Dissecting Multimodality in VideoQA Transformer Models by Impairing Modality Fusion

Ishaan Singh Rawal [1 2] Alexander Matyasko [1 2] Shantanu Jaiswal [1 2] Basura Fernando [1 2] Cheston Tan [1 2 3]

Abstract

While VideoQA Transformer models demonstrate competitive performance on standard benchmarks, the reasons behind their success are not fully understood. Do these models capture the rich multimodal structures and dynamics from video and text jointly? Or are they achieving high scores by exploiting biases and spurious features? Hence, to provide insights, we design *QUAG* (QUadrant AveraGe), a lightweight and non-parametric probe, to conduct dataset-model combined representation analysis by impairing modality fusion. We find that the models achieve high performance on many datasets without leveraging multimodal representations. To validate QUAG further, we design *QUAG-attention*, a less-expressive replacement of self-attention with restricted token interactions. Models with QUAG-attention achieve similar performance with significantly fewer multiplication operations without any finetuning. Our findings raise doubts about the current models' abilities to learn highly-coupled multimodal representations. Hence, we design the *CLAVI* (Complements in LAnguage and VIdeo) dataset, a stress-test dataset curated by augmenting real-world videos to have high modality coupling. Consistent with the findings of QUAG, we find that most of the models achieve near-trivial performance on CLAVI. This reasserts the limitations of current models for learning highly-coupled multimodal representations, that is not evaluated by the current datasets (project page: https://dissect-videoqa.github.io).

[1]Centre for Frontier AI Research, Agency for Science, Technology & Research, Singapore [2] Institute of High Performance Computing, Agency for Science, Technology & Research, Singapore [3]Institute for Infocomm Research, Agency for Science, Technology & Research, Singapore. Correspondence to: Ishaan Singh Rawal <rawal_ishaan_singh@cfar.a-star.edu.sg>.

1. Introduction

Multimodal learning with videos and language presents a challenge, given their shared sequential nature but 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) (Zhong et al., 2022) are difficult as they necessitate the model to acquire accurate representations of both modalities and establish meaningful connections between them. Transformers have demonstrated exceptional performance on VideoQA benchmarks (Zhong et al., 2022). But does the good performance of Transformers on current VideoQA benchmarks necessarily mean that they learn to faithfully represent, leverage, understand, and reason about 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 leveraging and aligning the modalities (Geirhos et al., 2020). For example, seemingly spatio-temporal tasks, like some action classification problems, are shown to be solved without focusing much on temporal representations (Sevilla-Lara et al., 2021; Kowal et al., 2022). Similarly, in VideoQA, recent works report that the datasets contain specific biases (Buch et al., 2022; Lei et al., 2023). However, most studies on biases are restricted to isolated analyses of either the models or the datasets. This raises question: **Do VideoQA models learn to jointly leverage the modalities to achieve competitive performance on current benchmarks?**

To answer these questions, we propose QUAG (QUadrant AveraGe), a lightweight and non-parametric probe to systematically gauge the reliance of a finetuned model's performance on joint multimodal representations. QUAG impairs modality fusion by block-averaging attention weights. We apply QUAG on multiple dataset-model combinations, and consistently find that the models manage to achieve high performance on the benchmarks without leveraging multimodal representations. This finding is concerning because high performance on established benchmarks should be ideally indicative of coupled multimodal understanding. We

confirm this by achieving similar performance by replacing self-attention with QUAG-attention, a restricted variant of self-attention, without any finetuning. This might be because the models might be leveraging the spurious biases in the datasets (shortcuts), that weaken the modality coupling (we refer the degree of interdependence or interaction between different modalities as modality coupling (Xie & Liu, 2007; Hu et al., 2022; Liu et al., 2023). Then, **how do VideoQA models perform in highly-coupled multimodal settings?**

Thus, we create CLAVI (Complements in LAnguage and VIsion), a stress-test dataset with high multimodal coupling. It is created automatically by augmenting real-world videos and template-generated temporal questions (Figure 3). We observe that nearly all fine-tuned models exhibit high accuracy on shortcut instances (that is, lower multimodal coupling) in CLAVI, but most have near-trivial performance on highly-coupled multimodal instances. This reasserts the limitations of current models for learning highly-coupled multimodal representations, that is not tested in the current benchmarks.

In summary, our contributions are *(i)* We design QUAG, a systematic method for assessing the relative contribution of various multimodal components in the modality fusion stage; *(ii)* Using QUAG and QUAG-attention, we demonstrate that high performance on established VideoQA benchmarks is not completely representative of coupled multimodal understanding; and *(iii)* We develop CLAVI, a new stress-test for VideoQA with high multimodal coupling. Overall, QUAG and CLAVI demonstrate that the current VideoQA models are not proficient at learning highly-coupled multimodal representations, a crucial aspect that is not systematically tested in the current benchmarks.

2. Do VideoQA Models Learn to Jointly Leverage the Modalities?

Transformer architecture has become the de facto model for solving the VideoQA task in recent years. However, since they lack the relevant intrinsic inductive biases, they must learn it from the data (Xu et al., 2021). For an ideal multimodal model and dataset, high accuracy should be achievable only by aligning and leveraging both – the text (question) and the visual (video) modalities. Despite this, datasets often contain biases that may not necessitate multimodal understanding. However, the presence of spurious features in a dataset does not always imply that the model is exploiting them (Murali et al., 2023). Hence, we introduce QUAG for combined dataset-model representation analysis.

In the context of multimodal Transformers, we posit that modality fusion layers enable multimodal understanding by progressively attending to informative tokens within and be-

tween modalities. QUAG systematically ablates the effects of multimodal attention by augmenting self-attention in the fusion layer. For instance, for a given dataset, if the model heavily relies only on unimodal information, ablating the unimodal component of modality fusion should significantly decrease performance. Also, this decrease in performance should be more pronounced than the decrease from ablating the crossmodal component of attention. QUAG achieves such specific modality interaction ablation through block averaging, as explained below.

2.1. 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 $F_{\mathcal{V}} : \mathcal{V} \to \mathbb{R}^{l_{\mathcal{V}} \times d}$, text encoder $F_{\mathcal{T}} : \mathcal{T} \to \mathbb{R}^{l_{\mathcal{T}} \times d}$, a multimodal fusion module $M : (F_{\mathcal{V}}(\mathcal{V}), F_{\mathcal{T}}(\mathcal{T})) \to \mathbb{R}^{(l_{\mathcal{V}} + l_{\mathcal{T}}) \times d}$, and a classification layer, 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 model.

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. $\boldsymbol{X}_{\mathcal{V}\mathcal{T}} = [F_{\mathcal{V}}(\mathcal{V}) \parallel F_{\mathcal{T}}(\mathcal{T})] \in \mathbb{R}^{(l_{\mathcal{V}} + l_{\mathcal{T}}) \times d}$ is the input for M, where $\parallel$ is concatenation operator for the token dimension. Since QUAG is applied at inference time only, we assume the VideoQA model to be frozen.

2.2. QUAG: Ablation of Modality Interactions

Shortcuts are the spurious features learned by a given model on a given dataset (Murali et al., 2023). Along this axis, we use QUAG to pinpoint the exact failure modes in the datasets 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 $\boldsymbol{A}_i = \sigma(\boldsymbol{Q}_i \boldsymbol{K}_i^\top)$ is the attention matrix (we denote the softmax operation by σ and 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:

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

$$T_1 = \left[\begin{array}{c|c} A^1_{\mathcal{VV}} & A^1_{\mathcal{VT}} \\ \hline A^1_{\mathcal{TV}} & A^1_{\mathcal{TT}} \end{array} \right] \left[\begin{array}{c} V_{1\mathcal{V}} \\ \hline V_{1\mathcal{T}} \end{array} \right]$$

where $A^1_{u_1 u_2}$ represents the quadrant of A_1 corresponding to $(Q_{1u_1} K^\top_{1u_2})$. Note that we skip layer normalization in the discussion for simplicity. We can simplify T_1 as:

$$T_1 = \left[\begin{array}{c} A^1_{\mathcal{VV}} V_{1\mathcal{V}} + A^1_{\mathcal{VT}} V_{1\mathcal{T}} \\ \hline A^1_{\mathcal{TV}} V_{1\mathcal{V}} + A^1_{\mathcal{TT}} V_{1\mathcal{T}} \end{array} \right] \qquad (1)$$

Following the quadrant partitioning scheme of A_1 in M_1, we define quadrants for all the attention matrices in the downstream fusion layers.

Next, we define row-wise average-and-replace operator $\mathcal{R}$ that operates on a quadrant. $\mathcal{R}$ replaces all the values in a given partitioned row with the respective mean value. Note that the values in the other quadrants are unaffected. Given a matrix Z of size $p \times q$ and let W denote the location of the quadrant of 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, $\mathcal{R}$ is formally defined as:

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

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

$$\phi(A_i, [s_1, s_2, \cdots, s_m]) := (\mathcal{R}_{s_1} \circ \mathcal{R}_{s_2} \cdots \circ \mathcal{R}_{s_m}(A_i)),$$

where $S = [s_1, s_2, \cdots, s_m]$ is a list of quadrants such that $\forall s \in S : s \in \{\mathcal{TT}, \mathcal{TV}, \mathcal{VT}, \mathcal{VV}\}$, $\mathcal{R}_{s_i}(A_i)$ is short-hand for $\mathcal{R}(A_i, s_i)$, A_i is the attention matrix of fusion layer M_i. Note that $\mathcal{TT}$ refers to the quadrant $A^i_{\mathcal{TT}}$ of A_i and so on. Refer Figure 1 for an illustrative example. We define QUAG operation on a layer i as a modification of the existing attention matrix, that is, $A_i \leftarrow \phi(A_i, [s_1, s_2, \cdots, s_m])$. Incorporating QUAG into the existing model pipeline is straightforward and we provide the code in the Appendix A.4. Since we will be applying the QUAG operator successively on all the layers of M, for brevity, we denote $\Phi(M, n, S) = \forall_{i \in [1, \cdots, n]} A_i \leftarrow \phi(A_i, S)$. Note that QUAG is **light-weight**, **non-parametric**, requires **no fine-tuning** and operates at **inference time** for combined dataset-model analysis.

2.3. Short-circuit Operations

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

1) $S = [\mathcal{VV}, \mathcal{TT}]$: As evident from Eqn. 1, in the token-mixing step, the QUAG operation $A_1 \leftarrow \phi(A_1, [\mathcal{VV}, \mathcal{TT}])$ results in scaling the average values of $V_{1\mathcal{V}}$ and $V_{1\mathcal{T}}$ in the upper and lower blocks of T_1 respectively. Mathematically, we can represent it as $\phi(A_1, [\mathcal{VV}, \mathcal{TT}])_{\mathcal{VV}} V_{1\mathcal{V}} = \mu_{\mathcal{VV}} \mathbf{1} V_{1\mathcal{V}}$, where $\mu_{\mathcal{VV}}$ is a column vector containing row-wise average values of $A_{\mathcal{VV}}$ and $\mathbf{1}$ is a row vector of ones. Clearly, $\mathbf{1} V_{1\mathcal{V}}$ is equivalent to averaging across the video features. Similarly, $\mathcal{TT}$ features are averaged in the lower block. Note that the crossmodal attention weights, $A^1_{\mathcal{VT}}$ and $A^1_{\mathcal{TV}}$ are unchanged, and only the unimodal token mixing is *short-circuited*. We name such a fusion block as unimodal average conformable. And when applied to all the fusion blocks, the operation is known as **unimodal short-circuiting**. We provide a toy example in §A.1.

Theorem 2.1. *Unimodal short-circuiting produces unimodal average conformable features in all the fusion blocks.*

Proof. We use mathematical induction to prove that $\Phi(M, n, [\mathcal{VV}, \mathcal{TT}])$ makes all the n fusion blocks in M unimodal average conformable.

We have already explained the base case above.

Let us assume that the induction hypothesis is true for first $L - 1$ fusion blocks in M. That is, given that $\Phi(M, L - 1, [\mathcal{VV}, \mathcal{TT}])$ is unimodal average conformable, we want to show that $A_L \leftarrow \phi(A_L, [\mathcal{VV}, \mathcal{TT}])$ will make the output of M unimodal average conformable.

Because of skip connections in transformers, X_L (the input of M_L) can be decomposed as a function of $(X_{L-1} + T_{L-1})$. Note that while the function is non-linear, it is applied pointwise. That is, there is no token mixing and hence, the conformability is not affected. By the inductive hypothesis, we know that T_{L-1} is unimodal average conformable. We can recursively decompose X_{L-1} and show that the only unconformable component is the input, X_0. As shown in the base case, ϕ operated on a fusion block with input X_0 makes it unimodal average conformable. Hence, given $\Phi(M, L - 1, [\mathcal{VV}, \mathcal{TT}])$ is unimodal average conformable, $A_L \leftarrow \phi(A_L, [\mathcal{VV}, \mathcal{TT}])$ will make M_L unimodal average conformable. Hence, overall, $\Phi(M, n, [\mathcal{VV}, \mathcal{TT}])$ makes M unimodal average conformable. $\square$

Effectively, $\Phi(M, n, [\mathcal{VV}, \mathcal{TT}])$ bypasses the effect of video-video attention and text-text attention. We proved above that unimodal token-mixing is reduced to scaling the average of the modalities. That is, 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.

2) $S = [\mathcal{VT}, \mathcal{TV}]$: Parallel to unimodal short-circuiting, $\phi(A_1, [\mathcal{VT}, \mathcal{TV}])$ is equivalent to scaling the average val-

Figure 1. Illustrative toy example of row-wise average-and-replace operation or $\mathcal{R}(\boldsymbol{A}, \mathcal{VV})$, where $\boldsymbol{A}$ is the input attention matrix (left). The cells are colored as per their quadrants ($\mathcal{VV}$: red, $\mathcal{VT}$: yellow, $\mathcal{TV}$: blue, $\mathcal{TT}$: green). We apply $\mathcal{R}$ operator on the $\mathcal{VV}$ quadrant (highlighted in yellow) to replace the values with the respective row-wise average value (right).

ues of $\boldsymbol{V}_{1\mathcal{T}}$ and $\boldsymbol{V}_{1\mathcal{V}}$ in the upper and lower blocks of $\boldsymbol{T}_1$ respectively. While the crossmodal attention in video-text is reduced to uniform attention, the unimodal components of attention are unaffected. 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 information.

3) $\boldsymbol{S} = [\mathcal{VV}, \mathcal{TV}]$: This is equivalent to ablating 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, $\boldsymbol{S} = [\mathcal{TT}, \mathcal{VT}]$ leads to **text short-circuiting**. Singular values and their properties have been used to study the information richness of attention matrices. For example, Gui & Xiao (2023) use top-k singular values to measure the richness of attention matrix. We prove that video and text short-circuiting reduces the upper bound on the rank (the number of non-zero singular values) of the attention matrix. This in turn reduces the maximum representation power of the attention operation (Bhojanapalli et al., 2020).

Lemma 2.2. *For an attention matrix $\boldsymbol{A}$ of size $(l_\mathcal{V} + l_\mathcal{T}) \times (l_\mathcal{V} + l_\mathcal{T})$, the rank, ρ, after video short-circuiting is bounded as: $\rho(\phi(\boldsymbol{A}, [\mathcal{VV}, \mathcal{TV}])) \leq l_\mathcal{T} + 2$.*

We provide the proof in the §A.1. Note that the similar bound exists for text short-circuiting.

2.4. QUAG-attention

Along with the assessment of multimodal understanding, QUAG enables a detailed analysis of token mixing for identifying the sub-optimality of learned representations. Sub-optimality occurs if the modality fusion process doesn't effectively capture the information within each modality along with the complementary information in the other modality, even with computationally expensive operations like self-

attention. Hence, we propose QUAG-attention to rely only on the sub-optimal representations learnt by the model by restricted token mixing, without finetuning.

QUAG-attention, inspired by text and video short-circuiting operations, is a replacement of self-attention in fusion module that performs attention calculation and token-mixing on already short-circuited sequences. That is, QUAG-attention calculates attention using already averaged key tokens and applies on averaged value tokens, thus effectively pruning the number of keys and values (as opposed to block-averaging post attention calculation in QUAG).

As proved in Theorem 2.1, consistent averaging of partitions maintains their average conformability. Extending this idea for QUAG-attention, we average the partition block corresponding to the modality in each of the fusion layers before transforming the tokens into keys and values. For example, if the input of the model is $\boldsymbol{X}_{\mathcal{VT}} = [F_\mathcal{V}(\mathcal{V}) \,\|\, F_\mathcal{T}(\mathcal{T})]$, then consistently averaging the upper partition block (corresponding to the video modality), before key and value transformation would result in video-average QUAG-attention. QUAG-attention can be applied to only text, video or both the modalities. It reduces the number of key and value tokens from $(l_\mathcal{V} + l_\mathcal{T})$ to either $(l_\mathcal{T} + 1)$ (text-average), $(l_\mathcal{V} + 1)$ (video-average) or 2 (text-video-average) in all the multimodal fusion blocks while the number of query tokens remain the same. Each averaged token in QUAG-attention is representative of multiple tokens. This might cause the average tokens to have less contribution in the output of softmax operation. Hence, we use scale the softmax function (similar to proportional attention (Bolya et al., 2023) as $\boldsymbol{A} = \sigma(\frac{\boldsymbol{QK}^\top}{\sqrt{d}} + log(\boldsymbol{S}))$, where $\boldsymbol{S}$ is a row vector containing the effective size of each token (that is, the number of tokens the average token is representing)).

Similar to QUAG, QUAG-attention also require no finetuning or additional parameters. High accuracy with QUAG-average is a concern because that would mean that the model never learnt to leverage that modality effectively, yet manages to achieve high scores on the multimodal benchmark.

3. Simulation Study

The amount of modality coupling affects the kind of representations the model learns. We want to analyze the effects of degree of coupling between modalities on the learnt representations using QUAG. However, accurately characterizing the distribution of multimodal data is an intractable problem. Therefore, it is not possible to quantify the degree of coupling between the modalities. Hence, inspired by Huang et al. (2021), we propose a simple method of generating multimodal data with known degree of coupling.

Data Generation: Consider two modalities – t and v with dimensionality 100. Each sample contains 15 tokens from

each modality. Below is the sample generation process:

Step 1: Generate $m_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for $i = 1, 2$, where m_i is of size $[15 \times 100]$. Let $z = [m_1 \| m_2]$.

Step 2: Generate the 30-dimensional output vector y by taking weighted mean across the feature dimension. So, $[y]_i = \frac{1}{100} \sum_{j=1}^{100} p_j \cdot [z]_{ij}$, where $p_j = 10 \cdot \sin \frac{j\pi}{202}$ and $i = 1, 2, \cdots, 30$. We choose weighted mean over mean, because the mean value tends to zero.

Step 3: Generate t and v. We define the following transformations: $t \leftarrow (1 - \alpha) \cdot m_1 - \alpha \cdot m_2$ and $v \leftarrow (1 - \alpha) \cdot m_2 - \alpha \cdot m_1$, where $\alpha \in (0, 0.5)$

Where α is the crossmodal coupling coefficient. Increasing α increases the degree of crossmodal token mixing in the inputs. We use **modality coupling** to jointly refer to unimodal and crossmodal coupling in the proceeding sections. For simplicity we consider cross-modality coupling to be linear but it might be more complex for real data.
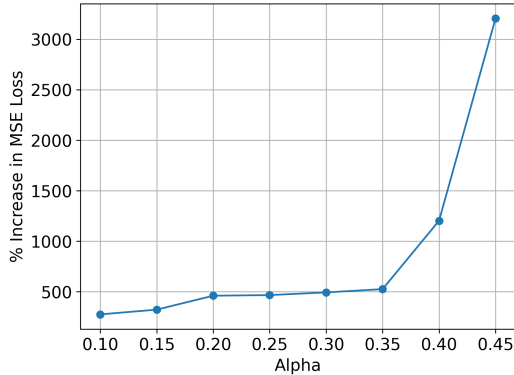

Figure 2. Result of the simulation study. The plot of percentage increase in the mean squared loss after crossmodal short-circuiting (y) versus α, the crossmodal coupling coefficient (x).

We train a simple transformer model on the crossmodal data for different values of α. The model and training details are provided in §A.5. As shown in Figure 2, we find that as α increases, the test error after crossmodal short-circuiting also increases. This is expected because increasing α would necessitate the model to rely increasingly on crossmodal information. Therefore, ablating crossmodal interactions with QUAG has increasingly deleterious effect. **This validates QUAG as a method of impairing multimodal fusion**. Using this understanding, in the next section, we apply QUAG to the models trained on real-world datasets to analyze the modality interactions in the model.

4. Combined Dataset-Model Analysis

Ideally, for optimal multimodal representations learnt on a dataset, we expect dramatic drops in all the short-circuiting

operations. However, datasets contain biases and the model may or may not learn them. Hence, we perform combined dataset-model analysis with QUAG to find the reliance of a model on the biases of a dataset through performance drops under short-circuiting operations.

We evaluate QUAG and QUAG-attention on JustAsk (Yang et al., 2021) and FrozenBiLM (Yang et al., 2022) models. We evalaute it on the following datasets ActivityNet-QA (Yu et al., 2019), MSRVTT-QA (Xu et al., 2017), and NeXT-QA (Xiao et al., 2021) We also report results on the ATP-Hard subset of NeXT-QA (Buch et al., 2022) that contains a higher concentration of temporally challenging data requiring multi-frame understanding. We provide the implementation details in §A.6

4.1. Analysis

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.

(1) 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. This means that FrozenBiLM consistently has strong reliance unimodal interactions and the text modality. However, the model learns to leverage crossmodal representations only for ActivityNet-QA and MSRVTT-QA and not NeXT-QA and ATP-Hard.

(2) On the other hand, the performance of JustAsk model does not drop by more than 1% for any of the datasets under any short-circuiting operation. That is, JustAsk model does not learn to align and fuse the modalities across the datasets.

While the relative performance drop in the *classical* language-only and video-only settings for JustAsk and FrozenBiLM on ActivityNet-QA, MSRVTT-QA and ATP-H is similar, QUAG, with combined dataset-model analysis, points out the differences in their representations.

The diagnostic results from QUAG provide confidence in achieving efficiency gains without significant performance drops by using QUAG-attention. We find that applying QUAG-attention on FrozenBiLM and JustAsk reduces the number of multiplication operations by **13.6%** and **68.0%**

Table 1. Short-circuit (SC) and QUAG-attention 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 (*: video-average for FrozenBiLM and video-text-average for JustAsk; [†]: % decrease in multiplication operations due to QUAG-attention; ↓: drop ≥ 2% with respect to baseline (the higher drop, the better).

	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
QUAG-atten*	43.0	45.8	55.6	55.9	38.0	41.0	53.5	44.1
ΔMultiplication Ops[†]	**13.6%**				**68.0%**			

respectively, for a less than 1% drop in performance consistently for all the datasets. This raises serious concerns because this would mean that the models can learn to *hack* their way around the accuracy metrics by leveraging shortcuts. Therefore, high accuracy on current datasets is not representative of multimodal understanding because the datasets contain exploitable biases that weaken the modality coupling.

5. Can VideoQA Models Learn Highly-Coupled Multimodal Representations?

In the previous sections we found that the models rely on different features on different datasets (Section 4). However, unlike the simulated datasets in Section 3, we do not know the degree of coupling between and within the modalities for real world datasets. Then, **how can we test the performance of current VideoQA models in tightly-coupled multimodal settings?** The results from QUAG and QUAG-attention suggest that the current VideoQA datasets might have many biases and hence might not be a good test of multimodal understanding. Therefore, we build CLAVI, a simple VideoQA stress-test dataset by augmenting real world videos and their associated questions to get their respective temporal complements, and exhaustively cover all the possible combinations, ensuring high multimodal coupling and coverage. CLAVI is not positioned to replace existing datasets but rather to supplement them for enhancing the understanding of VideoQA models.

5.1. CLAVI: Testing Through Complements

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 sin-

gle action (beginning/end) or the occurrence sequence for a pair of non-overlapping actions (before/after). CLAVI allows for systematic testing of coupled multimodal understanding through balanced video and question temporal complements. We use question templates to automatically curate the question-answer pairs from the temporal grounding annotations of Charades-STA (Gao et al., 2017). To create temporal complements in the question domain, we replace *before* with *after* and *beginning* with *end* and vice versa. Further, we create temporal video complements by swapping only the action-segments in the video (Figure 3).

Many stress-test and diagnostic datasets have artefacts or discontinuous features (Wang et al., 2023b; Zhang et al., 2020; Bomatter et al., 2021), like the two points of discontinuity in the complement video of CLAVI. However, we ensure that it does not serve as a bias by maximizing the coverage of the dataset by exhaustively considering all the compositions of complements in video and question domains. This is also verified in our results in the following section where we test models with low, medium and high frame sampling rates on CLAVI. The exhaustivity also ensures tighter coupling within and between the modalities, as explained below.

We briefly explain the design principle of CLAVI. We choose temporal sequence complements to test coupled multimodal understanding because it requires **unimodal understanding** within the modalities (sensitive to the sequence of *(i)* frames in the video; *(ii)* objects, verbs and temporal phrases in the question) as well as **crossmodal understanding** (relating the sequence of actions in the video with that of the question). This also makes temporal ordering as one of the **fundamental** elements of VideoQA. Using yes-no questions with balanced negative instances allows us to have questions that are **unambiguous**, and answers that are **mutually exclusive** and **equally informative** to not be eliminated by prior biased knowledge. We deliberately
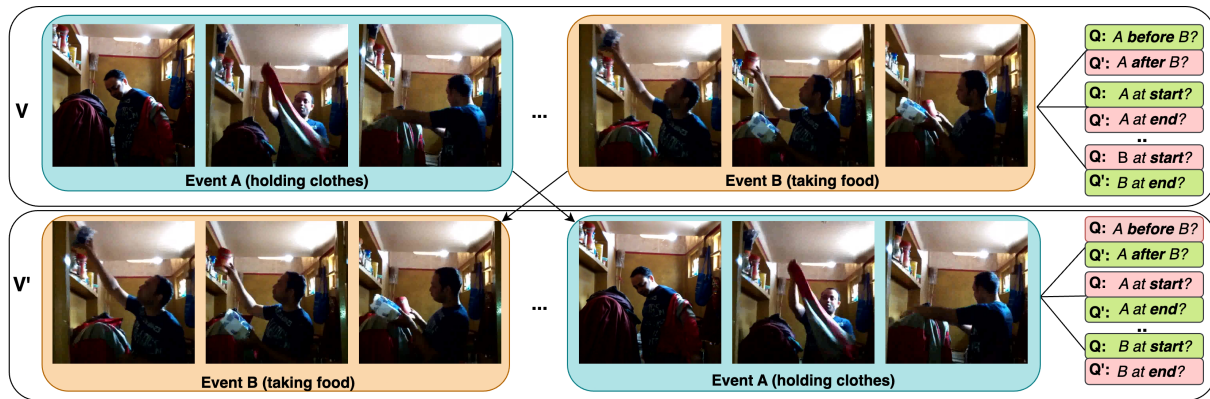
Figure 3. Illustrative example of the creation of CLAVI. In the original video (V), the action "holding clothes" (Event A; blue pane) follows "taking food" (Event B; brown pane). To create a complement video (V'), we swap the action segments without manipulating the segment separating them. The questions (Q), along with their complement (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 Table 2.

Table 2. List of questions and their complements in CLAVI for the illustrative example in Fig. 3 (E: Existence, BE:Beginning/End, BA:Before/After, and NC:Negative Control). NC questions contain an action ("washing mirror", here) that never occurs in the video and hence, the answer is always "no". For brevity, we present 4 (out of 8) NC questions for BA type; comprehensive list in §B.3.

Type	Question (Q)	Complement Question (Q')
E	Was someone holding clothes? Was someone taking food?	
E-NC	Was someone washing mirror?	
BE	Was the person holding clothes at the **beginning**? Was the person taking food at the **end**?	Was the person holding clothes at the **end**? Was the person taking food at the **beginning**?
BA	Did holding clothes happen **before** taking food? Did taking food happen **after** holding clothes?	Did holding clothes happen **after** taking food? Did taking food happen **before** holding clothes?
BA-NC	Did holding clothes happen **before** washing mirror? Did taking food happen **after** washing mirror?	Did holding clothes happen **after** washing mirror? Did taking food happen **before** washing mirror?

maintain a simple design for question templates and answer vocabulary that excludes other abilities such as language comprehension, commonsense reasoning, and long-term memory to facilitate **isolated analysis and testing** of joint multimodal understanding. Also, we ensure that the dataset size is sufficiently large, as compared to the existing datasets, so that the models do not overfit (§B.2).

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 questions containing actions that do not occur in the video (that is, the answer is always "no") for E and BA types as shown in Table 2. Answering the negative control does not require understanding temporal cues. Hence, it serves the dual purpose of sanity check of learning and a baseline for learning by temporal shortcuts (low coupling instances). We remove the bias against *beginning* and *end* by randomly extending the boundaries of the action-segments in the video (detailed curation process in §B.1).

We want to evaluate the sensitivity of the model to the temporal cues in language and video separately. Hence, we define consistent **video-consistent accuracy (CAcc$_\mathcal{V}$)** and **text-consistent accuracy (CAcc$_\mathcal{T}$)**.

$$\text{CAcc}_\mathcal{V} = \mathbb{1}[(y_{(V,Q)} == \hat{y}_{(V,Q)}) \wedge (y_{(V',Q)} == \hat{y}_{(V',Q)})]$$
$$\text{CAcc}_\mathcal{T} = \mathbb{1}[(y_{(V,Q)} == \hat{y}_{(V,Q)}) \wedge (y_{(V,Q')} == \hat{y}_{(V,Q')})]$$

where y and $\hat{y}$ are the ground truth and predictions, with the video and question indicated in the subscript and $\mathbb{1}$ is the indicator function. Note that ($'$) indicates the complement in CLAVI. We average the consistent accuracies over the entire dataset and report it separately for the **control subset** (E, E-NC, and BA-NC question types) and the **complement subset** (BE and BA question types). The coupling between the modalities is lower in the control subset since can be answered by even ignoring the temporality in video and text while answering the complement subset (higher modality coupling) necessitates joint multimodal understanding.

7

5.2. Experiment and Analysis

We finetune and evaluate 4 models: JustAsk (Yang et al., 2021) (640 frames), FrozenBiLM (Yang et al., 2022) (10 frames), Singularity-Temporal (Lei et al., 2023) (12 frames) and All-In-One+ (Wang et al., 2023a) (3 frames) on CLAVI using the official finetuning instructions (§B.5). We do not report zero-shot results because the models almost always answer either of the two answers, which is a well known phenomenon in multimodal QA (Guo et al., 2022). We follow the same experimental settings as discussed in Section 4. To account for class imbalance in the answers, we use balanced accuracy for validation and testing.

The results are summarized in Table 3. All the models have greater than 70% balanced accuracy. Next, we do a finegrained analysis of the performance.

Text and video consistent accuracies are greater than 90% for the control subset for all the models. This is because, unlike the complement subset, the control subset requires less 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 complement 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 complement subset, except for FrozenBiLM. Overall, **most models have trivial performance on high-multimodal coupling subset**.

No dataset can be free of biases. Therefore, while low performance on CLAVI is representative of poor coupled multimodal understanding, high performance doesn't necessarily guarantee it. How can we be sure that FrozenBiLM is not learning spurious shortcuts on CLAVI? We find that the video-average QUAG-attention on FrozenBiLM cause the $CAcc_\mathcal{T}$-complement and $CAcc_\mathcal{V}$-complement to drop to **23%** and **3.6%** respectively. That is, the performance on the complement subset significantly drops under multimodal impairment. However, $CAcc_\mathcal{T}$-control and $CAcc_\mathcal{V}$-control values increase to **98.6%** and **99.2%** respectively, perhaps because QUAG-attention promotes reliance on shortcuts, and the control subset can be solved easily by shortcuts. These results confirm FrozenBiLM's high reliance on multimodal representations for its higher performance.

Beyond the consistency accuracy metrics we can use CLAVI for diverse representation analyses. As an example, we present a qualitative representation sensitivity analysis for FrozenBiLM in §B.7. We align the attention matrices for complement pairs and find that the representations of correctly answered complement pairs are more distinct than the wrongly answered pairs to support our understanding.

6. Related Work

Dataset Biases: Works in NLP (Papadimitriou et al., 2022; Sinha et al., 2021), vision (Brendel & Bethge, 2019) and vision-language (Yuksekgonul et al., 2023) 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 (Pham et al., 2021; Yuksekgonul et al., 2023; Kafle & Kanan, 2017; Sevilla-Lara et al., 2021). For VideoQA, MovieQA (Tapaswi et al., 2016) and TVQA (Lei et al., 2018) datasets are biased towards plot understanding or dialogue comprehension (Winterbottom et al., 2020). Biases are not always immediately apparent; for example, Social-IQ (Zadeh et al., 2019) contains sentiment-biased annotations (Gat et al., 2021). Moreover, statistical regularities like answer length, answer frequency (Goyal et al., 2017; Agrawal et al., 2016) and co-occurrence (Dancette et al., 2021; Manjunatha et al., 2019; Subramanian et al., 2019) introduce spurious features. Overall, these biases allow the models learn shortcuts (Geirhos et al., 2020) that circumvent multimodal reasoning (Chao et al., 2018; Ye & Kovashka, 2021). While synthetic VideoQA benchmarks such as VQuAD (Gupta et al., 2022), CLEVRER (Yi et al., 2020) have been carefully curated to mitigate many biases, they are unable to capture the intricate dynamics of the real world. Recently proposed Preception Test (Pătrăucean et al., 2023), while comprehensive, does not contain diagnostic metrics that penalize the effect of shortcut temporal learning. 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 (Cadène et al., 2019; Clark et al., 2019), Lei et al. (2023) and Winterbottom et al. (2020) achieve state-of-the-art performance with simple models by leveraging VideoQA dataset shortcuts in the model. ATP (Buch et al., 2022) demonstrates single frame bias by re-training the models with an informative frame-selection module to achieve competitive performance. Perceptual Score (Gat et al., 2021) quantifies modality bias in terms of relative performance drop under modality-permutation operation. QUAG combines these ideas to evaluate the dependence of models on shortcuts for circumventing multimodal understanding in terms of performance drop under multimodal representation collapse. Unlike others, it assists in identifying sub-optimal representations in a combined model-dataset approach at test time.

Leveraging Complements: We share our motivation for developing CLAVI with VQA-CP (Agrawal et al., 2018):

Table 3. Test performance in terms of our consistent accuracy metric (CAcc) on CLAVI after finetuning. The control subset can be answered by ignoring temporality, and hence has low multimodal coupling while the complement subset has high multimodal coupling (red color denotes near-trivial performance on complement subset and green denotes high performance (the more the better)).

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

that iid train-test splits in the presence of strong priors leads to learning via shortcuts. However, rather than reducing the bias by mining new complementary image instances, CLAVI weakens prior of multimodal understanding with synthesized balanced video-question temporal hard-negatives. Momeni et al. (2023) and Wang et al. (2023b) have employed hard-negatives for improving verb-understanding in VideoQA models. CLAVI is inspired by Bagad et al. (2023) stitch video clips to improve the temporal understanding of video-language models. Unlike previous works, we (i) introduce atemporal question types, and (ii) propose a new diagnostic metric (consistent accuracy) that are needed for understanding multimodality in VideoQA models.

Multimodal Fusion Interpretability and Visualization: Liang et al. (2024) and Liang et al. (2023a) analyze multimodal fusion interactions along the dimensions of response, information, and mechanics. The closest alignment of QUAG and CLAVI is at the interface of multimodal fusion response and mechanics. Previous works have quantified the presence or absence of specific kinds of modality interactions through the study of datasets (Dancette et al., 2021), models (Chefer et al., 2021), projections onto simpler models (Hessel & Lee, 2020; Wörtwein et al., 2022) and visualization studies (Liang et al., 2023b; Aflalo et al., 2022; Wang et al., 2022). However, using QUAG, unlike other methods, we perform a combined dataset-model analysis without any additional parameters or finetuning.

7. Conclusion

In this work, we perform a rigorous analysis of VideoQA models, focusing on multimodal representations. We introduced QUAG, a coupled dataset-model approach, to conduct a systematic analysis of learned multimodal representations. It provides deep insights into *how* the models infer and *why* the models fail. We found that VideoQA models do not necessarily learn to align and fuse the information both – within

and between the modalities from multimodal datasets. Using this understanding, we developed QUAG-attention and exposed the sub-optimality of VideoQA models. Since the datasets might not have high coupling, we proposed CLAVI, a stress-test dataset for testing highly-coupled multimodal understanding in VideoQA models. With the simple task of temporal ordering we find that most of the current models are unable to jointly infer from text and video modalities.

All our proposed approaches – QUAG, QUAG-attention and CLAVI are simple, compute-friendly and generic to be extended to any combination of modalities, datasets and models. Our thorough and systematic dataset-model combined representation analysis asserts the inability of VideoQA models to learn highly-coupled multimodal settings, that is not evaluated by current datasets.

Acknowledgements

Impact Statement

Datasets in machine learning have a history of containing unintentional biases like race, gender, age along with safety and privacy concerns (Birhane & Prabhu, 2021; Peng et al., 2021; Hirota et al., 2022). We curate CLAVI from existing and popular Charades (Sigurdsson et al., 2016) dataset because it is well-studied, licensed for academic usage, and collected in controlled settings with consent. Further the owners of Charades ensure the anonymity and privacy of the participants. Also, for automatically generated questions, we ensure to keep the question templates gender neutral. One of the central goals of our work was to reassert the brittleness in multimodal models by presenting a combined dataset-model centric interpretable representation learning approach through QUAG and CLAVI. We hope our work

galvanizes the research community further to not just blindly trust the accuracy score on benchmarks but thoroughly investigate the potential biases that are *(1)* present in the dataset and *(2)* are learned by the models.

References

Aflalo, E., Du, M., Tseng, S., Liu, Y., Wu, C., Duan, N., and Lal, V. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 21374–21383. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02072. URL https://doi.org/10.1109/CVPR52688.2022.02072.

Agrawal, A., Batra, D., and Parikh, D. Analyzing the Behavior of Visual Question Answering Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1955–1960, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1203. URL https://aclanthology.org/D16-1203.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4971–4980. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00522. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html.

Bagad, P., Tapaswi, M., and Snoek, C. G. 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*, pp. 2503–2516, 2023.

Bhojanapalli, S., Yun, C., Rawat, A. S., Reddi, S., and Kumar, S. Low-Rank Bottleneck in Multi-head Attention Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 864–873. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bhojanapalli20a.html.

Birhane, A. and Prabhu, V. U. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token Merging: Your ViT But Faster. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Bomatter, P., Zhang, M., Karev, D., Madan, S., Tseng, C., and Kreiman, G. When Pigs Fly: Contextual Reasoning in Synthetic and Natural Scenes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 255–264. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00032. URL https://doi.org/10.1109/ICCV48922.2021.00032.

Brendel, W. and Bethge, M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkfMWhAqYQ.

Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., and Niebles, J. C. Revisiting the "Video" in Video-Language Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 2907–2917. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00293. URL https://doi.org/10.1109/CVPR52688.2022.00293.

Cadène, R., Dancette, C., Ben-younes, H., Cord, M., and Parikh, D. RUBi: Reducing Unimodal Biases for Visual Question Answering. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 839–850, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/51d92be1c60d1db1d2e5e7a07da55b26-Abstract.html.

Castro, S., Wang, R., Huang, P., Stewart, I., Ignat, O., Liu, N., Stroud, J., and Mihalcea, R. 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)*, pp. 2925–2940, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.209. URL https://aclanthology.org/2022.acl-long.209.

Chao, W.-L., Hu, H., and Sha, F. 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)*, pp. 431–441, New Orleans, Louisiana, 2018. Association for Com-

putational Linguistics. doi: 10.18653/v1/N18-1040. URL https://aclanthology.org/N18-1040.

Chefer, H., Gur, S., and Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 387–396. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00045. URL https://doi.org/10.1109/ICCV48922.2021.00045.

Clark, C., Yatskar, M., and Zettlemoyer, L. 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)*, pp. 4069–4082, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1418. URL https://aclanthology.org/D19-1418.

Crouse, D. F. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. doi: 10.1109/TAES.2016.140952.

Dancette, C., Cadène, R., Teney, D., and Cord, M. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 1554–1563. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00160. URL https://doi.org/10.1109/ICCV48922.2021.00160.

Gao, J., Sun, C., Yang, Z., and Nevatia, R. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5277–5285. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.563. URL https://doi.org/10.1109/ICCV.2017.563.

Gat, I., Schwartz, I., and Schwing, A. G. Perceptual Score: What Data Modalities Does Your Model Perceive? In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 21630–21643, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b51a15f382ac914391a58850ab343b00-Abstract.html.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov 2021. ISSN 0001-0782. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Gui, A. and Xiao, H. HiFi: High-information attention heads hold for parameter-efficient model adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8521–8537, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.475. URL https://aclanthology.org/2023.acl-long.475.

Guo, Y., Nie, L., Cheng, Z., Tian, Q., and Zhang, M. Loss Re-Scaling VQA: Revisiting the Language Prior Problem From a Class-Imbalance View. *IEEE Transactions on Image Processing*, 31:227–238, 2022. doi: 10.1109/TIP.2021.3128322.

Gupta, V., Patro, B. N., Parihar, H., and Namboodiri, V. P. VQuAD: Video Question Answering Diagnostic Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 282–291, January 2022.

Hao, J., Sun, H., Ren, P., Wang, J., Qi, Q., and Liao, J. 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*, pp. 130–147. Springer, 2022.

Hessel, J. and Lee, L. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62. URL https://aclanthology.org/2020.emnlp-main.62.

Hirota, Y., Nakashima, Y., and Garcia, N. Gender and Racial Bias in Visual Question Answering Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1280–1292, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533184. URL https://doi.org/10.1145/3531146.3533184.

Hu, F., Weinstein, S. M., Baller, E. B., Valcarcel, A. M., Adebimpe, A., Raznahan, A., Roalf, D. R., Robert-Fitzgerald, T. E., Gonzenbach, V., Gur, R. C., Gur, R. E., Vandekar, S., Detre, J. A., Linn, K. A., Alexander-Bloch, A., Satterthwaite, T. D., and Shinohara, R. T. Voxel-wise intermodal coupling analysis of two or more modalities using local covariance decomposition. *Human Brain Mapping*, 43(15):4650–4663, June 2022.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. What Makes Multi-Modal Learning Better than Single (Provably). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 10944–10956, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5aa3405a3f865c10f420a4a7b55cbff3-Abstract.html.

Jahagirdar, S., Mathew, M., Karatzas, D., and Jawahar, C. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4441–4450, 2023.

Jia, B., Lei, T., Zhu, S.-C., and Huang, S. Ego-TaskQA: Understanding Human Tasks in Egocentric Videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/161c94a58ca25bafcaf47893e8233deb-Paper-Datasets_and_Benchmarks.pdf.

Kafle, K. and Kanan, C. An Analysis of Visual Question Answering Algorithms. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1983–1991. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.217. URL https://doi.org/10.1109/ICCV.2017.217.

Kowal, M., Siam, M., Islam, M. A., Bruce, N. D. B., Wildes, R. P., and Derpanis, K. G. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New*

Orleans, LA, USA, June 18-24, 2022, pp. 13979–13989. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01361. URL https://doi.org/10.1109/CVPR52688.2022.01361.

Lei, J., Yu, L., Bansal, M., and Berg, T. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167.

Lei, J., Berg, T., and Bansal, M. 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)*, pp. 487–507, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.29. URL https://aclanthology.org/2023.acl-long.29.

Liang, P. P., Cheng, Y., Salakhutdinov, R., and Morency, L.-P. Multimodal Fusion Interactions: A Study of Human and Automatic Quantification. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, pp. 425–435, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9798400700552. doi: 10.1145/3577190.3614151. URL https://doi.org/10.1145/3577190.3614151.

Liang, P. P., Lyu, Y., Chhablani, G., Jain, N., Deng, Z., Wang, X., Morency, L.-P., and Salakhutdinov, R. MultiViz: Towards Visualizing and Understanding Multimodal Models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=i2_TvOFmEml.

Liang, P. P., Zadeh, A., and Morency, L.-P. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ACM Comput. Surv.*, 2024. ISSN 0360-0300. doi: 10.1145/3656580. URL https://doi.org/10.1145/3656580. Just Accepted.

Liu, Y., Huang, Y., Wang, S., Lu, W., and Wu, H. Modality Coupling for Privacy Image Classification. *IEEE Transactions on Information Forensics and Security*, 18: 4843–4853, 2023. doi: 10.1109/TIFS.2023.3301414.

Manjunatha, V., Saini, N., and Davis, L. S. Explicit Bias Discovery in Visual Question Answering Models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 9562–9571. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00979. URL

http://openaccess.thecvf.com/content_CVPR_2019/html/Manjunatha_Explicit_Bias_Discovery_in_Visual_Question_Answering_Models_CVPR_2019_paper.html.

Meyer, Jr, C. D. Generalized inverses and ranks of block matrices. *SIAM Journal on Applied Mathematics*, 25(4): 597–602, 1973.

Momeni, L., Caron, M., Nagrani, A., Zisserman, A., and Schmid, C. Verbs in action: Improving verb understanding in video-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15533–15545, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.01428. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01428.

Murali, N., Puli, A. M., Yu, K., Ranganath, R., and kayhan Batmanghelich. Beyond Distribution Shift: Spurious Features Through the Lens of Training Dynamics. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=Tkvmt9nDmB.

Otani, M., Nakashima, Y., Rahtu, E., and Heikkilä, J. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0306.pdf.

Papadimitriou, I., Futrell, R., and Mahowald, K. 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)*, pp. 636–643, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.71. URL https://aclanthology.org/2022.acl-short.71.

Peng, K., Mathur, A., and Narayanan, A. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/077e29b11be80ab57e1a2ecabb7da330-Paper-round2.pdf.

Pham, T., Bui, T., Mai, L., and Nguyen, A. 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*, pp. 1145–1160, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.98. URL https://aclanthology.org/2021.findings-acl.98.

Pătrăucean, V., Smaira, L., Gupta, A., Continente, A. R., Markeeva, L., Banarse, D., Koppula, S., Heyward, J., Malinowski, M., Yang, Y., Doersch, C., Matejovicova, T., Sulsky, Y., Miech, A., Frechette, A., Klimczak, H., Koster, R., Zhang, J., Winkler, S., Aytar, Y., Osindero, S., Damen, D., Zisserman, A., and Carreira, J. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HYEGXFnPoq.

Sevilla-Lara, L., Zha, S., Yan, Z., Goswami, V., Feiszli, M., and Torresani, L. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 535–544, 2021.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 510–526. Springer, 2016.

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. 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*, pp. 2888–2913, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.230. URL https://aclanthology.org/2021.emnlp-main.230.

Subramanian, S., Singh, S., and Gardner, M. Analyzing compositionality in visual question answering. In *Visually Grounded Interaction and Language Workshop*, 2019.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. MovieQA: Understanding Stories in Movies through Question-Answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4631–4640. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.501. URL https://doi.org/10.1109/CVPR.2016.501.

Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K. Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., and Shou, M. Z. All in One: Exploring Unified Video-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference*

on *Computer Vision and Pattern Recognition (CVPR)*, pp. 6598–6608, June 2023a.

Wang, X., He, J., Jin, Z., Yang, M., Wang, Y., and Qu, H. M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2022. doi: 10.1109/TVCG.2021.3114794.

Wang, Z., Blume, A., Li, S., Liu, G., Cho, J., Tang, Z., Bansal, M., and Ji, H. Paxion: Patching Action Knowledge in Video-Language Foundation Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20729–20749, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/420492060687ca7448398c4c3fa10366-Paper-Conference.pdf.

Winterbottom, T., Xiao, S., McLean, A., and Moubayed, N. A. On Modality Bias in the TVQA dataset. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0476.pdf.

Wörtwein, T., Sheeber, L., Allen, N., Cohn, J., and Morency, L.-P. Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4681–4696, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.344.

Wu, B., Yu, S., Chen, Z., Tenenbaum, J., and Gan, C. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5ef059938ba799aaa845e1c2e8a762bd-Paper-round2.pdf.

Xiao, J., Shang, X., Yao, A., and Chua, T. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00965. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00965.

Xiao, J., Zhou, P., Chua, T.-S., and Yan, S. 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, pp. 39–58, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20058-8. doi: 10.1007/978-3-031-20059-5_3. URL https://doi.org/10.1007/978-3-031-20059-5_3.

Xie, L. and Liu, Z.-Q. A coupled HMM approach to video-realistic speech animation. *Pattern Recognition*, 40(8):2325–2340, 2007. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2006.12.001. URL https://www.sciencedirect.com/science/article/pii/S0031320306004973. Part Special Issue on Visual Information Processing.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pp. 1645–1653, 2017. doi: 10.1145/3123266.3123427. URL https://doi.org/10.1145/3123266.3123427.

Xu, Y., ZHANG, Q., Zhang, J., and Tao, D. ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28522–28535, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/efb76cff97aaf057654ef2f38cd77d73-Paper.pdf.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 1666–1677. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00171. URL https://doi.org/10.1109/ICCV48922.2021.00171.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 124–141, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/00d1f03b87a401b1c7957e0cc785d0bc-Paper-Conference.pdf.

Ye, K. and Kovashka, A. A Case Study of the Shortcut Effects in Visual Commonsense Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3181–3189, May 2021. doi: 10.1609/aaai.v35i4.16428. URL https://ojs.aaai.org/index.php/AAAI/article/view/16428.

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. CLEVRER: collision events for video representation and reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9127–9134, Jul. 2019. doi: 10.1609/aaai.v33i01.33019127. URL https://ojs.aaai.org/index.php/AAAI/article/view/4946.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It? In *The Eleventh International Conference on Learning Representations* , 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Zadeh, A., Chan, M., Liang, P. P., Tong, E., and Morency, L. Social-IQ: A question answering benchmark for artificial social intelligence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8807–8817. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00901. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.

Zhang, M., Tseng, C., and Kreiman, G. Putting Visual Object Recognition in Context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12982–12991. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01300. URL https://doi.org/10.1109/CVPR42600.2020.01300.

Zhong, Y., Ji, W., Xiao, J., Li, Y., Deng, W., and Chua, T.-S. Video Question Answering: Datasets, Algorithms and Challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6439–6455, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.432.

A. QUAG

A.1. Toy Example

Consider the toy example in Fig 4. The left-most matrix is the input matrix. As per the definition of ϕ, we can write, $\phi(Z, [\mathcal{TT}, \mathcal{VV}]) = \mathcal{R}_{\mathcal{TT}} \circ \mathcal{R}_{\mathcal{VV}}(Z)$ (which we define as unimodal short-circuiting operation). We demonstrate the successive application of $\mathcal{R}$ operator in the example. Note that the padding is ignored; this is equivalent to applying $\mathcal{R}$ to the padding-free sub-partition of the quadrant. Also, as illustrated in the example, since the quadrants cannot overlap, the sequence of application of $\mathcal{R}$ does not matter.
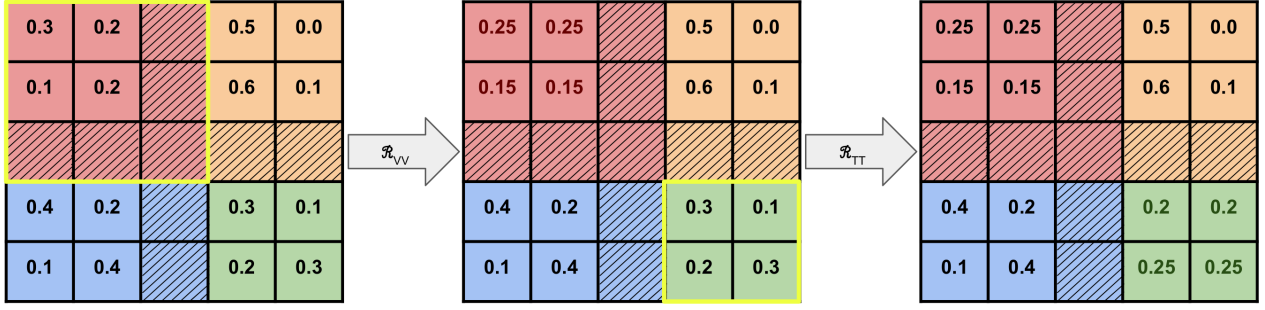

Figure 4. Toy example of $\phi(Z, [\mathcal{TT}, \mathcal{VV}])$, where Z is the input (left-most matrix) $\mathcal{R}$ is the row-wise average and replace operator and hatching denotes padding. The quadrants that are operated on are highlighted in bright yellow box. Note that $L_{\mathcal{V}} = 3$ and $L_{\mathcal{T}} = 2$, such that video embeddings are pre-concatenated to question embeddings (as in the main manuscript). The cells are colored as per their quadrants ($\mathcal{VV}$: red, $\mathcal{VT}$: yellow, $\mathcal{TV}$: blue, $\mathcal{TT}$: green)

A.2. Proof of Lemma 2.2

Lemma 2.2. *For an attention matrix $\boldsymbol{A}$ of size $(l_{\mathcal{V}} + l_{\mathcal{T}}) \times (l_{\mathcal{V}} + l_{\mathcal{T}})$, the rank, ρ, after video short-circuiting is bounded as:* $\rho(\phi(\boldsymbol{A}, [\mathcal{VV}, \mathcal{TV}])) \leq l_{\mathcal{T}} + 2$.

Proof. The attention matrix $\boldsymbol{A}$ can be written in terms of its partitions as: $\begin{pmatrix} A_{\mathcal{VV}} & A_{\mathcal{VT}} \\ A_{\mathcal{TV}} & A_{\mathcal{TT}} \end{pmatrix}$. Also, $\rho(\boldsymbol{A}) \leq l_{\mathcal{V}} + l_{\mathcal{T}}$ Then,

$$\phi(\boldsymbol{A}, [\mathcal{VV}, \mathcal{TV}]) = \begin{pmatrix} \begin{pmatrix} \mu_1 & \cdots & \mu_1 \\ . & \cdots & . \\ . & \cdots & . \\ \mu_{l_{\mathcal{V}}} & \cdots & \mu_{l_{\mathcal{V}}} \end{pmatrix} & A_{\mathcal{VT}} \\ \begin{pmatrix} \mu_{l_{\mathcal{V}}+1} & \cdots & \mu_{l_{\mathcal{V}}+1} \\ . & \cdots & . \\ . & \cdots & . \\ \mu_{l_{\mathcal{T}}+l_{\mathcal{V}}} & \cdots & \mu_{l_{\mathcal{T}}+l_{\mathcal{V}}} \end{pmatrix} & A_{\mathcal{TT}} \end{pmatrix}$$, where μ_i is the mean of the ith row of the partition.

Assuming $\mu_1 \neq 0$, we apply the following row operation on the matrix: $R_j \leftarrow R_j - \frac{\mu_j}{\mu_1} R_1$, for $j = 2, 3, \cdots, (l_{\mathcal{V}} + l_{\mathcal{T}})$, where R_k represents k^{th} row of $\phi(\boldsymbol{A}, [\mathcal{VV}, \mathcal{TV}])$. The resulting matrix is:

$$\begin{pmatrix} \begin{pmatrix} \mu_1 & \cdots & \mu_1 \\ . & \cdots & . \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{pmatrix} & \tilde{A}_{\mathcal{VT}} \\ \begin{pmatrix} 0 & \cdots & 0 \\ . & \cdots & . \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{pmatrix} & \tilde{A}_{\mathcal{TT}} \end{pmatrix}$$

Rearranging the partitions, we can write,

$$
\begin{pmatrix}
\begin{pmatrix} \mu_1 & \cdots & \mu_1 \end{pmatrix} & \hat{A}_{\mathcal{VT}} \\
\begin{pmatrix} 0 & \cdots & 0 \\ . & \cdots & . \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{pmatrix} & \hat{A}_{\mathcal{TT}}
\end{pmatrix}
\equiv
\begin{pmatrix}
[\hat{A}_{\mathcal{VV}}]_{1\times l_\mathcal{V}} & [\hat{A}_{\mathcal{VT}}]_{1\times l_\mathcal{T}} \\
[\hat{A}_{\mathcal{TV}}]_{(l_\mathcal{V}+l_\mathcal{T}-1)\times l_\mathcal{V}} & [\hat{A}_{\mathcal{TT}}]_{(l_\mathcal{V}+l_\mathcal{T}-1)\times l_\mathcal{T}}
\end{pmatrix}
\equiv \hat{A}
$$

where $[.]_{a\times b}$ denotes that the shape of the partition is $a \times b$.

Using the rank-inequality result of Meyer (1973), we can write:

$$
\rho(\hat{A}) \leq \rho(\hat{A}_{\mathcal{VV}}) + \rho(\hat{A}_{\mathcal{TV}}) + \rho(\hat{A}_{\mathcal{VT}}) + \rho(\hat{A}_{\mathcal{TT}} - \hat{A}_{\mathcal{TV}}\hat{A}_{\mathcal{VV}}^{-}\hat{A}_{\mathcal{VT}}) \tag{2}
$$

where $\hat{A}_{\mathcal{VV}}^{-}$ denotes the Moore-Penrose pseudo-inverse of the partition. Since $\hat{A}_{\mathcal{TV}}$ is a zero matrix, $\rho(\hat{A}_{\mathcal{TV}}) = 0$, and $\rho(\hat{A}_{\mathcal{TT}} - \hat{A}_{\mathcal{TV}}\hat{A}_{\mathcal{VV}}^{-}\hat{A}_{\mathcal{VT}}) = \rho(\hat{A}_{\mathcal{TT}}) \leq l_\mathcal{T}$ (we assume that $l_\mathcal{V} > 1$ for all practical purposes). Also, $\rho(\hat{A}_{\mathcal{VV}}) = 1$, and $\rho(\hat{A}_{\mathcal{VT}}) \leq 1$. Plugging these values in Equation (2), we get, $\rho(\hat{A}) \leq 1 + 0 + 1 + l_\mathcal{T} \implies \rho(\hat{A}) \leq l_\mathcal{T} + 2$.

Since row operations do not change column rank, $\rho(\hat{A}) = \rho(\phi(A, [\mathcal{VV}, \mathcal{TV}]))$. Therefore, $\rho(\phi(A, [\mathcal{VV}, \mathcal{TV}])) \leq l_\mathcal{T} + 2$

$\square$

A.3. Attention Map Visualization

We provide a visualization example of the attention values before and after short-circuting operations in Figure 5.

A.4. 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.

```
1  def self_attention(inputs, mask, dim_model, l_v, l_t, quads):
2      # Inputs:
3      #   inputs: Tensor of shape (batch_size, sequence_length, dim_model)
4      #   mask: Tensor of shape (batch_'size, sequence_length)
5      #   dim_model: Dimension of the model (e.g., 512)
6      #   l_v: int    maximum length of video tokens
7      #   l_t: int    maximum length of question tokens
8      #   quads: list containing elements from {'VV', 'VT', 'TV', 'TT'}
9      query = linear_transform_query(inputs)
10     key = linear_transform_key(inputs)
11     value = linear_transform_value(inputs)
12     attention_scores = compute_attention_scores(query, key, mask)
13     apply_quag(attention_scores, mask, l_v, l_t, quads)
14     attended_output = apply_attention_scores(attention_scores, value)
15     return attended_output
16
17 def compute_attention_scores(query, key, mask):
18     scaled_dot_product = dot_product(query, key) / sqrt(dim_model)
19     attention_scores = softmax(scaled_dot_product + (1 - mask) * -1e9)
20     return attention_scores
21
22 def apply_quag(attention_scores, mask, l_v, l_t, quads):
23     if 'VV' is in quads:
24         replace_with_rowwise_average(attention_scores[:, :l_v, :l_v], mask[:, :l_v, :
       l_v])
25     if 'VT' is in quads:
26         replace_with_rowwise_average(attention_scores[:, :l_v, -l_t:], mask[:, :l_v, -
       l_t:])
27     if 'TV' is in quads:
28         replace_with_rowwise_average(attention_scores[:, -l_t:, :l_v], mask[:, -l_t:, :
       l_v])
```
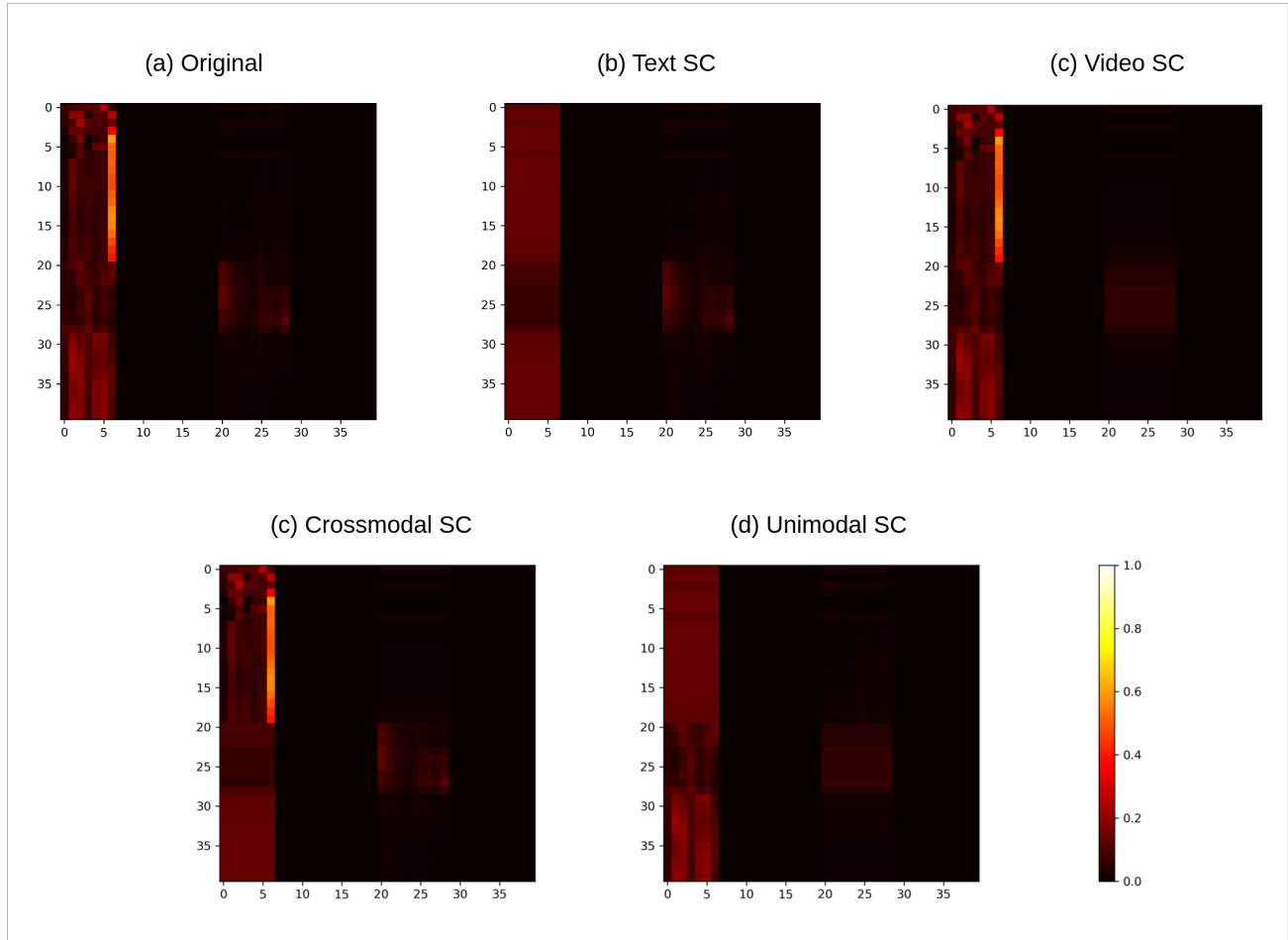
Figure 5. Visualization of the first attention head, as a heatmap, from the second layer of JustAsk model with $l_\mathcal{V} = 20$ and $l_\mathcal{T} = 20$. Note that here the text embeddings are pre-concatenated to the video embedding in the input. The lengths of the video and text tokens are 9 and 7 respectively. The text and video tokens are individually padded to length 20 each. We visualize *(a)* the original attention values and *(b)-(d)* after short-circuiting (SC) operations.

```
29      if 'TT' is in quads:
30          replace_with_rowwise_average(attention_scores[:, -l_t:, -l_t:], mask[:, -l_t:,
        -l_t:])
31
32  def replace_with_rowwise_average(scores, mask):
33      rowwise_sum = sum(scores, axis=-1)
34      rowwise_mean = rowwise_sum / sum(mask, axis=-2)
35      expanded_rowwise_mean = expand_dims(rowwise_mean, axis=-1)
36      replace_elements(scores, expanded_rowwise_mean)
37
38  def apply_attention_scores(attention_scores, value):
39      attended_output = dot_product(attention_scores, value)
40      return attended_output
```

Next, we provide the code for QUAG-attention. QUAG-attention modifies the existing self-attention block in the fusion module by replacing the block with the block average. We also demonstrate the normalizing the softmax function so that the each single average sequence is representative of the constituent sequences.

```
1  def quag_attention(inputs, mask, dim_model, l_v, l_t, type):
2      # Inputs:
```

```
3       #    inputs: Tensor of shape (batch_size, sequence_length, dim_model)
4       #    mask: Tensor of shape (batch_size, sequence_length)
5       #    dim_model: Dimension of the model (e.g., 512)
6       #    l_v: int    maximum length of video tokens
7       #    l_t: int    maximum length of question tokens
8       #    type: one of 'text', 'video', 'text-video'
9       query = linear_transform_query(inputs)
10      avg_input = compute_avg_input(inputs, l_v, l_t, type)
11      key = linear_transform_key(avg_input)
12      value = linear_transform_value(avg_input)
13      mask = apply_mask(mask, l_v, l_t, type)
14      scaled_dot_product = compute_scaled_dot_product(query, key, dim_model, mask)
15      attention_scores = softmax(scaled_dot_product)
16      attended_output = apply_attention_scores(attention_scores, value)
17      return attended_output
18
19  def compute_avg_input(inputs, l_v, l_t, type):
20      if type == "video":
21          avg_upper_block = sum(inputs[:, :l_v, :], axis=-2)
22          avg_upper_block = expand_dims(avg_upper_block, axis=1)
23          avg_input = concatenate((avg_upper_block, inputs[:, :-l_t, :]), axis=1)
24      elif type == "text":
25          avg_lower_block = sum(inputs[:, :-l_t, :], axis=-2)
26          avg_lower_block = expand_dims(avg_lower_block, axis=1)
27          avg_input = concatenate((inputs[:, :l_v, :], avg_lower_block), axis=1)
28      elif type == "text-video":
29          avg_upper_block = sum(inputs[:, :l_v, :], axis=-2)
30          avg_upper_block = expand_dims(avg_upper_block, axis=1)
31          avg_lower_block = sum(inputs[:, :-l_t, :], axis=-2)
32          avg_lower_block = expand_dims(avg_lower_block, axis=1)
33          avg_input = concatenate((avg_upper_block, avg_lower_block), axis=1)
34      return avg_input
35
36  def apply_mask(mask, l_v, l_t, type):
37      mask = expand_dims(mask, axis=-1)
38      mask = tile(mask, [1, 1, sequence_length])
39
40      if "video" in type:
41          video_length = sum(mask[:, :l_v, 0], axis=1)
42          video_length = expand_dims(video_length, axis=-1)
43          scaled_dot_product[:, :, 0] = scaled_dot_product[:, :, 0] * log(video_length)
44          upper_mask = ones(mask.shape[0], mask.shape[1], 1)
45          mask = concatenate((upper_mask, mask[:, :, l_v:]), axis=-1)
46
47      if "text" in type:
48          text_length = sum(mask[:, :-l_t, 0], axis=1)
49          text_length = expand_dims(text_length, axis=-1)
50          scaled_dot_product[:, :, -1] = scaled_dot_product[:, :, -1] * log(text_length)
51          lower_mask = ones(mask.shape[0], mask.shape[1], 1)
52          mask = concatenate((mask[:, :, :-l_t], lower_mask), axis=-1)
53
54      return mask
55
56  def compute_scaled_dot_product(query, key, dim_model, mask):
57      scaled_dot_product = dot_product(query, key) / sqrt(dim_model)
58      return scaled_dot_product
59
60  def apply_attention_scores(attention_scores, value):
61      attended_output = dot_product(attention_scores, value)
62      return attended_output
```

Table 4. 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
QUAG-attention	28.1	18.5	4.9	64.1	33.6	25.2	34.7	48.0	36.6

A.5. Training and Data Details for Simulation Study

Using the dataset generation strategy as described in the main paper, we sample 40,000 data points (24,000: training, 8,000: validation, 8,000: testing). We train a 4-layer transformer model, with learnable modality encoding and sinusoidal position encoding, having a dimensionality of 100. Following conventions, we set the hidden dimension to four times the hidden dimension (400). To prevent overfitting, we add dropouts in the embedding, attention, and penultimate layers. We Adam optimizer with a learning rate of 0.001 for 2000 epochs. The training batch size was 1024.

A.6. Experiment Details for Real World Data

All our experiments were performed on 4 NVIDIA A5000 GPUs. We use the official open-source code of the models on GitHub and modify only the self-attention modules. We use the official evaluation code and checkpoints. For NeXT-QA, we use the official dataset and finetune the models with the default parameters

As mentioned in the main manuscript, 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 (Xiao et al., 2022). 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.

A.7. Finegrained Accuracies

A.8. JustAsk Model

We present the fine-grained performance of JustAsk on the discussed datasets in Tables 4, 5, 6, and 7

Table 5. 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
QUAG-attention	35.1	83.5	51.1	38.6	50.2	82.1

Table 6. 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
QUAG-attention	50.8	52.0	65.1

Table 7. 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
QUAG-attention	44.2	43.9

A.9. FrozenBiLM Model

We present the fine-grained performance of FrozenBiLM on the discussed datasets in Tables 8, 9, 10, and 11

A.10. Additional Results

We evaluated QUAG on All-in-one model and find that, as the authors claim, the model utilizes both – unimodal and cross-modal modality interactions. The results are summarized in Table 12.

B. CLAVI

B.1. Dataset Creation

We curate CLAVI by leveraging Charades-STA (`https://prior.allenai.org/projects/data/charades/license.txt`) (Gao et al., 2017), 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

Table 8. 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
QUAG-attention	28.9	22.3	6.0	74.4	35.0	27.3	37.3	54.1	41.1

Table 9. 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
QUAG-attention	39.9	86.2	58.1	42.7	55.2	81.1

Table 10. 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
QUAG-attention	55.9	55.8	54.1

Table 11. 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
QUAG-attention	55.3	56.7

Table 12. Short-circuit (SC) results for All-in-one+ model on ActivityNet-QA (A-QA), and MSRVTT-QA (M-QA) datasets.

	All-in-one+	
	ActivityNet-QA	MSRVTT-QA
Acc	41.9	43.1
text_only	23.5	20.8
vid_only	14.2	4.2
SC: unimodal	11.4	3.8
SC: crossmodal	20.6	27.6
SC: video	19.2	12.7
SC: text	5.6	7.3

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 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 (Hao et al., 2022; Otani et al., 2020), the lengths of the extension boundaries are selected randomly. However, since inter-action separation can affect their recognition (Bagad et al., 2023), we constrain 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 complementary video, we swap the boundary-extended action regions as shown in Figure 3. 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 complements (unlike manipulations like video reversal (Wang et al., 2023b)). 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.

B.2. Comparison with Existing Datasets

We provide a comparison of size of CLAVI with established VideoQA datasets in Table 13.

B.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**: *holding clothes* **B**: *taking food* **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 complement (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").

Table 13. Comparison of CLAVI with other other VideoQA datasets sorted in the reverse order of recency.

Dataset	Number of (V,Q,A) samples
MSRVTT-QA (Xu et al., 2017)	243K
ActivityNet-QA (Yu et al., 2019)	58K
Social-IQ QA (Zadeh et al., 2019)	7.5K
NeXT-QA (Xiao et al., 2021)	52K
iVQA (Yang et al., 2021)	10K
STAR (Wu et al., 2021)	60K
EgoTaskQA (Jia et al., 2022)	40K
FIBER (Castro et al., 2022)	28K
NewsQA (Jahagirdar et al., 2023)	8.6K
CLAVI (Ours)	**114K**

E-Type:

Q : Was someone holding clothes?

Q : Was someone taking food?

E-Type (negative control):

Q : Was someone washing mirror?

BE-Type

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

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

Q : Was the person taking food at the **end**?

Q': Was the person taking food at the **beginning**?

BA-Type

Q : Did holding clothes happen **before** taking food?

Q': Did holding clothes happen **after** taking food?

Q : Did taking food happen **after** holding clothes?

Q': Did taking food happen **before** holding clothes?

BA-Type (negative-control)

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?

Q': Did washing mirror happen **before** taking food?

Q': Did washing mirror happen **after** taking food?

Q': Did taking food happen **before** washing mirror?

Q': Did taking food 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 holding clothes?

> **Q :** Was someone taking food?

E-Type (negative control):

> **Q :** Was someone washing mirror?

BE-Type

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

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

> **Q :** Was the person taking food at the **end**?

> **Q':** Was the person taking food at the **beginning**?

BA-Type

> **Q :** Did holding clothes happen **before** taking food?

> **Q':** Did holding clothes happen **after** taking food?

> **Q :** Did taking food happen **after** holding clothes?

> **Q':** Did taking food happen **before** holding clothes?

BA-Type (negative-control)

> **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?

> **Q':** Did washing mirror happen **before** taking food?

> **Q':** Did washing mirror happen **after** taking food?

> **Q':** Did taking food happen **before** washing mirror?

> **Q':** Did taking food happen **after** washing mirror?

B.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 **7.7 ± 3.42** sec. The average video length is **19.95 ± 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. 6.

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 7.

B.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 14.
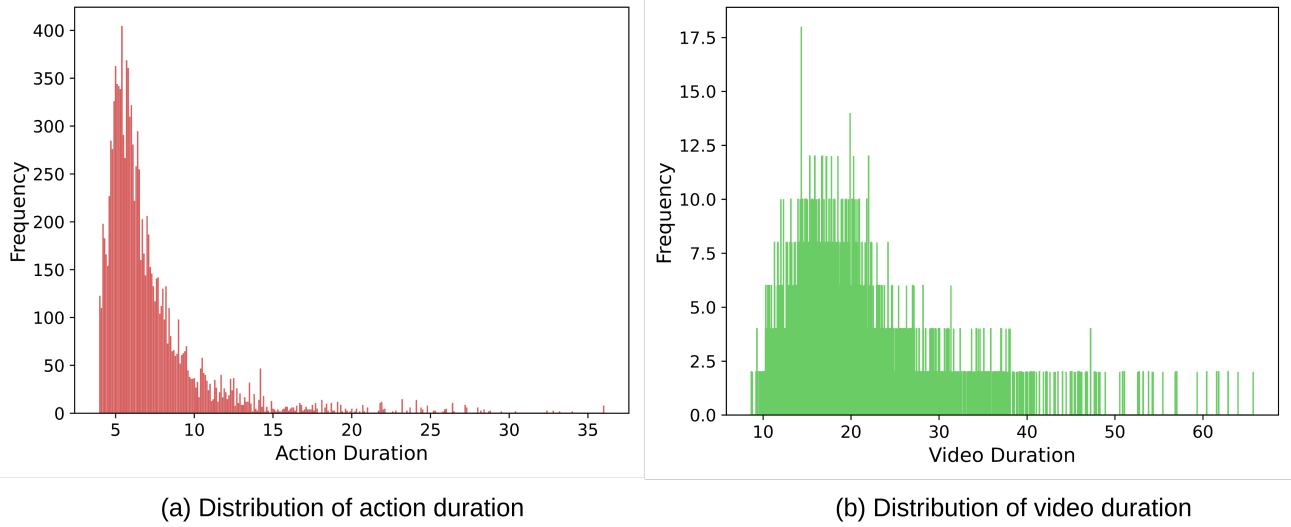
(a) Distribution of action duration

(b) Distribution of video duration

Figure 6. Distribution of length of (a) action and (b) video durations

Table 14. Hyperparameters and checkpoint details of CLAVI finetuning experiment

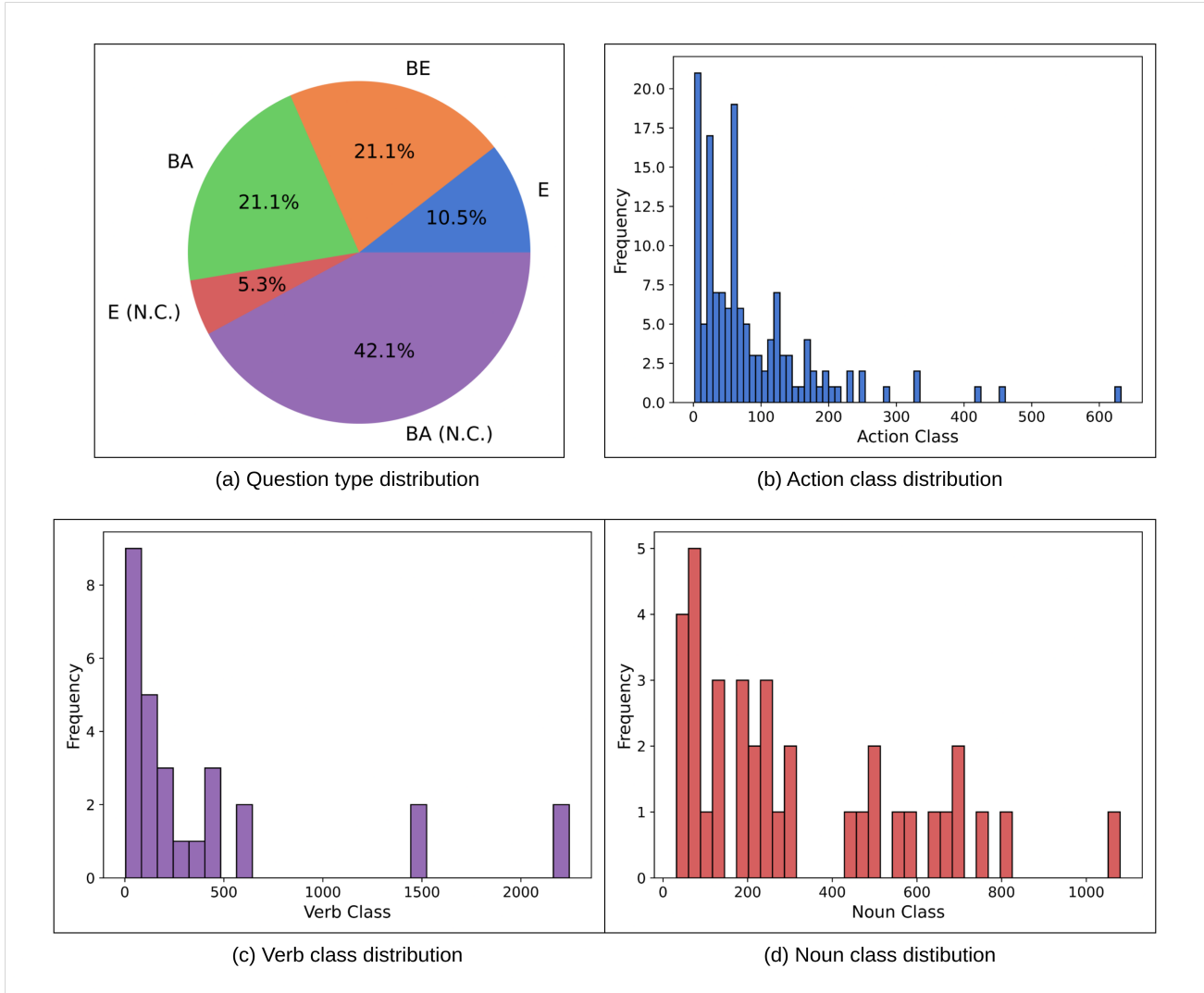| Model	Checkpoint	Epochs	LR
JustAsk	HowToVQA69M, WebVidVQA3M	20	1.00E-05
FrozenBiLM	WebVid10M	20	5.00E-05
All-In-One+	Webvid, YT-Temporal-180M, HowTo100M, CC3M, CC12M, COCO, Visual Genome, SBU Captions	10	1.00E-04
Singularity-T	COCO, Visual Genome, SBU Captions, CC3M, CC12M, WebVid	20	1.00E-05

Figure 7. 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.

Table 15. Fine-grained performance (% of accuracy) on CLAVI for question (Q) and complement question (Q'), video (V) and complement 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

B.6. Fine-grained Accuracies

In Table 15 we provide error bars for the finetuning experiments. The experiments were performed thrice on the same hardware with the same set of hyperparameters.

B.7. Representation Sensitivity Analysis

CLAVI can be used for diverse analyses to understand and interpret the joint multimodal representations in VideoQA models. We present one such analysis here. We want to find out the difference in representations between correctly and wrongly-answered complement pairs. Ideally, the complement pairs should have distinctly dissimilar representations to be answered correctly.

We use L2 norm as the distance metric. For CLAVI, we construct complements by augmenting the sequence of the frames (video complements) or replacing before/after and beginning/end (text complement). Hence, we cannot directly compute the distance between the attention matrices of the complements because they contain different tokens (text complement) or different order of the same tokens (video complement). We solve this by finding token correspondence between the complement pairs for each layer and head. By treating each attention matrix as a graph, we model the matrix alignment problem to finding the node correspondence between two isomorphic weighted directed complete graphs. Node correspondence between two graphs can be viewed as an instance of a linear sum assignment problem. That is, we want to learn a permutation transformation so that the two attention matrices as similar. We define similarity as negative of L2 distance. We solve this using modified Jonker-Volgenant algorithm as described by (Crouse, 2016).

We plot the histogram of L2 distance (averaged over heads and layers) for BA-type video and question complements in Figure 8. As expected, we find that if the answer is correct, then the average L2 distance is generally higher (skewed towards right). The mean and variance values L2 mean distribution of the correctly and incorrectly answered complement pairs is summarized in Table 16. We find that the correct predictions have higher mean and lower variance than the incorrectly inferred complement pairs. These findings validate that the model in indeed learning joint multimodal representations rather than creating its illusion.
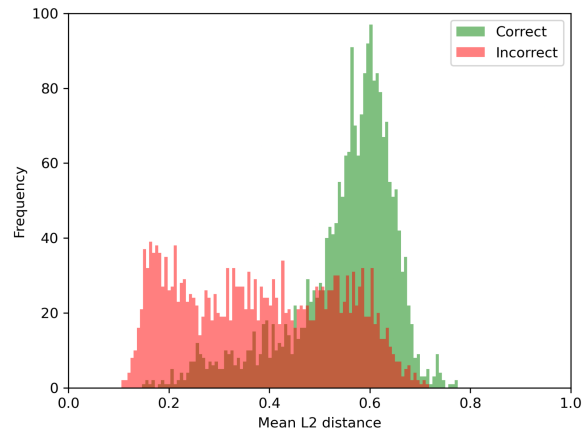
C. Limitations and Future Work

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

Table 16. Statistics of L2 distance values between aligned attention matrices of BA-type CLAVI questions, averaged over all heads and layers for FrozenBiLM. We report the statistics separately for correctly and incorrectly answered consistent complement predictions.

Type	Consistent Prediction	Mean	Variance
Video complement	Correct	0.70	0.02
	Incorrect	0.50	0.03
Text complement	Correct	0.55	0.01
	Incorrect	0.38	0.02
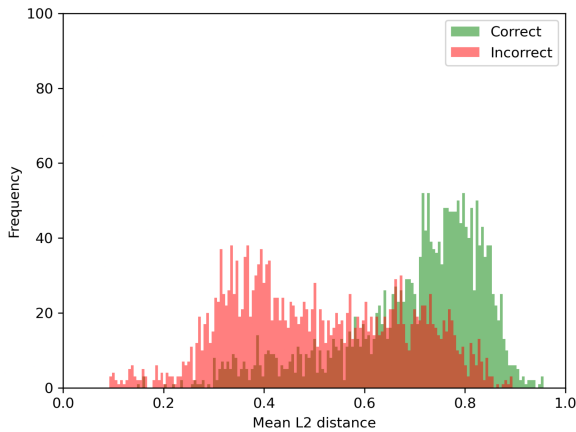
(a) Distribution of L2 distance for video complements (b) Distribution of L2 distance for text complements

Figure 8. Histogram plots of mean l2 distance between complement BA-type pairs for (a) video and (b) text complements for FrozenBiLM predictions (note that green is consistently correct; that is both the pairs in the complements are correctly answered, Similarly, red is consistently incorrect; that is at least one of the instance from the complement pair is incorrectly answered).

natural questions that include temporal referring expressions with similar balanced doubly-negative strategy.

C.1. Datasheet

In this section we provide a more detailed documentation of the dataset with the intended uses. We base ourselves on the datasheet proposed by Gebru et al. (2021)

C.1.1. MOTIVATION

- **For what purpose was the dataset created?** CLAVI is curated to diagnose and stress-test the joint multimodal understanding in VideoQA models. It uses temporal complements in video and question domains to assess if the models are able to learn the temporal structure from both the modalities.

- **Who created the dataset and on behalf of which entity?** CLAVI was curated by a team of researchers from the Agency For Science, Technology and Research (A*STAR), Singapore, comprising of Ishaan Singh Rawal, Alexander Matyasko, Shantanu Jaiswal, Basura Fernando and Cheston Tan.

- **Who funded the creation of the dataset?** Agency For Science Technology and Research (A*STAR), Singapore

C.1.2. COMPOSITION

- **What do the instances that comprise the dataset represent?** Each instance in CLAVI comprises of a video_id (from Charades-STA (freely available online)), question, question type, and answer ("yes" or "no"). Due to licensing issues, we do not release the videos but have provided the scripts for curating the dataset from Charades-STA.

- **How many instances are there in total?** CLAVI consists of 6,018 videos composing of 3,830 training and 2,188 testing videos. Each video is associated with 19 question-answer pairs, hence 114,342 data-points (72,770 training and 41,572 testing).

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** The videos and question answer pairs in CLAVI are generated by manipulating real-world real-world videos from Charades-STA. In theory, we can generate more instances with more videos with temporal annotations. We will release the code to generate the true and counterfactual video and question instances.

- **What data does each instance consist of?** Each instance in CLAVI comprises of a video_id (from Charades-STA), question, question type, and answer ("yes" or "no").

- **Is there a label or target associated with each instance?** Yes, each question is associated with a "type" label depending on the type of the question (types described in the main manuscript).

- **Is any information missing from individual instances?** No, all the instances have complete information the corresponding attributes.

- **Are relationships between individual instances made explicit?** Yes, the *video_name* attribute if of the form XXXXXXXX_1 for the original video segment and XXXXXXXX_2 for the counterfactual video segment, where XXXXXXXX is a unique 8-digit video id. The relationship between counterfactual questions is tabulated in the README file of the dataset.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** We provide the split files which are curated from the original split files of Charades.

- **Are there any errors, sources of noise, or redundancies in the dataset?** No. The owners of Charades do not report any known errors. And since our data is generated by machine, we do not expect any errors. For unforeseen errors in temporal annotation boundaries in the original dataset, we eliminate it by selecting the segments where the actions of interest do not occur in the immediate neighbourhood (detailed in the main manuscript).

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?** This dataset provides video IDs from the Charades dataset under their Non-Commercial license.

- **Does the dataset contain data that might be considered confidential?** No. We curate our dataset from publicly and non-commercially available Charades dataset.

- **Does the dataset identify any sub-populations (e.g., by age, gender)?** No. While it is possible to identify gender from Charades temporal captions, we do not use it in the curation of CLAVI. We only use neutral pronoun *someone*.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** No. The owners of Charades dataset have anonymized the subject information.

- **Does the dataset contain data that might be considered sensitive in any way?** No. The owners of Charades dataset ensure this and we curate CLAVI from Charades.

C.1.3. COLLECTION PROCESS

- **How was the data associated with each instance acquired?** Each sample of CLAVI associates with a question, answer (yes/no) and video-id from Charades dataset.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** We design a template-based VideoQA generation process to generate data each instance from Charades.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** Not applicable

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** Yes. We have outlined the process of filtering the data in detail in the appendix.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** Not applicable.

- **Over what timeframe was the data collected?** Our dataset is generated from Charades. We generate the dataset from February 2023 to June 2023.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** Not Applicable.

- **Does the dataset relate to people?** Yes. The Charades dataset contains videos of humans performing actions and we use it to curate CLAVI under their non-commercial license. However, we do not use the information pertaining to humans anyway.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** No. Our video data is from Charades under their Non-Commercial license. Charades Homepage: https://prior.allenai.org/projects/charades.

- **Were the individuals in question notified about the data collection?** Not applicable. We curate our dataset from Charades and the original owners have ensured this.

- **Did the individuals in question consent to the collection and use of their data?** Not applicable. We curate our dataset from Charades and the original owners have ensured this.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** Not applicable.

C.1.4. PREPROCESSING/CLEANING/LABELING

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)** No.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** No. The raw data (Charades) is distributed under Non-Commerical license.

- **Is the software that was used to preprocess/clean/label the data available?** We automatically curated the dataset using Python3 and shell-scripting, both of which are acessible and widely used.

C.1.5. USES

- **Has the dataset been used for any tasks already?** Video Question Answering.

- **Is there a repository that links to any or all papers or systems that use the dataset?** No.

- **What (other) tasks could the dataset be used for?** Video-Text and Text-Video retrieval.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** No.

- **Are there tasks for which the dataset should not be used?** No, we do not foresee any such usage as of now.

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** CLAVI is an academic dataset for public non-commercial use.

- **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** The scripts to curate the dataset will be released on GitHub.

- **When will the dataset be distributed?** Latest by the official paper acceptance.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** Yes. The dataset will be released under GPL3.0 license and terms of usage will be outlined on the dataset hosting website along with the license and the required scripts.

  Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

  Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

C.1.6. MAINTENANCE

- **Who will be supporting/hosting/maintaining the dataset?** Ishaan Singh Rawal and Cheston Tan

  How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The owners can be contacted via email: {rawal_ishaan_singh, cheston_tan}@cfar.a-star.edu.sg

- **Is there an erratum?** Not yet. Errata, if any, will be communicated through Github.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Updates, if any, will be clearly mentioned on GitHub.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** No.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** There are no older versions of the datasets at the current moment. However, we plan to appropriately version the curation scripts to ensure reproducability.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Yes, we will provide the necessary code files with the dataset. We will host the scripts to curate the datasets on Github, welcoming open-source contributions.