Are VideoQA Models Truly Multimodal?

Ishaan Singh Rawal[1,2] **Shantanu Jaiswal**[1,2] **Basura Fernando**[1,2] **Cheston Tan**[1,2,3]

[1]Institute of High Performance Computing, Agency for Science, Technology & Research, Singapore
[2]Centre for Frontier AI Research, Agency for Science, Technology & Research, Singapore
[3]Institute for Infocomm Research, Agency for Science, Technology & Research, Singapore
rawal_ishaan_singh@cfar.a-star.edu.sg

Abstract

While VideoQA Transformer models demonstrate competitive performance on standard benchmarks, the reasons behind their success are not fully understood. Do these models jointly capture and leverage the rich multimodal structures and dynamics from video and text? Or are they merely exploiting shortcuts to achieve high scores? Hence, we design *QUAG* (QUadrant AveraGe), a lightweight and non-parametric probe, to critically analyze multimodal representations. QUAG facilitates combined dataset-model study by systematic ablation of model's coupled multimodal understanding during inference. Surprisingly, it demonstrates that the models manage to maintain high performance even under multimodal impairment. This indicates that the current VideoQA benchmarks and metrics do not penalize models that find shortcuts and discount joint multimodal understanding. Motivated by this, we propose *CLAVI* (Counterfactual in LAnguage and VIdeo), a diagnostic dataset for coupled multimodal understanding in VideoQA. CLAVI consists of temporal questions and videos that are augmented to curate balanced counterfactuals in language and video domains. We evaluate models on CLAVI and find that all models achieve high performance on multimodal shortcut instances, but most of them have very poor performance on the counterfactual instances that necessitate joint multimodal understanding. Overall, we show that many VideoQA models are incapable of learning multimodal representations and that their success on standard datasets is an illusion of joint multimodal understanding [1].

1 Introduction

Multimodal learning with videos and language is challenging, despite the shared sequential nature of these modalities, due to their distinct underlying structures. That is, videos exhibit spatio-temporal dynamics in the pixel space, whereas language representation is composed of the syntax and semantics of word sequences. Hence, tasks like Video Question Answering (VideoQA) [2] are difficult as they necessitate the model to acquire accurate representations of both the modalities and establish meaningful connections between them. Transformers have demonstrated exceptional performance on VideoQA benchmarks [2]. But does the good performance of Transformers on current VideoQA benchmarks necessarily mean that they learn to faithfully represent, leverage, understand and reason the modalities? Or do the current benchmarks and metrics fail to robustly evaluate the models for their multimodal understanding?

This is a valid concern because deep learning models can learn shortcuts to achieve good performance without faithfully representing underlying modalities [3]. For example, seemingly

[1]The extended version of the paper is under review: https://arxiv.org/abs/2306.08889 [1].

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

spatio-temporal tasks, like some action classification problems, are shown to be solved without focusing much on temporal representations [4, 5]. Similarly, in VideoQA, recent works report that the datasets contain specific biases [6, 7]. However, these works are restricted to isolated analyses of either the models or the datasets. This raises questions: **Are the models actually learning to jointly leverage and understand the modalities?**

To answer these questions, we propose QUadrant AveraGe (QUAG), a lightweight and non-parametric probe to systematically gauge the reliance of a finetuned model's performance on joint multimodal representations. We apply QUAG on multiple dataset-model combinations, and consistently find that the models manage to achieve high performance on the benchmarks without relying on specific multimodal interactions. This finding is concerning because high performance on established benchmarks should be ideally indicative of coupled multimodal understanding. This raises a follow-up question – **How then can we diagnose coupled multimodal understanding?**

Thus, we create Counterfactual in LAnguage and VIsion (CLAVI), a diagnostic benchmark to robustly assess joint multimodal understanding in VideoQA models. CLAVI contains automatically generated balanced temporal counterfactuals in both question and video domains to accurately test if the models can jointly understand temporal cues in the question (temporal prepositions and adverbs) and the video (order of frames) domains (Figure 2). We find that finetuned models have high-accuracy on shortcut instances in CLAVI, but have poor performance on the counterfactual instances that require coupled multimodal understanding.

In summary, our contributions are *(i)* we develop QUAG, a systematic method to identify sub-optimalities in joint multimodal representations and show that high performance on established VideoQA benchmarks is not representative of faithful coupled multimodal understanding, and *(ii)* we develop CLAVI, a new diagnostic benchmark that contains balanced temporal counterfactuals in videos and questions to confidently disambiguate the contributions of shortcuts in joint multimodal learning to benchmark the models. Overall, QUAG and CLAVI provide holistic dataset-model insights that reveal the illusion of multimodal understanding in VideoQA models.

2 Related work

Dataset Biases: Works in NLP [8–10], vision [11] and vision-language [12, 13] demonstrate that models can achieve high performance without even understanding the sequence of the embeddings. This is partly because the current benchmarks have unintended biases that could potentially be exploited by models to learn shortcuts; hence accuracy is not always a faithful metric [5, 12, 14, 15]. For VideoQA task, MovieQA [16] and TVQA [17] datasets are biased towards plot understanding or dialogue comprehension [18]. Biases are not always immediately apparent; for example, Social-IQ [19] contains sentiment-biased annotations [20]. Moreover, statistical regularities like answer length, answer frequency [21, 22] and co-occurrence [23–25] introduce spurious features. Overall, these biases allow the models learn shortcuts [3] that circumvent multimodal reasoning [26, 27]. We curate CLAVI by systematically augmenting real-world videos to faithfully represent the complexity of the physical world while controlling the biases to confidently evaluate multimodal temporal understanding.

Shortcut Learning: Tangential to the bias amelioration methods [28, 29], Lei et al. [7] and Winterbottom et al. [18] achieve state-of-the-art performance with simple models by leveraging VideoQA dataset shortcuts in the model. ATP [6] demonstrates single frame bias by re-training the models with an informative frame-selection module to achieve competitive performance. QUAG unifies these ideas to evaluate the dependence of models on shortcuts for circumventing multimodal understanding in terms of performance drop under multimodal representation collapse. Recent works have employed counterfactual instances for improving the performance of VideoQA models. Momeni et al. [30] and Wang et al. [31] have employed hard-negatives for improving verb-understanding in VideoQA models. Bagad et al. [32] stitch pairs of unrelated videos to improve the temporal understanding of video-language models. However, CLAVI synthesizes counterfactual video from the same video, thus qualifies as a more robust benchmark.

3 Do VideoQA models learn to jointly leverage the modalities?

We posit that coupled multimodal understanding is enabled in the fusion layers by progressively attending to the informative tokens within and between the modalities. Hence, we design QUAG to systematically ablate the effects of multimodal attention. It impairs the joint multimodal representations in the pretrained model by systematically block-averaging the attention weights to attend to all tokens uniformly at inference time. Based on the targeted modality-interactions, we define special cases of QUAG, collectively called short-circuit operations, and analyze the performance drop.

3.1 Video question answering setup

In VideoQA, the task is to predict the correct answer given a video-question tuple, $(\mathcal{V}, \mathcal{T})$. A VideoQA model consists of a vision encoder $\boldsymbol{F}_{\mathcal{V}} : \mathcal{V} \rightarrow \mathbb{R}^{l_{\mathcal{V}} \times d}$, text encoder $\boldsymbol{F}_{\mathcal{T}} : \mathcal{T} \rightarrow \mathbb{R}^{l_{\mathcal{T}} \times d}$, and a multimodal fusion module $M : (\boldsymbol{F}_{\mathcal{V}}(\mathcal{V}), \boldsymbol{F}_{\mathcal{T}}(\mathcal{T})) \rightarrow \mathbb{R}^{(l_{\mathcal{V}}+l_{\mathcal{T}}) \times d}$, where $l_{\mathcal{V}}$ and $l_{\mathcal{T}}$ are the maximum input sequence lengths of video and text modalities respectively and d is the dimensionality of the fusion module. Consider M as a composition of n attention-based multimodal fusion blocks, $M = M_n \circ M_{n-1} \circ \cdots M_1$. Each fusion block consists of attention, normalization, and token-mixing modules. For our analysis, we consider M to be composed of self-attention transformer blocks. That is, query, key, and value are the transformations of the same input sequence. $\boldsymbol{X}_{\mathcal{V}\mathcal{T}} = [\boldsymbol{F}_{\mathcal{V}}(\mathcal{V}) \parallel \boldsymbol{F}_{\mathcal{T}}(\mathcal{T})] \in \mathbb{R}^{(l_{\mathcal{V}}+l_{\mathcal{T}}) \times d}$ is the input for M, where $\parallel$ is concatenation operator. Since QUAG operates at inference time, we assume the VideoQA model to be finetuned and frozen.

3.2 QUAG: Ablation of modality interactions

Shortcuts are the spurious features learned by a given model on a given dataset [33]. Along this axis, we use QUAG to pinpoint the exact failure modes in the dataset representations learned by the models. Let $\boldsymbol{X}_{i-1}$ denote the input of the fusion block M_i and let $(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i)$ be its query, key, and value transformations and $\boldsymbol{X}_0 = \boldsymbol{X}_{\mathcal{V}\mathcal{T}}$. Then, the token-mixing operation is given by $\boldsymbol{T}_i = \boldsymbol{A}_i \boldsymbol{V}_i$, where $\boldsymbol{A}_i = softmax(\boldsymbol{Q}_i \boldsymbol{K}_i^\top)$ is the attention matrix (we omit the scaling factor $\sqrt{d}$ for readability). For $\boldsymbol{Q}_{1u}$, $\boldsymbol{K}_{1u}$, and $\boldsymbol{V}_{1u}$ to denote the query, key, and value projections of modality u for the first fusion block, M_1, we can simplify, $\boldsymbol{A}_1$ and $\boldsymbol{T}_1$ in terms of their partition blocks, referred to as quadrants henceforth, as:

$$\boldsymbol{A}_1 = softmax\left(\left[\begin{array}{c|c} \boldsymbol{Q}_{1\mathcal{V}} \boldsymbol{K}_{1\mathcal{V}}^\top & \boldsymbol{Q}_{1\mathcal{V}} \boldsymbol{K}_{1\mathcal{T}}^\top \\ \hline \boldsymbol{Q}_{1\mathcal{T}} \boldsymbol{K}_{1\mathcal{V}}^\top & \boldsymbol{Q}_{1\mathcal{T}} \boldsymbol{K}_{1\mathcal{T}}^\top \end{array}\right]\right) \quad \text{and} \quad \boldsymbol{T}_1 = \left[\begin{array}{c|c} \boldsymbol{A}_{\mathcal{V}\mathcal{V}}^1 & \boldsymbol{A}_{\mathcal{V}\mathcal{T}}^1 \\ \hline \boldsymbol{A}_{\mathcal{T}\mathcal{V}}^1 & \boldsymbol{A}_{\mathcal{T}\mathcal{T}}^1 \end{array}\right] \left[\begin{array}{c} \boldsymbol{V}_{1\mathcal{V}} \\ \hline \boldsymbol{V}_{1\mathcal{T}} \end{array}\right]$$

where $\boldsymbol{A}_{u_1 u_2}^1$ represents the quadrant of $\boldsymbol{A}_1$ corresponding to $(\boldsymbol{Q}_{1u_1} \boldsymbol{K}_{1u_2}^\top)$. Note that we skip layer normalization layers in the discussion for simplicity. Hence, we can simplify and write $\boldsymbol{T}_1$ as:

$$\boldsymbol{T}_1 = \left[\begin{array}{c} \boldsymbol{A}_{\mathcal{V}\mathcal{V}}^1 \boldsymbol{V}_{1\mathcal{V}} + \boldsymbol{A}_{\mathcal{V}\mathcal{T}}^1 \boldsymbol{V}_{1\mathcal{T}} \\ \hline \boldsymbol{A}_{\mathcal{T}\mathcal{V}}^1 \boldsymbol{V}_{1\mathcal{V}} + \boldsymbol{A}_{\mathcal{T}\mathcal{T}}^1 \boldsymbol{V}_{1\mathcal{T}} \end{array}\right] \tag{1}$$

We follow the same partition quadrants, as defined for $\boldsymbol{A}_1$ in M_1, for $\boldsymbol{A}_j$ in the downstream fusion layer M_j and denote the quadrants as $\boldsymbol{A}_{u_1 u_2}^j$. Next, we define row-wise average and replace operator $\mathcal{R}$ that operates on a quadrant of a matrix to replace the values in the quadrant with the mean value of the respective partitioned-row. Note that the values in the other quadrants are unaffected. Given a matrix $\boldsymbol{Z}$ of size $p \times q$ and let W denote the location of the quadrant of $\boldsymbol{Z}$ with indices $(p_1^W \cdots p_2^W) \times (q_1^W \cdots q_2^W)$. We use $[\ .\ ]_{ij}$ to index the element in row i and column j. Then,

$$[\mathcal{R}(\boldsymbol{Z}, W)]_{ij} = \begin{cases} \sum_{k=q_1^W}^{q_2^W} \frac{[\boldsymbol{Z}]_{ik}}{q_2^W - q_1^W + 1} & i \in \{p_1^W, \cdots, p_2^W\} \text{ and } j \in \{q_1^W, \cdots, q_2^W\} \\ [\boldsymbol{Z}]_{ij} & \text{otherwise} \end{cases}$$

We can now formally define the QUAG operator, ϕ, as:

$$\phi(\boldsymbol{A}_i, \boldsymbol{V}_i, [s_1, s_2, \cdots, s_m]) = (\mathcal{R}_{s_1} \circ \mathcal{R}_{s_2} \cdots \circ \mathcal{R}_{s_m}(\boldsymbol{A}_i))\boldsymbol{V}_i$$

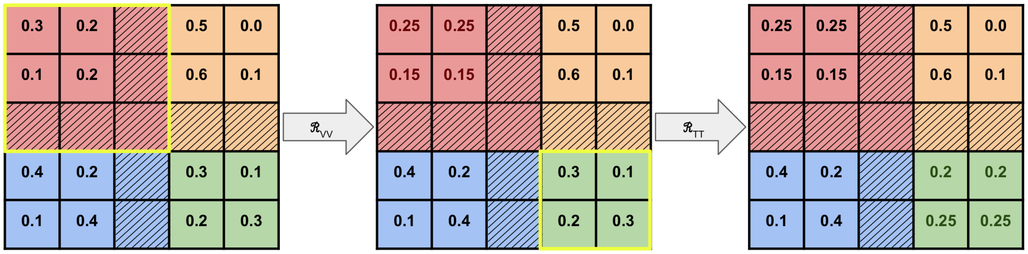


Figure 1: Illustrative toy example of of unimodal short-circuiting or $\phi(\boldsymbol{Z}, [\mathcal{TT}, \mathcal{VV}])$, where $\boldsymbol{Z}$ is the input attention matrix (left-most in the figure), $\mathcal{R}$ is the row-wise average and replace operator and hatching denotes padding. The quadrants that are operated on are highlighted in bright yellow box. Note that $l_{\mathcal{V}} = 3$ and $l_{\mathcal{T}} = 2$ for the model and the video embeddings are pre-concatenated with the question embeddings. As shown in the figure, we apply $\mathcal{R}$ successively to replace the values in the quadrant with the respective row-wise average value. The cells are colored as per their quadrants ($\mathcal{VV}$: red, $\mathcal{VT}$: yellow, $\mathcal{TV}$: blue, $\mathcal{TT}$: green).

where $S = [s_1, s_2, \cdots, s_m]$ is a list of quadrants such that $\forall s \in S : s \in \{\mathcal{TT}, \mathcal{TV}, \mathcal{VT}, \mathcal{VV}\}$, $\mathcal{R}_{s_i}(\boldsymbol{Z})$ is short-hand for $\mathcal{R}(\boldsymbol{Z}, s_i)$, $\boldsymbol{A}_i$ and $\boldsymbol{V}_i$ are the attention and value matrices of M_i respectively. Note that $\mathcal{TT}$ refers to the quadrant corresponding to $A^i_{\mathcal{TT}}$ (independent of the index $1 \leq i \leq n$ of A), similarly $\mathcal{TV}$ refers to the quadrant corresponding to $A^i_{\mathcal{TV}}$, and so on. In implementation, we re-adjust the quadrant boundaries to ignore the padded elements. Refer Figure 1 for an illustrative example. Incorporating QUAG in the existing model pipeline is very easy and we provide the code in the Appendix A.1.2. Since we apply the same QUAG operator successively on all the layers of M (explained in the next section), for brevity, we denote $\Phi(M, S) = \forall_{1 \leq i \leq n}\ \phi(\boldsymbol{A}_i, \boldsymbol{V}_i, S)$. Note that ϕ, and hence, Φ is independent of the order of elements in S.

3.3 Short-circuit operations

As QUAG is a generic method of probing multimodal fusion, we consider some special cases based on the value of S below. We call these operations collectively as short-circuiting operations:

1) $\boldsymbol{S} = [\boldsymbol{\mathcal{VV}}, \boldsymbol{\mathcal{TT}}]$: $\phi(\boldsymbol{A}_1, \boldsymbol{V}_1, [\mathcal{VV}, \mathcal{TT}])$ is equivalent to scaling the average values of $\boldsymbol{V}_{1\mathcal{V}}$ and $\boldsymbol{V}_{1\mathcal{T}}$ in the upper and lower blocks of $\boldsymbol{T}_1$ respectively (as evident from Eqn. 1). Hence, in the upper block, video queries faithfully attend over text keys but uniformly over video keys. Likewise, text queries attend faithfully over video queries but uniformly over text queries in the lower block. We illustrate this operation in Figure 1 and call such a fusion block to be unimodal average conformable. Having explained the trivial case, in Appendix A.1.1 we prove by induction that $\Phi(M, [\mathcal{VV}, \mathcal{TT}])$ makes the overall fusion module, M, unimodal conformable. That is, it bypasses the effect of video-video attention and text-text attention. We term this as **unimodal short-circuiting**.

2) $\boldsymbol{S} = [\boldsymbol{\mathcal{VT}}, \boldsymbol{\mathcal{TV}}]$: Parallel to unimodal short-circuiting, $\phi(\boldsymbol{A}_1, \boldsymbol{V}_1, [\mathcal{VT}, \mathcal{TV}])$ is equivalent to scaling the average values of $\boldsymbol{V}_{1\mathcal{T}}$ and $\boldsymbol{V}_{1\mathcal{V}}$ in the upper and lower blocks of $\boldsymbol{T}_1$ respectively. Video and text queries faithfully attend to video and text keys respectively while crossmodal attention in video-text is reduced to uniform attention. We term this effect as **crossmodal short-circuiting**. It is complementary to unimodal short-circuiting and assesses the importance of inter-modality token-mixing. It probes if the models actually learns by fusing the information between the two modalities or is it largely driven by unimodal biases within the modalities.

3) $\boldsymbol{S} = [\boldsymbol{\mathcal{VV}}, \boldsymbol{\mathcal{TV}}]$: This is equivalent to removing the effect of individual of video keys, resulting in averaging the components of video modality in the upper and lower blocks of all $\boldsymbol{T}_i$. We call this **video short-circuiting**. Similarly, $\boldsymbol{S} = [\boldsymbol{\mathcal{TT}}, \boldsymbol{\mathcal{VT}}]$ leads to **text short-circuiting**.

Table 1: Short-circuit (SC) accuracies for JustAsk and FrozenBiLM models on ActivityNet-QA (A-QA), MSRVTT-QA (M-QA), NeXT-QA (N-QA) and ATP-Hard (ATP-H) datasets.

	FrozenBiLM				JustAsk			
	A-QA	M-QA	N-QA	ATP-H	A-QA	M-QA	N-QA	ATP-H
Baseline	43.6	46.6	55.8	55.7	38.7	41.8	53.8	44.0
Language-only	32.2	33.2	55.7	55.8	28.2	29.9	42.2	42.0
Video-only	0.1	0.0	20.2	20.1	2.6	6.7	39.1	23.0
SC: unimodal	2.4	1.0	19.8	21.4	38.5	41.5	53.6	43.6
SC: crossmodal	32.3	32.8	56.0	55.6	38.3	41.3	53.5	44.3
SC: video	43.1	45.7	55.8	55.7	38.2	41.3	53.4	44.3
SC: text	1.4	1.0	20.5	21.1	38.6	41.5	53.7	43.6

3.4 Experiment and Analysis

We evaluate QUAG on JustAsk [34] and FrozenBiLM [35] models. For datasets, we select ActivityNet-QA [36], MSRVTT-QA [37] and NeXT-QA [38] and ATP-Hard subset of NeXT-QA [6]. We provide implementation details in Appendix A.1.3.

The results are shown in Table 1. For comparison to the unperturbed model, we specify the baseline, language-only (without video input) and video-only (without text input) accuracies. The high performance in language-only setting relative to the baseline is indicative of strong unimodal bias towards language. However, these metrics do not provide any information about the exact nature and degree of the sub-optimal representations learned by the models on the dataset, hence we use QUAG.

The performance of FrozenBiLM on ActivityNet-QA and MSRVTT-QA drops by over 10% with crossmodal short-circuiting, and by 40% with both unimodal and text short-circuiting. Furthermore, the drop is less than 1% under video short-circuiting. However, for NeXT-QA and ATP-Hard, the performance of FrozenBiLM drops to chance level (20%) under text and unimodal short-circuiting operations but hardly drops with video and text short-circuiting. Parallelly, the performance of JustAsk model does not drop by more than 1% for any of the datasets under any short-circuting operation.

This means that FrozenBiLM consistently does not rely on the core features of the video modality and has a strong reliance on text-modality. Further, for NeXT-QA and ATP-Hard, the model does not leverage any crossmodal interactions. However, for ActivityNet-QA and MSRVTT-QA, it leverages some crossmodal interactions (video (query) and text (key) only). On the other hand, JustAsk model does not learn to fuse the modalities across the datasets and relies largely on the text-modality. Note that while the relative performance drop in the *classical* language-only and video-only settings for JustAsk and FrozenBiLM models on ActivityNet-QA and MSRVTT-QA is similar, QUAG points out the differences in their sub-optimal representations.

Our findings reveal that the models can achieve high performance on multimodal benchmarks without leveraging multimodal representations. However, this raises serious concerns because models can learn to *hack* their way around the accuracy metrics for leveraging shortcuts. The supposedly multimodal datasets contain biases and the evaluation metrics do not penalize shortcut learning and provide a false confidence about the abilities of the model. This raises the follow-up question – **How can we confidently benchmark multimodal understanding in VideoQA?**

4 Does multimodal sub-optimality stems from dataset biases?

Sub-optimality in model representations and shortcut learning can stem from a combination of facets like dataset biases, model architecture [39], optimization method [40], learning paradigm [41] etc. Hence, to ablate the effect of dataset biases we curate CLAVI, a diagnostic dataset with temporal counterfactuals in questions and videos that necessitates joint multimodal understanding and penalizes simple shortcut learning. CLAVI is not positioned to replace existing datasets but rather to supplement them, enhancing the understanding of VideoQA models. We finetune the

Figure 2: Illustrative example of the creation of CLAVI. In the original video (V), the action "turning on a light" (Event A; blue pane) follows "holding clothes" (Event B; brown pane). To create a counterfactual video (V'), we swap the action segments without manipulating the segment separating them. The questions (Q), along with their counterfactual (Q'), are curated for each of the videos. Note that the color of the question panel reflects the correct answer (green for "yes", pink for "no"). We provide the list of questions in Appendix A.2.3.

VideoQA models on CLAVI with the prescribed model architecture and training recipe to study and diagnose the representational prowess of the pretrained models.

4.1 CLAVI: Diagnosing through counterfactuals

CLAVI consists of 6,018 videos and 114,342 questions (72,770 train and 41,572 test). It contains simple yes-no questions to probe the absolute temporal location of a single action (beginning/end) or the occurrence sequence for a pair of non-overlapping actions (before/after). CLAVI allows for systematic benchmarking and diagnosis of joint multimodal understanding through the lens of balanced video and question temporal counterfactuals. We use question templates to automatically curate the question-answer pairs from the temporal grounding annotations of Charades-STA [42]. To create temporal counterfactuals in the question domain, we replace *before* with *after* and *beginning* with *end* and vice versa. Further, we create temporal counterfactuals in the video domain by swapping only the action-segments in the video as shown in Figure 2. We exhaustively consider all the compositions of temporal counterfactuals in video and question domains to create balanced negative instances for systematic assessment of multimodal understanding in videos.

Based on the temporal cue in the question, CLAVI contains three question types – **Existence (E)**, **Beginning/End (BE)** and **Before/After (BA)**. Further, we define negative control (**NC**) questions containing actions that do not occur in the video (that is, the answer is always "no") for E and BA types (Appendix A.2.3). Answering the negative control does not require understanding temporal cues in language and video (detailed curation process in Appendix A.2.1).

We define consistent accuracy metrics for evaluating performance on CLAVI. Given a question, if the model predicts the answers correctly for both – the video and its corresponding counterfactual video, it is called video-consistent. Similarly, for a given video, if the model correctly answers a question and it corresponding counterfactual question, it is called text-consistent. The proportion of video and question consistent predictions are reported as **video-consistent accuracy** ($\mathbf{CAcc}_{\mathcal{V}}$) and **text-consistent accuracy** ($\mathbf{CAcc}_{\mathcal{T}}$) respectively. We report the consistent accuracies separately for the **control subset** (E, E-NC, and BA-NC question types) and the **counterfactual subset** (BE and BA question types). The control subset can be answered by leveraging shortcuts while answering the counterfactual subset necessitates joint multimodal understanding.

4.2 Experiment and Analysis

We finetune JustAsk [34], FrozenBiLM [35], Singularity-Temporal [7] and All-In-One+ [43] on CLAVI (Appendix A.2.5). To account for class imbalance in the answers, we use balanced accuracy for validation and testing. The results are summarized in Table 2. All the models have greater than 70% balanced accuracy. At first, it might give an illusion of good multimodal understanding in VideoQA models. However, the consistent accuracy metrics demystify the illusion.

Table 2: Test performance (% accuracy) on CLAVI after finetuning.

Metric	JustAsk (Yang et al., 2021)	FrozenBiLM (Yang et al., 2022)	Singularity-T (Lei et al., 2023)	All-In-One+ (Wang et al., 2023)
Balanced Acc	72.2 ± 0.2	80.5 ± 0.1	76.8 ± 0.5	73.9 ± 0.1
$\mathrm{CAcc}_{\mathcal{V}}$	50.6 ± 0.3	74.0 ± 0.1	47.2 ± 1.1	49.6 ± 0.5
$\mathrm{CAcc}_{\mathcal{T}}$	50.3 ± 0.1	75.5 ± 0.1	47.0 ± 1.0	49.5 ± 0.3
$\mathrm{CAcc}_{\mathcal{V}}$-control	98.0 ± 0.2	93.2 ± 0.2	92.7 ± 2.0	98.1 ± 0.5
$\mathrm{CAcc}_{\mathcal{T}}$-control	98.2 ± 0.2	93.7 ± 0.2	93.5 ± 1.9	98.2 ± 0.7
$\mathrm{CAcc}_{\mathcal{V}}$-counter	**3.6 ± 0.1**	54.1 ± 0.2	**1.7 ± 0.2**	**1.2 ± 0.3**
$\mathrm{CAcc}_{\mathcal{T}}$-counter	**2.4 ± 0.1**	57.2 ± 0.2	**0.5 ± 0.2**	**0.8 ± 0.1**

Text and video consistent accuracies are greater than 90% for the control subset for all the models. This is because, unlike the counterfactual subset, the control subset does not requires coupled understanding. That is, the model can answer it correctly by simple shortcuts – irrespective of the context of the negative control action in the question and the location of the object and/or the action in the video. However, for achieving high consistent accuracies on the counterfactual subset, the model needs to jointly understand the order of the events and the temporal cues in the question along with the order of the events in the video. We get significantly lower consistent accuracies (less than 4%) for the counterfactual subset, except for FrozenBiLM.

Overall, this means that many **models are able to exploit shortcuts but unable to learn joint multimodal representations**. This is consistent with our results from QUAG that the VideoQA models gain high performance not through multimodal understanding but by leveraging shortcuts. We hope that QUAG and CLAVI galvanizes the community to not just assess the models for their accuracy but also understand the representations they are learning.

References

[1] Ishaan Singh Rawal, Shantanu Jaiswal, Basura Fernando, and Cheston Tan. Revealing the illusion of joint multimodal understanding in videoqa models. *arXiv preprint arXiv:2306.08889*, 2023.

[2] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.432. URL `https://aclanthology.org/2022.emnlp-main.432`.

[3] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[4] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13999–14009, 2022.

[5] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 535–544, 2021.

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

[7] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.29. URL https://aclanthology.org/2023.acl-long.29.

[8] Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. When classifying grammatical role, bert doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, 2022.

[9] Jack Hessel and Alexandra Schofield. How effective is bert without word ordering? implications for language understanding and data privacy. 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 204–211, 2021.

[10] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, 2021.

[11] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019. URL https://openreview.net/pdf?id=SkfMWhAqYQ.

[12] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

[13] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021.

[14] Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, 2021.

[15] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017.

[16] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[17] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167.

[18] T. Winterbottom, S. Xiao, A. McLean, and N. Al Moubayed. On modality bias in the TVQA dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[19] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[20] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1203. URL https://aclanthology.org/D16-1203.

[23] Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1574–1583, October 2021.

[24] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019.

[25] Sanjay Subramanian, Sameer Singh, and Matt Gardner. Analyzing compositionality in visual question answering. *ViGIL@ NeurIPS*, 7, 2019.

[26] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 431–441, 2018.

[27] Keren Ye and Adriana Kovashka. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3181–3189, 2021.

[28] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

[29] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1418. URL https://aclanthology.org/D19-1418.

[30] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15579–15591, October 2023.

[31] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models, 2023.

[32] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516, 2023.

[33] Nihal Murali, Aahlad Manas Puli, Ke Yu, Rajesh Ranganath, and Kayhan Batmanghelich. Shortcut learning through the lens of early training dynamics. *arXiv preprint arXiv:2302.09344*, 2023.

[34] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 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, 2021.

[35] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022.

[36] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[37] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[38] 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 (CVPR)*, pages 9777–9786, June 2021.

[39] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

[40] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[41] Xiao Liu, Ankur Sikarwar, Joo Hwee Lim, Gabriel Kreiman, Zenglin Shi, and Mengmi Zhang. Reason from context with self-supervised learning. *arXiv preprint arXiv:2211.12817*, 2022.

[42] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[43] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023.

[44] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 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, 2021.

[45] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *Advances in Neural Information Processing Systems*, 2022.

[46] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[47] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 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, 2022.

[48] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 130–147. Springer, 2022.

[49] Mayu Otani, Yuta Nakahima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *The British Machine Vision Conference (BMVC)*, 2020.

[50] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[51] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022.

[52] Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan Stroud, and Rada Mihalcea. FIBER: Fill-in-the-blanks as a challenging video understanding evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2925–2940, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.209. URL https://aclanthology.org/2022.acl-long.209.

[53] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4441–4450, 2023.

A Appendix

A.1 QUAG

A.1.1 Inductive Proof of Short-circuiting

Having explained the trivial case, we prove by induction that $\Phi(M, [\mathcal{VV}, \mathcal{TT}])$ leads to unimodal average conformability of all the component fusion blocks in M. Consider a block $M_j \in M$ such that $j > 1$. We want to show that unimodal average conformability of first $\{M_0, M_1, \cdots, M_{j-1}\}$ blocks using $\forall_{1 \leq i \leq j-1}\ \phi(\boldsymbol{A}_i, \boldsymbol{V}_i, [\mathcal{VV}, \mathcal{TT}])$ implies $\phi(\boldsymbol{A}_j, \boldsymbol{V}_j, [\mathcal{VV}, \mathcal{TT}])$ will make M_j unimodal average conformable. The input of M_j can be decomposed into non-linear and linear (from the residual connection that skips the feed-forward layer of M_{j-1}) projections of $\boldsymbol{T}_{j-1} + M_{j-2} \circ M_{j-3} \cdots \circ M_1(\boldsymbol{X}_{\mathcal{VT}}) + \boldsymbol{X}_{\mathcal{VT}}$. Hence, when $\{M_0, M_1, \cdots, M_{j-1}\}$ are unimodal average conformable, $\boldsymbol{X}_{\mathcal{VT}}$ is the only non-conformable component. And we have shown in the trivial case that $\phi(\boldsymbol{A}_1, \boldsymbol{V}_1, [\mathcal{VV}, \mathcal{TT}])$ makes M_1 conformable, hence M_j is also unimodal average conformable under ϕ.
Ultimately, $\Phi(M, [\mathcal{VV}, \mathcal{TT}])$ bypasses the effect of video-video attention and text-text attention. We prove that unimodal token-mixing is reduced to scaling the average of the modalities. We term this as **unimodal short-circuiting**. It ablates unimodal representations to analyze their dependence on the performance of the models. Since the following cases can be proved similarly using induction, we skip the proofs for conciseness.

A.1.2 Code

Below is the implementation of QUAG as an augmentation of the existing self-attention function. We use row-wise average and replace operation in each if-clause statements, while ignoring the padding, to ablate the effect of the quadrant.

```
def self_attention(inputs, mask, dim_model, l_v, l_t, quads):
    # Inputs:
    #   inputs: Tensor of shape (batch_size, sequence_length, dim_model)
    #   mask: Tensor of shape (batch_‘size, sequence_length)
    #   dim_model: Dimension of the model (e.g., 512)
    #   l_v: int    maximum length of video tokens
    #   l_t: int    maximum length of question tokens
    #   quads: list containing elements from {’VV’, ’VT’, ’TV’, ’TT’}
```

```
    query = linear_transform_query(inputs)
    key = linear_transform_key(inputs)
    value = linear_transform_value(inputs)
    attention_scores = compute_attention_scores(query, key, mask)
    apply_quag(attention_scores, mask, l_v, l_t, quads)
    attended_output = apply_attention_scores(attention_scores,
   value)
    return attended_output

def compute_attention_scores(query, key, mask):
    scaled_dot_product = dot_product(query, key) / sqrt(dim_model)
    attention_scores = softmax(scaled_dot_product + (1 - mask) *
   -1e9)
    return attention_scores

def apply_quag(attention_scores, mask, l_v, l_t, quads):
    if 'VV' is in quads:
        replace_with_rowwise_average(attention_scores[:, :l_v, :
   l_v], mask[:, :l_v, :l_v])
    if 'VT' is in quads:
        replace_with_rowwise_average(attention_scores[:, :l_v, -
   l_t:], mask[:, :l_v, -l_t:])
    if 'TV' is in quads:
        replace_with_rowwise_average(attention_scores[:, -l_t:, :
   l_v], mask[:, -l_t:, :l_v])
    if 'TT' is in quads:
        replace_with_rowwise_average(attention_scores[:, -l_t:, -
   l_t:], mask[:, -l_t:, -l_t:])

def replace_with_rowwise_average(scores, mask):
    rowwise_sum = sum(scores, axis=-1)
    rowwise_mean = rowwise_sum / sum(mask, axis=-2)
    expanded_rowwise_mean = expand_dims(rowwise_mean, axis=-1)
    replace_elements(scores, expanded_rowwise_mean)

def apply_attention_scores(attention_scores, value):
    attended_output = dot_product(attention_scores, value)
    return attended_output
```

A.1.3 Experiment Details

All our experiments were performed on 4 NVIDIA A5000 GPUs. We use the official checkpoints and code of JustAsk [website] and FrozenBiLM [website]. For all the experiments with JustAsk, we use the checkpoints of the model pretrained on HowToVQA69M and WebVidVQA3M. For FrozenBiLM, we use the WebVid10M-pretrained checkpoint for all our experiments. Since QUAG operates at inference time, we do not need to perform any training. Since the model owners do not report results on NeXT-QA, we finetune the models with the official recipe to achieve performance similar to that independently reported by others [47]. While FrozenBiLM can also take subtitles as the input, for fair comparison, we do not pass it in any of the experiments. We provide the hardware details in the main manuscript. For NeXT-QA, we use the official dataset and finetune the models with the default parameters

A.1.4 Finegrained Accuracies

A.1.5 JustAsk Model

We present the fine-grained performance of JustAsk on the discussed datasets in Tables 3, 4, 5, and 6

A.1.6 FrozenBiLM Model

We present the fine-grained performance of FrozenBiLM on the discussed datasets in Tables 7, 8, 9, and 10

Table 3: Fine-grained performance of JustAsk on ActivityNet-QA

Config	Motion	Spatial	Temp	Y/N	Color	Obj	Loc	Num	Other
Baseline	30.6	19.9	4.9	64.2	34.7	26.7	35.5	48.9	36.8
Lang-only	1.4	9.1	4.3	51.8	28.7	23.0	16.6	46.9	29.1
Vid-only	20.3	0.9	1.8	0.0	0.0	1.6	1.3	0.0	0.7
SC: unimodal	30.1	19.1	4.9	63.9	33.6	26.4	36.8	48.4	37.0
SC: crossmodal	28,0	18.9	4.8	64.7	34.7	25.8	35.5	48.5	36.4
SC: text	30.4	19.3	5.0	64.1	34.0	26.4	35.5	46.7	37.2
SC: video	28.6	18.8	4.5	64.3	34.6	25.5	35.5	48.4	36.1

Table 4: Fine-grained performance of JustAsk on MSRVTT-QA

Config	What	How	Color	Where	Who	When
Baseline	35.8	83.7	51.7	39.4	51.3	82.3
Lang-only	24.3	83.3	43.4	30.5	37.1	72.3
Vid-only	8.5	0.0	3.5	0.4	3.0	10.1
SC: unimodal	35.6	83.3	51.8	39.8	50.8	82.3
SC: crossmodal	35.35	83.75	51.98	39.8	50.8	81.8
SC: text	35.7	83.2	51.8	39.0	50.8	82.1
SC: video	35.4	83.8	51.8	39.8	50.7	81.6

Table 5: Fine-grained performance of JustAsk on NeXT-QA

Config	Causal	Temporal	Descriptive
Baseline	50.8	52.8	65.0
Lang-only	39.5	44.3	47.1
Vid-only	39.2	37.9	44.0
SC: unimodal	50.5	52.5	65.3
SC: crossmodal	50.8	51.8	65.0
SC: text	50.7	52.7	65.0
SC: video	50.7	52.1	65.0

Table 6: Fine-grained performance of JustAsk on ATP-Hard subset of NeXT-QA

Config	Causal	Temporal
Baseline	44.4	43.4
Lang-only	41.2	43.1
Vid-only	23.5	22.3
SC: unimodal	43.2	43.3
SC: crossmodal	44.2	44.4
SC: text	43.7	43.4
SC: video	44.3	44.4

A.2 CLAVI

A.2.1 Dataset Creation

We curate CLAVI by leveraging Charades-STA (`https://prior.allenai.org/projects/data/charades/license.txt`) [42], containing 9,848 videos of humans performing actions based on a short script written by composing predefined vocabulary that describe multiple daily actions. The videos are annotated with the start and end times of each action. The action category, the start, and the end of each action segment are referred to as the *action tuple*. Each video may contain more than two action tuples. We select pairs of action tuples based on the uniqueness of the action category and complete exclusivity (that is no overlap between the occurrence of the actions). In a given selected pair of action tuples, the two actions along with

Table 7: Fine-grained performance of FrozenBiLM on ActivityNet-QA

Config	Motion	Spatial	Temp	Y/N	Color	Obj	Loc	Num	Other
Baseline	30.1	22.5	6.4	75.6	34.6	27.7	37.1	55.8	41.6
Lang-only	2.6	10.5	4.8	63.3	32.3	23.9	16.6	44.7	31.6
Vid-only	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0
SC: unimodal	0.0	0.1	0.1	8.3	0.0	0.0	0.0	1.3	0.5
SC: crossmodal	1.8	11.1	3.88	64.5	32.7	21.7	16.8	46.0	32.1
SC: text	0.0	0.1	0.1	4.4	0.1	0.3	0.0	1.2	0.3
SC: video	28.8	21.8	6.5	75.1	34.3	29.3	36.0	55.3	41.0

Table 8: Fine-grained performance of FrozenBiLM on MSRVTT-QA

Config	What	How	Color	Where	Who	When
Baseline	40.5	87.2	57.9	41.5	56.6	81.4
Lang-only	27.3	83.6	50.0	35.8	41.2	77.6
Vid-only	0.0	0.0	0.0	0.0	0.0	0.0
SC: unimodal	0.7	0.0	1.2	0.8	1.8	0.2
SC: crossmodal	27.1	83.4	50.9	32.9	41.1	66.3
SC: text	0.3	0.0	0.8	0.0	2.8	0.0
SC: video	39.8	85.5	58.8	41.9	55.4	80.9

Table 9: Fine-grained performance of FrozenBiLM on NeXT-QA

Config	Causal	Temporal	Descriptive
Baseline	56.0	56.1	54.5
Lang-only	55.9	56.1	54.2
Vid-only	20.7	19.1	20.9
SC: unimodal	19.7	21.1	17.3
SC: crossmodal	56.1	56.5	54.3
SC: text	20.0	21.6	19.9
SC: video	56.1	56.1	54.5

Table 10: Fine-grained performance of FrozenBiLM on ATH-Hard subset of NeXT-QA

Config	Causal	Temporal
Baseline	55.2	56.3
Lang-only	55.5	56.2
Vid-only	20.0	20.1
SC: unimodal	20.7	22.5
SC: crossmodal	54.9	56.6
SC: text	20.2	22.3
SC: video	55.3	56.3

the inter-action region constitute the video segment. We ensure that the two action categories in the pair are distinct. Additionally, to address temporal boundary ambiguities in the annotations, we filter out segments where either of the selected action classes occurs in close proximity to the segment boundaries

We also extend the boundaries of the two actions in the pair. We define two boundary extensions – out-extension and in-extension. The out-extension encompasses regions that are not a part of the selected segment but extend outwards in both directions into the original video. Similarly, in-extension extends inwards into the inter-action segment. To avoid temporal position bias [48, 49], the lengths of the extension boundaries are selected randomly. However, since inter-action separation can affect their recognition [32], we constraint the inter-action separation in the original and the corresponding negative video to be the same. That is, the sum of out-extension boundaries is always equal to the sum of in-extension boundaries.

We trim each boundary-extended contiguous segment from the original video to curate a positive video instance. To create the counterfactual video, we swap the boundary-extended action regions as shown in Figure 2. Note that the region between the boundary-extended actions is unaffected. Swapping operation preserves the actions but only alters their chronology, and can be applied independently to question negatives (unlike manipulations like video reversal [31]). This independence provides fine-grained control to create a balanced benchmark for comprehensive analysis.

We create three types of questions using pre-defined templates and action-class annotations:

1) **Existence (E) type**: The E-type questions for both the action classes follow the template *"Was someone ⟨A⟩?"*, where *⟨A⟩* is one of two action classes in video. We use it as a positive control to verify if the model is able to correctly recognize the action classes. We use the exact same question for negative video instance as well, totalling to 4 control (questions, video, answer) instances for a Charades-extracted video segment.

2) **Beginning/End (BE) type**: BE type questions the absolute location of the action in the video. The question is of the form, *"Was the person ⟨A⟩ at the {beginning/end}?"* where *⟨A⟩* is one of two action classes in the video, and we select one of *beginning* and *end*. Hence, for a given video and its negative, we have, in total, 8 instances of BE (questions, video, answer) tuples combined. Note that the answer for a given BE question is complemented in the negative video.

3) **Before/After (BA) type**: BA type comprises of questions on the relative order of occurrence of actions. The question is of the form *"Did ⟨A1⟩ happen {after/before} ⟨A2⟩?"*, where *⟨A1⟩* and *⟨A2⟩* are the selected action classes. We consider all the permutations of action classes. Hence, we have a total of 8 instances of BA type (questions, video, answer) tuples per extracted video. Similar to BE type, the answer is complemented in the negative video.

Further, we add negative controls for E and BA type questions. A negative control action is an action that does not occur in the video. Since we want to probe only for temporal understanding, we keep the negative control action-class easy to detect by randomly selecting an action-class that does not contain any of the objects or actions in the original video. Hence, answering the negative control does not require understanding temporal cues in language and video and can be answered by object elimination. It serves the dual purpose of sanity check of learning and a baseline for learning by temporal shortcuts. The answer of negative control questions is always false. This adds two E type and sixteen BA type negative control questions for the video and its negative combined. Hence, including the negative control questions, each video in CLAVI is associated with 19 questions: 2 E, 4 BE, 4 BA, 1 E negative control and 8 BA negative controls. The ratio of "yes":"no" answers is 6:13.

A.2.2 Comparison with Existing Datasets

We provide a comparison of size of CLAVI with established VideoQA datasets in Table 11.

A.2.3 Comprehensive List of Questions

We provide a comprehensive list of the questions for the example presented in Fig 2 of the main paper. We define the actions as: **A**: *turning on light* **B**: *holding clothes* **C**: *washing mirror*, where action A occurs before action B in the original video and action C does not occur anywhere in the original video.

Enlisted below are the questions and its negatives (Q and Q' respectively) for the video (V) (that is event A occurs after event B):Note that the color of the panel is representative of the answer of the question (red: “no”, green: “yes”).

E-Type:

Q : Was someone turning on light?

Q : Was someone holding clothes?

Table 11: Comparison of CLAVI with other other VideoQA datasets sorted in the reverse order of recency.

Dataset	Number of (V,Q,A) samples
MSRVTT-QA [37]	243K
ActivityNet-QA [36]	58K
Social-IQ QA [19]	7.5K
NeXT-QA [38]	52K
iVQA [44]	10K
STAR [50]	60K
EgoTaskQA [51]	40K
FIBER [52]	28K
NewsQA [53]	8.6K
CLAVI (Ours)	**114K**

E-Type (negative control):

Q : Was someone washing mirror?

BE-Type

Q : Was the person turning on light at the **beginning**?

Q': Was the person turning on light at the **end**?

Q : Was the person holding clothes at the **end**?

Q': Was the person holding clothes at the **beginning**?

BA-Type

Q : Did turning on light happen **before** holding clothes?

Q': Did turning on light happen **after** holding clothes?

Q : Did holding clothes happen **after** turning on light?

Q': Did holding clothes happen **before** turning on light?

BA-Type (negative-control)

Q': Did washing mirror happen **before** turning on light?

Q': Did washing mirror happen **after** turning on light?

Q': Did turning on light happen **before** washing mirror?

Q': Did turning on light happen **after** washing mirror?

Q': Did washing mirror happen **before** holding clothes?

Q': Did washing mirror happen **after** holding clothes?

Q': Did holding clothes happen **before** washing mirror?

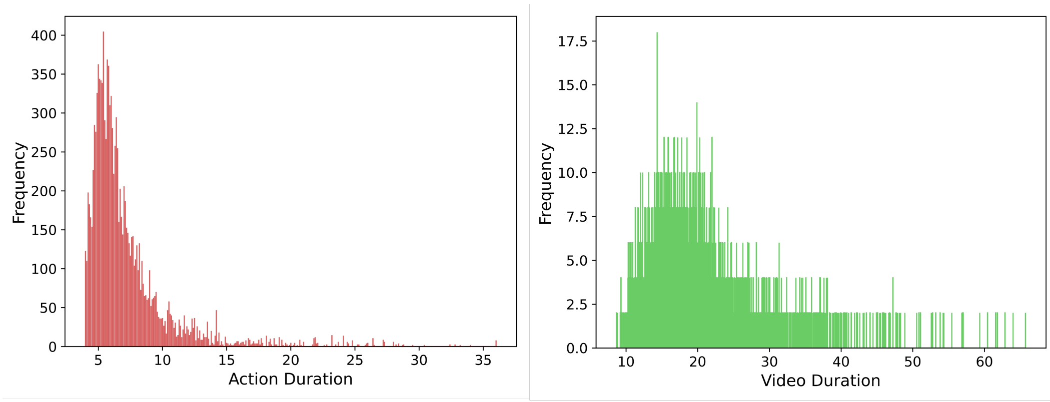
Q': Did holding clothes happen **after** washing mirror?

Enlisted below are the questions and its negatives (Q and Q' respectively) for the negative video instance (V') (that is event B occurs after event A).

E-Type:

Q : Was someone turning on light?

Q : Was someone holding clothes?



(a) Distribution of action duration

(b) Distribution of video duration

Figure 3: Distribution of length of (a) action and (b) video durations

E-Type (negative control):

Q : Was someone washing mirror?

BE-Type

Q : Was the person turning on light at the **beginning**?

Q': Was the person turning on light at the **end**?

Q : Was the person holding clothes at the **end**?

Q': Was the person holding clothes at the **beginning**?

BA-Type

Q : Did turning on light happen **before** holding clothes?

Q': Did turning on light happen **after** holding clothes?

Q : Did holding clothes happen **after** turning on light?

Q': Did holding clothes happen **before** turning on light?

BA-Type (negative-control)

Q': Did washing mirror happen **before** turning on light?

Q': Did washing mirror happen **after** turning on light?

Q': Did turning on light happen **before** washing mirror?

Q': Did turning on light happen **after** washing mirror?

Q': Did washing mirror happen **before** holding clothes?

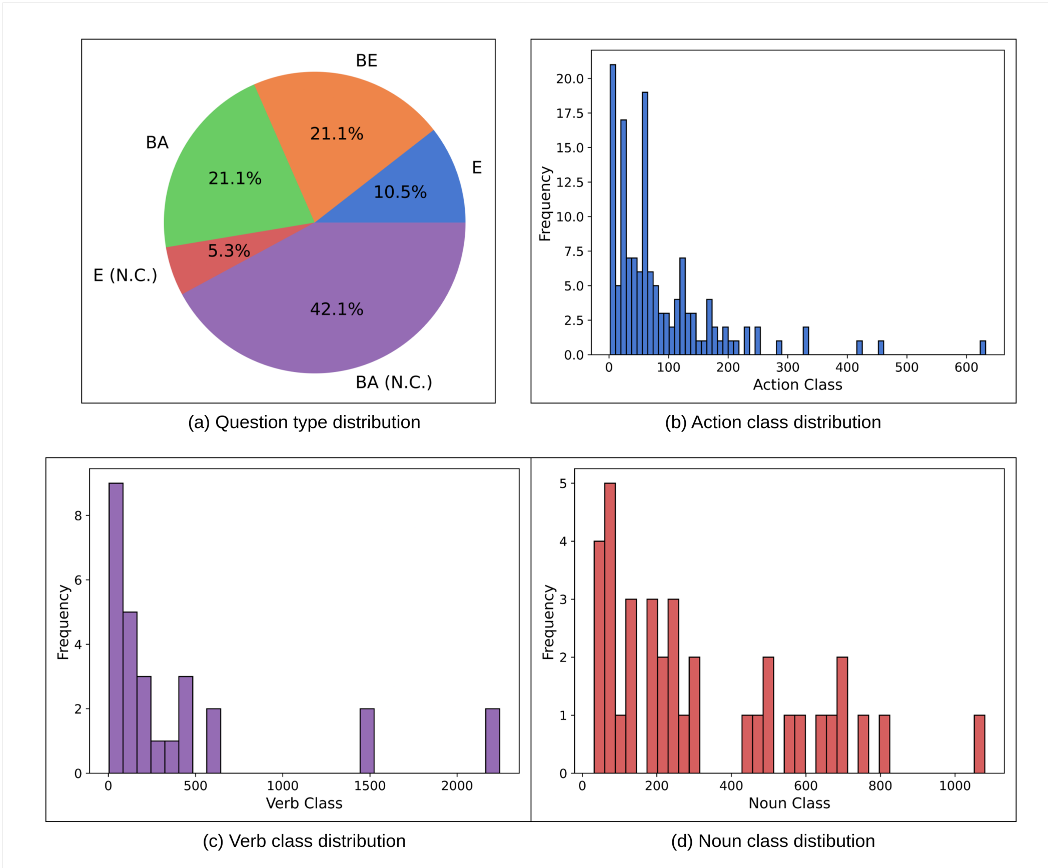
Q': Did washing mirror happen **after** holding clothes?

Q': Did holding clothes happen **before** washing mirror?

Q': Did holding clothes happen **after** washing mirror?

A.2.4 Dataset Metrics

The duration of individual action in CLAVI lies in the range [4.0 sec, 36.0 sec]; the average length of action is $\mathbf{7.7} \pm \mathbf{3.42}$ sec. The average video length is $\mathbf{19.95} \pm \mathbf{7.34}$ secs and the range is [8.67 sec, 65.73 sec]. We plot the distribution of the action and video durations in Fig. 3.



(a) Question type distribution

(b) Action class distribution

(c) Verb class distribution

(d) Noun class distibution

Figure 4: Metrics of the dataset (a) distribution of question types (same for training and testing set), (b) histogram plot of frequencies of action classes (c) histogram plot of frequencies of verb classes (d) histogram plot of frequencies of noun classes.

CLAVI consists of **141** unique action classes. Each action class is composed of noun (objects) and verb. There are **37** unique noun classes and **28** unique verb classes. We show the frequency distributions of action, verb and noun classes in Fig 4.

A.2.5 Experiment Details

As mentioned in the main manuscript, we use the official checkpoints, finetuning code and hyper-parameters of JustAsk [website], FrozenBiLM [website] , Singularity-Temporal [website], and All-in-one+ [website]. For JustAsk, we use the checkpoint of the model pretrained on HowToVQA69M and WebVidVQA3M. For FrozenBiLM, we use the WebVid10M-pretrained checkpoint. All-in-one+ is pretrained on eight datasets comprising of both images and videos (videos: Webvid, YT-Temporal-180M, HowTo100M and images: CC3M, CC12M, COCO, Visual Genome, SBU Captions). Singularity-Temporal is pretrained on a 17.28M images and video subset (images: COCO, Visual Genome, SBU Captions, CC3M, CC12M and videos: WebVid). We have depicted the finetuning details in Table 12.

A.2.6 Fine-grained Accuracies

In Table 13 we provide error bars for the finetuning experiments. The experiments were performed thrice on the same hardware with the same set of hyperparameters.

Table 12: Hyperparameters and checkpoint details of CLAVI finetuning experiment

Model	Checkpoint	Epochs	LR
JustAsk	HowToVQA69M, WebVidVQA3M	20	1.00E-05
FrozenBiLM	WebVid10M	20	5.00E-05
All-In-One+	Webvid, YT-Temporal-180M, HowTo100M, CC3M, CC12M, COCO, Visual Genome, SBU Captions	10	1.00E-04
Singularity-T	COCO, Visual Genome, SBU Captions, CC3M, CC12M, WebVid	20	1.00E-05

Table 13: Fine-grained performance (% of accuracy) on CLAVI for question (Q) and counterfactual question (Q'), video (V) and counterfactual video (V') (Note: N.C. refers to Negative Control)

V/V'	Question	Q/Q'	JustAsk	FrozenBiLM	Singularity-T	All-in-one+
V	E-type	Q	89.55 ± 0.01	87.51 ± 0.00	90.75 ± 0.03	86.08 ± 2.59
	E-type (N.C.)	-	75.28 ± 0.02	88.66 ± 0.00	79.16 ± 0.03	69.34 ± 11.72
	BE-type	Q	69.80 ± 0.07	69.15 ± 0.01	98.23 ± 0.01	99.31 ± 0.84
		Q'	30.58 ± 0.07	73.25 ± 0.01	1.87 ± 0.01	0.73 ± 0.84
	BA-type	Q	27.81 ± 0.02	56.88 ± 0.01	62.55 ± 0.09	25.82 ± 5.49
		Q'	72.31 ± 0.02	86.79 ± 0.01	37.23 ± 0.09	74.31 ± 0.84
	BA-type (N.C.)	-	98.23 ± 0.00	96.79 ± 0.00	93.72 ± 0.03	98.44 ± 1.02
V'	E-type	Q	89.17 ± 0.01	86.96 ± 0.01	90.58 ± 0.02	86.03 ± 2.66
	E-type (N.C.)	Q	76.10 ± 0.03	88.45 ± 0.01	79.04 ± 0.03	69.17 ± 11.26
	BE-type	Q	30.18 ± 0.07	73.61 ± 0.01	1.80 ± 0.01	0.76 ± 1.00
		Q'	69.88 ± 0.07	70.00 ± 0.02	98.28 ± 0.01	99.12 ± 1.02
	BA-type	Q	71.61 ± 0.02	85.43 ± 0.01	38.00 ± 0.08	74.24 ± 5.12
		Q'	28.34 ± 0.02	54.44 ± 0.00	62.15 ± 0.07	25.90 ± 4.93
	BA-type (N.C.)	-	98.51 ± 0.00	96.87 ± 0.00	93.51 ± 0.03	98.46 ± 1.04

A.3 Limitations and Future Work

Our dataset is intentionally simple, so as to focus the benchmark only on simple temporal sequence understanding, which preempts spatio-temporal referential understanding. We plan to include more complex temporal organizations of action classes like containment and partial-overlap that are defined using prepositions like *during* and *while* in future work. As the current state-of-the-art models catch-up to our benchmark, our future plan is to curate a more complex dataset with more natural questions that include temporal referring expressions with similar balanced doubly-negative strategy.