# A VIDEO IS NOT WORTH A THOUSAND WORDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As we become increasingly dependent on vision language models (VLMs) to answer questions about the world around us, there is a significant amount of research devoted to increasing both the difficulty of video question answering (VQA) datasets, and the context lengths of the models that they evaluate. The reliance on large language models as backbones has lead to concerns about potential text dominance, and the exploration of interactions between modalities is underdeveloped. How do we measure whether we're heading in the right direction, with the complexity that multi-modal models introduce? We propose a joint method of computing both feature attributions and modality scores based on Shapley values, where both the features and modalities are arbitrarily definable. Using these metrics, we compare 6 VLM models of varying context lengths on 4 representative datasets, focusing on multiple-choice VQA. In particular, we consider video frames and whole textual elements as equal features in the hierarchy, and the multiple-choice VQA task as an interaction between three modalities: video, question and answer. Our results demonstrate a dependence on text and show that the multiple-choice VQA task devolves into a model's ability to ignore distractors.

## 1 INTRODUCTION

Since the advent of pre-trained large language models (LLMs) with strong reasoning capabilities, vision language models (VLMs) have rapidly become a catchall system for users desiring to interact with multi-modal models, primarily because they can be queried in a similar manner to humans. VLMs are frequently modelled as a paired vision encoder and a pre-trained LLM, where visual and text tokens are projected into the same input space. With larger and more powerful models, VLMs are now created for visual understanding, either specialising in video or additionally allowing multi-image input. These models place the onus to reason well on the LLM, assuming that so long as the features are well-aligned, the model will understand the relationship between the modalities. We explore this hypothesis by determining the degree that each modality feature contributes to a model's response, and hence, whether modality preferences present themselves in current approaches to video question answering (VQA)—a common benchmark task for video language models.

Our effort stems from an intuition that video is not being reliably integrated into VLMs, motivated by a variety of related work. Recently, Deng et al. (2025) showcased a blind reliance on text when modalities disagree with eachother. Using Shapley values (Shapley, 1997), the modality preference of image/text tasks was investigated by Parcalabescu & Frank (2023; 2025), showing that VLMs are less self-consistent than LLMs. The separate, but related, problem of video redundancy has been investigated by works including Buch et al. (2022), which aimed to determine "what can be understood from a single image", by training a probe to select single frames capable of answering questions intended to be temporally complex. Similarly, Price & Damen (2020) showed that uni-modal video model performance for action recognition could be improved by removing "distractor" frames, calculated via Shapley values.

Where previous works have used token level features for calculating attributions and accuracy-based heuristics for determining modality preference (Parcalabescu & Frank, 2023; Goldshmidt & Horovicz, 2024), we instead propose an attribution method for arbitrarily grouped features that can be used to calculate modality scores. We consider whole frames and text elements (words, numbers, etc.) as individual features, as these are the smallest unit of meaningful information from a human perspective and this incorporates any model specific encoding/tokenisation while reducing the complexity of

the input. Conducting this analysis across a range of VQA datasets, we jointly investigate current capabilities of open source multi-modal models to properly integrate video and text.

In summary, our contributions are as follows: (i) we expand upon feature attribution methods built on Shapley values, being the first to bring it to the video/text domain; (ii) propose a method of calulating modality scores independent of model accuracy and modality length; (iii) benchmark 6 VLM models on 4 datasets; (iv) determine that video as an entire modality is being under-utilised by models—frames are consistently undervalued compared to text and (v) demonstrate that video contribution (and dataset difficulty) can be increased simply by adding more multiple-choice answers.

## 2 Related Work

**Interpretability** Early interpretability approaches focused on uni-modal models, e.g., vision: (Fan et al., 2022; Koh et al., 2020; Wong et al., 2021; Smilkov et al., 2017), audio (Parekh et al., 2024; Mishra et al., 2017; Becker et al., 2018), and language: (Ribeiro et al., 2016a; Bau et al., 2017; Chen et al., 2024a; Shi et al., 2023; Xie et al., 2024b). Works focused on gradient based methods (Sundararajan et al., 2017; Binder et al., 2016; Selvaraju et al., 2017) and attention-based methods concurrent with the rise of transformers (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Others employed feature attribution methods (Goldshmidt & Horovicz, 2024; Price & Damen, 2020) using Shapley values (Shapley, 1997) from game theory, which calculates the contribution of players within a co-op game. Shapley values have been employed as a black box approach, requiring only perturbations to the input and examining the output to interpret the model. Specifically, SHapley Additive exPlanation (SHAP) (Lundberg & Lee, 2017), unified Shapley values such that given a prediction, each input feature is assigned an importance value.

With the rise in popularity of vision language models, recent work has also focused on interpretability across multi-modal models (Stan et al., 2024; Chen et al., 2024d; Aflalo et al., 2022; Chefer et al., 2021; Liu et al., 2024; Ramesh & Koh, 2022; Swamy et al., 2023). DIME (Lyu et al., 2022)—itself an extension of LIME (Ribeiro et al., 2016b)—interprets models via the pertubation of input features to affect the model outputs. It first disentangles the model into uni-modal and multi-modal contributions before generating intepretable visualisations via applying LIME. Similar to our work, MM-SHAP (Parcalabescu & Frank, 2023) extends SHAP for multiple modalities, in this case for image and text, to measure contributions of image patches and words in text to the final decision made by the model. PixelSHAP (Goldshmidt, 2025), was proposed to instead model the contributions of feature groups of pixels related to the same concept within the input image, thus allowing for a finer granularity than attributions of image patches. Our work lies within the feature attribution paradigm, extending SHAP for the task of VQA enabling both frame attribution and modality dependence for video, question, and answer inputs.

**Modality Preference** Concurrent with interpretability approaches, Huang et al. (2022) theorise that "During joint training, multiple modalities will compete with each other [and only a subset of modalities will be utilised] ... with other modalities failing to be explored." Many works have reached similar conclusions by exploring modality dominance. These works have focused on: pertubation of input modalities (Park et al., 2025; Wu et al., 2022)/input features (Parcalabescu & Frank, 2023; Frank et al., 2021) to show that one modality is dominant; finding similar images which showcase the gap between vision and vision-language (Tong et al., 2024); few shot evaluation of language models (Chen et al., 2024b); or hard negatives/foiling examples (Parcalabescu et al., 2022; Gat et al., 2021; Deng et al., 2025). Previous work has overwhelmingly shown that models frequently exhibit a preference to one or more modalities, which is commonly language within a vision language model. Others have proposed solutions to the modality bias, typically by introducing datasets that require models to understand and utilise information from both modalities (Goyal et al., 2017; Parcalabescu et al., 2022; Chen et al., 2024c) or via training regimes that balance the utilisation of each modality (Yang et al., 2024c; Leng et al., 2024; Wang et al., 2024a; Xiao et al., 2025; Pi et al., 2024; Deng et al., 2024). Closest to our work, Park et al. (2025) propose the Modality Importance Score (MIS) to discover modality bias for VQA, showing that current models do not effectively utilise information across multiple modalities. They conclude that the datasets they evaluated for VQA do not demand multi-modal reasoning, with 90–95% of questions requiring only a single modality or are modality agnostic. Our metrics differ from MIS in two ways: Firstly, our metrics can provide feature level attribution *in addition* to modality preference. Secondly, our metrics can provide the degree at which a feature is positive/negative/neutral rather than as ternary outcomes.

# 3 METHOD

## 3.1 SHAPLEY VALUES

Typically, deep learning models take a large set of input features as input and distill them into a one-dimensional confidence (or logit) in any given output class (or token), making it exceptionally difficult to identify the degree of impact that each feature has. Shapley values were introduced to the field of game theory in Shapley (1997), as a solution concept for the fair distribution of contributions to a cooperative game. More formally, given a set of $N = \{1, 2, \ldots, n\}$ players, the function $r : 2^N \to \mathbb{R}$ maps from all possible subsets of $N$ (the power set) to a value (or reward). If a coalition of players is represented as $S$, then $r(S)$ represents the total contribution of that coalition.

**Definition 3.1 (The Shapley value)** *The Shapley value of player $i$, $\varphi_i(r)$ can be defined as:*

$$\varphi_i(r) = \frac{1}{n!} \sum_{P \in S_n} (r(P_i \cup i) - r(P_i))$$

*where $S_n$ is the set of all possible permutations (or orderings) of $N$ and $P_i$ is the set of players that precede player $i$ in a given permutation.*

Informally, this formula represents taking every possible order that the $N$ features can be added to the coalition and averaging the marginal contributions of the $i$th feature.

Shapley values are the only payment rule to satisfy the following four properties of: **Efficiency**: the sum of contributions across all players is equal to the model output; **Symmetry**: if two players contribute equally then they will be assigned an equal score; **Linearity**: the Shapley value of the sum of two (or more) reward functions is the sum of their individual Shapley values; and **Null Player**: if a player does not contribute then it will be assigned a score of 0.

It should be clear that Shapley values are a convincing candidate as a method to compute local feature attributions of a model, by which we mean: the impact that the features of a single input instance will have on a model's prediction. We now consult the work of Lundberg & Lee (2017) to provide a unified framework for using Shapley values as an attribution method for any arbitrary machine learning model. Assuming we have some prediction model $f$, we define a new explanation model $g$, which takes a simplified input $x'$. This simplified input is mapped to the real input $x$ by some mapping function $h_x(\cdot)$, depending on $x$.

**Definition 3.2 (Additive feature attribution)** *An additive feature attribution method has explanation model:*

$$g(x') = \phi_0 + \sum_{i=1}^{M} (\phi_i \times x'_i)$$

*where $x' \in \{0, 1\}^M$, $M$ is the number of simplified features and $\phi_i \in \mathbb{R}$ is the attribution of each simplified feature.*

Through this definition, we can arbitrarily group a model's features at any scale, e.g., images instead of pixels or words instead of text tokens. Zeros in the simplified feature vector $x'$ represent masking the entire grouped feature out, and ones represent keeping the full grouped feature in the input. Subsequently, the following three properties are desirable for an additive feature attribution method: **Local Accuracy**: when the original model is approximated, its output should match the original output; **Missingness**: if the feature is missing it won't be given any attribution; and **Consistency**: if a feature's contribution is constant/increases due to a model change, it's attibution won't decrease. According to Lundberg & Lee (2017) the attribution satisfying these properties, is the Shapley value.

## 3.2 MULTI-MODAL SHAPLEY VALUES

In our case, the "game" is multiple-choice VQA, and each "player" is a frame or word in the input. Calculating exact Shapley values is computationally infeasible for large numbers of features, as the number of possible coalitions scales in a factorial manner, therefore it is typically approximated. In this paper, we use a Monte Carlo method to approximate Shapley values from the SHAP library (Lundberg, 2018).

We'll refer to the VLM model as $f$ (whose logits will be used as reward $r$), taking multi-modal video and text input $(v, t)$, which generates an arbitrary number of text tokens. The video modality $v$ is a sequence of $n_v$ video frames (or video features) and the text modality $t$ is a sequence of $n_t$ textual elements. Textual elements include individual words, numbers or punctuation. We choose to represent atomic elements rather than sub-word tokens to match human perspective. We can transform the multi-modal features $(v, t)$ to simplified features $x' \in \{0, 1\}^{n_v+n_t}$ by defining a mapping function that retains the multi-modal feature $(v, t)_i$ if $x'_i = 1$ and masks it out if $x'_i = 0$. In the video modality, by masking we simply zero out all of the pixels/features of the masked image. Conversely, in the text modality, masking is represented by changing the textual element to be whitespace.

To update the Shapley values, we need to retrieve a valid reward from the model in relation to the input. We opt to use the logits of the predicted text tokens, as it does not require internal information from the model and can be returned along with these tokens. In multiple-choice VQA, each question has $n_c$ corresponding choices (or classes) for the model to pick from, where we can jointly calculate the Shapley values for all classes by simply retrieving the respective logits. We label the answers from A–E, skipping letters if the question does not provide 5 choices and query the model to pick a letter. In cases where the language model does not respect requests to respond with only a single letter token, we first check whether a letter from A–E appears in the output. If it does, we index this token, otherwise we take the first token in the output sequence.

### 3.3 METRICS

In this subsection we'll explain the quantitative metrics that we'll use to highlight differences between the Shapley values for the question modalities: video (V), question (Q), and answers (A)[1]. The reason for differentiating between questions and answers arises from an intuition that in multiple-choice, the semantic purpose of the questions is different enough to that of the answers. Given a model and a dataset of VQA-tuples—containing a video, a question, and the corresponding answers—these metrics are calculated over the Shapley values for all of these tuples. While the Shapley values are divided by class $c$, in practice we'd like to split them based on whether they are the ground truth or negative. Let $\varphi_i^{\text{gt}} \in \mathbb{R}$ refer to the Shapley value of multi-modal element $i$ for the ground truth class and $\varphi^{\text{gt}}$ be the list of these Shapley values. Furthermore, let $n_v$, $n_q$ and $n_a$ be the number of video, question and answer elements respectively. The scale of a model's logits can vary, so to normalise a VQA-tuple's Shapley values, we divide by the maximum absolute value to normalise the values in the interval of $[-1, 1]$ while, importantly, preserving 0 values: $\hat{\varphi}_i^c = \varphi_i^c / |\max_i \varphi_i^c|$.

The following metrics are defined per VQA-tuple and will be averaged over the dataset. For our first quantitative metric, we want to measure the magnitude of *contribution* of each VQA-tuple modality, which we call Modality Contribution (MC). We calculate this contribution by dividing the total magnitude per modality by the sum of the total magnitudes of each modality. Consequently, the modality specific contributions are the proportion of this total:

$$\text{MC}_\text{V} = \frac{\sum_{i=1}^{n_v} |\hat{\varphi}_i^{\text{gt}}|}{\sum |\hat{\varphi}_i^{\text{gt}}|}, \quad \text{MC}_\text{Q} = \frac{\sum_{i=n_v+1}^{n_v+n_q} |\hat{\varphi}_i^{\text{gt}}|}{\sum |\hat{\varphi}_i^{\text{gt}}|}, \quad \text{MC}_\text{A} = \frac{\sum_{i=n_v+n_q+1}^{n_v+n_q+n_a} |\hat{\varphi}_i^{\text{gt}}|}{\sum |\hat{\varphi}_i^{\text{gt}}|}$$

The second quantitative metric measures the *average contribution* of each feature within each modality, which we refer to as Per-Feature Contribution (PFC). When the number of features in a modality is significantly high, that modality can become over-represented, leading to a large sum of many smaller Shapley values. Taking the mean of the magnitudes per modality helps to avoid the imbalance caused by differing numbers of features. To simplify the formula, we define the halfway variable $M$ to represent the mean Shapley value of a modality:

$$M_\text{V} = \frac{\sum_{i=1}^{n_v} |\hat{\varphi}_i^{\text{gt}}|}{n_v}, \quad M_\text{Q} = \frac{\sum_{i=n_v+1}^{n_v+n_q} |\hat{\varphi}_i^{\text{gt}}|}{n_q}, \quad M_\text{A} = \frac{\sum_{i=n_v+n_q+1}^{n_v+n_q+n_a} |\hat{\varphi}_i^{\text{gt}}|}{n_a}$$

Then the Per-Feature Contributions are the proportion of these summed means:

$$\text{PFC}_\text{V} = \frac{M_\text{V}}{M_\text{V} + M_\text{Q} + M_\text{A}}, \quad \text{PFC}_\text{Q} = \frac{M_\text{Q}}{M_\text{V} + M_\text{Q} + M_\text{A}}, \quad \text{PFC}_\text{A} = \frac{M_\text{A}}{M_\text{V} + M_\text{Q} + M_\text{A}}$$

[1]We assume for this paper that we have 3 modalities in a specific order, but these metrics extend trivially to any number of modalities.

# 4 RESULTS

## 4.1 MODELS AND DATASETS

**Models** We choose 6 VLM models to examine using our metrics defined previously to cover several aspects: Firstly, to have a variety of context lengths. Secondly, to compare two-stream encoder approaches to VLMs which use LLM decoders. Thirdly, to investigate how models might have changed over time, evaluating older vs. newer approaches. According to these requirements, we select the following: **FrozenBiLM** (Yang et al., 2022): A compact, early example of a two-stream VLM model that uses pre-extracted CLIP (Radford et al., 2021) features. We use it with 10 frames. **InternVideo** (Wang et al., 2022): An early two-stream foundation video model. We use it, again, with 10 frames. **VideoLLaMA2** (Cheng et al., 2024): A short context VLM using Mistral (Jiang et al., 2023) as the LLM decoder designed for question answering/captioning and to improve spatio-temporal reasoning. We use the maximum of 16 frames. **LLaVA-Video** (Zhang et al., 2024): A medium context VLM using Qwen2 (Yang et al., 2024a) as the LLM decoder, trained with a curated dataset and carefully intruction tuned. We use 64 frames. **LongVA** (Zhang et al., 2025b): A long context open-source model, also using Qwen2 as the LLM decoder. We use 128 frames (the maximum that we could fit into GPU memory). **VideoLLaMA3** (Zhang et al., 2025a): An update to VideoLLaMA2 using Qwen2.5-7B (Yang et al., 2024b) as the LLM decoder, trained on a longer context with a focus on efficiency. We used the maximum of 180 frames.

**Datasets** We select 4 datasets to evaluate under the following requirements; Firstly, to cover both popular and new datasets. Secondly, showcasing first-person (egocentric) and third-person (exocentric) viewpoints and, finally, to include long and short video contexts. We took subsets (necessary due to the excessive time it would take to calculate Shapley values for entire datasets) from the following 4 VQA datasets: **EgoSchema** (Mangalam et al., 2023): Egocentric VQA dataset intended to be unanswerable without viewing all of the video. We take a subset of 50 questions from the set released with ground truths. **HD-EPIC** (Perrett et al., 2025): Egocentric VQA dataset collected within the kitchen domain, ranging from questions about single images to multi-video queries spanning several hours. We take a subset of 60 questions, with 2 from each of the 30 question types. **MVBench** (Li et al., 2024): General VQA dataset curated from several other datasets to create a multiple-choice benchmark across different tasks. We take a subset of 60 questions, 3 from each question type. **LVBench** (Wang et al., 2024b): General VQA dataset for very long video understanding, where many questions are asked about small collection of lengthy YouTube videos. We take a subset of 60 questions, 10 from each question type.

## 4.2 WHAT DOES THE CONTRIBUTION METRIC SHOW?

In table 1, we first show the Modality Contribution and Per-Feature Contribution for each model/-dataset combination along with the corresponding accuracy. The cells are highlighted to show high (**blue**) and low (**red**) contribution scores. Note, for all of these coloured tables, values of $1/3$ would represent balanced contributions across the three modalities.

**Under-representation of Video** For all methods but VideoLLaMA3, video is consistently under-represented in the Modality Contribution, indicating that the modality as a whole is consistently contributing less to the decision of the models. VideoLLaMA3 on the other hand, shows strong contributions from video, especially for LVBench, where the entire dataset comprises of $\sim 1$ hour long videos. Looking at the Per-Feature Contribution values, we see that video is still consistently underrepresented among the three modalities. However, for long context models, video shows vastly reduced contributions, meaning that per-frame the Shapley values are much smaller than their text feature counterparts. Video as a whole modality is clearly still highly relevant, but this is evidence that the Shapley values of its individual frames are more centered around zero, and that the model's attention to them is much less guided than for the text.

**Importance of Question vs. Answer** The question is particularly important for FrozenBiLM and InternVideo because the answers are independently queried as multiple binary questions for these models—they do not need to discriminate between answers. For the stronger models, the question is often undervalued compared to the answers, indicating that the model cares less about the specifics of the question, and more about discriminating between the possible answers. This follows recent design of VQA datasets to use hard negatives answers to ensure results are not text-biased, i.e., Perrett et al. (2025); Xie et al. (2024a); Chen et al. (2024c).

Table 1: MC and PFC for each modality in the VQA-tuple. Calculated based on the Shapley values for the ground truth logit averaged across all VQA-tuples. Here **blue** scores relate to large magnitudes of Shapley values, regardless of their sign, while **red** scores relate to values close to 0.

(a) EgoSchema

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.09 | 0.33 | 0.58 | 0.31 | 0.46 | 0.23 | 0.20 |
| IV | 0.20 | 0.36 | 0.44 | 0.51 | 0.36 | 0.13 | 0.36 |
| VL2 | 0.11 | 0.18 | 0.71 | 0.30 | 0.33 | 0.36 | 0.56 |
| L-V | 0.15 | 0.15 | 0.70 | 0.16 | 0.36 | 0.48 | 0.72 |
| LVA | 0.12 | 0.13 | 0.75 | 0.07 | 0.36 | 0.57 | 0.48 |
| VL3 | 0.30 | 0.14 | 0.56 | 0.14 | 0.40 | 0.46 | 0.70 |

(b) HD-EPIC

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.09 | 0.56 | 0.36 | 0.19 | 0.56 | 0.24 | 0.22 |
| IV | 0.34 | 0.51 | 0.15 | 0.55 | 0.39 | 0.07 | 0.15 |
| VL2 | 0.17 | 0.25 | 0.58 | 0.27 | 0.26 | 0.47 | 0.28 |
| L-V | 0.19 | 0.26 | 0.55 | 0.17 | 0.32 | 0.51 | 0.35 |
| LVA | 0.18 | 0.22 | 0.61 | 0.13 | 0.28 | 0.59 | 0.35 |
| VL3 | 0.36 | 0.21 | 0.43 | 0.17 | 0.34 | 0.48 | 0.35 |

(c) MVBench

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.13 | 0.49 | 0.38 | 0.16 | 0.47 | 0.37 | 0.40 |
| IV | 0.26 | 0.51 | 0.23 | 0.32 | 0.48 | 0.20 | 0.42 |
| VL2 | 0.19 | 0.27 | 0.54 | 0.17 | 0.26 | 0.57 | 0.62 |
| L-V | 0.23 | 0.26 | 0.52 | 0.06 | 0.30 | 0.64 | 0.65 |
| LVA | 0.14 | 0.24 | 0.63 | 0.02 | 0.27 | 0.71 | 0.45 |
| VL3 | 0.40 | 0.26 | 0.34 | 0.05 | 0.41 | 0.54 | 0.65 |

(d) LVBench

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.17 | 0.44 | 0.40 | 0.23 | 0.46 | 0.32 | 0.30 |
| IV | 0.34 | 0.44 | 0.22 | 0.42 | 0.43 | 0.15 | 0.25 |
| VL2 | 0.23 | 0.23 | 0.54 | 0.23 | 0.27 | 0.50 | 0.27 |
| L-V | 0.26 | 0.21 | 0.53 | 0.09 | 0.30 | 0.61 | 0.42 |
| LVA | 0.30 | 0.17 | 0.52 | 0.05 | 0.29 | 0.66 | 0.35 |
| VL3 | 0.47 | 0.16 | 0.37 | 0.08 | 0.34 | 0.58 | 0.48 |

**Dataset Comparisons** According to the Per-Feature Contribution, across the datasets, video is consistently more important for EgoSchema and HD-EPIC than it is for LVBench and MVBench. The video content of these two egocentric datasets is much more diverse (i.e. camera pose, occlusion and complexity of actions/sequences) than for exocentric datasets, evidenced by the relative increase in each frame's importance. Answer Per-Feature Contributions are particularly large for MVBench and LVBench because the answers are often shorter for these datasets, leading to a high density of relevant information, whereas EgoSchema's whole sentences are more akin to natural language and contain more realistic textual distractors, present in its high answer Modality Contributions.

### 4.3 How Does Masking Input Affect Accuracy?

In table 2, we compare the accuracy of vanilla input, no input and masked modalities. The datasets are generally not completely answerable blind, i.e. with the video modality being masked, which reassures the intuition that these benchmarks have been developed to not be trivial for a language-only model to solve. However, it's possible to get between 30% and 50% on EgoSchema and MVBench without video—well above the random performance of 20% and 29%—meaning that the combination of the question and answers is pushing the model towards the ground truth. Masking out the question has a surprisingly small effect on performance, but this follows from the low question Modality Contribution values presented earlier, as it is frequently undervalued. HD-EPIC is the hardest to answer without question, suggesting its complexity is important, as the question content of the other datasets is usually less specialised to a specific domain and much shorter. Overall, the under-reliance on question leads us to believe this is clear evidence of a basic limitation of multiple-choice VQA: with just the video and the answer the model appears to be able to discriminate between the answers. In particular, sometimes a model will gain performance when the question is completely removed, suggesting that the task is less about truthfully meeting the requirements of a question, and more about picking amongst a set of information-dense answers in a similar fashion to a video-text matching task. As one would expect, masking out the answer generally gives the largest drop in performance compared to the other modalities with many models being reduced to random performance.

### 4.4 Answer Replacement

To determine the extent to which these trends are biased by the number of answers typically used in multiple-choice questions, we experiment with injected negatives in fig. 1, by automatically creating

Table 2: *How does masking modalities affect performance?* We mask either all input features ("All"), or each modality separately and compare to baseline performance. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy compared to baseline.

| Model | Masking | EgoSchema | HD-EPIC | MVBench | LVBench |
|---|---|---|---|---|---|
| FrozenBiLM | None | 0.20 | 0.22 | 0.40 | 0.30 |
| | All | -0.02 | +0.00 | -0.23 | +0.03 |
| | Video | +0.00 | -0.03 | -0.05 | -0.02 |
| | Question | +0.02 | -0.02 | -0.03 | +0.00 |
| | Answer | -0.02 | +0.00 | -0.23 | +0.03 |
| InternVideo | None | 0.36 | 0.15 | 0.42 | 0.25 |
| | All | -0.18 | +0.07 | -0.25 | +0.08 |
| | Video | -0.14 | +0.07 | -0.10 | +0.05 |
| | Question | +0.06 | +0.02 | +0.02 | +0.03 |
| | Answer | -0.18 | +0.07 | -0.25 | +0.08 |
| VideoLLaMA2 | None | 0.56 | 0.28 | 0.62 | 0.27 |
| | All | -0.36 | +0.00 | -0.30 | -0.02 |
| | Video | -0.18 | +0.00 | -0.23 | +0.00 |
| | Question | +0.02 | -0.13 | -0.13 | +0.05 |
| | Answer | -0.34 | +0.03 | -0.22 | -0.03 |
| LLaVA-Video | None | 0.72 | 0.35 | 0.65 | 0.42 |
| | All | -0.54 | -0.17 | -0.48 | -0.08 |
| | Video | -0.26 | -0.12 | -0.23 | -0.07 |
| | Question | +0.04 | -0.05 | -0.07 | -0.12 |
| | Answer | -0.58 | -0.17 | -0.28 | -0.32 |
| LongVA | None | 0.48 | 0.35 | 0.45 | 0.35 |
| | All | -0.28 | -0.17 | -0.15 | -0.05 |
| | Video | -0.10 | -0.08 | -0.05 | -0.05 |
| | Question | -0.06 | -0.12 | -0.02 | +0.05 |
| | Answer | -0.30 | -0.17 | -0.12 | -0.17 |
| VideoLLaMA3 | None | 0.70 | 0.35 | 0.65 | 0.48 |
| | All | -0.52 | -0.23 | -0.48 | -0.15 |
| | Video | -0.26 | -0.05 | -0.33 | -0.20 |
| | Question | -0.06 | -0.07 | -0.08 | -0.08 |
| | Answer | -0.48 | -0.23 | -0.40 | -0.18 |

new annotations for VQA-tuples. We introduce two types of answer replacement: "Easy" where the negatives are simply rotated in from one randomly selected question, and "New-$x$" where $x$ new negatives are randomly added from across questions (increasing the number of options) and positions are shuffled. MVBench is excluded as it does not fix the number of options, and "New" negatives are only valid for HD-EPIC and LVBench if the question type is the same as the original negatives.

When testing the "Easy" negatives, the accuracy rises drastically as it becomes easy for the model to match context between the positive answer and the question. As the number of extra answers increases from 5 to 20 in the "New" case, the contribution of answer features decreases while the contribution of video and question features increases. To see whether this translates into model performance, we include modality masking tables in appendix E.8. For example, with VideoLLaMA3 (when adding 10 answers and masking video) that performance drops by $40\%$ and $15\%$ on EgoSchema and LVBench respectively, while when masking questions the drop is $6\%$ and $18\%$ respectively. This indicates that merely adding more multiple-choice answers can drastically alter the contribution of under-represented modalities. Finally, when evaluating model performance on these new VQA-tuples, we see that the accuracy consistently decreases until "New-15" where the drop is often less pronounced.

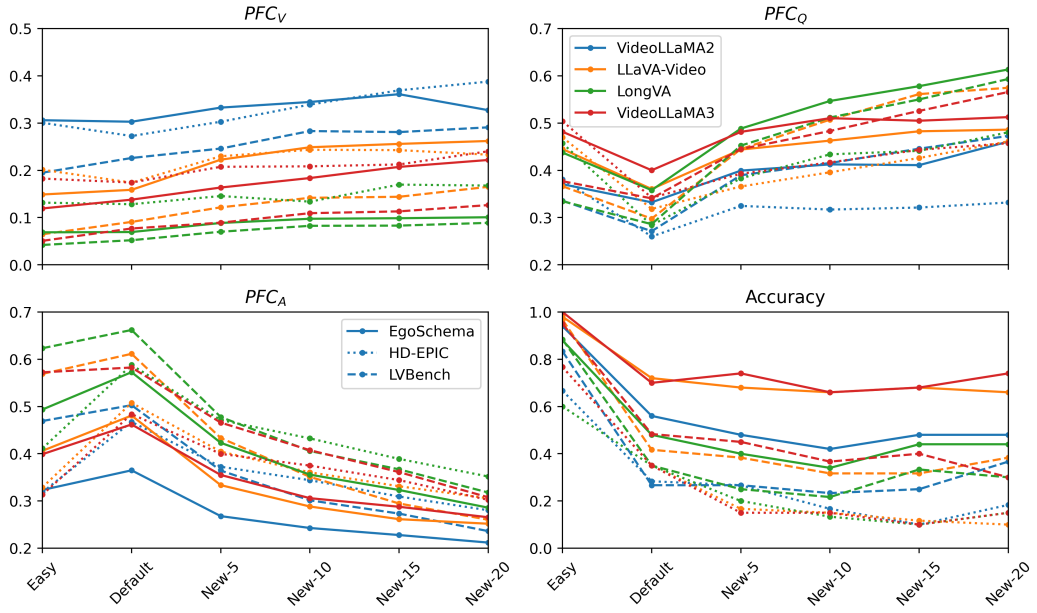
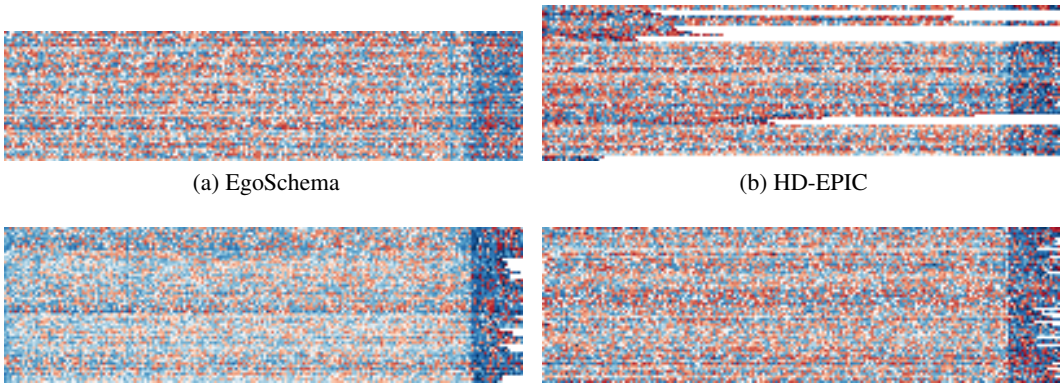



Figure 1: Per-Feature Contribution and accuracy as new negative answers are injected into the VQA-tuples, varying from easiest to hardest.



(a) EgoSchema (b) HD-EPIC

(c) MVBench (d) LVBench

Figure 2: Matrix of Shapley values per subset, where each row, left-to-right, represents the features of a VQA-tuple. Rows are truncated to a maximum of 200 features.

### 4.5 Qualitative Results

We visualise the overall distribution of attributions for each VQA-tuple (row) using VideoLLaMA3, in fig. 2. We truncate this figure for readability given the variable length of the questions and answers. The magnitude of the Shapley values are much larger towards the right hand side of each heatmap, representing the question and answer attributions. This stark boundary is where the video frames end and the text features begin, demonstrating that the video modality contribution is much less than the question/answer. Whilst there are many peaks in the questions and answers, most frames within the video tend to have a similar contribution. Reassuringly, the values for the video frame attributions do not show strong temporal bias, i.e., there are no similar values within each column, yet they are not unstructured, with many examples of several temporally consecutive frames having similar signs. However, MVBench video is typically more positive and slightly skewed to the first frame, showing that latter parts of the video do not contribute as much as in the other datasets.

Next, we provide an example VQA-tuple from EgoSchema evaluated using VideoLLaMA3 in fig. 3 (more examples in appendix F.2). Even though the displayed frames are the 16 most relevant, we

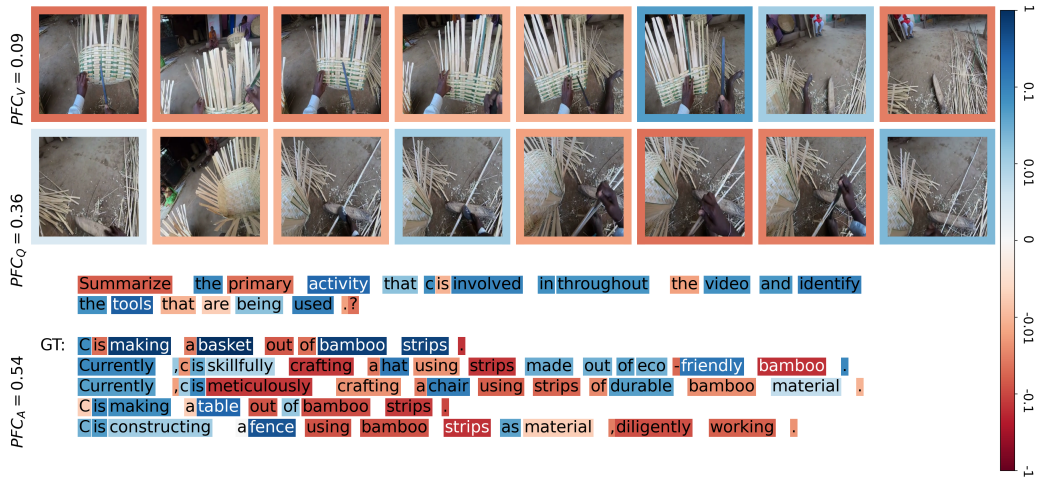



Figure 3: Qualitative figure of an example from EgoSchema evaluated using VideoLLaMA3. For brevity, we select the 16 most important frames, ranked by the magnitude of their Shapley values. Here **blue** represents positively attributed inputs whereas **red** represents negatively attributed inputs.

note that the frames show much smaller contributions, either **positively** or **negatively**, compared to the questions and the answers. This can be seen in the contribution scores, with $\text{PFC}_\text{V} = 0.09$, whereas $\text{PFC}_\text{Q} = 0.36$ and $\text{PFC}_\text{A} = 0.54$. We find no strong trend between +ve/-ve attribution score of each frame and its contents—frames depicting extremely similar content (e.g., the 5th and 6th or 12th and 14th frames) can vary greatly. Interestingly, the relevant nouns that vary in the false answers ("hat", "chair", "table" and "fence") all pull the model towards the ground truth, while "bamboo strips" (appearing in all answers) pushes away from the ground truth, providing evidence of discriminators in the text being employed to choose amongst multiple-choice answers.

## 5 Limitations and Future Work

Whilst within this work we aim for a thorough suite of benchmark datasets (of which we could only use subsets) and VLMs, we recognise this is not an exhaustive collection of both, which could be expanded for future work. As the number of coalitions to calculate true Shapley values scales factorially and we have limited computational resources, we approximate them for our results, using 5000 coalitions across all experiments (ablated in appendix C). As well as this, we focus entirely on multiple-choice VQA so that Shapley contributions are based on reward from the ground truth instead of the generated token, where it would be interesting if it is possible to study open-ended VQA in the same manner. Pushing this interpretability framework further by including more modalities like audio or 3D data will be a relevant focus for future research as multi-modal models grow.

## 6 Conclusion

In this paper, we have proposed a joint method of both feature attribution and modality scoring based on the Shapley values of VLMs for VQA. In general, our metrics indicate that VLMs under-represent video compared to the question or answers. By defining a modality metric normalised by the modality length, we find that there is significant divergence between individual frame and text contributions. Furthermore, we demonstrated that the VQA task is limited in properly evaluating multi-modal understanding, as strong accuracy can be achieved without even being presented with a question. We also use the framework to explore a scenario where trivially adding multiple-choice options *beyond* the typical $4/5$ improves dataset difficulty, and find corresponding increases in video feature contribution and dependence of the video modality as a consequence. We hope that our paradigm can be used to better understand and develop future multi-modal understanding in a flexible and interpretable manner, whether it be to examine inputs case-by-case, benchmark entire datasets or test whether models are reliable, unbiased by their impressive improvements in accuracy.

## REFERENCES


Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *CVPR*, pp. 21374–21383. IEEE, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pp. 2425–2433. IEEE Computer Society, 2015.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pp. 3319–3327. IEEE Computer Society, 2017.

Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *CoRR*, abs/1807.03418, 2018.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *ICANN (2)*, volume 9887 of *Lecture Notes in Computer Science*, pp. 63–71. Springer, 2016.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *CVPR*, pp. 2907–2917. IEEE, 2022.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pp. 387–396. IEEE, 2021.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *AAAI*, pp. 17754–17762. AAAI Press, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024b.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *EMNLP (Findings)*, pp. 16449–16469. Association for Computational Linguistics, 2024c.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *NeurIPS*, 2024d.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024.

Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Object-centric representation learning for video question answering. In *IJCNN*, pp. 1–8. IEEE, 2021.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *CVPR*, pp. 3867–3876. Computer Vision Foundation / IEEE, 2025.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y. Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. In *NeurIPS*, 2024.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pp. 1999–2007. Computer Vision Foundation / IEEE, 2019.

Quanfu Fan, Donghyun Kim, Chun-Fu Chen, Stan Sclaroff, Kate Saenko, and Sarah Adel Bargal. Temporal relevance analysis for video action models. *CoRR*, abs/2204.11929, 2022.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In *NeurIPS*, 2024.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *EMNLP (1)*, pp. 9847–9857. Association for Computational Linguistics, 2021.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pp. 6576–6585. Computer Vision Foundation / IEEE Computer Society, 2018.

Itai Gat, Idan Schwartz, and Alexander G. Schwing. Perceptual score: What data modalities does your model perceive? In *NeurIPS*, pp. 21630–21643, 2021.

Roni Goldshmidt. Attention, please! pixelshap reveals what vision-language models actually focus on. *CoRR*, abs/2503.06670, 2025.

Roni Goldshmidt and Miriam Horovicz. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. *CoRR*, abs/2407.10114, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6325–6334. IEEE Computer Society, 2017.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, pp. 11021–11028. AAAI Press, 2020.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 2022.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL-HLT (1)*, pp. 3543–3556. Association for Computational Linguistics, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *AAAI*, pp. 11101–11108. AAAI Press, 2020.

Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pp. 11109–11116. AAAI Press, 2020.

Junyeong Kim, Minuk Ma, Trung X. Pham, Kyungsu Kim, and Chang D. Yoo. Modality shifting attention network for multi-modal video question answering. In *CVPR*, pp. 10103–10112. Computer Vision Foundation / IEEE, 2020.

Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video QA. In *AAAI*, pp. 13171–13179. AAAI Press, 2021.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pp. 13872–13882. IEEE, 2024.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pp. 22195–22206. IEEE, 2024.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, pp. 8658–8665. AAAI Press, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL https://api.semanticscholar.org/CorpusID:964287.

Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, Wenge Rong, and Zhang Xiong. Multimodal contrastive transformer for explainable recommendation. *IEEE Trans. Comput. Soc. Syst.*, 11(2):2632–2643, 2024.

Scott Lundberg. Welcome to the SHAP documentation — SHAP latest documentation, 2018. URL https://shap.readthedocs.io/en/latest/.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pp. 4765–4774, 2017.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. DIME: fine-grained interpretations of multimodal models via disentangled local explanations. In *AIES*, pp. 455–467. ACM, 2022.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL (1)*, pp. 12585–12602. Association for Computational Linguistics, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Saumitra Mishra, Bob L. Sturm, and Simon Dixon. Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, pp. 537–543, 2017.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *ACL (1)*, pp. 4032–4059. Association for Computational Linguistics, 2023.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *ICLR*. OpenReview.net, 2025.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *ACL (1)*, pp. 8253–8280. Association for Computational Linguistics, 2022.

Jayneel Parekh, Sanjeel Parekh, Pavlo Mozharovskyi, Gaël Richard, and Florence d'Alché-Buc. Tackling interpretability in audio classification networks with non-negative matrix factorization. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:1392–1405, 2024.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *AAAI*, pp. 19821–19829. AAAI Press, 2025.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *CVPR*, pp. 15526–15535. Computer Vision Foundation / IEEE, 2021.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *CVPR*, pp. 23901–23913. Computer Vision Foundation / IEEE, 2025.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. In *ECCV (33)*, volume 15091 of *Lecture Notes in Computer Science*, pp. 382–398. Springer, 2024.

Will Price and Dima Damen. Play fair: Frame attributions in video models. In *ACCV (5)*, volume 12626 of *Lecture Notes in Computer Science*, pp. 480–497. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Krithik Ramesh and Yun Sing Koh. Investigation of explainability techniques for multimodal transformers. In *AusDM*, volume 1741 of *Communications in Computer and Information Science*, pp. 90–98. Springer, 2022.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pp. 1135–1144. ACM, 2016a.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *CoRR*, abs/1606.05386, 2016b.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626. IEEE Computer Society, 2017.

Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *CVPR*, pp. 16877–16887. Computer Vision Foundation / IEEE, 2021.

L. Shapley. *7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317.*, pp. 69–79. Princeton University Press, Princeton, 1997. ISBN 9781400829156. doi: 10.1515/9781400829156-012. URL https://doi.org/10.1515/9781400829156-012.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, pp. 18221–18232. IEEE, 2024.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models. In *XAI4CV*, pp. 8182–8187, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. Multimodn - multimodal, multi-task, interpretable modular networks. In *NeurIPS*, 2023.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pp. 4631–4640. IEEE Computer Society, 2016.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pp. 9568–9578. IEEE, 2024.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *EMNLP*, pp. 8078–8088. Association for Computational Linguistics, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035, 2024b.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *CoRR*, abs/2212.03191, 2022.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV (85)*, volume 15143 of *Lecture Notes in Computer Science*, pp. 396–416. Springer, 2024c.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP/IJCNLP (1)*, pp. 11–20. Association for Computational Linguistics, 2019.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11205–11216. PMLR, 2021.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, 2024.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 2022.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pp. 9777–9786. Computer Vision Foundation / IEEE, 2021.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback. In *AAAI*, pp. 25543–25551. AAAI Press, 2025.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *ECCV (1)*, volume 15059 of *Lecture Notes in Computer Science*, pp. 39–57. Springer, 2024a.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *ICLR*. OpenReview.net, 2024b.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024b.

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

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. BERT representations for video question answering. In *WACV*, pp. 1545–1554. IEEE, 2020.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. In *ICLR*. OpenReview.net, 2024c.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*. OpenReview.net, 2025.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023.

Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *SIGIR*, pp. 829–832. ACM, 2017.

Zheng-Jun Zha, Jiawei Liu, Tianhao Yang, and Yongdong Zhang. Spatiotemporal-textual co-attention network for video question answering. *ACM Trans. Multim. Comput. Commun. Appl.*, 15(2s): 53:1–53:18, 2019.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *CoRR*, abs/2501.13106, 2025a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Trans. Mach. Learn. Res.*, 2025, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *CoRR*, abs/2410.02713, 2024.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. Uncovering the temporal context for video question answering. *Int. J. Comput. Vis.*, 124(3):409–421, 2017.

# APPENDIX

In the technical appendices we provide further related work for Video Question Answering in appendix A; more details about the mathematical background of Shapley values in appendix B; additional implementation details in appendix C; statistics of each of the dataset subsets in appendix D; further quantitative results in appendix E; and qualitative results in appendix F.

## A FURTHER RELATED WORK

**Video Question Answering** Video question answering (VQA) was an extension of the visual question answering task (Antol et al., 2015), from images to videos (Tapaswi et al., 2016; Zhu et al., 2017). Initially, VQA methods utilised a two-stream encoder approach (Zha et al., 2019; Ye et al., 2017; Yang et al., 2020; Seo et al., 2021; Park et al., 2021; Li et al., 2019; Kim et al., 2021; 2020; Jiang & Han, 2020; Jiang et al., 2020; Huang et al., 2020; Gao et al., 2018; Fan et al., 2019; Dang et al., 2021; Wang et al., 2024c) to answer questions, which has progressed to multi-modal large language models (MLLMs) able to reason further about the visual and textual content (Ye et al., 2025; Li et al., 2023; Maaz et al., 2024; Ye et al., 2023). Recently, VQA has become a common benchmark for MLLMs with many datasets being recently proposed to test various aspects of methods' understanding. There has been a growing trend among VQA datasets and methods on longer videos (Wang et al., 2024b; Wu et al., 2024; Song et al., 2024; Fang et al., 2024) as well as constructing datasets that require strong multi-modal understanding (Xiao et al., 2021; Perrett et al., 2025; Mangalam et al., 2023; Chen et al., 2024c;b; Xie et al., 2024a). In this work, we investigate 6 VLMs, ranging from two-stream approaches (Yang et al., 2022; Wang et al., 2022) to MLLMs (Cheng et al., 2024; Zhang et al., 2024; 2025b;a) across 4 recent, challenging datasets (Perrett et al., 2025; Mangalam et al., 2023; Wang et al., 2024b; Li et al., 2024).

## B MATHEMATICAL BACKGROUND

We first expand on the properties mentioned within the main paper for Shapley values (Shapley, 1997) and SHAP (Lundberg & Lee, 2017) below.

### B.1 SHAPLEY VALUE PROPERTIES

**Property B.1 (Efficiency)**

$$\sum_{i \in N} \varphi_i(r) = r(N)$$

This property requires that the sum of the contributions across all players is equal to the model output.

**Property B.2 (Symmetry)** *If* $r(S \cup i) = r(S \cup j) \quad \forall S \subseteq N \setminus \{i, j\}, \forall i, j \in N$ *such that* $i \neq j$*, then:*

$$\varphi_i(r) = \varphi_j(r)$$

This property requires that two players that contribute equally will be assigned the same score.

**Property B.3 (Linearity)** *If we have two different reward functions* $v$ *and* $w$*:*

$$\varphi_i(r + t) = \varphi_i(r) + \varphi_i(t) \quad \textit{and} \quad \varphi_i(ar) = a\varphi_i(r) \quad \forall a \in \mathbb{R}$$

The linearity property requires that the Shapley value of the sum of two (or more) reward functions is the same as the sum of their individual Shapley values.

**Property B.4 (Null player)** *If* $r(S \cup \{i\}) = r(S) \quad \forall S \in N \setminus \{i\}$*:*

$$\varphi_i(r) = 0$$

Finally, the Null player property requires that if a player does not contribute then they get assigned a score of 0.

### B.2 SHAP PROPERTIES

**Property B.5 (Local accuracy)**

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M}(\phi_i \times x'_i)$$

Local accuracy requires that if a model is approximated, then the output of the approximated model should match the original output.

**Property B.6 (Missingness)**

$$x'_i = 0 \implies \phi_i = 0$$

The Missingness property requires that if a feature is missing then it won't be given any attribution.

**Property B.7 (Consistency)** *Let $\mathbf{e}_i$ be the $M$ sized vector of 1s with a 0 at position $i$ and $\odot$ the Hadamard (elementwise) product. For any two models $f$ and $f'$ with attributions $\phi_i$ and $\phi'_i$ respectively, if:*

$$f'(h_x(z')) - f'(h_x(z' \odot \mathbf{e}_i)) \geq f(h_x(z')) - f(h_x(z' \odot \mathbf{e}_i)) \quad \forall z' \in \{0,1\}^M$$

*then $\phi'_i \geq \phi_i$*

Finally, the consistency property requires that if a feature's contribution stays constant or is increased due to a change in the model, then its attribution won't decrease.

**Definition B.1 (Additive feature attribution)** *An additive feature attribution method has explanation model:*

$$g(x') = \phi_0 + \sum_{i=1}^{M}(\phi_i \times x'_i)$$

*where $x' \in \{0,1\}^M$, $M$ is the number of simplified features and $\phi_i \in \mathbb{R}$ is the attribution of each simplified feature.*

**Theorem B.1** *If we have explanation model $g$ defined as in Definition 3.2, then the only attribution satisfying Properties B.5 to B.7 is the Shapley value. In other words:*

$$\phi_i = \varphi_i(f)$$

From the SHAP (Lundberg & Lee, 2017) paper, only the Shapley value satisfies all of the above properties via Additive feature attribution.

## C IMPLEMENTATION DETAILS

FrozenBiLM and InternVideo were evaluated on a single 1080Ti GPU. VideoLLaMA2, LLaVA-Video, LongVA and VideoLLaMA3 were evaluated on a single GH200 GPU. All experiments took a total of $\sim 1800$ node hours/7200 GPU hours. For EgoSchema, MVBench and LVBench, frames are sampled uniformly from the single source videos (at the default framerate). However, as HD-EPIC takes multi-video input, we instead concatenate all videos from the VQA-tuple and then uniformly sample frames from this sequence. We pre-processed HD-EPIC into 1 FPS videos first.

In the name of reproducibility, these are the model checkpoints we used:

- FrozenBiLM (Yang et al., 2022) - CLIP ViT-L-14 and pre-trained on WebVid10M + How2QA

- InternVideo (Wang et al., 2022) - CLIP ViT-L-14 and pre-trained on MSRVTT
- VideoLLaMA2 (Cheng et al., 2024) - VideoLLaMA2-7B-16F
- LLaVA-Video (Zhang et al., 2024) - LLaVA-Video-7B-Qwen2
- LongVA (Zhang et al., 2025b) - LongVA-7B
- VideoLLaMA3 (Zhang et al., 2025a) - VideoLLaMA3-7B

As well as this, all code, dataset subsets, results and model checkpoints used in our experiments will be released after the review period to encourage reproducibility and usage of the framework.

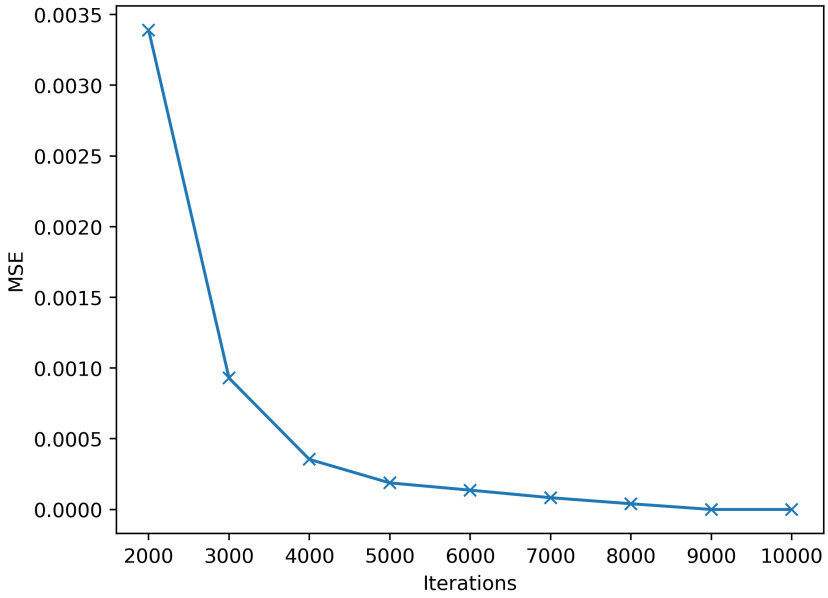



Figure 4: Plot of the MSE (mean squared error) of Shapley values at varying iterations for Frozen-BiLM on the longest EgoSchema question. Error is calculated against values for $10,000$ iterations.

To select $5000$ iterations as the value for our experiments, we ablated the parameter by varying it and selecting the point with a sensible error trade-off. In fig. 4 we have the results of the ablation and can see that $5000$ iterations is past the elbow of the curve and where the gradient of the error begins to flatten. As such, we selected the value for all experiments.

## D Subset Statistics

Table 3: Statistics of the subsets used for evaluation.

| Dataset | # VQA-tuples | Avg. Video Length | Avg. Question Length | Avg. Answer Length |
|---|---|---|---|---|
| EgoSchema | 50 | 180.00s | 24.36 words | 108.66 words |
| HD-EPIC | 60 | 1180.84s | 21.17 words | 56.75 words |
| MVBench | 60 | 16.09s | 13.32 words | 15.25 words |
| LVBench | 60 | 3800.13s | 10.25 words | 28.80 words |

We provide additional details and statistics of the subsets we used for the calculation of the Shapley values and the evaluation. Table 3, contains the number of VQA-tuples, as well as the average video length (of all unique videos corresponding to the subset), question length, and answer length.

# E QUANTITATIVE RESULTS

We provide additional results of experiments in the appendix. Firstly, we demonstrate that we can calculate Shapley values for open-ended VQA in appendix E.1. Secondly, we compare Shapley based rankings of frames to those generated by Gemini in appendix E.2. Then, we investigate the Modality Contribution and Per-Feature Contribution metrics for false logits in appendix E.3. After this, we showcase how masking *negatively* contributing features affects the results in appendix E.4, and the same for *positively* contributing features in appendix E.5. We visualise how video contributions vary across video context length in appendix E.6, and plot the distribution of Shapley values across all methods and datasets to compare how they differ in appendix E.7 Finally, we demonstrate how injecting new answers affects the masked performance in appendix E.8.

## E.1 OPEN-ENDED VISUAL QUESTION ANSWERING

Table 4: MC and PFC for each modality in the VQA-tuple. Calculated based on the Shapley values for the ground truth text for open-ended EgoSchema and averaged across all VQA-tuples. Here **blue** scores relate to large magnitudes of Shapley values, regardless of their sign, while **red** scores relate to values close to 0.

| | Modality Contribution | | Per-Feature Contribution | |
|---|---|---|---|---|
| | V | Q | V | Q |
| L-V | 0.38 | 0.62 | 0.21 | 0.79 |
| VL3 | 0.69 | 0.31 | 0.25 | 0.75 |

Throughout the main paper we only discuss multiple-choice VQA because of the fact that its evaluation is simpler and it has a greater potential for textual bias. However, to demonstrate that our approach extends trivially to open-ended VQA, we include an example. In the open-ended scenario, the language model predicts a set of several output tokens, which then need to be compared to a piece of ground truth text. There is no longer a single logit to use as the reward function (and the set of logits of the generated tokens cannot be easily used because it is of variable length), so we decided to employ a captioning metric. We used ROUGE-L (Lin, 2004) (between the predicted and ground truth text) as the captioning metric. In table 4, we have the feature contributions for LLaVA-Video and VideoLLaMA3 on an open-ended version of EgoSchema where the answers have been removed from the input. We see that the ratio between video and question is extremely similar to that of the multiple-choice results and that the Per-Feature Contribution of video remains significantly smaller than the textual modality.

## E.2 GEMINI RANKING CORRELATION

```
You will be given frames from a video and a question with
multiple-choice answer options.
frame_0: {frame 0} ,... ,frame_n-1:{frame n-1}
Question: {question}
Options: {answer choices}
You do not need to answer the question; return a comma
separated list of frame ids in the order of their importance
for answering the above question. Order the frame ids by
importance, do not leave them in chronological order. Respond
with exactly 180 frame ids, returning only this comma
separated list and excluding all other textual output.
```

Listing 1: Example system prompt for generating ranks for a VQA-tuple with 180 input frames.

To determine how much the framewise Shapley value contributions relate to common sense understanding of video, we compare them to a baseline generated from Gemini queries. In particular, we first ask Gemini to rank all of the input frames that would be sampled for a given model and dataset pair. The system prompt for querying Gemini is shown in listing 1. Then we calculate

the Spearman's correlation between this ranking and the ranking obtained by sorting the frames by absolute Shapley value. Plotting the correlation for each VQA-tuple in fig. 5, we see their distributions for several model/dataset combinations. As the modal correlation is always very close to 0, there is little correlation between the two ranking systems, which aligns with our observations that the top 16 most influential frames often disagree with common sense reasoning. The models are often focusing on frames that are irrelevant for answering the question.

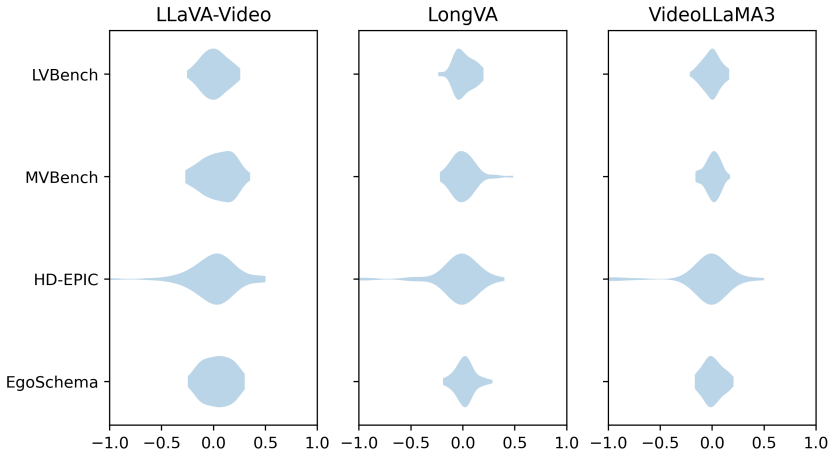



Figure 5: Violin plots of the Spearman's correlations between the Gemini rankings and the Shapley value rankings for the frames of VQA-tuple questions.

### E.3 Contributions for False Logits

Table 5: MC and PFC for each modality in the VQA-tuple. Calculated based on the Shapley values for the *false logits* averaged across all VQA-tuples. Here **blue** scores relate to large magnitudes of Shapley values, regardless of their sign, while **red** scores relate to values close to 0.

(a) EgoSchema

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.08 | 0.27 | 0.66 | 0.30 | 0.41 | 0.29 | 0.20 |
| IV | 0.22 | 0.37 | 0.41 | 0.54 | 0.35 | 0.11 | 0.36 |
| VL2 | 0.11 | 0.20 | 0.69 | 0.30 | 0.35 | 0.34 | 0.56 |
| L-V | 0.17 | 0.17 | 0.66 | 0.17 | 0.40 | 0.43 | 0.72 |
| LVA | 0.18 | 0.16 | 0.66 | 0.10 | 0.43 | 0.48 | 0.48 |
| VL3 | 0.37 | 0.14 | 0.49 | 0.17 | 0.43 | 0.40 | 0.70 |

(b) HD-EPIC

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.09 | 0.54 | 0.38 | 0.20 | 0.55 | 0.25 | 0.22 |
| IV | 0.33 | 0.50 | 0.17 | 0.53 | 0.39 | 0.08 | 0.15 |
| VL2 | 0.19 | 0.28 | 0.53 | 0.29 | 0.29 | 0.41 | 0.28 |
| L-V | 0.22 | 0.29 | 0.49 | 0.19 | 0.36 | 0.45 | 0.35 |
| LVA | 0.22 | 0.23 | 0.55 | 0.14 | 0.32 | 0.55 | 0.35 |
| VL3 | 0.41 | 0.22 | 0.37 | 0.19 | 0.40 | 0.42 | 0.35 |

(c) MVBench

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.11 | 0.52 | 0.37 | 0.13 | 0.50 | 0.36 | 0.40 |
| IV | 0.27 | 0.52 | 0.21 | 0.33 | 0.49 | 0.18 | 0.42 |
| VL2 | 0.23 | 0.28 | 0.50 | 0.20 | 0.28 | 0.52 | 0.62 |
| L-V | 0.25 | 0.28 | 0.47 | 0.07 | 0.34 | 0.59 | 0.65 |
| LVA | 0.18 | 0.28 | 0.54 | 0.02 | 0.33 | 0.64 | 0.45 |
| VL3 | 0.44 | 0.26 | 0.30 | 0.06 | 0.44 | 0.51 | 0.65 |

(d) LVBench

| | Modality Contribution | | | Per-Feature Contribution | | | Acc |
|---|---|---|---|---|---|---|---|
| | V | Q | A | V | Q | A | |
| FBLM | 0.15 | 0.45 | 0.40 | 0.21 | 0.47 | 0.32 | 0.30 |
| IV | 0.34 | 0.44 | 0.22 | 0.42 | 0.44 | 0.14 | 0.25 |
| VL2 | 0.26 | 0.27 | 0.47 | 0.25 | 0.32 | 0.43 | 0.27 |
| L-V | 0.30 | 0.27 | 0.43 | 0.10 | 0.40 | 0.49 | 0.42 |
| LVA | 0.39 | 0.19 | 0.42 | 0.07 | 0.35 | 0.58 | 0.35 |
| VL3 | 0.54 | 0.17 | 0.29 | 0.10 | 0.41 | 0.49 | 0.48 |

Within the main paper we focused on contributions based on ground truth logits in section 4.2, whereas here in the appendix we provide an exploration into how contributions differ based on the false logits. As there are several false logits, instead of a single ground truth logit, we average the Shapley values across the false logits. Table 5, highlights these results. Overall, we see similar trends between the two tables across all datasets and models: namely that video is important as an entire

modality for VideoLLaMA3, but that per-frame contributions remain low for long context models. As well as this, the question remains under-represented compared to the answers.

### E.4 Masking Negative Contributions

Table 6: *How does masking the input based upon **negative** Per-Feature Contributions affect accuracy?* We mask all negative features across the entire input (joint) or each modality separately. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy.

| Model | Masking | EgoSchema | HD-EPIC | MVBench | LVBench |
|---|---|---|---|---|---|
| FrozenBiLM | None | 0.20 | 0.22 | 0.40 | 0.30 |
| | All | +0.14 | +0.02 | +0.10 | +0.00 |
| | Video | +0.06 | +0.03 | +0.10 | +0.10 |
| | Question | +0.00 | +0.07 | +0.08 | +0.03 |
| | Answer | +0.14 | +0.02 | +0.10 | +0.00 |
| InternVideo | None | 0.36 | 0.15 | 0.42 | 0.25 |
| | All | +0.10 | -0.03 | -0.03 | +0.05 |
| | Video | +0.04 | +0.08 | +0.02 | +0.08 |
| | Question | +0.08 | +0.00 | +0.05 | +0.07 |
| | Answer | +0.10 | -0.03 | -0.03 | +0.05 |
| VideoLLaMA2 | None | 0.56 | 0.28 | 0.62 | 0.27 |
| | All | +0.10 | +0.08 | +0.05 | +0.03 |
| | Video | +0.02 | +0.07 | -0.02 | +0.22 |
| | Question | +0.06 | +0.07 | +0.07 | +0.12 |
| | Answer | +0.10 | +0.08 | +0.05 | +0.03 |
| LLaVA-Video | None | 0.72 | 0.35 | 0.65 | 0.42 |
| | All | +0.08 | +0.03 | +0.08 | +0.02 |
| | Video | +0.06 | +0.05 | +0.07 | +0.20 |
| | Question | +0.08 | +0.03 | +0.13 | +0.02 |
| | Answer | +0.08 | +0.03 | +0.08 | +0.02 |
| LongVA | None | 0.48 | 0.35 | 0.45 | 0.35 |
| | All | +0.18 | +0.12 | +0.10 | +0.03 |
| | Video | +0.10 | +0.02 | +0.00 | +0.17 |
| | Question | +0.16 | -0.02 | +0.10 | +0.13 |
| | Answer | +0.18 | +0.12 | +0.10 | +0.03 |
| VideoLLaMA3 | None | 0.70 | 0.35 | 0.65 | 0.48 |
| | All | +0.10 | +0.03 | +0.05 | -0.02 |
| | Video | +0.12 | +0.15 | +0.07 | +0.12 |
| | Question | +0.02 | +0.12 | +0.07 | +0.10 |
| | Answer | +0.10 | +0.03 | +0.05 | -0.02 |

In the main paper in section 4.3 we demonstrated the effect upon performance when masking entire modalities. Now, in table 6 we mask individual features if their Shapley values are negative. For the sake of fairness, only the ground truth answer can be masked to ensure that the masking does not just greedily remove all of the false answer text (thus trivialising the multiple-choice). The intuition here is to verify the extent to which the Shapley values help inform the potential accuracy ceiling of the model. Generally, all models gain performance when distractors are masked. For EgoSchema, HD-EPIC and MVBench, the biggest increases tend to be answer, question, then video. In LVBench, masking video becomes more important, likely because the context is always long, and the question may only be relevant to a small portion of frames. VideoLLaMA3 gains a significant performance boost from masking video in EgoSchema, HD-EPIC and LVBench but requires masking (on average) 35%, 39% and 40% of frames respectively, compared to 28% for MVBench which only gains 7% accuracy. To exploit the video modality to its best ability, even the strongest model we test needs to remove more than a third of the frames. This points toward better frame sampling methods to be an important consideration for future models. Overall, we see a tendency for accuracy to be explained well by the Shapley values.

### E.5 Masking Positive Contributions

Table 7: *How does masking the input based upon **positive** Per-Feature Contributions affect accuracy?* We mask all negative features across the entire input (joint) or each modality separately. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy.

| Model | Masking | EgoSchema | HD-EPIC | MVBench | LVBench |
|---|---|---|---|---|---|
| FrozenBiLM | None | 0.20 | 0.22 | 0.40 | 0.30 |
| | All | +0.14 | +0.03 | +0.10 | +0.00 |
| | Video | -0.02 | -0.02 | -0.08 | -0.07 |
| | Question | +0.04 | +0.02 | -0.13 | +0.00 |
| | Answer | +0.14 | +0.03 | +0.10 | +0.00 |
| InternVideo | None | 0.36 | 0.15 | 0.42 | 0.25 |
| | All | +0.10 | -0.03 | -0.03 | +0.05 |
| | Video | +0.02 | +0.00 | -0.03 | +0.00 |
| | Question | +0.00 | +0.05 | +0.02 | -0.03 |
| | Answer | +0.10 | -0.03 | -0.03 | +0.05 |
| VideoLLaMA2 | None | 0.56 | 0.28 | 0.62 | 0.27 |
| | All | +0.06 | +0.02 | +0.00 | +0.05 |
| | Video | +0.00 | -0.10 | -0.10 | -0.08 |
| | Question | -0.02 | -0.12 | -0.20 | -0.07 |
| | Answer | +0.06 | +0.02 | +0.00 | +0.05 |
| LLaVA-Video | None | 0.72 | 0.35 | 0.65 | 0.42 |
| | All | +0.02 | -0.07 | +0.07 | +0.00 |
| | Video | -0.26 | -0.17 | -0.10 | -0.20 |
| | Question | -0.02 | -0.17 | -0.20 | -0.13 |
| | Answer | +0.02 | -0.07 | +0.07 | +0.00 |
| LongVA | None | 0.48 | 0.35 | 0.45 | 0.35 |
| | All | +0.04 | -0.02 | +0.00 | -0.07 |
| | Video | -0.12 | -0.10 | -0.07 | -0.12 |
| | Question | -0.22 | -0.23 | -0.17 | -0.05 |
| | Answer | +0.04 | -0.02 | +0.00 | -0.07 |
| VideoLLaMA3 | None | 0.70 | 0.35 | 0.65 | 0.48 |
| | All | +0.02 | +0.05 | +0.05 | -0.10 |
| | Video | -0.06 | -0.08 | -0.02 | -0.12 |
| | Question | -0.02 | -0.13 | -0.15 | -0.15 |
| | Answer | +0.02 | +0.05 | +0.05 | -0.10 |

Here we instead mask individual features if their Shapley values are positive, to see how the accuracy is affected if *all* features are distractors. Again, for the sake of fairness, only the ground truth answer can be masked. Table 7 shows the results. Compared to the results in the main paper, we see that masking the video or question decreases the accuracy significantly, confirming that features contributing positively are required for strong performance. However, masking the answers actually results in small improvements in accuracy, likely because the masking of the ground truth answer makes this option stand out and allows the language model to exploit the difference.

### E.6 Video Contribution vs. Video Context

We visualise the relationship between video Per-Feature Contribution ($\text{PFC}_\text{V}$) and video context length in fig. 6. HD-EPIC was used for this because it contains videos with the highest variance in context length. The Pearson correlation between these two variables is $-0.36$, suggesting a slight negative correlation, which means that as the video input increases in length, the less each frame contributes to the output.

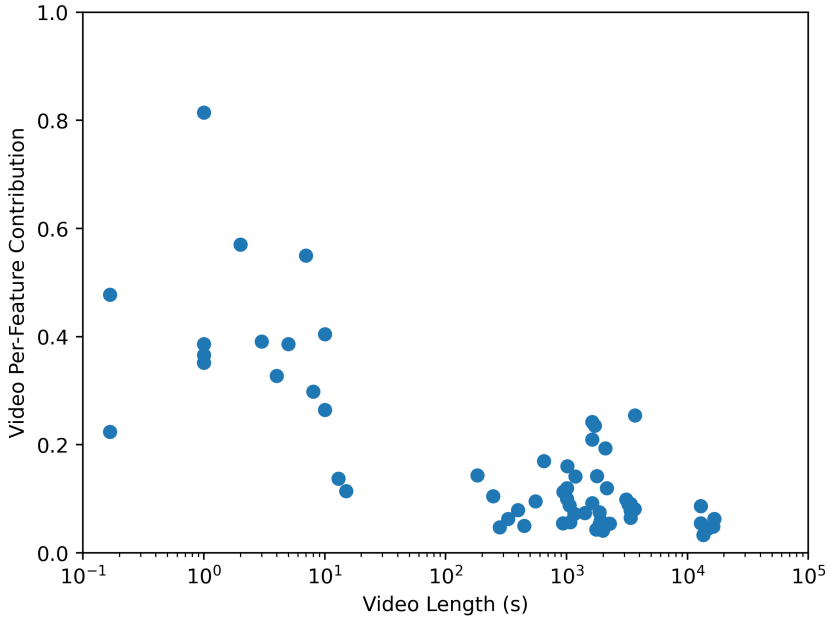



Figure 6: $PFC_V$ plotted against video context length in seconds for the HD-EPIC subset we used. The x-axis is log base 10 scale.

### E.7 Distributions of Shapley Values

We showcase the per modality distributions of Shapley values across all models and datasets as violin plots in fig. 7. We plot a violin of the set of Shapley values for all of the video, question and answer features separately for a given dataset subset. As we move down the plots, we see that the height of the violin for video decreases significantly, and that it increases for the answers. On the other hand, the violins for the question presents similar distributions throughout. MVBench and LVBench demonstrate larger answer contributions than EgoSchema and HD-EPIC, further indicating that these datasets are skewing attention towards discriminating between multiple-choice answers. Overall, we see that the larger models utilising LLMs as backbones tend to be biased towards the text modalities.

### E.8 Answer Replacement Masking

In this subsection we plot the accuracy of models (similarly to section 4.3) when masking modalities for the datasets with injected negatives from answer replacement. We see in table 8 that "Easy" answer replacement significantly improves performance in all scenarios as it becomes more trivial for the model to discern the true negative. Then, in table 9, table 10, table 11 and table 12 we have the performance on each of the "New" replacement types. In general, we see that adding these extra negatives often increases the accuracy lost when masking out video, particularly for EgoSchema and LVBench. Masking out the question can also have an increased effect on performance, especially for LVBench. The differences are significantly less pronounced for HD-EPIC, likely because the unmasked performance drops as low as $10\%$. To summarise, these tables show support for our findings in section 4.4 that injecting new negatives can have an effect on the contributions of the video and question modalities.

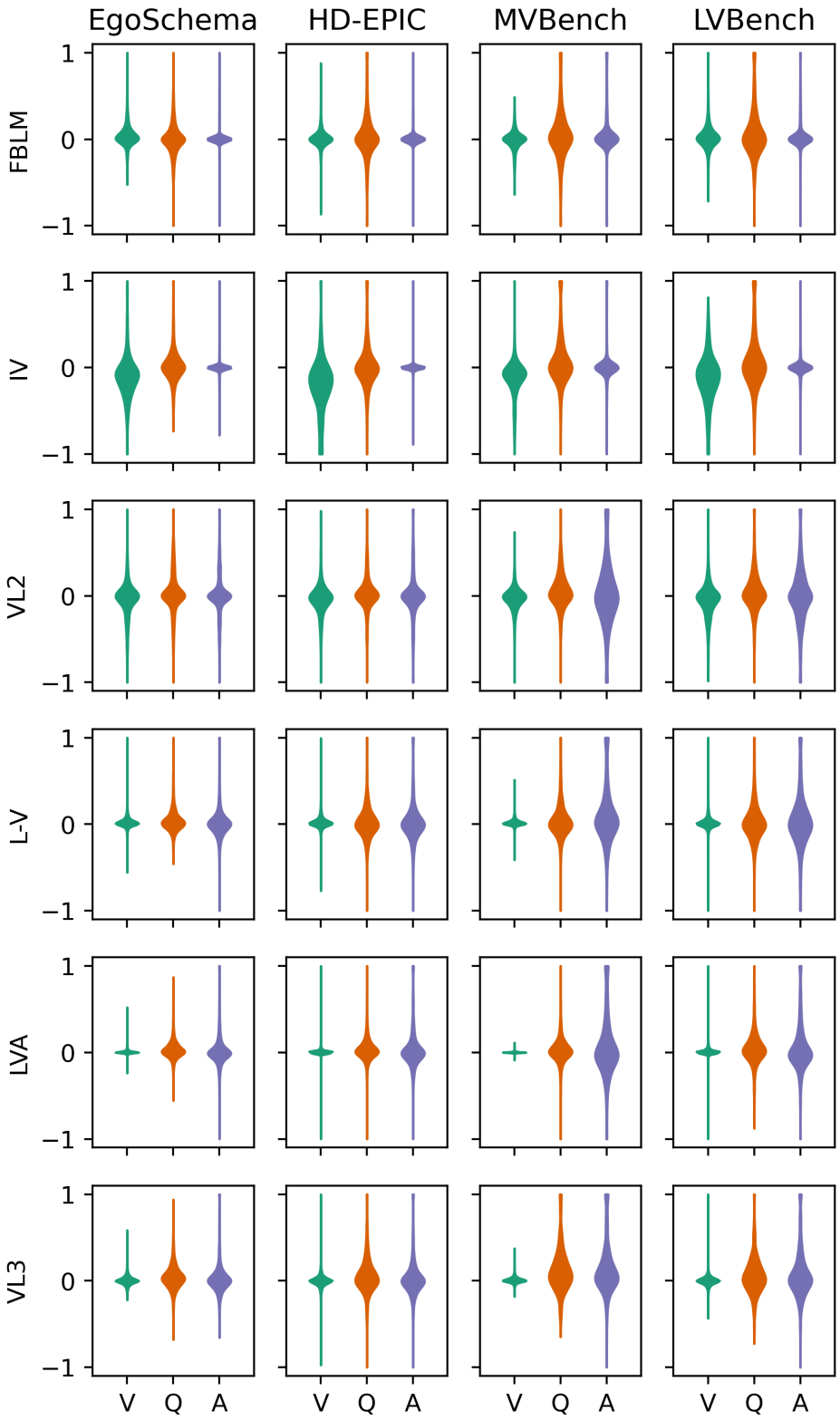



Figure 7: Violin plots of the Shapley values per modality across all models and datasets, for the ground truth logits.

Table 8: Modality masking performance for "Easy" answer replacement. We mask either all input features ("All"), or each modality separately and compare to baseline performance. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy compared to baseline.

| Model | Masking | EgoSchema | HD-EPIC | LVBench |
|---|---|---|---|---|
| VideoLLaMA2 | None | 0.94 | 0.67 | 0.83 |
| | All | -0.68 | -0.25 | -0.58 |
| | Video | -0.40 | -0.03 | -0.10 |
| | Question | -0.24 | -0.47 | -0.35 |
| | Answer | -0.68 | -0.25 | -0.58 |
| LLaVA-Video | None | 0.98 | 0.77 | 0.97 |
| | All | -0.80 | -0.35 | -0.60 |
| | Video | -0.36 | 0.02 | -0.15 |
| | Question | -0.04 | -0.42 | -0.30 |
| | Answer | -0.80 | -0.35 | -0.60 |
| LongVA | None | 0.88 | 0.60 | 0.88 |
| | All | -0.70 | -0.27 | -0.45 |
| | Video | -0.34 | -0.03 | -0.05 |
| | Question | -0.14 | -0.33 | -0.17 |
| | Answer | -0.70 | -0.27 | -0.45 |
| VideoLLaMA3 | None | 1.00 | 0.77 | 0.95 |
| | All | -0.82 | -0.52 | -0.58 |
| | Video | -0.40 | -0.02 | -0.03 |
| | Question | -0.10 | -0.45 | -0.22 |
| | Answer | -0.82 | -0.52 | -0.58 |

Table 9: Modality masking performance for "New-5" answer replacement. We mask either all input features ("All"), or each modality separately and compare to baseline performance. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy compared to baseline.

| Model | Masking | EgoSchema | HD-EPIC | LVBench |
|---|---|---|---|---|
| VideoLLaMA2 | None | 0.48 | 0.18 | 0.37 |
| | All | -0.48 | -0.15 | -0.35 |
| | Video | -0.28 | -0.10 | -0.10 |
| | Question | 0.02 | -0.05 | -0.20 |
| | Answer | -0.48 | -0.15 | -0.35 |
| LLaVA-Video | None | 0.66 | 0.10 | 0.38 |
| | All | -0.56 | -0.03 | -0.32 |
| | Video | -0.40 | 0.02 | -0.22 |
| | Question | -0.04 | -0.02 | -0.18 |
| | Answer | -0.56 | -0.03 | -0.32 |
| LongVA | None | 0.44 | 0.15 | 0.30 |
| | All | -0.38 | -0.10 | -0.25 |
| | Video | -0.22 | -0.07 | -0.15 |
| | Question | -0.12 | -0.08 | -0.07 |
| | Answer | -0.38 | -0.10 | -0.25 |
| VideoLLaMA3 | None | 0.74 | 0.15 | 0.30 |
| | All | -0.74 | -0.07 | -0.27 |
| | Video | -0.58 | -0.05 | -0.03 |
| | Question | -0.08 | -0.02 | -0.13 |
| | Answer | -0.74 | -0.07 | -0.27 |

Table 10: Modality masking performance for "New-10" answer replacement. We mask either all input features ("All"), or each modality separately and compare to baseline performance. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy compared to baseline.

| Model | Masking | EgoSchema | HD-EPIC | LVBench |
|---|---|---|---|---|
| VideoLLaMA2 | None | 0.48 | 0.10 | 0.25 |
| | All | -0.44 | -0.05 | -0.23 |
| | Video | -0.26 | 0.00 | -0.05 |
| | Question | -0.06 | 0.05 | -0.10 |
| | Answer | -0.44 | -0.05 | -0.23 |
| LLaVA-Video | None | 0.68 | 0.12 | 0.32 |
| | All | -0.62 | -0.08 | -0.20 |
| | Video | -0.48 | 0.02 | -0.17 |
| | Question | -0.08 | -0.05 | -0.10 |
| | Answer | -0.62 | -0.08 | -0.20 |
| LongVA | None | 0.44 | 0.10 | 0.33 |
| | All | -0.38 | -0.03 | -0.28 |
| | Video | -0.30 | -0.02 | -0.15 |
| | Question | -0.16 | -0.05 | -0.10 |
| | Answer | -0.38 | -0.03 | -0.28 |
| VideoLLaMA3 | None | 0.68 | 0.10 | 0.40 |
| | All | -0.58 | -0.05 | -0.37 |
| | Video | -0.40 | 0.02 | -0.15 |
| | Question | -0.06 | -0.03 | -0.18 |
| | Answer | -0.58 | -0.05 | -0.37 |

Table 11: Modality masking performance for "New-15" answer replacement. We mask either all input features ("All"), or each modality separately and compare to baseline performance. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy compared to baseline.

| Model | Masking | EgoSchema | HD-EPIC | LVBench |
|---|---|---|---|---|
| VideoLLaMA2 | None | 0.42 | 0.17 | 0.23 |
| | All | -0.40 | -0.10 | -0.18 |
| | Video | -0.18 | -0.05 | -0.05 |
| | Question | 0.02 | -0.03 | -0.07 |
| | Answer | -0.40 | -0.10 | -0.18 |
| LLaVA-Video | None | 0.66 | 0.15 | 0.32 |
| | All | -0.54 | -0.10 | -0.28 |
| | Video | -0.34 | -0.07 | -0.13 |
| | Question | 0.00 | -0.05 | -0.08 |
| | Answer | -0.54 | -0.10 | -0.28 |
| LongVA | None | 0.34 | 0.13 | 0.22 |
| | All | -0.26 | 0.02 | -0.17 |
| | Video | -0.22 | 0.03 | -0.07 |
| | Question | -0.04 | -0.08 | 0.02 |
| | Answer | -0.26 | 0.02 | -0.17 |
| VideoLLaMA3 | None | 0.66 | 0.15 | 0.37 |
| | All | -0.56 | -0.08 | -0.25 |
| | Video | -0.34 | 0.00 | -0.13 |
| | Question | -0.04 | -0.03 | -0.15 |
| | Answer | -0.56 | -0.08 | -0.25 |

Table 12: Modality masking performance for "New-20" answer replacement. We mask either all input features ("All"), or each modality separately and compare to baseline performance. "None" represents the vanilla accuracy. Here **green**/**red** refers to an increase/decrease in accuracy compared to baseline.

| Model | Masking | EgoSchema | HD-EPIC | LVBench |
|---|---|---|---|---|
| VideoLLaMA2 | None | 0.48 | 0.27 | 0.27 |
| | All | -0.38 | -0.15 | -0.18 |
| | Video | -0.12 | -0.10 | 0.00 |
| | Question | 0.08 | -0.07 | -0.07 |
| | Answer | -0.38 | -0.15 | -0.18 |
| LLaVA-Video | None | 0.68 | 0.17 | 0.38 |
| | All | -0.60 | -0.07 | -0.28 |
| | Video | -0.48 | -0.08 | -0.15 |
| | Question | 0.02 | -0.05 | -0.08 |
| | Answer | -0.60 | -0.07 | -0.28 |
| LongVA | None | 0.40 | 0.20 | 0.25 |
| | All | -0.32 | 0.00 | -0.18 |
| | Video | -0.24 | -0.03 | -0.07 |
| | Question | -0.04 | -0.07 | 0.00 |
| | Answer | -0.32 | 0.00 | -0.18 |
| VideoLLaMA3 | None | 0.74 | 0.15 | 0.45 |
| | All | -0.64 | -0.05 | -0.35 |
| | Video | -0.48 | 0.03 | -0.18 |
| | Question | -0.08 | -0.02 | -0.15 |
| | Answer | -0.64 | -0.05 | -0.35 |

# F QUALITATIVE RESULTS

Here, we provide further qualitative results of our experimental results. First, in appendix F.1, we showcase all remaining heatmaps across all datasets. Next, we highlight further examples of entire VQA-tuple visualisations in appendix F.2. Finally, we also provide wordclouds showcasing Shapley values of specific words in the subsets in appendix F.3.

## F.1 HEATMAPS OF SHAPLEY VALUES

Expanding from the example in the main paper for VideoLLaMA3, we present the heatmaps across all other methods and datasets. FrozenBiLM in fig. 8, InternVideo in fig. 9, VideoLLaMA2 in fig. 10, LLaVA-Video in fig. 11, and LongVA can be found in fig. 12. Note that we truncate these figures for readability in a similar fashion for the heatmaps in the main paper. We find similar trends across all models in which there is a clear separation between videos and question/answer values. However, for VideoLLaMA2 and InternVideo we find that the video frames are distinguished often via a negative Shapley value, rather than a lower magnitude. Similarly to the results in the main paper for VideoLLaMA3, there is no strong temporal bias across the features (i.e. similar columns of values across the questions). For long context models, we again see that individual frames have low contributions.

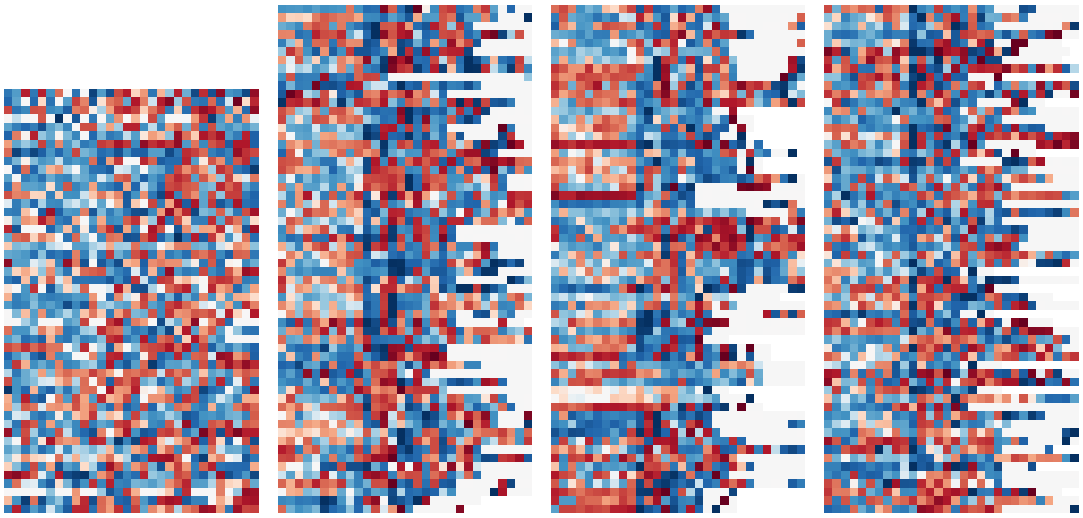
## F.2 FURTHER QUALITATIVE EXAMPLES

We showcase further qualitative examples from VideoLLaMA3 across all datasets including EgoSchema in fig. 13, HD-EPIC in fig. 14, MVBench in fig. 15, and LVBench in fig. 16. We see similar trends to the example shown in the main paper in which VideoLLaMA3 will attribute negative scores to words in negative answers that match those in the ground truth answer (i.e. "paintbrush" in fig. 13). We also see a tendency for the sign of the frame contributions to not correspond with common sense. For example, in fig. 13, frames containing the mentioned "cup of water" are actually negatively attributed, which is strange considering it is the topic of the question. In fig. 15 the frame attributions appear to make more sense, as the model is positively influenced by early frames of the object at the start of motion and latter frames of the object at the end of motion, largely disregarding more intermediate frames. Although this is sensible, the question can essentially be solved solely by identifying the translation of the object between two frames, meaning it requires little temporal context. Therefore, these valid attributions are indicative of the simplicity of the underlying dataset. Finally, in fig. 14 we notice that none of the selected frames actually contain the ground truth object (the Jack of Spades card). Checking the sampled frames, we found that a frame containing this object is sampled in the input, but that its attribution is close to zero, providing more evidence that the model is not necessarily being guided by relevant frames. Overall, the video frames also have the same tendency to show lower peaks than the question/answer and the $\text{PFC}_\text{V}$ scores continue to be less than $0.1$ across all the examples.

## F.3 WORD CLOUDS OF SHAPLEY VALUES

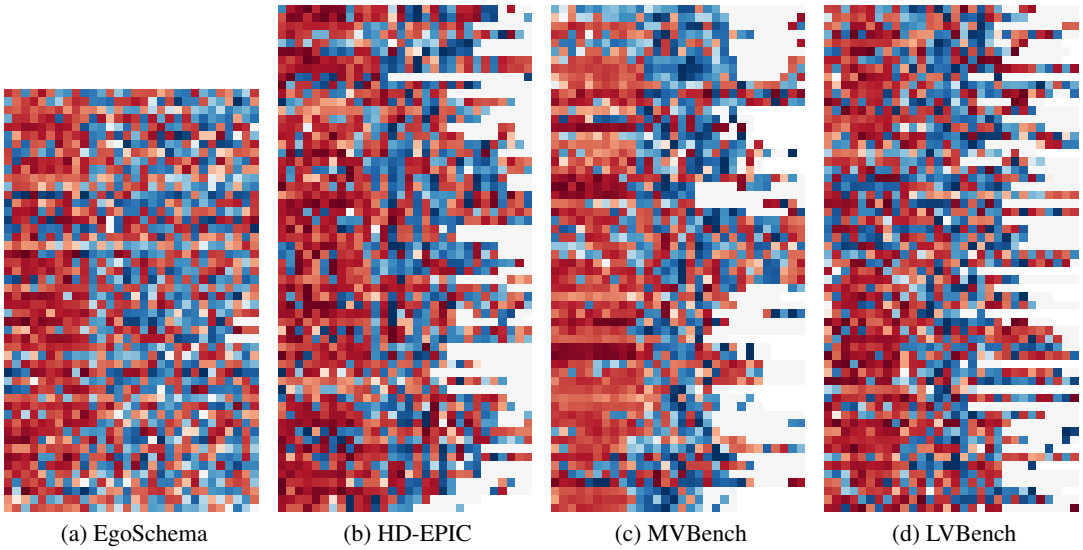
In this section, we explore how words are attributed based on their frequency within each of the dataset subsets. Specifically, we plot word clouds for each method on each dataset, combining all question and answers, so that the size of the word is proportional to its frequency. The colour of each word is calculated as the word's average Shapley value (for the ground truth logit) across all of its instances, with **blue** representing positively attributed words and **red** representing negatively attributed words.

The wordcloud for FrozenBiLM is in fig. 17, InternVideo is in fig. 18, VideoLLaMA2 is in fig. 19, LLaVA-Video is in fig. 20, LongVA is in fig. 21, and VideoLLaMA3 is in fig. 22. Stronger models, such as LLaVA-Video and VideoLLaMA3, tend to assign common words with a positive attribution, though sometimes "video" is seen as negative. Otherwise, we see dataset specific objects, actions, and adjectives get high attribution values. For example, "cylinder" and "cube" in MVBench and "left", "right", and "counter" in HD-EPIC. Apart from VideoLLaMA2, most words tend to have a positive attribution score, again showcasing the preference for the question and answer modalities over the video modality. In general, we fail to see any obvious biases towards specific types of words (verbs or nouns).



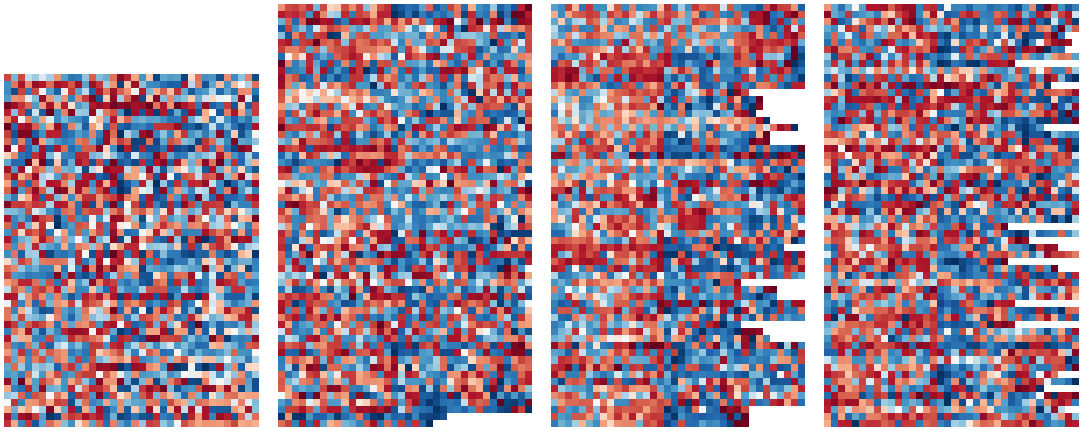
(a) EgoSchema (b) HD-EPIC (c) MVBench (d) LVBench

Figure 8: Matrix of Shapley values per subset for FrozenBiLM, where each row, left-to-right, represents the features of a VQA-tuple. Rows are truncated to a maximum of 30 features with the first 10 values representing video frames.



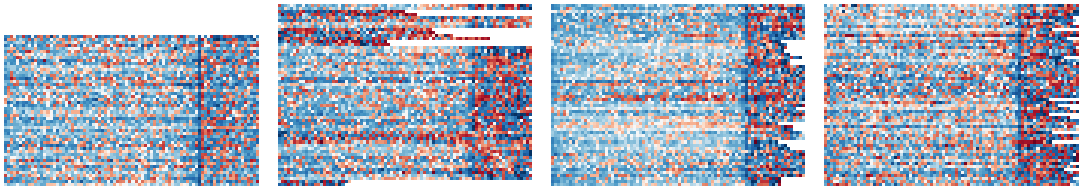
(a) EgoSchema (b) HD-EPIC (c) MVBench (d) LVBench

Figure 9: Matrix of Shapley values per subset for InternVideo, where each row, left-to-right, represents the features of a VQA-tuple. Rows are truncated to a maximum of 30 features with the first 10 values representing video frames.



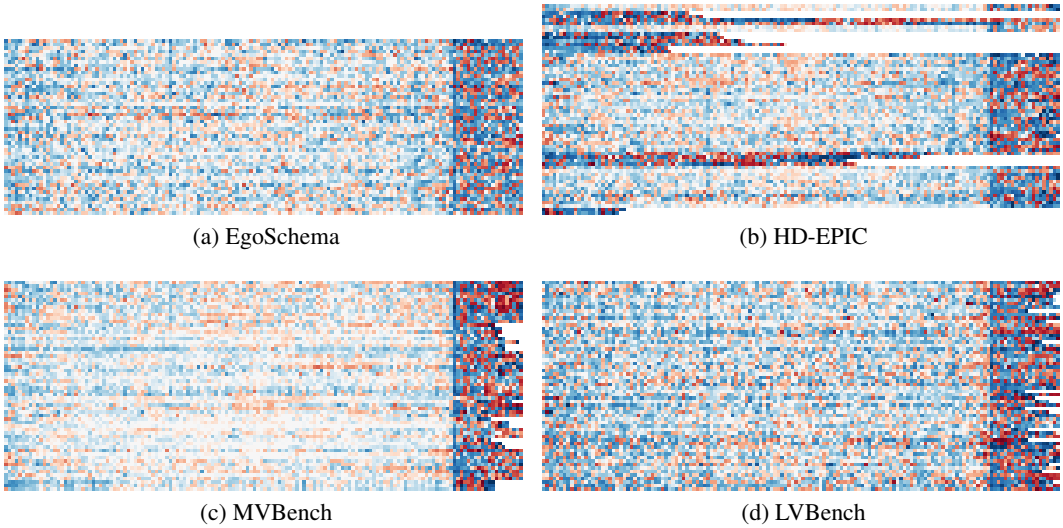
(a) EgoSchema (b) HD-EPIC (c) MVBench (d) LVBench

Figure 10: Matrix of Shapley values per subset for VideoLLaMA2, where each row, left-to-right, represents the features of a VQA-tuple. Rows are truncated to a maximum of 36 features with the first 16 values representing video frames.



(a) EgoSchema (b) HD-EPIC (c) MVBench (d) LVBench

Figure 11: Matrix of Shapley values per subset for LLaVA-Video, where each row, left-to-right, represents the features of a VQA-tuple. Rows are truncated to a maximum of 84 features with the first 64 values representing video frames.



(a) EgoSchema (b) HD-EPIC

(c) MVBench (d) LVBench

Figure 12: Matrix of Shapley values per subset for LongVA, where each row, left-to-right, represents the features of a VQA-tuple. Rows are truncated to a maximum of 148 features with the first 128 values representing video frames.

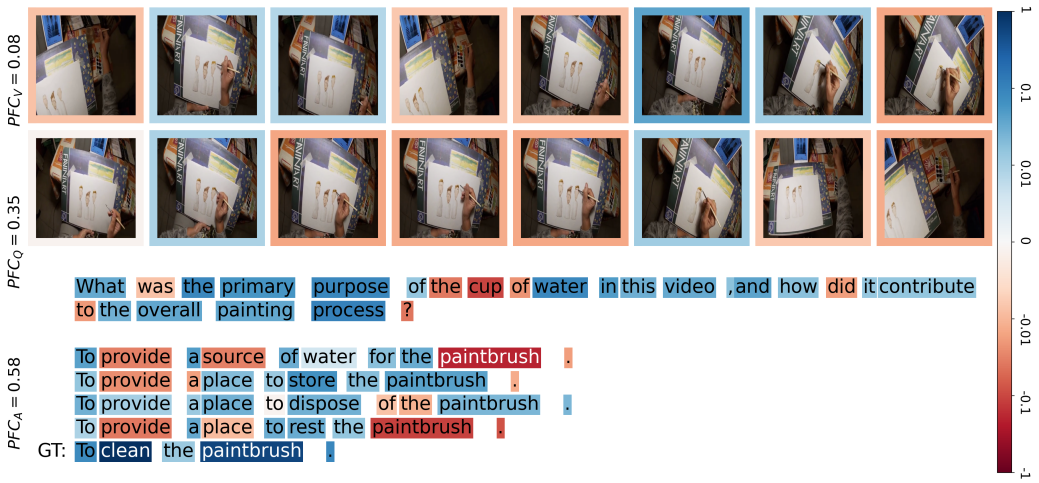



Figure 13: Qualitative figure of an example from EgoSchema evaluated using VideoLLaMA3. For brevity, we select the 16 most important frames, ranked by the magnitude of their Shapley values. Here **blue** represents positively attributed inputs whereas **red** represents negatively attributed inputs.


Figure 14: Qualitative figure of an example from HD-EPIC evaluated using VideoLLaMA3. For brevity, we select the 16 most important frames, ranked by the magnitude of their Shapley values. Here **blue** represents positively attributed inputs whereas **red** represents negatively attributed inputs.

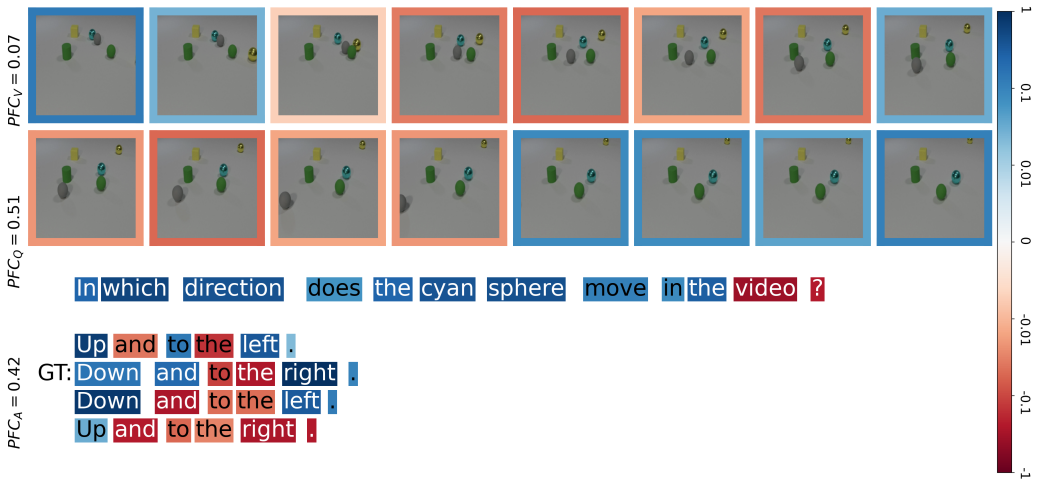



Figure 15: Qualitative figure of an example from MVBench evaluated using VideoLLaMA3. For brevity, we select the 16 most important frames, ranked by the magnitude of their Shapley values. Here **blue** represents positively attributed inputs whereas **red** represents negatively attributed inputs.

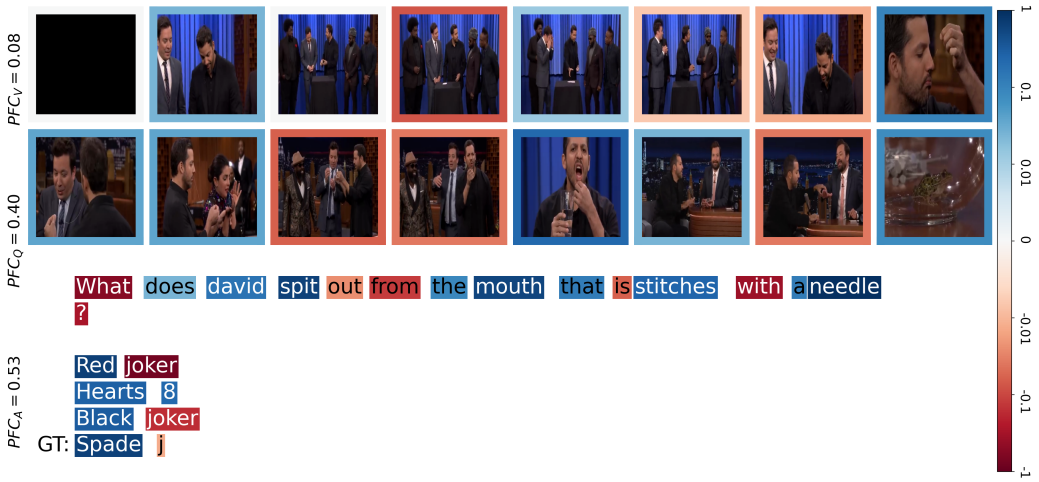
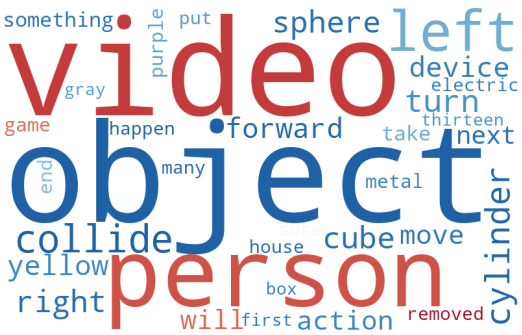



Figure 16: Qualitative figure of an example from LVBench evaluated using VideoLLaMA3. For brevity, we select the 16 most important frames, ranked by the magnitude of their Shapley values. Here **blue** represents positively attributed inputs whereas **red** represents negatively attributed inputs.


(a) EgoSchema


(b) HD-EPIC



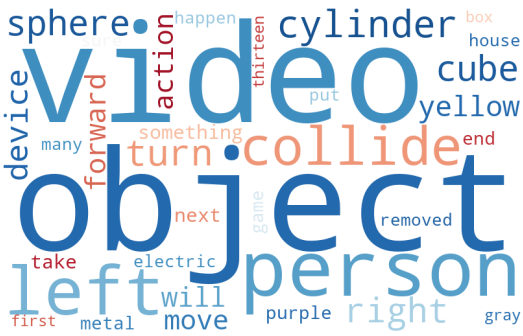

(c) MVBench


(d) LVBench

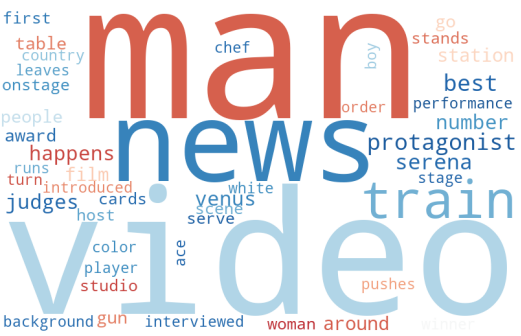
Figure 17: FrozenBiLM word clouds showing the frequency of the words within the dataset subset via their size and the attribution by their colour, **blue** for positive attribution and **red** for negative.


(a) EgoSchema

(b) HD-EPIC

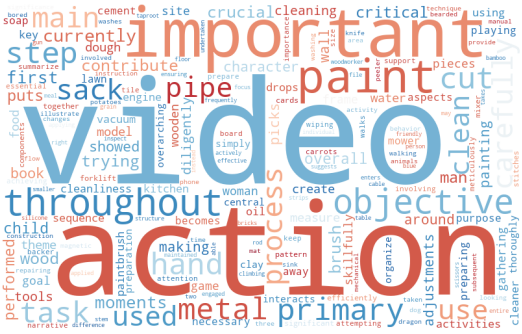



(c) MVBench




(d) LVBench

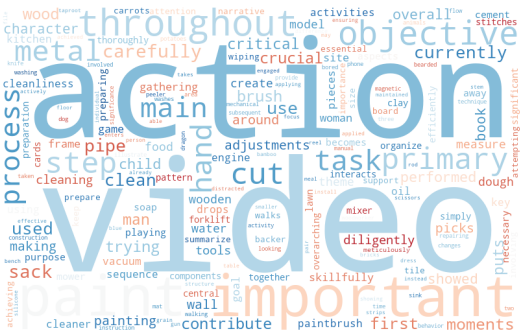
Figure 18: InternVideo word clouds showing the frequency of the words within the dataset subset via their size and the attribution by their colour, **blue** for positive attribution and **red** for negative.




(a) EgoSchema (b) HD-EPIC

(c) MVBench (d) LVBench

Figure 19: VideoLLaMA2 word clouds showing the frequency of the words within the dataset subset via their size and the attribution by their colour, **blue** for positive attribution and **red** for negative.



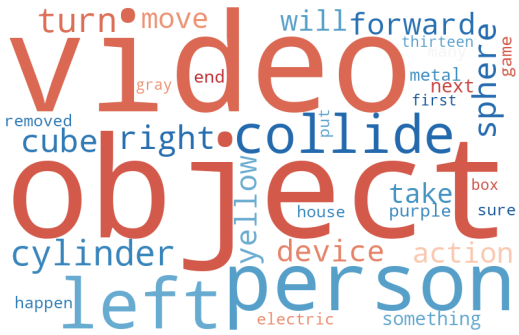

(a) EgoSchema (b) HD-EPIC

(c) MVBench (d) LVBench

Figure 20: LLaVA-Video word clouds showing the frequency of the words within the dataset subset via their size and the attribution by their colour, **blue** for positive attribution and **red** for negative.

(a) EgoSchema

(b) HD-EPIC



(c) MVBench

(d) LVBench

Figure 21: LongVA word clouds showing the frequency of the words within the dataset subset via their size and the attribution by their colour, **blue** for positive attribution and **red** for negative.

(a) EgoSchema

(b) HD-EPIC

(c) MVBench

(d) LVBench

Figure 22: VideoLLaMA3 word clouds showing the frequency of the words within the dataset subset via their size and the attribution by their colour, **blue** for positive attribution and **red** for negative.