

Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models

Jean Park¹, Kuk Jin Jang¹, Basam Alasaly², Sriharsha Mopidevi², Andrew Zolensky¹, Eric Eaton¹, Insup Lee¹, Kevin Johnson^{1,2}

¹Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA

²Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

{hlpark, jangkj, zolensky, eeaton, lee}@seas.upenn.edu,

{basam.alasaly, sriharsha.mopidevi, kevin.johnson1}@penntest.net

Abstract

Multimodal large language models (MLLMs) can simultaneously process visual, textual, and auditory data, capturing insights that complement human analysis. However, existing video question-answering (VidQA) benchmarks and datasets often exhibit a bias toward a single modality, despite the goal of requiring advanced reasoning skills that integrate diverse modalities to answer the queries.

In this work, we introduce the modality importance score (MIS) to identify such bias. It is designed to assess which modality embeds the necessary information to answer the question. Additionally, we propose an innovative method using state-of-the-art MLLMs to estimate the modality importance, which can serve as a proxy for human judgments of modality perception. With this MIS, we demonstrate the presence of unimodal bias and the scarcity of genuinely multimodal questions in existing datasets. We further validate the modality importance score with multiple ablation studies to evaluate the performance of MLLMs on permuted feature sets. Our results indicate that current models do not effectively integrate information due to modality imbalance in existing datasets. Our proposed MLLM-derived MIS can guide the curation of modality-balanced datasets that advance multimodal learning and enhance MLLMs' capabilities to understand and utilize synergistic relations across modalities.

1 Introduction

In recent years, trends in AI development have leaned towards multimodal models, particularly multimodal large language models (MLLMs), as many complex problems necessitate the integration of diverse modalities to achieve more accurate and comprehensive reasoning. Video question answering (VidQA) stands out as a particularly challenging task, requiring the integration of various modalities along with complex spatial and temporal reasoning (Xiao et al. 2021). As such, this task serves as a vital benchmark for assessing the vision-language understanding capabilities of AI systems.

In recent years, several VidQA benchmarks have been developed to train and evaluate the capabilities of MLLMs in these areas (Yu et al. 2019; Gupta et al. 2022). However, a fundamental question remains: Are these models genuinely

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Video: Lady in the floral top and jean jacket is bleeding from her side

0:00 0:09 0:91

Subtitle:

00:00:02,257 --> 00:00:04,384 (Thirteen:)What the hell happened? We got to get you to a hospital.

00:00:06,461 --> 00:00:07,792 (Thirteen:)It's more complicated than that. We need to...

00:00:04,459 --> 00:00:06,393 No, no, you're a doctor, just stitch me up.

00:00:07,896 --> 00:00:09,830 (Darrien:)The cops will be waiting for me at the hospital.

"Stitch me up" implies she is hurt and bleeding

Lady doesn't want to go to hospitals because she wants to avoid cops

(a) Video and subtitle (TVQA)

Modality-Agnostic Correct

Q1 : Why is 13 worried when she is talking to the lady in the floral top and jean jacket

- (a) 13 is worried because the lady is going to tell about 13's illness
- (b) 13 is worried because the lady is having severe headaches
- (c) **13 is worried because the lady is bleeding from her side**
- (d) 13 is worried because the lady became unconscious
- (e) 13 is worried because the lady won't stop crying

Complementary

Q2 : Why is 13 worried?

- (a) Because lady in the **jean jacket** needed help and wanted to go to the hospital.
- (b) Because lady in the grey cotton shirt needed help **but did not want to go the hospital.**
- (c) **Because lady in the jean jacket needed help but did not want to go the hospital.**
- (d) Because lady in the grey cotton shirt needed help and wanted to go the hospital.
- (e) Because lady in the grey cotton shirt **wanted to avoid cops.**

(b) Example questions

Figure 1: Example of a video clip with multimodal questions demonstrating different modality importance. Q_1 is answerable using either subtitle or video information, while Q_2 requires integrating information from both modalities. (Sec. 3.2)

integrating information from various sources, or are they simply leveraging biases inherent in the datasets? Our observations suggest that many existing benchmarks are limited in their ability to assess this integration. The questions

often tend to be biased toward a single modality, or *modality bias*, lacking the complexity that would require genuine multimodal integration. For instance, the video question Q_1 depicted in Fig. 1b can be answered using only the video alone or the subtitles alone. Although having redundant information across modalities may be beneficial for learning cross-modal relationship, it doesn't fully represent the complexity of real-world multimodal reasoning tasks. As illustrated in Q_2 from Fig. 1b, some multimodal questions require integrating distinct pieces of information from the text (not wanting to go to the hospital) and from the video (material of clothing) to accurately deduce the answer. Unfortunately, such questions that demand genuine integration of multiple modalities are notably scarce in current datasets.

To address these limitations, we need a method that quantitatively assesses modality bias in questions. To this end, we introduce a novel **modality importance score (MIS)**, which evaluates the extent to which each modality contributes to answering a given question. Using this score, we perform a comprehensive assessment of modality bias in existing VidQA benchmarks. Our analysis reveals significant limitations in current datasets and highlights the need for more balanced and challenging multimodal questions.

Our main contributions are as follows:

- We propose a novel modality importance score (MIS) and a method that leverages multimodal large language models (MLLMs) to estimate the MIS. We show that this approach could serve as a proxy for human judgements of modality perception.
- Using the proposed modality importance score, we demonstrate the existence of a unimodal bias and the scarcity of truly multimodal questions in current multimodal datasets.
- We evaluate several state-of-the-art multimodal models on questions with permuted features for modalities with low importance scores. The results reveal that current multimodal models do not optimally combine information from different sources due to modality imbalance in existing multimodal datasets.

By addressing these limitations in VidQA benchmarks, our work aims to advance the field of multimodal AI, pushing towards models that can genuinely integrate information across modalities to perform complex reasoning tasks more effectively.

2 Related Work

2.1 Video Question Answering

Video question answering (VidQA) is a well-explored field in AI, presenting the challenge of integrating multimodal input from videos, understanding temporal and causal relations, and selecting the correct answer (Lei et al. 2019). Many recent VidQA models are pretrained on large datasets using contrastive learning objectives (Kim et al. 2021), masked language modeling (Fu et al. 2021), and other techniques to learn joint representations and improve spatio-temporal understanding (Zhao et al. 2017; Jiang et al. 2020). These models are subsequently fine-tuned on downstream

tasks, such as open-ended or multiple choice video-question answering (Wang et al. 2023), video-text retrieval (Luo et al. 2020), and video captioning (Fu et al. 2023).

In this study, we focus on four approaches that have been developed to utilize both subtitle and video information for video question answering. **Merlot Reserve** (Zellers et al. 2022) is pretrained to predict either the correct text or audio snippet hidden by a MASK token, given uniformly sampled images from a video. Its architecture includes pretrained encoders for each modality input and a joint encoder trained with a contrastive spanning objective. **FrozenBiLM** (Yang et al. 2022a) employs a frozen bidirectional language model trained on web-scale multimodal data. **Llama-VQA** (Ko et al. 2023) builds upon the Llama model, incorporating additional learnable parameters through the Flipped-VQA framework. This approach leverages the LLM's prior knowledge of temporal and causal reasoning. **MiniGPT4-Video** (Ataallah et al. 2024) is an open-source multimodal large language model designed for video-language tasks. Its training process involves pretraining using either Llama2 or Mistral on video-text pairs consisting of frame sequences and subtitles appended to a pre-defined prompt. In addition, other VidQA approaches utilize captions and videos, such as VindLu or MMFT-BERT, and MSAN (Cheng et al. 2023; Khan et al. 2020; Kim et al. 2020). Additional tasks and approaches outside the scope of this study can be found in a survey by Zhong et al. (2022).

While these models show improved performance by integrating language and video inputs for video understanding, a critical issue remains: they are trained on datasets that have questions with modality bias. This bias raises the question of whether these models can leverage both modalities for each question and context and whether they are biased in their ability to leverage either modality as appropriate. Our research examines whether current models can effectively identify and use the most relevant modality, even with irrelevant information. Our findings reveal limitations in their ability to perform this task optimally.

2.2 VidQA Datasets and Benchmarks

Several notable datasets and benchmarks have been proposed for multiple-choice VidQA approaches.

TVQA The TVQA dataset (Lei et al. 2018) comprises over 150K question-answer pairs derived from 21,793 clips across six TV shows. These clips average 76 seconds, with each question providing a localized timestamp indicating where the answer can be found within the clip.

In TVQA's test-public set, human accuracy varied across different modality combinations: 61.96% for video-only, 73.03% for subtitles-only, and 89.41% for both. While the authors interpret this result as evidence for the necessity of both visual and textual understanding, we propose an alternative perspective. We hypothesize that many questions in the dataset contain redundant information across both video and subtitle sources rather than requiring the integration of information from distinct sources. Furthermore, we believe this result insufficiently captures how questions depend on different modalities or their combinations.

LifeQA The LifeQA dataset (Castro et al. 2020) comprises 2.3K questions derived from 275 real-life YouTube videos. These videos were recorded by individuals in uncontrolled environments, capturing meaningful visual and linguistic interactions. The human performance on this dataset varied significantly: when given only video, participants achieved 48.5% accuracy; with audio alone, accuracy rose to 63.4%; and with all modalities combined, accuracy peaked at 90.6%. Interestingly, these results contradict the authors’ Venn diagram (Castro et al. (2020), Fig. 3) categorization of LifeQA questions by answer type. Their categorization suggests that over 60% of questions are visual-based, while only 29% are speech-based, with the remaining questions (10%) requiring both modalities. This distribution seems at odds with the observed human performance across different modality combinations. We argue this discrepancy suggests that the authors’ categorization of answer types may have been based on the perceived nature of the question rather than actual modality dependency. This method may be less accurate, as some questions labeled as “Visual” like “Where are they located?” might also be answered based on dialogue or background sounds.

AVQA The AVQA (Audio-Visual Question Answering) dataset (Yang et al. 2022b), derived from the VGG Sound dataset (Chen et al. 2020), contains over 57K question-answer pairs derived from 57K real-life videos focusing on object-generated sounds rather than human speech. It was designed to require information from both audio and visual modalities for most questions, to ensure that relying on just one modality would be insufficient or ambiguous for an accurate answer. However, the annotators who designed the questions also categorized the question types. Similar to LifeQA, this approach could introduce bias, as annotators might focus on the perceived modality requirements rather than objectively assessing whether relevant information is present in each modality.

2.3 Modality Contribution in Multimodal Tasks

The concept of quantifying modality contributions in multimodal tasks was explored in perceptual score paper (Gat, Schwartz, and Schwing 2021). They introduced a “perceptual score” to measure a model’s reliance on specific input modalities or subsets. Their method involved removing the influence of a modality M from the set of all modalities and measuring the resulting change in accuracy.

Others, such as Yang et al. (2024), revealed that multimodal models often prefer certain modalities, leading to less robust performance when a modality is missing or perturbed. Their research showed that models tend to rely on one specific modality even when trained on multiple modalities, demonstrating vulnerability to unimodal attacks. To address this issue, they introduced Certifiable Robust Multi-modal Training, a method designed to mitigate the influence of the model’s modality preference and regulate essential components to improve its robustness.

While such works aim to analyze models’ bias towards specific modalities and suggest solutions for reliable and robust performance, our work focuses on quantifying the

modality contribution in the dataset, specifically in multiple-choice VidQA datasets. We identify modality bias in these datasets and provide a more fine-grained categorization of question types. This approach aims to guide the development of more balanced datasets, a crucial first step toward enabling multimodal models to utilize modalities effectively.

3 Method

3.1 Modality Importance Score

Intuition. Understanding the contribution of each modality is crucial in multi-modal question-answering tasks. Our goal is to distinguish between questions answerable by a single modality, those with redundant signals from multiple modalities, and those requiring integration of modalities.

Consider the scenario in Figure 1, with two input modalities: video and subtitles (audio in the form of text). Three input combinations are possible: video alone, subtitle alone, and video + subtitle. The importance of a modality, such as video, can be quantified by estimating the increase in accuracy when video is present in the input combination (video, video+subtitle) relative to when it is not (subtitle).

In Figure 1b, the question Q_1 is an example where accuracy does not increase when the video is added. From the phrase, “stitch me up” in the subtitle, one can reasonably infer that the lady is likely bleeding. The video confirms this fact by displaying a bleeding lady, but adds redundant signals rather than providing essential new details. In contrast, question Q_2 , exemplifies a multimodal question that cannot be answered correctly with a single modality. When considering only the video, two answer choices, (a) and (c), become confusing as both mention the correct visual detail “lady in the jean jacket”. Similarly, with only subtitles, three plausible answer choices are given (b), (c), and (e). The question requires integrating information from both modalities for an accurate response. We formalize this intuition by defining the modality importance score.

Definition. Given an input question q_i , its corresponding ground truth label y_i , and a set of source modalities $M = \{m_1, m_2, \dots, m_k\}$, we denote combinations of modalities in M as the power set of M excluding the \emptyset , $\mathcal{P}(M) \setminus \emptyset$. We first define the performance measurement function as:

$$perf(q_i | M') = \frac{\sum_{S \subseteq M'} \mathbb{1}[A_S^i = y_i]}{|M'|}, \quad (1)$$

where M' is a subset of modalities defined as $M' \subseteq \mathcal{P}(M) \setminus \emptyset$, and $|M'|$ is the cardinality. $\mathbb{1}[A_S^i = y_i]$ is the response accuracy function we use to measure the performance in VidQA tasks defined as,

$$\mathbb{1}[A_S^i = y_i] = \begin{cases} 1 & \text{if } A_S^i = y_i \\ 0 & \text{if } A_S^i \neq y_i \end{cases}. \quad (2)$$

This is an indicator function that returns 1 if the answer for question q_i , A_S^i , obtained using a subset of modalities S matches the ground truth, and 0 otherwise. While our current performance measurement function $perf(q_i | M')$ considers only response accuracy, it can be generalized to incorporate

other performance metrics. Finally, the **Modality Importance Score (MIS)** for a single modality m_j and question q_i , is defined as:

$$\text{MIS}_{m_j}^i = \text{perf}(q_i | M_j^+) - \text{perf}(q_i | M_j^-), \quad (3)$$

where $M_j^+ = \{S \subseteq \mathcal{P}(M) \setminus \{\emptyset, \{m_j\}\} : m_j \in S\}$ are all the non-empty subsets of modalities that must include m_j excluding the singleton set containing only m_j and $M_j^- = \{S' \subseteq \mathcal{P}(M) \setminus \{\emptyset, \{m_j\}\} : m_j \notin S'\}$ are all non-empty subsets of modalities that exclude m_j .

This formulation captures two key aspects of modality importance. The $\text{perf}(q_i | M_j^+)$ calculates the average accuracy across all subsets of modalities that include m_j and at least one other element from the set of modalities in M , capturing how well m_j contributes in combination with other modalities. The $\text{perf}(q_i | M_j^-)$ computes the average accuracy across all subsets that exclude m_j . The difference measures the overall impact of including m_j versus excluding it. Note that our intention is to compute the modality importance score for a single modality m_j and not a set of multiple modalities; however, it is trivial to expand the definitions of M^+ and M^- to include or exclude combinations of multiple modalities.

Response Accuracy			MIS _{Vid}	MIS _{Sub}
Vid	Sub	Vid + Sub		
0	0	0	0	0
0	1	0	-1	0
1	0	0	0	-1
1	1	0	-1	-1
0	0	1	1	1
0	1	1	0	1
1	0	1	1	0
1	1	1	0	0

Table 1: Modality Importance Score for Two Individual Modalities : Video (Vid), Subtitle (Sub)

Table 1 illustrates modality importance scores for response accuracies of three modality combinations. The scores can be interpreted as follows: Positive MIS indicate that the modality embeds a signal contributing to the answer beyond other modalities. Negative MIS suggest that the modality adds interference of conflicting information, potentially masking another modality’s contribution. An MIS of 0 implies that the modality doesn’t contribute additional information beyond other modalities.

Note that the MIS reflects a modality’s relative contribution compared to others, not its absolute ability to answer a question. For instance, if the subtitle alone can answer a question, the video’s MIS may be 0, indicating no additional contribution, and vice versa. In such cases, the modality subset with both modalities might have MIS of 0, reflecting their redundancy rather than their inability to answer the question.

MLLM-derived Modality Importance Score To estimate the modality importance for questions in dataset D , we

can leverage the capabilities of MLLMs for scalability purposes. This approach is applicable to datasets with $|M| \geq 2$ modalities.

For each combination, we prompt the MLLM to select the most plausible answer choice given the provided input combination. We compare the model’s response accuracy across different input combinations and quantify the relative importance of each modality according to (3).

This approach provides insights into the distribution of critical information across modalities in multimodal question-answering tasks. Previous approaches (Gat, Schwartz, and Schwing 2021), used random permutation to simulate the removal of a modality’s influence due to the complexity of altering trained models. Our approach does not require permutation as MLLMs allow for more direct manipulation of input modalities. Although our MIS metric can quantify each individual modality’s contribution when more than two modalities are present, current MLLMs typically support only images and text. Hence, for this study, we compute modality importance providing three distinct input combinations to the MLLM: subtitle only, video only, and both modalities together.

3.2 Categorizing Question Types with MIS

Unimodal-bias questions Using the MIS, we can identify unimodal-biased questions. If $\text{MIS}_{m_k}^i \leq 0 \leq \text{MIS}_{m_j}^i$, $\text{MIS}_{m_k}^i \neq \text{MIS}_{m_j}^i \forall m_k \in M$ where $m_k \neq m_j$, we classify question q_i as m_j -biased. Such questions can be answered using only m_j , but cannot be answered correctly using any other single modality m_k . For instance, with video and subtitle modalities, video-biased questions can manifest in two ways. First, correct answers might be obtained whenever the video modality is included, but using only subtitles leads to incorrect answers due to their irrelevance. Alternatively, the video alone might yield correct answers, but combining video and subtitles could result in incorrect answers. In this latter case, the MIS for subtitles becomes negative, indicating interference.

Modality-agnostic vs Complementary questions In addition to identifying unimodal-biased questions, we use MIS to provide a more fine-grained categorization of questions. This categorization helps our understanding of multimodal questions and the relationships between different modalities in answering them.

Modality-agnostic Question As shown in Table 1 rows 1 and 8, there are cases where the same MIS is obtained regardless of which the subset of modalities, correct or incorrect. We define these as **modality-agnostic questions**, where $\forall m_j \in M, \text{MIS}_{m_j}^i = 0$. We further divide modality-agnostic questions into two subcategories:

- Modality-agnostic correct questions:
 $\forall S \subseteq \mathcal{P}(M), \mathbb{1}[A_S^i = y_i] = 1$
- Modality-agnostic incorrect questions:
 $\forall S \subseteq \mathcal{P}(M), \mathbb{1}[A_S^i = y_i] = 0$

Complementary Questions As illustrated in row 5 of Table 1, there exist questions where no single modality can

strongly determine the answer and signals from multiple modalities can be combined to determine the correct answer. We define these questions as **complementary questions**, where $\forall m_j \in M, MIS_{m_j} > 0$. In this case, all modalities contribute to answering the question correctly when combined with other modalities.

Note that in the case of only two modalities, complementary questions cannot be answered correctly unless both modalities are utilized. For scenarios with more than two modalities, complementary questions may involve varying contributions from each modality.

4 Evaluation

4.1 Experimental Setup and Overview

Estimating modality importance score For our experiments, we utilized GPT-4 Turbo (OpenAI et al. 2024), one of the top-performing MLLMs that supports both image and text inputs. We prompted the model to select the correct answer by providing the question, answer choices, and the corresponding modality combination under evaluation. Specific constraints and image extraction were applied to account for GPT-4 Turbo’s token limitations and allow longer video clips. Detailed information about our prompts and process can be found in Appendix (Park et al. 2024).

Datasets We evaluated three VidQA datasets, each containing both video and subtitle/audio components. For TVQA (Lei et al. 2018) and LifeQA (Castro et al. 2020), we use transcripts/subtitles provided by the dataset. AVQA (Yang et al. 2022b) does not provide transcripts, but we use the audio labels from VGG Sound dataset (Chen et al. 2020) as the subtitle. Due to the large number of questions, we limited evaluation to the validation or test sets. For TVQA and AVQA, we uniformly sampled 1,019 and 796 questions, respectively, representing approximately 6-10% of the total questions. For LifeQA, we evaluated the entire test set of 372 questions.

VidQA Models Our study evaluates four multimodal VidQA models, listed in Table 3, capable of processing both visual and textual (audio captions or subtitle) inputs to answer multiple-choice questions. We use the MLLM-derived MIS to identify unimodal-biased questions. Our feature permutation experiments show how effectively these models integrate and utilize information across different modalities.

4.2 Human Study Validation of MLLM-derived Modality Importance

To assess human perception of modality importance, we employed a split-group methodology involving four participants, each evaluating 197 TVQA questions. The detailed methodology is in Appendix A.3 (Park et al. 2024), along with Figure 5 depicting the study and Table 4 showing accuracy distributions across confidence levels. Our study aimed to validate the alignment between MLLM-derived MIS and human perception of modality importance. The evaluation yielded a substantial inter-annotator agreement (Fleiss’ kappa: 0.76) for questions answered with both modalities, with an average accuracy of 87.8%.

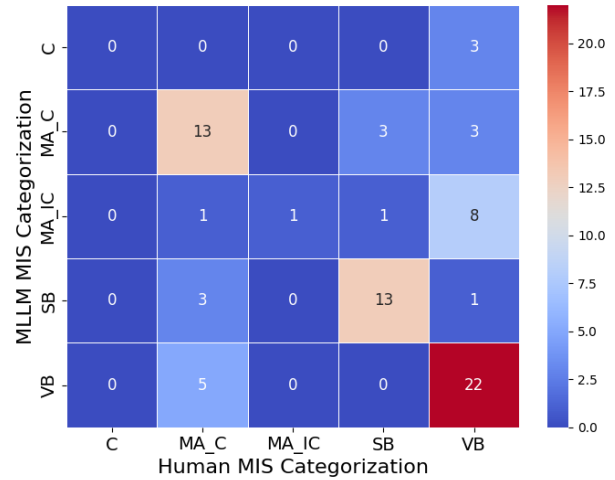


Figure 2: Question categorization based on human study vs MLLM-derived MIS

For our analysis, we focused on questions that showed unanimous agreement per modality. For these questions annotators were either all correct or all incorrect. As shown in Fig. 2, this method revealed a moderate alignment between human perception-based and MLLM-derived categorizations for three types of questions: modality-agnostic correct, subtitle-biased, and video-biased. This suggests that when human annotators are clearly in agreement, their judgments often align with the MLLM’s assessments.

Under this categorization based on human scores, we were unable to identify any complementary questions from the evaluated subset of questions. This observation suggests that questions whose answer relies on information from both modalities might indeed be scarce in the multimodal VidQA dataset. This finding highlights a potential limitation in current multimodal datasets.

4.3 Evaluation of Modality Bias in VidQA Datasets

In this section, we analyze the distribution of question types based on MLLM-derived MIS.

TVQA The results, reported in Table 2, support our assumption that many questions in TVQA would be modality-agnostic correct. About 35% of the questions were classified as modality-agnostic correct, while only 2% were identified as complementary, requiring information from both modalities. We had 7% of questions that were modality-agnostic incorrect. As shown in Figure 2, GPT has limited visual understanding compared to humans, as 8 out of 11 modality-agnostic incorrect questions were actually video-biased. While the subtitle does not provide relevant information for these questions, GPT fails to extract or comprehend details from the sequence of images. Consequently, the model consistently incorrect regardless of input modality. Overall, the results show a potential discrepancy between the dataset’s intended multimodal nature and the actual distribution of question types.

Question Type	TVQA	LifeQA	AVQA
SB	224 (22.0%)	74 (19.9%)	39 (4.9%)
VB	345 (33.9%)	125 (33.6%)	93 (11.7%)
C	21 (2.1%)	9 (2.4%)	5 (0.6%)
MA _C	357 (35.1%)	135 (36.3%)	625 (78.5%)
MA _{IC}	71 (7.0%)	29 (7.8%)	32 (4.0%)
None	1 (0.1%)	0 (0.0%)	4 (0.5%)
Total	1,019	372	796

Table 2: Distribution of Question Types based on MIS Across Different Datasets : Subtitle-biased (SB), Video-biased (VB), Complementary (C), Modality-agnostic Correct (MA_C), Modality-agnostic Incorrect (MA_{IC})

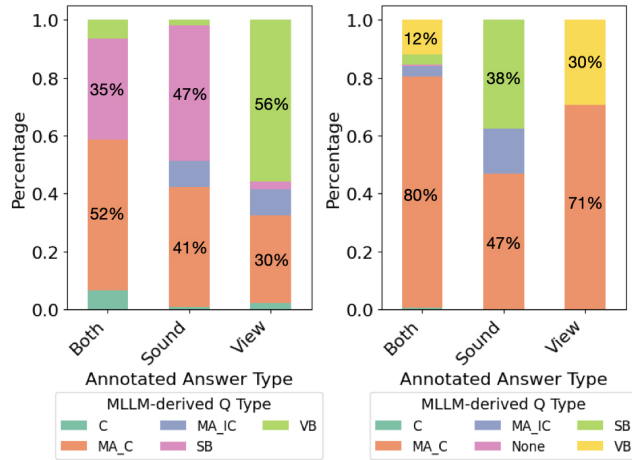


Figure 3: Proportion of MIS based Question types per Annotated Answer Type (Left: LifeQA, Right: AVQA)

LifeQA The distribution of question types based on our MIS categorization shown in Table 2 revealed that modality-agnostic correct questions formed the largest category, accounting for approximately 36% of the dataset. Video-biased questions followed closely, comprising 33% of the dataset, and subtitle-biased questions accounted for 19.9%. Less than 10% of questions were modality-agnostic incorrect for “Sound” and “View” types. For “View” types, we found out that GPT-4’s had limitations in identifying image details. For “Sound” types, errors were primarily due to insufficient information in the provided automated captions. The low percentage of complementary questions (2%) indicates that most questions in the LifeQA dataset can be answered using a single modality or are modality-agnostic.

The left side of Figure 3 shows the relationship between our MIS-based question categories and the annotated answer types. For “Sound” answer types, 46.8% were classified as subtitle-biased, aligning with the annotated type. However, a significant 41% were categorized as modality-agnostic. This suggests that many questions annotated as language-dependent can actually be answered with

all modalities. Similarly, for “View” answer type questions, while the majority were video-biased, a significant number were modality-agnostic correct. These observations indicate that our categorization generally aligns with human-annotated answer types. Moreover, the significant proportion of modality-agnostic correct questions in both “Sound” and “View” answer types suggests that many questions may not be single modality-dependent. See Appendix A.4 (Park et al. 2024) for examples.

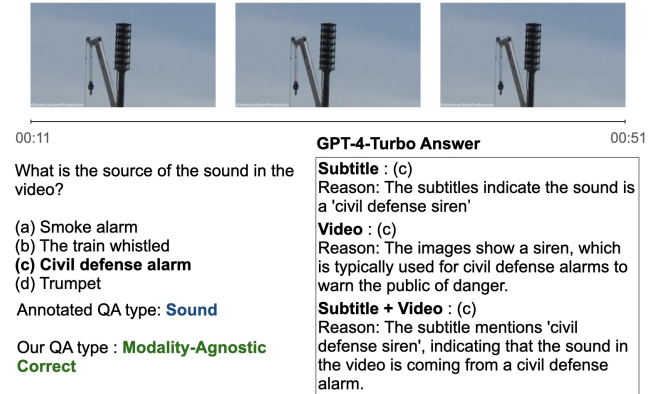


Figure 4: Example from AVQA where annotated answer type is different from our categorization. For this video, the subtitle is “civil defense siren”.

AVQA Table 2 depicts our analysis of AVQA. Our analysis found that the distribution of question types contradicts the dataset’s original design intention of requiring both modalities to answer accurately. 78.5% of 796 questions were modality-agnostic correct questions. This implies that many questions in this dataset are answerable using any single modality, as shown in Figure 4. Only a small fraction of questions, approximately 0.6%, were complementary. Additional examples can be found in Appendix A.4 (Park et al. 2024).

The right side of Figure 3 shows patterns similar to LifeQA. Based on our categorization, questions annotated with the “Sound” answer consisted of 37.5% subtitle-biased questions and no video-biased questions. Similarly, the “Video” answer type questions showed a high number of video-biased questions (29.4%) and no subtitle-biased questions.

Summary In summary, our study demonstrates that the MLLM-derived MIS and question categorization align well with human perception of modality relevance. This approach shows that many seemingly single-modality questions are modality-agnostic correct, indicating the presence of redundant signals across modalities. Although our sampling method prevented us from definitively proving dataset-wide unimodal bias, our approach shows significant potential in identifying such biases and highlighting the scarcity of truly multimodal questions requiring sophisticated information integration from multiple modalities.

	Subtitle-biased			Video-biased		
	Orig.	SP (Δ)	VP (Δ)	Orig.	SP (Δ)	VP (Δ)
Merlot R*	91.5 \pm 0.0	32.2 \pm 3.8 (-59.3)	87.4 \pm 1.9 (-4.1)	71.9 \pm 0.0	72.0 \pm 1.5 (+0.1)	43.2 \pm 5.0 (-28.7)
FrozenBiLM	95.5 \pm 0.0	31.3 \pm 4.3 (-64.2)	96.3 \pm 0.3 (+0.8)	75.4 \pm 0.0	73.4 \pm 2.7 (-1.9)	41.5 \pm 4.4 (-33.9)
Llama-VQA	95.1 \pm 0.0	37.3 \pm 1.8 (-57.8)	94.3 \pm 0.0 (-0.8)	56.9 \pm 0.0	56.1 \pm 0.3 (-0.8)	47.5 \pm 1.5 (-9.4)
MiniGPT4*	61.4 \pm 0.2	35.9 \pm 3.6 (-25.5)	58.7 \pm 3.5 (-2.8)	42.4 \pm 0.8	40.9 \pm 2.0 (-1.5)	38.6 \pm 3.2 (-3.9)
Average	85.9 \pm 0.0	34.2 \pm 3.4 (-51.7)	84.2 \pm 1.5 (-1.7)	61.6 \pm 0.2	60.6 \pm 1.6 (-1.0)	42.7 \pm 3.0 (-19.0)

Table 3: Accuracy (%) comparison after feature permutation with five random seeds (Orig: Original, SP: Subtitle permuted, VP: Video permuted, Merlot R*: Merlot Reserve, MiniGPT4* : MiniGPT4-Video). All models except for MiniGPT4-Video were fine-tuned on TVQA dataset.

4.4 Multimodal Model Evaluation

Using the MIS, we partition the TVQA questions into those that exhibit bias towards subtitles or video content to assess the multimodal capability of models. We perform feature permutation experiments to evaluate how well the models focus on information relevant to each question type.

The results presented in Table 3 demonstrate the effectiveness of MIS in capturing unimodal bias across different models. We observe that permuting features with low MIS leads to a significantly smaller decrease in accuracy than permuting features with high MIS. For instance, with the subtitle-biased question, “*Why did Marshall think they should have their marriage waiting period waived?*” First, we permute the less important video features by providing the correct subtitle with the wrong images from a different TV show. Then, we permuted the more important feature by providing the wrong subtitle with the correct images. If our MIS effectively categorizes the questions, we would expect the model to perform well in the former case but fail in the latter. This expectation aligned with our results, as the average decrease in accuracy between low-MIS and high-MIS feature permutations was 34%, considering both subtitle and video-biased questions.

Our evaluation reveals several key insights. First, the significant decrease in accuracy between low and high-importance feature permutations confirms that our modality importance score effectively identifies unimodal-biased questions. Second, models generally show degraded performance on video-biased questions than subtitle-biased ones. This difference suggests a limitation in understanding visually relevant features across the evaluated models. This may be due to the prevalence of subtitle-biased and modality-agnostic questions in the original TVQA datasets. Although we were unable to determine the total number of unimodal-biased questions in the TVQA dataset, we can infer from human performance on the TVQA test set. In the original TVQA results, human accuracy with subtitles exceeded that with video by 11%, encompassing both subtitle-biased and modality-agnostic questions. Consequently, we hypothesize that models were trained to focus more on subtitles than video. This is also supported by our observation that permuting video in video-biased questions resulted in a lower accuracy decrease than permuting subtitles in subtitle-biased questions. Our additional experiments in Ap-

pendix A.6 (Park et al. 2024) further validate this hypothesis, showing that even when both modalities contain informative signals, models predominantly rely on textual information. Lastly, when we permuted features with low importance scores, all models showed decreased accuracy except FrozenBiLM with the subtitle modality and Merlot Reserve with video modality. This observation indicates that most models struggle to optimally combine information from different modalities, even when one modality is deemed less important for a given question. These findings highlight the challenges in multimodal learning and the need for improved strategies in integrating information across modalities.

5 Discussion and Limitations

The main limitation of our study is the use of a single MLLM, though a small-scale verification with multiple MLLMs in Appendix A.7 (Park et al. 2024) supports our claim that most questions are modality-agnostic, with few complementary. Our approach is also constrained by the MLLM’s visual processing limitations, likely affecting the categorization of some video-biased questions. Future studies should factor MLLM performance into MIS computation for more robust bias assessment and a more comprehensive evaluation of modality importance in multimodal datasets.

6 Conclusion

Our findings reveal a significant challenge in the field of multimodal AI: current Video Question Answering datasets may not be optimally enabling multimodal reasoning. Our novel method for assessing the relative importance of different modalities, the MLLM-derived MIS, shows that across three VidQA benchmarks, a substantial 89.8% to 94.8% of questions can be answered using a single modality or are modality-agnostic. Only 0.6% to 2% require genuine multimodal integration. Our analysis shows that our MLLM-derived MIS correlates with the human perception of modality importance, suggesting its potential for guiding the scalable curation of more balanced datasets. Based on these findings, we propose two future directions: creating new benchmarks that include complementary questions to properly train and evaluate multimodal reasoning, and developing models with dynamic modality integration mechanisms (Kim et al. 2020) to effectively combine information across modalities.

Acknowledgments

This research was partially supported by the National Institutes of Health (NIH) under award #DP1-LM014558 (PI: Johnson, 09/01/2023-07/31/2028) for the project “Helping Doctors Doctor: Using AI to Automate Documentation and ‘De-Automate’ Health Care,” the National Science Foundation (#NSF-1915398), the Army Research Office MURI (#W911NF-20-1-0080), the Institute for Translational Medicine and Therapeutics, the National Center for Advancing Translational Sciences of the NIH (#UL1TR001878), and by the Collaborative Research in Trustworthy AI for Medicine grant from ASSET at the University of Pennsylvania.

References

- Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhu, D.; Ding, J.; and Elhoseiny, M. 2024. MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens. *arXiv preprint arXiv:2404.03413*.
- Castro, S.; Azab, M.; Stroud, J.; Noujaim, C.; Wang, R.; Deng, J.; and Mihalcea, R. 2020. LifeQA: A Real-life Dataset for Video Question Answering. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4352–4358. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.
- Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.
- Cheng, F.; Wang, X.; Lei, J.; Crandall, D.; Bansal, M.; and Bertasius, G. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10739–10750.
- Fu, T.-J.; Li, L.; Gan, Z.; Lin, K.; Wang, W. Y.; Wang, L.; and Liu, Z. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.
- Fu, T.-J.; Li, L.; Gan, Z.; Lin, K.; Wang, W. Y.; Wang, L.; and Liu, Z. 2023. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22898–22909.
- Gat, I.; Schwartz, I.; and Schwing, A. 2021. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34: 21630–21643.
- Gupta, V.; Patro, B. N.; Parihar, H.; and Namboodiri, V. P. 2022. Vquad: Video question answering diagnostic dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 282–291.
- Jiang, J.; Chen, Z.; Lin, H.; Zhao, X.; and Gao, Y. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11101–11108.
- Khan, A. U.; Mazaheri, A.; Lobo, N. D. V.; and Shah, M. 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*.
- Kim, J.; Ma, M.; Pham, T.; Kim, K.; and Yoo, C. D. 2020. Modality Shifting Attention Network for Multi-Modal Video Question Answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10103–10112.
- Kim, S.; Jeong, S.; Kim, E.; Kang, I.; and Kwak, N. 2021. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13171–13179.
- Ko, D.; Lee, J.; Kang, W.-Y.; Roh, B.; and Kim, H. 2023. Large Language Models are Temporal and Causal Reasoners for Video Question Answering. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4300–4316. Singapore: Association for Computational Linguistics.
- Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.
- Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2019. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*.
- Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.
- OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.;

- Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaf-
tan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar,
N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.;
Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz
Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.;
Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike,
J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.;
Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Mal-
facini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin,
B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.;
McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta,
A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.;
Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.;
Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Nee-
lakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O’Keefe, C.;
Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo,
G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.;
Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov,
M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass,
M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl,
E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond,
C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez,
H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.;
Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Sel-
sam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker,
S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.;
Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher,
N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak,
N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng,
E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Val-
lone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang,
J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.;
Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff,
M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Work-
man, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo,
S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.;
Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and
Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.
- Park, J.; Jang, K. J.; Alasaly, B.; Mopidevi, S.; Zolensky, A.;
Eaton, E.; Lee, I.; and Johnson, K. 2024. Assessing Modal-
ity Bias in Video Question Answering Benchmarks with
Multimodal Large Language Models. arXiv:2408.12763.
- Wang, J.; Ge, Y.; Yan, R.; Ge, Y.; Lin, K. Q.; Tsutsui, S.;
Lin, X.; Cai, G.; Wu, J.; Shan, Y.; et al. 2023. All in one:
Exploring unified video-language pre-training. In *Proceed-
ings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition*, 6598–6608.
- Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-
qa: Next phase of question-answering to explaining tempo-
ral actions. In *Proceedings of the IEEE/CVF conference on
computer vision and pattern recognition*, 9777–9786.
- Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C.
2022a. Zero-Shot Video Question Answering via Frozen
Bidirectional Language Models. In *NeurIPS*.
- Yang, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.;
and Zhu, W. 2022b. Avqa: A dataset for audio-visual ques-
tion answering on videos. In *Proceedings of the 30th ACM
international conference on multimedia*, 3480–3491.
- Yang, Z.; Wei, Y.; Liang, C.; and Hu, D. 2024. Quantify-
ing and Enhancing Multi-modal Robustness with Modality
Preference. arXiv:2402.06244.
- Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao,
D. 2019. Activitynet-qa: A dataset for understanding com-
plex web videos via question answering. In *Proceedings of
the AAAI Conference on Artificial Intelligence*, volume 33,
9127–9134.
- Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusu-
pati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. Mer-
lot reserve: Neural script knowledge through vision and lan-
guage and sound. In *Proceedings of the IEEE/CVF Confer-
ence on Computer Vision and Pattern Recognition*, 16375–
16387.
- Zhao, Z.; Yang, Q.; Cai, D.; He, X.; Zhuang, Y.; Zhao, Z.;
Yang, Q.; Cai, D.; He, X.; and Zhuang, Y. 2017. Video Ques-
tion Answering via Hierarchical Spatio-Temporal Attention
Networks. In *IJCAI*, volume 2, 8.
- Zhong, Y.; Xiao, J.; Ji, W.; Li, Y.; Deng, W.; and Chua, T.-S.
2022. Video question answering: Datasets, algorithms and
challenges. *arXiv preprint arXiv:2203.01225*.