates both contextual comprehension of the posed question and spatial-temporal compositional reasoning capability of the given video. To tackle this task, prevalent strategies used either graph neural network (GNN) (Jiang and Han, 2020; Xiao et al., 2022b,a; Peng et al., 2021; Guo et al., 2021; Seo et al., 2021) or transformer (Fu et al., 2022, 2021; Luo et al., 2020; Piergiovanni et al., 2022; Wang et al., 2022; Zellers et al., 2021; Li et al., 2020; Yang et al., 2021, 2022) to achieve such reasoning ability.

GNN-based approaches constructed graphs based on objects (Peng et al., 2021; Liu et al., 2021; Seo et al., 2021), frames (Jiang and Han, 2020; Liu et al., 2021; Guo et al., 2021), or clips (Jiang and Han, 2020; Xiao et al., 2022a,b) to handle the relationships between visual features and textual cues. And transformer-based approaches would employ a range of pre-training methods such as video-caption matching (Fu et al., 2022, 2021; Luo et al., 2020; Piergiovanni et al., 2022; Wang et al., 2022), locating captions to video segments (Zellers et al., 2021; Li et al., 2020), masked visual matching (Fu et al., 2022, 2021; Luo et al., 2020), or even direct pre-training on transformed question-answer pair data (Yang et al., 2021, 2022). These dedicated pre-trained objective are designed to enhance models' abilities for compositional reasoning.

However, recent research indicated that many existing work relied on superficial correlations between video-question pairs and answers (Li et al., 2022b,a). Some models even performed worse than answering questions by a single frame, as shown in (Buch et al., 2022; Lei et al., 2022). Moreover, a recent study (Lee et al., 2022) discovered that even pre-trained models struggle with correctly handling temporal information in videos. These findings suggest that models may not acquire knowledge from the correct regions during learning phase. We thus propose using priors in the image model to guide the video model in better locating spatial-temporal information from videos.

3 Method

Due to the inaccessibility to the ground-truth of causal frames, previous video QA work often encounters the "misfocus" issue, where machine erroneously focuses on irrelevant spatial and temporal contents. To address this, we propose a novel training pipeline: "Spatial distillation And Reliable Causal frame localization" (SpARC), which integrates spatial and temporal (causal frame localization) guidance to the video QA model. SpARC has two main steps. First, we extract causal frame prior and spatial knowledge from a well-trained image QA model (Section 3.1) for subsequent guidance. Second, we use this knowledge to guide video QA model during training. To provide temporal guidance, we use the causal frame knowledge to supply explicit indication of causal frames (Section 3.2.1). For spatial guidance, we distill the spatial knowledge to enhance video QA model's spatial reasoning capability (Section 3.2.2). An overview of the pipeline is shown in Figure 2. Note that our method only provides guidance in training phase. During inference, the video QA model would process the entire video without any explicit signal.

3.1 Extraction of Causal Frame Prior and Spatial Knowledge

To acquire knowledge from the well-trained image QA model, we feed each video frame into the image QA model $M_{\mathcal{I}}$ and use the resulting predictions as causal frame prior and spatial knowledge. For a given video-question pair, we input each frame (image) $\mathcal{I}_k$ and the question $\mathcal{Q}$ to obtain predicted probabilities for each answer candidate $p_k = M_{\mathcal{I}}(\mathcal{I}_k, \mathcal{Q})$. The predictions of all frames $\{p_1, p_2, ..., p_n\}$ are then used to identify causal frames and provide spatial guidance during the training phase of the video QA model. In open-ended QA datasets, it's typical to convert them to multi-choice QA by creating a global answer set. The answer candidates are collected from all answers in training data that appear more than once (Yang et al., 2021). Hence, we can still get the predicted probability in open-ended datasets.

3.2 Spatial-Temporal Guidance

3.2.1 Temporal Guided

As illustrated in Section 1, we can use the image QA model's prediction to generate pseudo ground-truth of frames that are crucial for answering the given question. In the following part, we will describe how our approach incorporates these causal frames in training phase of video model and handle the unavailability of ground-truth labels during inference in detail.

Guided Prediction. We use the predicted probability of the correct answer as a measure of the likelihood that the respective frame is a causal frame. Specifically, given image QA predictions $\{p_1, p_2, ..., p_n\}$ with corresponding features $\{f_1, f_2, ..., f_n\} = \mathcal{V}$ extracted from frames $\{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n\}$, we use a threshold t to determine whether a frame should be considered a causal frame. Let p_{ki} be the predicted probability of the k-th frame for the i-th answer candidate, and let the correct answer be the a-th answer candidate. The causal portion of the video input, denoted as $\mathcal{V}_c$, would be $\mathcal{V}_c = \{f_k \mid p_{ka} > t, \ \forall k = 1...n\} \subseteq \mathcal{V}$. We then input $\mathcal{V}_c$ into video QA model $M_{\mathcal{V}}$ to obtain "Guided Prediction" $M_{\mathcal{V}}(\mathcal{V}_c, \mathcal{Q})$.

We use "Guided Prediction" to ensure that model learns the reasoning capability by only relevant frames, thus avoiding misfocus issue and enhancing model's performance. To achieve this, we optimize "Guided Prediction" to the ground-truth $\mathcal{A}$ by cross-entropy:

$$\mathcal{L}_g = CrossEntropy(M_{\mathcal{V}}(\mathcal{V}_c, \mathcal{Q}), \ \mathcal{A}).$$

Consistency. In the previous part, we made the video QA model perform well when providing causal frames guidance. However, during inference, such explicit guidance is not available due to the lack of ground-truth answers. Consequently, the model can only perceive the entire video features $\mathcal{V}$, which may contain irrelevant frames. To ensure model's performance under this situation, we aim to make the prediction of entire video (called "Whole Video Prediction") $M_{\mathcal{V}}(\mathcal{V}, \mathcal{Q})$ be consistent with "Guided Prediction" $M_{\mathcal{V}}(\mathcal{V}_c, \mathcal{Q})$. We achieve this by minimizing the Kullback-Leibler divergence between these two predictions, which we term as consistency loss $\mathcal{L}_c$:

$$\mathcal{L}_c = KL(M_{\mathcal{V}}(\mathcal{V}, \mathcal{Q}), \ M_{\mathcal{V}}(\mathcal{V}_c, \mathcal{Q})).$$

This helps the model learn to identify and give less focus on irrelevant frames; thus ensures promising performance during inference and also improves model's temporal robustness.

3.2.2 Spatial Guided

Besides providing temporal guidance, we also utilize the superior spatial understanding of image model (compared to video model) (Lin et al., 2022; Kae and Song, 2020; Li et al., 2017). Our goal is to distill such spatial knowledge from the image model to the video model. In practice, it starts with sending a randomly selected encoded feature f_i from frame $\mathcal{I}_i$ to the video QA model, which means that the model can only perceive spatial information. Given the same frame-level inputs,

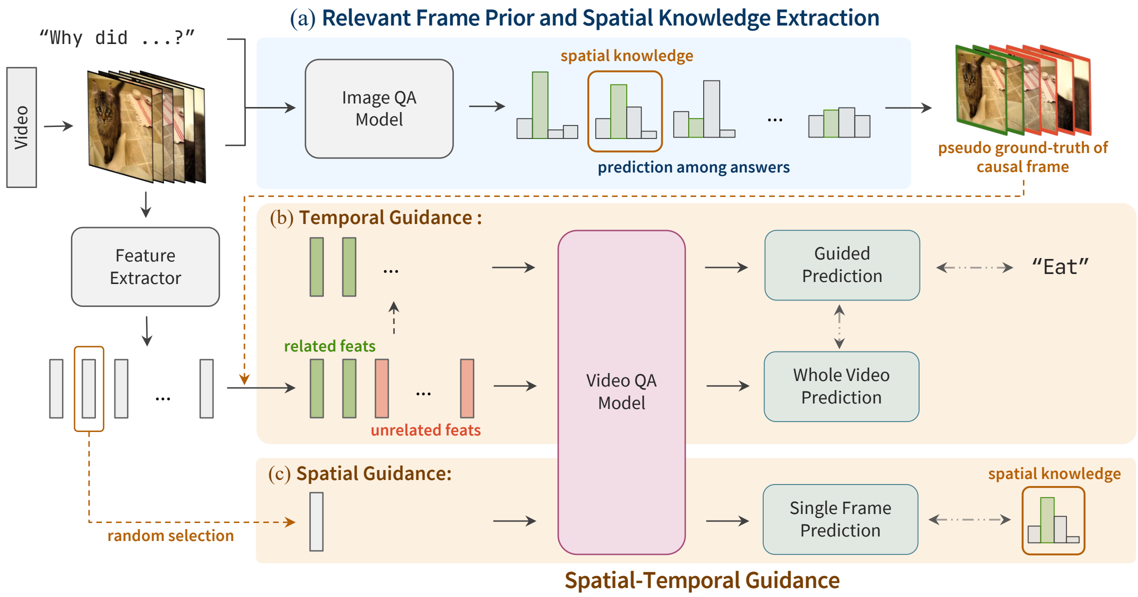

Figure 2: **Spatial distillation And Reliable Causal frame localization (SpARC).** Our novel training method (SpARC) provides spatial-temporal guidance to the video QA model. (a) We use an off-the-shelf image QA model to extract knowledge of causal frames and spatial understanding (Section 3.1). (b) For temporal guidance (Section 3.2.1), we use the pseudo ground-truth to identify causal frames and use them to obtain "Guided Prediction". Additionally, we use "Whole Video Prediction" to ensure temporal consistency of the prediction. (c) For spatial guidance (Section 3.2.2), we randomly select a frame to obtain "Single Frame Prediction" that should be similar to the image QA model's prediction for that frame.

the video QA model's prediction (called "Single Frame Prediction") $M_{\mathcal{V}}(\{f_i\}, \mathcal{Q})$ is expected to be similar to that of image QA teacher $M_{\mathcal{I}}(\mathcal{I}_i, \mathcal{Q})$. To successfully distill the spatial knowledge, we optimize video QA model by the following spatial-distillation loss $\mathcal{L}_s$:

$$\mathcal{L}_s = CrossEntropy(M_{\mathcal{V}}(\{f_i\}, \mathcal{Q}),\ M_{\mathcal{I}}(\mathcal{I}_i, \mathcal{Q})).$$

3.2.3 Training Objectives

We only train the video QA model and freeze other parts (*e.g.* , the feature extractor) in training phase. The final optimization target $\mathcal{L}_{total}$ combines the above three training targets and is represented as follows, where w_c and w_s are hyper-parameters standing for the weights of the consistency loss and spatial-distillation loss:

$$\mathcal{L}_{total} = \mathcal{L}_g + w_c \cdot \mathcal{L}_c + w_s \cdot \mathcal{L}_s.$$

4 Experiments

4.1 Preliminary

4.1.1 Quantification Result for Misfocus

To quantitatively illustrate the extent of the "*mis-focus*" issue in previous work, we conducted a small pilot experiment on the AGQA2.0 benchmark (Grunde-McLaughlin et al., 2022). For quantification purposes, we employed the previous hard selection work, IGV (Li et al., 2022b), as it allows for easier verification of changes in the selection of causal frames.

Specifically, we focused on questions containing the terms "before" or "after" and interchanged these terms ("before" to "after" and vice versa). If preceding work correctly captured question-relevant information in video, the chosen causal frames should have differed due to the shift in temporal emphasis. However, our findings reveal that 74.87% of instances had identical predicted causal frames. This provides compelling evidence that prior work is insensitive to questions involving temporal information and tend to focus on irrelevant video segments.

4.1.2 Capability of Image QA Model

To validate the reliability of causal frames provided by the image QA model, we visualize the predictions from ALBEF (Li et al., 2021) as shown in Figure 3. Examples (a) and (b) pertain to questions that necessitate an understanding of temporal relationships and causality, where the misfocus issue tends to occur. Our results demonstrate that the accurate predictions from the image QA model align well with the causal frames. Additionally, we showcase a qualitative result for a descriptive question (Figure 3 (c)), where the image QA model's

Figure 3: **Visualization of the image QA model's prediction.** We present examples of (a) temporal, (b) causal and (c) descriptive questions. In each example, the grayed-out frames represent non causal frames verified by humans. The prediction of the image QA model is shown below the image. These examples demonstrate that predictions of the image QA model can effectively guide the video QA model to focus on causal frames.

prediction remains satisfactory.

The qualitative results indicate that the image QA model can serve as a trustworthy pseudo ground-truth provider, supplying guidance for causal frames to the video QA model. Such guidance can solve the issue that previous approaches focused on unrelated frames due to a lack of ground-truth of causal frames, leading to a skeptical understanding of the video.

4.2 Settings

We present the benchmarks, video backbones, and settings employed in the subsequent sections. Detailed implementation settings such as hyperparameters used in training phase and time consume will be elaborated upon in the supplementary material.

4.2.1 Benchmarks

We evaluate the capability of SpARC by multiple video QA benchmarks: NExT-QA (Xiao et al., 2021), its ATP-hard subset (Buch et al., 2022), and AGQA2.0 (Grunde-McLaughlin et al., 2022). NExT-QA is a multi-choice benchmark that assesses videos' spatial, temporal, and descriptive aspects. The ATP-hard subset of NExT-QA contains spatial and temporal questions that have been manually verified to require information from multiple frames to answer correctly. AGQA2.0 is a large open-ended benchmark that necessitates spatial-temporal compositional reasoning. We report all the performance with accuracy(↑).

4.2.2 Video QA Models

We test efficiency of SpARC on several types of video QA backbones. These include GNN-based architecture (employing on HGA (Jiang and Han, 2020)) and transformer-based (Vaswani et al., 2017) architecture (employing on VGT (Xiao et al., 2022b)). In addition, due to the recent emergence of large-scale video-language pre-training, we also examine the efficacy of our work on pre-trained VGT (Xiao et al., 2022b).

The reason we select one model from each mainstream video QA architecture is that our approach is not specifically tailored to address particular challenges within each type of video QA model architectures. Therefore, by demonstrating the efficacy of our approach on an advanced model of each type, we can demonstrate that even SOTA approaches still encounter the misfocus issue and our method offers a solution to alleviate the issue.

4.2.3 Image QA Model

We consider image QA model as the knowledge source of causal frames and spatial understanding. Although different architectures of the knowledge source can be explored, we focus on using only ALBEF (Li et al., 2021) as our image QA model since the effect of different architectures is not critical for our approach.

In addition, to ensure the reliability of the pseudo ground-truth for causal frames, a fine-tuning process is required to avoid the domain shift between

Method	Causal	Temp.	Desc.	Total
Co-Mem (Gao et al., 2018)	45.85	50.02	54.38	48.54
HCRN (Le et al., 2020)	47.07	49.27	54.02	48.82
HME (Fan et al., 2019)	46.76	48.89	57.37	49.16
HGA (Jiang and Han, 2020)	48.13	49.08	57.79	50.01
IGV (Li et al., 2022b)	48.56	51.67	59.64	51.34
EIGV (Li et al., 2022a)	51.29	53.11	62.78	53.74
ATP (Buch et al., 2022)	53.10	50.20	**66.80**	54.30
VGT (Xiao et al., 2022b)	51.62	51.94	63.65	53.68
SpARC (w/ HGA)	52.95	53.52	64.70	55.06
SpARC (w/ VGT)	**53.47**	**53.93**	65.12	**55.52**

Table 1: **Comparison with prior SOTAs on NExT-QA benchmark.** SpARC (ours) surpasses previous non video-language pre-trained state-of-the-arts, particularly in the temporal and causal genre. (Model in brackets means the video backbone we use)

Method	Binary	Open	Total
PSAC (Li et al., 2019c)	48.87	31.63	40.18
HME (Fan et al., 2019)	48.91	31.01	39.89
HCRN (Le et al., 2020)	47.97	36.34	42.11
HGA* (Jiang and Han, 2020)	50.89	39.25	45.03
IGV* (Li et al., 2022b) (w/ HGA)	47.95	41.01	44.45
SpARC (w/ HGA)	**51.65**	**42.32**	**46.95**

Table 2: **Comparison with past SOTAs on AGQA2.0 benchmark.** The results show that SpARC (with HGA as video QA model) outperforms all prior non video-language pre-trained work. ($*$: the result was obtained by re-implementation using publicly available code)

image datasets and video datasets. For the detail fine-tune approach and the specific hyperparameters employed, please refer to the supplement.

Other Settings. During training, we also combine existing mixup augmentation (Zhang et al., 2017) and causal frames information provided by image QA model to enhance the video QA model's performance and robustness. The detail and its impact on pre-trained and non pre-trained model will be discussed in Section 4.5. The detail of our enhancement method and its impact compared to original augmentation, we'll discuss in supplement.

4.3 State-of-the-art Comparison

We primarily compare our approach to previous state-of-the-art methods without using video-language pre-training. Our method with VGT as the video model outperforms previous approaches in NExT-QA, especially in temporal and causal aspects, as shown in Table 1. Similarly, SpARC with HGA as the video QA backbone achieves superior results in AGQA2.0 compared to previous work, as demonstrated in Table 2. Notably, in

Method	Causal	Temporal	Total
ATP (Buch et al., 2022)	38.40	36.50	37.62
HGA (Jiang and Han, 2020)	43.30	45.30	44.12
EIGV (Li et al., 2022a)	44.68	43.96	44.38
VGT (Xiao et al., 2022b)	46.70	47.59	47.07
VGT-PT (Xiao et al., 2022b)	43.25	46.31	44.15
SpARC (w/ HGA)	45.65	**49.30**	47.16
SpARC (w/ VGT)	46.78	**49.30**	**47.82**
SpARC (w/ VGT-PT)	**46.93**	48.88	47.73

Table 3: **Comparison with previous SOTAs on ATP-hard set.** SpARC (ours) consistently improves upon the original training method across various video backbones and demonstrates superior performance compared to all previous work. (VGT-PT: pre-trained VGT model; model in brackets means the backbone we use)

Method	Causal	Temp.	Desc.	Total
HGA (Jiang and Han, 2020)	48.13	49.08	57.79	50.01
+ IGV (Li et al., 2022b)	48.56	51.67	59.64	51.34
+ EIGV (Li et al., 2022a)	51.29	53.11	62.78	53.74
+ SpARC (ours)	**52.95**	**53.52**	**64.70**	**55.06**
VGT (Xiao et al., 2022b)	51.62	51.94	63.65	53.68
+ IGV*	50.56	52.84	63.20	53.34
+ EIGV*	51.84	52.88	64.27	54.20
+ SpARC (ours)	**53.47**	**53.93**	**65.12**	**55.52**
VGT-PT (Xiao et al., 2022b)	52.78	54.54	**67.26**	55.70
+ IGV*	50.89	53.74	64.41	53.99
+ EIGV*	52.33	53.26	65.34	54.75
+ SpARC (ours)	**54.24**	**55.25**	66.62	**56.59**

Table 4: **Efficacy comparing to previous model-agnostic work.** Our method outperforms previous model-agnostic work across both non pre-trained and pre-trained video QA backbones. ($*$: results obtained through re-implementation by public code; VGT-PT: pre-trained VGT)

both benchmarks, our approach consistently outperforms the original performance of video QA backbones across all question types, demonstrating the effectiveness of our method in improving performance of video QA models.

We also evaluate SpARC on the ATP-hard subset of NExT-QA, which comprises questions requiring multi-frame information. As presented in Table 3, our approach surpasses previous work and consistently achieves superior results across various video QA backbones, even in pre-trained models. This demonstrates the effectiveness of SpARC in enabling models to handle temporal information. We observe that the pre-trained VGT performs worse than the non pre-trained one, likely due to the lack of temporal pre-training targets (Lee et al., 2022). However, SpARC can address this issue and rehandle multi-frame information accurately.

T	S	Aug.	Non-pretrained Backbone Causal	Temporal	Descriptive	Total	Pre-trained Backbone Causal	Temporal	Descriptive	Total
			51.62	51.94	63.65	53.68	52.78	54.54	67.26	55.70
✓			51.89	53.14	63.91	54.25	53.73	54.61	66.76	56.14
	✓		52.24	53.33	63.84	54.48	54.02	54.42	66.48	56.19
		✓	52.49	51.79	64.84	54.30	53.44	54.20	66.83	55.87
✓	✓		51.93	53.67	**65.84**	54.75	53.89	55.06	**67.54**	56.49
✓		✓	53.15	52.50	64.91	54.88	53.78	54.61	66.62	56.14
✓	✓	✓	**53.47**	**53.93**	65.12	**55.52**	**54.24**	**55.25**	66.62	**56.59**

Table 5: **Ablation study on both non pre-trained and pre-trained video-QA backbone (VGT).** The performance gain from each component in our pipeline. (T: temporal guidance, S: spatial guidance, Aug: use augmentation or not, ✓: the component or augmentation is used)

Our performance across these three sets supports the primary concept of our work: prior approaches failed to handle video-question relationships accurately due to a lack of focus on causal parts of videos, particularly in questions requiring temporal information (*i.e.* temporal relationship and causality questions). In contrast, our method can mitigate this problem and lead to better performance.

4.4 Analysis of Effectiveness and Applicability

We compare SpARC to previous model-agnostic approaches that enhance video QA model learning by improving the localization of causal frames. We incorporate these methods into both pre-trained and non pre-trained video backbones and present the results in Table 4.

In the case of non pre-trained models, both SpARC and previous approaches show improvements, but our method outperforms the previous ones. However, when we incorporate the methods into pre-trained backbones, SpARC is the only one that shows improvement. We speculate that previous methods suffer a performance drop because their selected frames are unsatisfactory and disrupt the video-language knowledge within the pre-trained backbone. In contrast, our method can offer proper guidance and enhance pre-trained models. These results show that SpARC provides better improvement and has broader applicability.

4.5 Ablation Studies

We conduct a comprehensive study to evaluate the efficiency of each component in our methods. The results, as presented in Table 7, demonstrate that incorporating either temporal or spatial guidance can improve model performance, regardless of whether it is a pre-trained or non pre-trained video QA backbone. Combining both guidance can further enhance performance as they complement each other. Additionally, the result shows that mixup (Zhang et al., 2017) augmentation can boost the performance of non pre-trained video QA model, while its effect on the pre-trained model is limited. This result is foreseeable since pre-trained model has already been exposed to a large amount of video-language data, meaning that the additional diversity of training input can only have a slight impact.

Moreover, we observe that most of the components in our method can elevate the performance of the model in questions related to temporal relationships and causality in both pre-trained and non pre-trained video QA backbones. However, these components do not perform as well in descriptive questions when incorporating to the pre-trained model. We attribute this to the pre-trained objectives in existing work, which mostly focuses on spatial information and does not handle temporal information adequately. Therefore, pre-trained models would still benefit from our method to correctly handle temporal information unless they have a dedicated training target.

5 Conclusion

Our novel model-agnostic training approach, "Spatial distillation And Reliable Causal frame localization" (SpARC), solves the misfocus problem encountered by existing methods that focus on irrelevant frames. We use an off-the-shelf image QA model to create pseudo ground-truth of causal frames, which explicitly guides the video QA model for better locating crucial information and addresses the misfocus issue. In addition, we leverage spatial knowledge in the image QA model to guide the video QA model for better spatial understanding. SpARC outperforms previous work on several benchmarks and shows consistent improvement across various video QA models, including pre-trained ones.

6 Limitation and Potential Social Impact

6.1 Limitation

A major limitation of our work is that it requires the use of an off-the-shelf image QA model with satisfactory performance. It doesn't have to be perfect, but its performance should not be significantly worse. While this limitation doesn't have a significant impact on most existing benchmarks, there are cases where it may make our approach challenging to implement. For instance, this could occur in scenarios where the language used in video QA is uncommon and it's difficult to find an off-the-shelf image QA model that aligns with that specific language.

6.2 Potential Social Impact

Our work enables video-language models to learn through guided processes, leading to a more accurate understanding of the relationship between video and language. This approach has the potential to inspire the development of methods rooted in our approach, ultimately leading to the creation of interpretable video QA models. These advancements would yield a positive impact if video QA services emerge in the future, as they could enhance the trustworthiness of such services and mitigate potential instances of discrimination.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927.

Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1989–1998.

Dongsheng Chen, Chaofan Tao, Lu Hou, Lifeng Shang, Xin Jiang, and Qun Liu. 2022. Litevl: Efficient video-language learning with enhanced spatial-temporal modeling. *arXiv preprint arXiv:2210.11929*.

Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2022. An empirical study of end-to-end video-language transformers with masked visual modeling. *arXiv preprint arXiv:2209.01540*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2022. AGQA 2.0: An updated benchmark for compositional spatio-temporal reasoning. *arXiv preprint arXiv:2204.06105*.

Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. 2021. Multi-scale progressive attention network for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 973–978.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116.

Andrew Kae and Yale Song. 2020. Image to video domain adaptation using web supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 567–575.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981.

Hsin-Ying Lee, Hung-Ting Su, Bing-Chen Tsai, Tsung-Han Wu, Jia-Fong Yeh, and Winston H Hsu. 2022. Learning fine-grained visual understanding for video question answering via decoupling spatial-temporal modeling. *arXiv preprint arXiv:2210.03941*.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2017. Attention transfer from web images for video recognition. In *Proceedings of the 25th ACM international conference on multimedia*, pages 1–9.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019a. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019c. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665.

Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. 2022a. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4714–4722.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022b. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Wei Lin, Anna Kukleva, Kunyang Sun, Horst Possegger, Hilde Kuehne, and Horst Bischof. 2022. Cycda: Unsupervised cycle domain adaptation to learn from image to video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 698–715. Springer.

Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. 2021. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1698–1707.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. *Advances in neural information processing systems*, 31.

Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. 2021. Progressive graph attention network for video question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2871–2879.

AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. 2022. Video question answering with iterative video-text co-tokenization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 76–94. Springer.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. 2021. Attend what you need: Motion-appearance synergistic networks for video question answering. *arXiv preprint arXiv:2106.10446*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. 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.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022a. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022b. Video graph transformer for video question answering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 39–58. Springer.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Learning to answer visual questions from web videos. *arXiv preprint arXiv:2205.05019*.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.

A Implementation Details

A.1 Fine-tuning Image Model

To ensure the reliability of the pseudo ground-truth for causal frames, we need to fine-tune the off-the-shelf image QA model due to the domain shift between image datasets and video datasets. The typical image-language models (Li et al., 2021; Bao et al., 2022) involve three components: a visual encoder (Dosovitskiy et al., 2020) that extracts visual information and represents it in a latent space, a text encoder (Vaswani et al., 2017) to transform semantic information into a latent space representation, and a visual-text interaction module that locates spatial information from the question to retrieve the correct answer. We freeze both encoders and only fine-tune the visual-text interaction module for a few epochs. This transfers the knowledge of the image QA model to the target dataset and helps prevent overfitting. Refer to Section A.4 for specific hyperparameters.

A.2 Temporal Guided Mixup

Inspired from (Li et al., 2022a), we enhance mixup (Zhang et al., 2017) by leveraging question-relevant information. In the original mixup augmentation, the input video-question pair $(\mathcal{V}, \mathcal{Q})$ and ground-truth answer $\mathcal{A}$ are transformed to $(\mathcal{V}^*, \mathcal{Q}^*)$ and $\mathcal{A}^*$. Our approach is to modify only the question-relevant input while keeping the non-relevant part

Hyperparameters	NExT-QA (HGA)	NExT-QA (VGT)	AGQA2.0 (HGA)
Learning Rate	10^{-4}	10^{-5}	10^{-4}
Training Epochs	60	10	10
Number of Frames	16	32	8
Batch Size	256	14	256
Using Augmentation	✓	✓	✗
α in Mixup	0.1	0.1	-
β in Mixup	0.1	0.1	-

Table 6: **Hyperparameters for all experiments.** Including NExT-QA (Xiao et al., 2021) benchmark (with HGA (Jiang and Han, 2020) and VGT (Xiao et al., 2022b) as video QA model) and AGQA2.0 (Grunde-McLaughlin et al., 2022) benchmark (with HGA).

intact. Since the model shouldn't rely on the information from the non-causal part of the video, the ground-truth remains $\mathcal{A}^*$.

Using the image QA priors, we split the video into a question-relevant part (causal) $\mathcal{V}_c$ and a question-irrelevant part (non-causal) $\mathcal{V}_n$. By modifying only the causal part of the video, the augmented video-question pair becomes $(\mathcal{V}_c^*, \mathcal{V}_n, \mathcal{Q}^*)$. With these augmented inputs $(\{\mathcal{V}_c^*, \mathcal{V}_n\}, \mathcal{Q}^*)$ and outputs $(\mathcal{A}^*)$, we then train the model using our SpARC pipeline. This variation of mixup augmentation is referred to as Temporal Guided Mixup (TGM). The effectiveness of TGM and the original mixup method is compared in Section B.2.

A.3 Frames Used for Knowledge Extraction

In video QA models that use a frame-based feature extractor (He et al., 2016; Xie et al., 2017), the frames sent to the image QA model for obtaining predicted probabilities would be identical to the frames used in the video QA model. However, some video QA approaches may employ a clip-based feature extractor. In this case, we choose the first frame of each clip and feed it into image model to obtain predicted probabilities, which serve as the knowledge for the corresponding clip. It is possible to raise concerns about the impact of spatial guidance in SpARC. However, since each clip has a very short duration (usually less than 1 second), it provides minimal temporal information. Hence, our spatial distillation approach would still work effectively under such circumstance.

A.4 Settings for Fine-tuning Image Model

As mentioned in Section A.1, we conduct fine-tuning on the cross-modal module of ALBEF (Li et al., 2021) before utilizing it as the knowledge source. Specifically, we fine-tune the model for 5 epochs using the Adam optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01 across all datasets. Regarding the input images, we uniformly sample 8 frames (in AGQA2.0 (Grunde-McLaughlin et al., 2022)) or 16 frames (in NExT-QA (Xiao et al., 2021)) for each video. These frames are then resized to a resolution of 384×384 before being processed by the model.

A.5 Settings for Training Video Models

To ensure a fair comparison, we adopt the same architecture configuration as the original setting in HGA (Jiang and Han, 2020) and VGT (Xiao et al., 2022b). Besides, we utilize the same input features provided by these works, which are publicly available. Regarding the loss setting, we set the weights of both consistency loss w_c and spatial-distillation loss w_s to 1 in all experiments. The training process utilizes the Adam optimizer, and the specific hyperparameters vary depending on the benchmarks and video QA models used. For a detailed list of the hyperparameters, please refer to Table 6. Note that the hyperparameters α, β represent the parameters used in the mixing ratio (sampled from Beta distribution) $\lambda \sim \text{Beta}(\alpha, \beta)$ for mixup augmentation (Zhang et al., 2017).

A.6 Computational Efficiency

We present the time cost of the NeXT-QA benchmark (Xiao et al., 2021). Extracting causal frame knowledge for the entire dataset using ALBEF (Li et al., 2021) takes 12 hours on a single NVIDIA GeForce RTX 3090. During training, we use a single NVIDIA Tesla P100, taking approximately 5 hours with VGT (Xiao et al., 2022b) as video QA backbone and around 6 hours with HGA (Jiang and Han, 2020). The inference time for the entire dataset on both video QA models is under 10 minutes on a single NVIDIA Tesla P100. It's worth noting that the time consumption may slightly vary based on CPU efficiency.

Figure 4: **Additional visualization of the image QA model's prediction.** In each example, the grayed-out frames represent non causal frames verified by humans. The prediction of the image QA model is shown below the image. (a) Location recognition question. (b) An example where the image QA model's predictions misalign with the actual causal frames.

B Additional Experimental Results

B.1 Additional Qualitative Results

We present additional visualizations of predictions from ALBEF (Li et al., 2021) in Figure 4. In addition to the questions discussed in the main paper, we include additional descriptive question that require understanding of locations (Figure 4 (a)). This examples also support the idea that the image QA model can provide reliable indications of causal frames.

Despite the overall positive results, Figure 4 (b) reveals that the predictions from image model may sometimes slightly deviate from the actual causal frames. However, even with this imperfection, the image model still directs the video model's attention to the first and third frames, which are crucial for answering the question. This example shows that despite occasional imperfect guidance of causal frames, the image model still provides valuable guidance to help the video model better handle spatial-temporal information.

B.2 Efficacy of Temporal Guided Mixup

We conduct ablation studies to compare the efficiency of Temporal Guided Mixup (TGM) and the original mixup augmentation (Zhang et al., 2017) when integrated into our spatial-temporal guided approach, SpARC. The studies are performed using both pre-trained and non pre-trained VGT models (Xiao et al., 2022b) as the video QA backbones.

The results, presented in Table 7, demonstrate that incorporating TGM yields slightly improved performance compared to the original mixup augmentation. This improvement is observed in both the pre-trained and non pre-trained video models. These findings indicate that our enhancement of the original mixup augmentation generates more diverse training samples and thus boosts the model's understanding of video information.

C Insights Behind our Method

C.1 Insights of Using Correct Answer

Some might wonder why we employ the correct answer to indicate the pseudo ground-truth of causal frames, as opposed to directly using ranking or applying a threshold to select frames with high confidence scores as causal frames. We illustrate the rationale behind our approach through the following example.

Consider a video where a man sits down, raises his hand, and stands up; a question asks “What does the man do before raising hand". The image model would assign high probabilities to “sit", “raise hand", and “stand" for the beginning, middle, and end frames respectively. Without ground-truth information, using methods like top-k or threshold by highest probability would hard to figure out causal part and lead to a misfocus. This example underscores the significance of employing ground-truth annotations for the identification of causal frames.

C.2 Why Using Hard Selection Guidance

To the best of our understanding, we pioneer the utilization of insights from an image QA model to

T	S	Aug.	Non-pretrained Backbone				Pre-trained Backbone			
			Causal	Temporal	Descriptive	Total	Causal	Temporal	Descriptive	Total
✓		normal	52.80	51.56	65.41	54.48	54.11	53.48	66.90	56.01
✓		TGM (ours)	53.15	52.50	64.91	**54.88**	53.78	54.61	66.62	**56.14**
✓	✓	normal	53.64	53.63	64.98	55.50	54.73	52.92	66.98	56.18
✓	✓	TGM (ours)	53.47	53.93	65.12	**55.52**	54.24	55.25	66.62	**56.59**

Table 7: **Ablation study of Temporal Guided Mixup (TGM).** We contrast the effectiveness of our improved augmentation (TGM) with the original mixup augmentation and our enhanced augmentation leads to a slight performance improvement. (T: temporal guidance, S: spatial guidance, Aug: use augmentation or not, ✓: the component is used, normal: original mixup, TGM: temporal guided mixup)

inform the learning process of a video QA model. There are plenty ways to utilize such knowledge prior, and among them, we choose to hard select causal frames to guide video model. This facilitates a more straightforward validation of the selected causal frames, providing qualitative support for our claims and approach. As we establish the viability of image QA model guidance, it lays the foundation for subsequent researchers to extend our work and explore further applications of such causal frame priors (*e.g.* soft guidance).