Read, Look or Listen? What's Needed for Solving a Multimodal Dataset

Netta Madvil Yonatan Bitton Roy Schwartz School of Computer Science, The Hebrew University, Jerusalem, Israel {netta.madvil,yonatan.bitton,roy.schwartz1}@mail.huji.ac.il

Abstract

The prevalence of large-scale multimodal datasets presents unique challenges in assessing dataset quality. We propose a two-step method to analyze multimodal datasets, which leverages a small seed of human annotation to map each multimodal instance to the modalities required to process it. Our method sheds light on the importance of different modalities in datasets, as well as the relationship between them. We apply our approach to TVQA, a video question-answering dataset, and discover that most questions can be answered using a single modality, without a substantial bias towards any specific modality. Moreover, we find that more than 70% of the questions are solvable using several different single-modality strategies, e.g., by either looking at the video or listening to the audio, highlighting the limited integration of multiple modalities in TVOA. We leverage our annotation and analyze the MERLOT Reserve model, finding that it struggles with image-based questions compared to text and audio, but also with auditory speaker identification. Based on our observations, we introduce a new test set that necessitates multiple modalities, observing a dramatic drop in model performance. Our methodology provides valuable insights into multimodal datasets and highlights the need for the development of more robust models.

1 Introduction

AI models are highly affected by their training data. As a result, understanding what's inside these datasets is important, both in order to improve the underlying models, and to mitigate their biases [Dodge et al., 2021]. Nonetheless, the scale of modern datasets makes such an analysis challenging. To tackle this task, previous work has primarily focused on understanding dataset characteristics, such as their outliers [Carlini et al., 2019], the learnability of different instances [Swayamdipta et al., 2020, Nam et al., 2022, Siddiqui et al., 2022], and the biases [Luccioni et al., 2023] and mislabels [Talukdar et al., 2021] they contain.

In this work we analyze a specific family of AI datasets—multimodal datasets [Tapaswi et al., 2015, Ye et al., 2017, Lei et al., 2018], which contain information from different modalities, such as text, images, and audio. An important question regarding these datasets is the relative importance of each modality and how it manifests within dataset instances.

We present a two-step method, which maps each instance in multimodal datasets to the subset of modalities required to process it. Our method relies on a small seed annotation step, which is later expanded to the full dataset using classification tools. For example, Fig. 1a shows an instance from TVQA [Lei et al., 2018], a video question-answering dataset, which contains a question regarding the clothes worn by one of the characters. This question could be solved by viewing the image, even without the audio or subtitles. In contrast, hiding the video and showing the other two modalities makes it impossible to solve. Our method allows analyzing multimodal datasets, while gaining

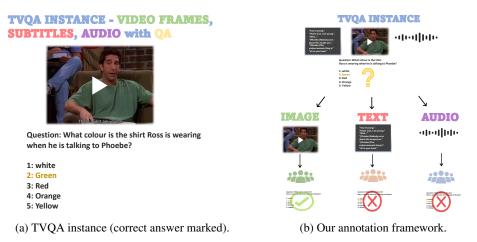


Figure 1: Our annotation framework for identifying the modalities required for processing multimodal instances. (a) a TVQA instance (b) our framework: we break each instance into three groups: audio, text, and image. Separate groups of crowdworkers try to answer the questions in each group. Here the workers are able to solve the question using image alone, but not using text or audio, indicating that it only requires the image.

insights into the underlying relationships and dependencies between the different modalities. It also allows assessing model capabilities on the different modalities.

We apply our approach to TVQA, observing a few interesting findings. First, we validate previous findings [Winterbottom et al., 2020], and show that 99% of the questions can be answered using a single modality. However, unlike that fully-automated work, which found a bias in the data towards the text modality, our method, which relies on human annotation, shows no substantial bias towards any modality. Second, we find that more than 70% of the questions are solvable by two or more modalities separately, and more than 15% using each of the three modalities. We then leverage our analysis to study a given model's performance on the TVQA dataset, by running the MERLOT Reserve model [Zellers et al., 2022] on different instances requiring different modalities. We find that this model generally struggles with image-based questions, but also with questions requiring audio speaker recognition.

Finally, based on our observations, we collect a new test set of 150 questions that cannot be answered using any single modality. We find that MERLOT Reserve performs dramatically worse on these questions compared to the original validation set (41% vs. 83% accuracy), suggesting that it struggles with questions that require multiple modalities. We hope these findings will inspire others to develop methods for training more robust multimodal models.

Contributions. In summary, our main contributions are as follows:

- A novel two-step method for mapping multimodal instances to the required modalities.
- An analysis of modality importance and manifestation within instances of the extended TVQA dataset, providing insights into the characteristics of the dataset.
- Assessment of MERLOT Reserve capabilities and biases on the TVQA dataset, revealing the model's performance with different modalities.
- A new challenging test containing questions that require multiple modalities.

2 Mapping Instances in Multimodal Datasets

Our goal is to map the instances in a given multimodal dataset to the modalities required for processing them. To accomplish this, we present a simple two-step annotation methodology. We first sample a subset of the dataset, and use human annotators to map each instance in it to the subset of modalities required to process it. We then train several classifiers, one per modality sub-group, on the collected annotations. We apply the resulting classifiers to the full dataset, resulting in a mapping of each instance to the a subset of the modalities it requires. We turn to describe both parts.

Collecting small seed annotations. We start by sampling a subset of the data, to be used for our seed annotation. We present human annotators with different subsets of modalities for each instance, recording their responses. We first collect annotations of a single modality and then gradually increase their number (e.g., both image and audio). This allows us to identify which subsets of modalities are sufficient for processing each instance.

More formally, consider a dataset D containing a set of modalities M of size |M|. We sample a subset $D' \subset D$. For an instance $i \in D'$ and modality $m \in M$, we mark the version of i that only contains information from modality m as i_m .¹ We then ask human annotators to label i_m for each instance $i \in D'$, and modality $m \in M$. Naturally, some instances i_m cannot be solved without access to some modalities. As a result, annotators are likely to perform at chance level in these cases, indicating that instance i is insolvable with modality m alone. In order not to contaminate the process, we divide the group of annotators between the different modalities, such that no annotator sees the same instance more than once with different modalities. After considering single modalities, if a sufficient portion of the instances cannot be solved using any single modality, we continue annotating them with pairs of modalities, and if necessary triples, quadruplets, etc. The resulting annotations allow us to analyze and characterize the underlying dataset, by asking questions such as which instances can be solved using individual modalities; which require more than a single modality to process; which can be processed by more than one modality; etc. Aggregating these annotations allows us to map the different regions of a given dataset, and visualize it.

As an example, consider Zellers et al. [2022]'s version of TVQA [Lei et al., 2018], which contains three modalities—audio, text, and image. For each modality m, participants are shown the m part of a scene (without access to the other modalities), and a multi-choice question. Fig. 1 shows Ross from the TV show "Friends" wearing a green shirt while talking to Phoebe. The corresponding question is "What colour is the shirt Ross is wearing when he is talking to Phoebe?", which can be answered using the visual signal, but not without it.

Expanding to the full dataset. We next extend our human annotations to the entire dataset. To do so, we train |M| classifiers, one for each modality m, on the annotations collected above, i.e., predicting whether or not a given instance is solvable using m only. To generate the features for our classifiers, we start by fine-tuning a model on the original training set of the given dataset (including all modalities). Next, we select all subset combinations of modalities ($2^{|M|}$ combinations), and for each one generate a version of the validation set with the given subset masked out. We then apply the trained model on the masked instances and extract the softmax layers' outputs obtained from each modality subset. We concatenate these output vectors to create an input vector of size $2^{|M|} * |L|$, where L is the label space. We then train a random forest [Ho, 1995] classifier for each modality, using these input features, and the annotated labels collected above. We apply the trained classifiers to the full validation set, and obtain a silver annotation.² See Fig. 2 for illustration.

3 A Case-study: TVQA

We apply our method to the TVQA video question answering dataset [Lei et al., 2018]. Below we describe our data collection (Section 3.1); how our results provide insights on both the TVQA dataset (Section 3.2); and a model trained on this dataset (Section 3.3).

3.1 Data Collection

TVQA contains 150K question-answer pairs over 6.5K video clips from six popular TV shows. The questions in the dataset cover a broad range of topics, including object recognition, scene understanding, and story comprehension. The dataset also includes a set of multiple-choice questions, where each question has five possible answer choices. The dataset was originally introduced as a two-modalities dataset (video frames and subtitles), but we consider Zellers et al. [2022]'s version, which includes a third modality, namely speech.

¹E.g., in a movie dataset, i_m could be the sound component of a given movie.

²A similar process can be applied to annotate the training set, for details see Appendix A.4.2.

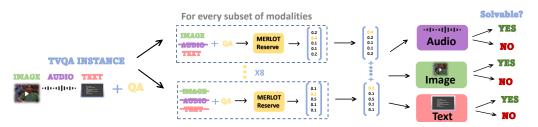


Figure 2: Illustration of our annotation of the full dataset, exemplified TVQA. An instance is mapped into eight $(=2^{|M|})$ different combinations of modalities. These subsets are fed to a model trained on the original TVQA training set, which computes the softmax output probabilities of possible answers. These probability vectors, sorted such that the gold label's probability is first', are then concatenated to form the input vector for each classifier, which is trained to predict the answerability of a given instance using a specific modality.

Humans annotations. We hire native English-speaking workers from Amazon Mechanical Turk to answer questions based on a specific modality of the input. We develop qualification tests to select high-quality annotators and divide them into three separate groups, one per modality. Following Chen et al. [2020] and Castro et al. [2020], we also require participants determine whether they think the question can be answered based solely on specific modality input. However, we find this approach to be less reliable—35% of the questions marked as unsolvable were, in fact, solvable. As a result, we rely on the annotators accuracy in order to determine whether a question is answerable using a subgroup of modalities, but use the yes/no information for monitoring the quality of their annotations.³ To prevent annotators from answering based on memory of already seeing that particular scene, we add a checkbox for them to indicate whether they had seen the scene before, and omit those answers.

We select a set of 650 examples from the validation set of TVQA, all from the TV show "Friends", containing 500 randomly selected examples, and 150 additional instances that are most "sensitive" to the model for each modality (See Appendix A.2). We break them down to 75%/15%/10% for training/validation/test, respectively. We additionally collect 150 examples from "House M.D", to serve as an out-of-distribution (OOD) test set. We show each annotator a single modality, and use five annotators for each (instance,modality) combination. The final label is determined by majority vote. See Appendix A.2 for more information about the annotation metrics and further analysis.

Expansion to the full dataset. For each of the TVQA modalities, we train a different classifier, each based on a frozen finetuned MERLOT Reserve model [Zellers et al., 2022], which achieves a relatively high level of accuracy (83%) on the original TVQA dataset. The classifiers are trained on our human annotated labels, with balanced class weights for each modality. An illustration of our three classifiers as applied to TVQA is presented in Fig. 2. We conduct a hyperparameter search over the basic parameters of Random Forest such as the number of trees in the forest and the maximum depth of each tree. We report test results on the top-performing model on our validation set. See Appendix A.3 for more details.

Classification results. Table 1 shows our results. Test results range between 74-81%, on average 12% higher than a majority baseline. Interestingly, our OOD test results are on par, and sometimes higher than the ID test results, indicating that our annotations generalize to other domains. We also test the effect of the training set size, by training the classifiers with 30%, 50% and 70% of our training data. Our results (Table 7 in Appendix A.3.2) show that training on as few as 30% of our original size yields performance comparable to our full training set. This finding not only streamlines the data collection process, but also suggests that strong results can be achieved without further annotation efforts.

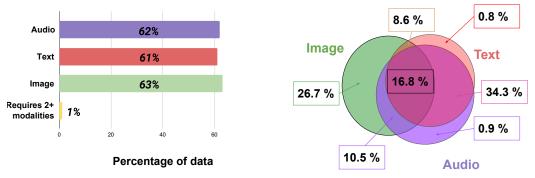
³Workers who consistently marked 'yes' or 'no' were excluded, as well as those who showed low agreement with the rest of the group.

		Image	,		Text			Audio	
Approach	Val	Test	OOD	Val	Test	OOD	Val	Test	OOD
Majority	72	72	61	61	61	69	69	69	71
Classifier	89	81	80	82	74	80	81	76	77

Table 1: Validation, test, and OOD accuracy of our classifiers for predicting the solvability of TVQA instances based on a single modality, compared to a majority baseline.

3.2 Dataset Analysis

We use our classifiers to analyze the validation set of the TVQA dataset. A similar analysis of the training data appears in Appendix A.4.2, and shows similar results. An analysis based on our seed annotation also yields similar results, see Appendix A.4.1.



(a) Proportions of questions answerable using a single modality (top three bars), as well as that of questions unanswerable using any single modality (final bar).

(b) Each circle in the graph represents a portion of the data answerable using a particular modality. Overlapping areas mark the proportions of instances solvable by either modality separately.

Figure 3: An analysis of the validation set of the TVQA dataset.

99% of the questions are solvable using a single modality. Fig. 3a shows the proportions of questions answerable with each modality, as well as the fraction of questions unanswerable using any single modality. Our results confirm previous results [Winterbottom et al., 2020] and show that almost 99% of the questions could be solved correctly using a single modality. As a result, we do not further annotate groups of more than one modality.

Questions are balanced across modalities. We observe that similar proportions of the questions could be solved by each of the different modalities—63%/61%/62% using image/text/audio, respectively. Prior work [Winterbottom et al., 2020] used partial-input models and identified a bias towards the text modality in this dataset. Our method, relying on a seed annotation, leads to a different conclusion and shows no evidence of such bias.

Many questions could be solved by more than one modality. The values in Fig. 3a indicate a large overlap between the instances answerable by the different modalities (as the sum of all bars largely exceeds 100%). To study this overlap, we plot (Fig. 3b) a Venn diagram, where each circle represents the partitions of instances answerable with a different modality, and overlapped areas represent instances solvable by either modality separately. Our results show that more than 70% of the instances are solvable using two or more modalities, and more than 15% using each of the three modalities, further indicating that many TVQA instances do not require integration of different modalities.

Large overlap between audio and text. We observe a high overlap between questions answerable using only text and those answerable using audio alone. This indicates that most questions focus

only on what is said in the scene (typically reflected in both the audio and the subtitles), rather than additional information, e.g., non-verbal sounds such as beeps, horns, or music. It also indicates few to no questions with blurry or noisy speaking that is hard to understand without the subtitles. In contrast, 26.7% of the questions are answerable using the image, but not the other modalities.

3.3 Analysis of Dataset/Model Interaction

We have so far presented a comprehensive analysis of the TVQA dataset, marking for each instance the modalities required for processing it. We now set to analyze a model trained on this dataset, in order to identify the modalities on which the model struggles with, and gain insights into the interplay between modalities when solving the task. We focus on MERLOT Reserve [Zellers et al., 2022], the only publicly available model finetuned on TVQA with all three modalities (to the best of our knowledge).⁴ We apply the model to several data splits based on our annotations, and describe our experiments and takeaways below.

Table 2: Accuracy of various versions of the MER-LOT Reserve model, each time masking different modalities (columns), and evaluated on different data splits of TVQA (rows). Columns: that modalities shown: (I)mage, (A)udio, (T)ext, A+T: audio and text. **base**:baseline model. The last block shows model performance on 'who' questions.

Data\Input	base	Ι	A	Т	A+T
All	81	57	68	69	_
Answerable-all	88	62	78	85	-
Only Image	74	75	_	_	_
Only audio-text	94	_	86	87	96
'who' questions	88	71	45	95	_

MERLOT Reserve reads better than it sees.

We first consider the finetuned MERLOT Reserve model, and perform inference with it on three versions of the validation set (**All** row in Table 2), each time masking 2/3 modalities, and leaving only one source of input (Image, Audio and Text columns). We compare these performance scores to the model's baseline performance (**base** column). Our results show that, unsurprisingly, model performance drops in all cases. However, this decrease is not uniform across modalities: using only the image modality leads to the largest drop (24%), compared to 12–13% for audio and text.

To further investigate this trend, we repeat this experiment with the questions answerable by each of the modalities (i.e., the center of Fig. 3b), which are considered somewhat easier, as they can be solved using multiple cues from the different modalities. Our results (Table 2, **Answerable-all**) show a similar trend for image and audio, but the text-only version is only 3% behind the baseline model. Our findings indicate that the model faces difficulty in using the image modality to answer questions. This aligns with previous work, which highlighted the underutilization of the image modality in multimodal tasks [Zhang et al., 2015, Goyal et al., 2016, Jabri et al., 2016, Hassantabar, 2018, Bitton et al., 2021, Dancette et al., 2021].

MERLOT Reserve struggles with image-based questions. We have so far shown that MERLOT Reserve struggles when given access only to the visual component. We next return to the full model (i.e., no masking, **base** column in Table 2), and examine whether this trend translates to difficulties in processing questions that require visual information. We consider questions labeled to be answerable by our classifiers using the image modality only, compared to those answerable only by either text or audio.⁵ Our results (middle block of Table 2) show a substantial gap (20%) in favor of the latter, indicating that the model struggles with questions that require the image modality, while it succeeds on those that are answerable by audio or text.

A training dynamics analysis. Results observed so far indicate that questions that require processing visual information are harder for MERLOT Reserve compared to those requiring text or audio. To further validate this hypothesis, we compute the training dynamics [Swayamdipta et al., 2020] of the TVQA dataset using MERLOT Reserve. This is an alternative method for mapping a dataset into different regions: *easy-to-learn*, *hard-to-learn* and *ambiguous*. We fine-tune the MERLOT Reserve model with the same parameters used in the original paper, training it for three epochs while calculating the mean and variance of the softmax output probability of the gold label for each instance.

⁴https://github.com/rowanz/merlot_reserve

⁵Due to the high overlap between audio and text observed in Section 3.2, we compare image-only questions to audio-text-only questions.

We then select the top 50% of our annotated instances with the highest variance (i.e., 50% most *ambiguous* instances in Swayamdipta et al. [2020]'s terminology). Table 3 presents the answerability proportions for each modality in the 50% most ambiguous questions, and compares them to the answerability proportions of all annotated data (from Fig. 7a). Interestingly, we notice a rise in the proportion of image-based questions in the 50% most ambiguous questions, whereas the proportion of audio/text-based questions decreases. As the 50% most ambiguous questions are considered more challenging [Swayamdipta et al., 2020], these results support our previous findings-MERLOT Reserve faces more difficulty in answering questions that require visual information compared to other modalities.

MERLOT Reserve's limitations in speaker recog- Table 3: Training dynamics analysis of the nition. We turn to further analyze the MERLOT Reserve model fine-tuned on TVQA, by evaluating its performance on a particular type of question: "who" questions, where the correct answer is one of the main characters in the TV show. E.g., the question "Who is Monica talking to when she is upset and crying?". To answer such questions, all three modalities—image, audio, and text-can be used, as the characters can be recognized through their looks, their voices, or their names as they appear in the subtitles. We therefore check whether the model is equally capable of using these modalities. We run the fine-tuned MERLOT Reserve model, and use masking to create image-only, audio-only, and text-only versions of the model to

validation set of TVQA. For each data portion (rows), we calculate the proportion of questions answerable by each modality on it. all: all annotated data, most ambig .: 50% most ambiguous examples. The prevalence of image-based ambiguous questions indicates the model's difficulty with the image modality.

Data	A (%)	T (%)	I (%)
all	63	61	63
most ambig.	48	43	83

these questions, as described above. As some "who" questions might be answerable by only one modality (e.g., if the target character does not speak during the scene), we only consider the proportion of "who" questions that are answerable using each modality according to our annotation. Our results (Table 2, last block), indicate that the audio modality has a substantially lower score compared to text and image when answering these questions. This indicates that the model struggles in speaker recognition. It is worth noting that in the text modality, the answer to "who" questions does not require any memory or learning of the characters since the character's name is usually explicitly written. In contrast, the image modality is similar to audio in this sense, where the model needs to recognize the main characters visually, based on its training. The gap between the audio and image modalities emphasizes the model's challenges in recognizing the main characters through speech, compared to visually.

4 A New Multimodal Test Set for TVOA.

Our analysis in Section 3.2 has shown that TVQA contains almost no questions that require more than one modality. To test the impact of this deficiency on models trained on TVQA, we crowdsource a set of 150 questions that require multiple modalities (see Appendix A.5 for information about the data collection). Another group of workers then filter out questions that are either not multimodal or insolvable. We observe that one approach used by our workers is to modify distractors in existing TVQA questions, in order to force the model to use multiple modalities.

For example, in Fig. 4, we have the question "What does Phoebe do after saying she has only had six drinks?". The original correct answer ("puts food in her mounth"), is the only option describing an action performed in the video, and as a result an image-only model is able to answer it correctly. Our annotators modify this question



Figure 4: An original TVOA question that relies on visual cues alone, along with improved distractors that now require both image and sound/text. The video frames appear in chronological order from left to right.

by adding distractors such as "smiles a big smile", an action Phoebe performs while saying something else. See Appendix A.5 for more examples.

We evaluate the pre-trained MERLOT Reserve model on the collected questions, observing that it performs substantially worse on them compared to its performance on the original validation set—41% vs. 83%. This indicates that the model struggles with questions that require more than one modality. In order to make models more robust, future work will involve collecting a larger set of such questions, both for training and evaluation.

5 Related Work

Dataset Analysis. Previous work has primarily focused on understanding dataset characteristics. Carlini et al. [2019] has focused on outlier analysis in datasets, while Siddiqui et al. [2022] and Swayamdipta et al. [2020] focused on identifying different subsets within datasets using training dynamics. Nam et al. [2022] focused on improving worst-group accuracy of datasets. Luccioni et al. [2023] explored social biases in text-to-image systems. De Silva et al. [2022] explored the value of OOD examples. Akiki et al. [2023] presents a qualitative analysis of large scale research datasets. Talukdar et al. [2021] has focused on identification and isolation of mislabelled data. Our work targeted a different element in large scale datasets—the importance of different modalities for each instance.

Multimodal Dataset Analysis. Previous work has primarily analyzed biases towards specific modalities in multimodal datasets by using existing models [Winterbottom et al., 2020, Bitton et al., 2021, Hendricks et al., 2021]. Specifically, Winterbottom et al. [2020] trained partial-input models and used them to analyze the different modalities used in TVOA. Our approach, which relies on a seed human annotation, allows us to reach different and more reliable conclusions (Section 3.2). Previous studies have employed human evaluation to discern biases towards specific modalities in their datasets [Lei et al., 2018, Antol et al., 2015, Tapaswi et al., 2016, Chao et al., 2017, Castro et al., 2020, Wang et al., 2021]. However, these investigations were limited to datasets containing only two modalities, and they did not expand their evaluation using automatic methods. Specifically, in the original TVQA paper, Lei et al. [2018] provided partial inputs to workers. Our study focuses on three modalities, and goes further by examining the interaction of single-modality questions, making further observations about the data. Talmor et al. [2021] and Alamri et al. [2019] created datasets with three modalities and used human evaluators for data assessment. Our human analysis differs as we specifically analyze the required modality for each instance and consider the integration of multiple modalities, unlike their approaches which either treated modalities as input or focused solely on dialog quality without considering modality integration for individual questions. Chen et al. [2020] created HybridQA, which contains two modalities, and relied on human assessments to analyze the answerability of questions across different modalities. As shown in Section 3.1, the reliability of human assessments can vary; they occasionally claim inability to answer based on a particular modality, when in fact, they can. By implementing an approach which includes testing human responses, we provide a more dependable evaluation of our dataset's modality distribution.

6 Limitations

Our analysis is based on automatic tools trained on relatively small amount of data (\sim 500 training instances, Section 3.1) to annotate 150K instances. The high costs of annotation prevents us from further expanding the initial annotation seed. Nonetheless, our analysis shows that our classifiers are fairly accurate (74–81%), that they generalize well to out-of-distribution, and finally, that more data doesn't necessarily improve performance (Section 3.1).

Our method is designed for classification-based tasks. Extending it to generation tasks is not straightforward, largely due to the challenges associated with evaluating human responses, which makes it hard to give binary solvable/insolvable labels to instances.

7 Conclusion

We presented a two-step method for analyzing multimodal datasets, contributing to the ongoing discourse on data quality assessment. We proposed an approach that leverages a small seed of human annotation to identify important modalities in a dataset. We applied this approach to analyze the TVQA dataset and the MERLOT Reserve model. Our findings reveal that almost all TVQA questions can be answered using only one modality at a time. Moreover, there's no specific bias towards any modality in the dataset, differing from previous research. Additionally, we demonstrated that the MERLOT Reserve model struggles with questions requiring the image modality but performs better with audio or text modality. We also highlighted the model's difficulty in speaker recognition. Finally, we collected 150 instances that require more than one modality to answer, and demonstrated that the model performs poorly on them. Our results enhance our understanding of dataset characteristics, as well as provide insights into the performance and limitations of AI models, highlighting the need for more robust multimodal modeling.

8 Acknowledgements

This work was supported in part by the Israel Science Foundation (grant no. 2045/21). We would like to extend our thanks to our colleagues from Roy Schwartz's lab at Huji for their valuable assistance with the pilot annotations. A special acknowledgment goes to Jeff Moskowitz for his continuous, thoughtful feedback, annotations, and support. We also express our appreciation to Eytan Siegel for his helpful feedback, as well as to Or Malka, Bar Madvil, Cheneil Clarke, and Adva Madvil for their annotations.

References

- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main. 98. URL https://aclanthology.org/2021.emnlp-main.98.
- Nicholas Carlini, Úlfar Erlingsson, and Nicolas Papernot. Distribution density, tails, and outliers in machine learning: Metrics and applications, 2019. arXiv:1910.13427.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Conference on Empirical Methods in Natural Language Processing*, 2020.
- Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022.
- Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. Metadata archaeology: Unearthing data subsets by leveraging training dynamics, 2022. arXiv: 2209.10015.
- Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models, 2023. arXiv:2303.11408.
- Arka Talukdar, Monika Dagar, Prachi Gupta, and Varun G. Menon. Training dynamic based data filtering may not work for nlp datasets. In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4631–4640, 2015.

- Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832, 2017.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering, 2018. arXiv:1809.01696.
- T Winterbottom, S Xiao, A McLean, and N Al Moubayed. On modality bias in the tvqa dataset. In *The British Machine Vision Conference*. DU, 2020.
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16354–16366, 2022.
- Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.
- Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.91. URL https://aclanthology.org/2020.findings-emnlp.91.
- Santiago Castro, Mahmoud Azab, Jonathan C. Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. Lifeqa: A real-life dataset for video question answering. In *International Conference on Language Resources and Evaluation*, 2020.
- Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5014–5022, 2015.
- 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. *International Journal of Computer Vision*, 127:398–414, 2016.
- Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, pages 727–739. Springer, 2016.
- Shayan Hassantabar. Visual question answering : Datasets, methods, challenges and oppurtunities, 2018. URL http://www.cs.princeton.edu/courses/archive/spring18/cos598B/ public/projects/LiteratureReview/COS598B_spr2018_VQAreview.pdf.
- Yonatan Bitton, Michael Elhadad, Gabriel Stanovsky, and Roy Schwartz. Data efficient masked language modeling for vision and language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3013–3028, 2021.
- Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1554–1563, 2021.
- Ashwin De Silva, Rahul Ramesh, Carey Priebe, Pratik Chaudhari, and Joshua T Vogelstein. The value of out-of-distribution data. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.
- Christopher Akiki, Odunayo Ogundepo, Aleksandra Piktus, Xinyu Zhang, Akintunde Oladipo, Jimmy Lin, and Martin Potthast. Spacerini: Plug-and-play search engines with pyserini and hugging face, 2023. arXiv:2302.14534.
- Lisa Anne Hendricks, John F. J. Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.
- Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *North American Chapter of the Association for Computational Linguistics*, 2017.
- Josiah Wang, Pranava Swaroop Madhyastha, Josiel Maimoni de Figueiredo, Chiraag Lala, and Lucia Specia. Multisubs: A large-scale multimodal and multilingual dataset. In *International Conference on Language Resources and Evaluation*, 2021.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021.
- Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings* of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7558–7567, 2019.

A Appendix

A.1 Dataset Supplementary Materials

- 1. Author statement: We bear all responsibility in case of violation of right in using our dataset.
- 2. License: Dataset is licensed under CC-BY 4.0 license https://creativecommons.org/ licenses/by/4.0/legalcode.
- 3. The data and code are included in the supplementary material folder and will be released at a later date.
- 4. Intended uses: Our aspiration is for researchers to use our newly created TVQA test set and the workers' annotations in order to evaluate their multimodal models.
- 5. TVQA Lei et al. [2018] and MERLOT Reserve Zellers et al. [2022] are both licensed under MIT License.

A.2 Human Annotation of TVQA

Workers from various modality groups participate in HITs as depicted in Figure 5. They receive clear instructions on how to provide their answers. For instance, the image group is presented with instructions similar to the one shown in Figure 6. The audio and text groups receive comparable instructions tailored to their respective modalities.

Qualification. The worker qualification process consists of two stages: an automatic stage followed by a manual stage. In the automatic stage, each modality group is assigned three HITs. Workers are required to answer questions based on the provided modality and indicate whether the question is answerable or not. Only workers who pass this initial stage proceed to the manual qualification process. In this stage, qualified workers annotate a batch of 10 HITs. Their annotations are evaluated based on agreement with other workers and ground truth. Workers who mark "seen in the past" for more than 30% of the questions are rejected. Additionally, the consistency between the signal of answering the question and the worker's response to "Is it possible to answer this question based on the specific modality?" is also assessed during evaluation.

Payment. We hire nearly 20 workers, each worker is compensated at a rate of \$14-16 per hour for their participation in the project. The total expenditure for this project amounts to approximately \$1500.

Selecting challenging instances. We select a set of 150 examples from the 'Friends' section of the TVQA validation set, which exhibit the highest sensitivity to each modality in the MERLOT Reserve model. This involved applying the model to the entire validation set and monitoring the probability of the gold label. We then run the model on inputs where each modality is masked. For each modality, we select the top 50 instances based on the largest decrease in the probability of the gold label when that modality is masked, compared to using all modalities as input. These instances may suggest questions that rely more heavily on a specific modality than others. Since there aren't many such instances for audio and text (due to the substantial overlap between these modalities in answering questions), this process makes the classifier training set more diverse for these modalities.

A.3 The Classifiers

A.3.1 Ablations

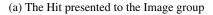
We experiment with various methods to recognize the data partitions that resulted in a decrease in performance when compared to human performance. These techniques highlight the importance of human annotations and model masking. Tables 4 to 6 show the performance of the different approaches.

1. Uni-modal models vs. human annotators.

We fine-tune the model on each subset of modalities five times using different seeds for each subset. During fine-tuning, all other modalities are masked. For instance, fine-tuning the model on the image modality involves masking audio and text in the video, allowing the







(b) The Hit presented to the Text group

	Why did Phoebe have earmuffs on when she was at her apartmen
	O Workers were using a jackhammer to fix the street.
	O Neighbors were playing loud music.
▶ ━━━━ → ◀। ▬ ≥	O The smoke alarm was going off.
	O The sound of the train was keeping her up.
	O A dog was barking.
Is it impossible to answer the question using the	e given audio?
⊖yes	
⊖ no	
I have seen it in the past	

(c) The Hit presented to the Audio group

Figure 5: HIT examples from Mechanical Turk showcasing the annotation of different modality groups in TVQA.

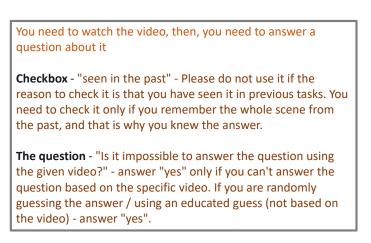


Figure 6: The Instructions presented to the image group when annotating the TVQA examples

model to receive only image frames as input for the question. For each subset and data point, we calculate the majority vote of the models to classify whether the example is solvable by that specific subset.

2. Single model probabilities vs. multiple masking modalities.

We train a classifier using random forest, with various hyperparameters, which takes as input only the probability of the model given all modalities.

3. MLP vs. random forest.

We train a classifier using MLP with various parameters, using the same inputs as our previous classifier.

4. With gold label vs. without gold label.

We train a classifier similar to the one described previously, but without modifying the input to incorporate the gold label of the original question.

Table 4: Different approache	s' performance f	or audio modality	prediction,	, including OOD	perfor-
mance (150 examples from '	House").				

Approach	Train	Val	Test	OOD
majority	-	69	69	71
random	-	50	50	50
1 - 5-models	67	65	80	79
2 - single probability	62	71	67	71
3 - mlp	74	77	77	74
4 - without gold label	67	77	73	78
our classifier	84	81	76	77

Table 5: Different approaches' performance for **text** modality prediction, including OOD performance (150 examples from "House").

Approach	Train	Val	Test	OOD
majority	-	61	61	69
random	-	50	50	50
1 - 5-models	70	66	65	77
2 - single probability	76	71	50	65
3 - mlp	78	78	77	78
4 - without gold label	89	76	74	80
our classifier	96	82	74	80

Table 6: Different approaches' performance for **image** modality prediction, including OOD performance (150 examples from "House").

Approach	Train	Val	Test	OOD
majority	-	72	72	61
random	-	50	50	50
1 - 5-models	66	60	68	65
2 - single probability	97	75	72	61
3 - mlp	77	83	79	75
4 - without gold label	95	88	85	79
our classifier	88	89	81	80

A.3.2 Training Size Analysis

The use of a small training size for the classifiers is beneficial as it simplifies the process of collecting data from workers, which can be both time-consuming and expensive. In this section, we show that

		Image	;		Text			Audio	
Train	Val	Test	OOD	Val	Test	OOD	Val	Test	OOD
30 %	83	85	83	80	71	80	78	83	69
50 %	85	84	81	81	82	78	80	79	79
70~%	84	82	84	82	79	79	79	79	75
100 %	89	81	80	82	74	80	81	76	77

Table 7: Accuracy of training the classifiers with various amounts of data.

even using a smaller training set size leads to roughly the same results. Table 7 displays the accuracy results for validation, test, and OOD when training the classifiers on randomly selected subsets of the training data. We conduct the same hyperparameter search on the validation set as done with our classifiers for each modality and amount of data. As shown in Table 7, increasing the data size does not substantially improve performance, indicating that there is no added benefit in collecting more data.

A.4 TVQA - Extra Analysis

A.4.1 Analysis of TVQA based on workers

To validate the findings from the analysis conducted on the entire validation set of TVQA, we perform similar experiments (as in Section 3.2) on the annotated data generated by workers. These experiments aim to assess the answerability of different modalities, evaluate the performance of the MERLOT Reserve model on questions solvable by various modalities, and replicate the model's difficulty with image modality.

The data splits resulting from the classifiers applied to the annotated data by workers are shown in Figure 7. These splits exhibit a similar trend to those observed in the validation data (Figure 3), indicating that the annotations are representative. Furthermore, we apply the experiments described Section 3.3 to the annotated data. The results, presented in Table 8, mostly replicate the findings from the full validation data.

Additionally, the dataset cartography results obtained on the collected annotations set, as shown in the first block of Table 9, were consistent with those obtained from the annotated data. This further supports the notion that the analyzed human annotations are representative and applicable to the entire TVQA validation set.



(a) The proportion of TVQA questions that could be answered using a single modality, with each bar representing a different modality. The final bar represents the proportion of data that is unanswerable using any single modality.

(b) Each circle in the graph represents a portion of the data that can be solved only using a particular modality. Overlapping regions indicate the partitions that can be answered by either modality separately.

Figure 7: An analysis of our collected annotations of the TVQA dataset.

Table 8: Accuracy scores of various partial-input models evaluated on different data splits derived from our collected annotations. Columns: Input modalities to the model, with others being masked (I:image, A:audio, T:text, I+A+T:all, A+T: audio and text). Rows: different splits evaluated in the study. Our analysis reveals a degradation in the accuracy of the image-only input model on various portions compared to other inputs or the full model. Moreover, the full model's performance on image-only questions is poor. We conclude that the model struggles with the image modality.

Data\Input	I+A+T	A+T	Ι	A	Т
All annotated data	81	_	52	61	66
Answerable-by-each	87		57	68	79
Only Image	74	_	66	_	
Only audio–text	89	90	-	86	

A.4.2 Extending the Analysis to the Training Data.

After establishing that our predictions are reasonably accurate, we turn to explore the TVQA training dataset. Applying the same method here is not possible, since the training data is used to fine-tune our MERLOT Reserve model. To address this, we randomly split the training data to two, and re-fine-tune MERLOT Reserve twice, one for each half of the dataset. Subsequently, we train three classifiers with input representations from each of these models on the labels provided by humans. By doing so, we can extract probabilities from the model that is not trained on a specific half of the training set and apply the relevant classifiers to predict the modalities that can answer each instance in that half. This approach enables us to more accurately identify the data splits of the training set as if it were validation data. The resulting splits of the training data, as determined by the classifiers applied to the training set, are shown in Figs. 8a and 8b. These splits are comparable to those observed in the workers' annotations analysis and to validation set, indicating a balanced distribution of the validation and training splits, which is valuable for generalizing the model. In an ideal scenario, it would be necessary to apply the classifiers to the test set of TVQA. However, since the gold labels for the test set are not available, and the classifiers relies on them, we are unable to perform this step.

Table 9: Training dynamics analysis of our annotated portion of TVQA. For each data portion presented in rows, we calculate the proportion of questions answerable by each modality on it. First Block: Row (1): all annotated data portion, row(2): 50% most ambiguous examples of the annotated data. Second Block: Row (3): all training portion, row(4): 50% most ambiguous examples of the training data. The prevalence of image-based questions over other modalities in the 50% most ambiguous questions indicates the model's difficulty with the image modality.

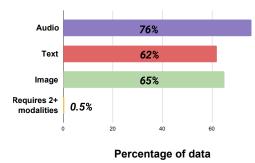
Data	A (%)	T (%)	I (%)
all annotated most ambig.	71	64	69
	58	51	84
all train	76	62	65
most ambig.	64	42	90

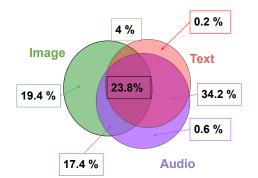
A.5 Creating Multi Modal Questions

A.5.1 Human Annotations

We recruit workers from Amazon Mechanical Turk (AMT) who are proficient in English. These workers are provided with explicit instructions on how to create the questions, as depicted in Fig. 9, to ensure clarity and consistency in their task.

Qualification. Our crowdworkers undergo a qualification test that includes a small set of questions to confirm their ability to generate valid and answerable multimodal questions. As part of this test,





(a) The graph displays the proportion of data that could be answered using different modalities such as audio, text, and image, with each bar representing a modality. The final bar represents the percentage of data that was unanswered by the workers for any modality.

(b) Each circle in the graph represents a portion of the data that can be solved only using a particular modality. The overlapping region of the circles indicates the partitions that can be answered by either modality separately.

Figure 8: An analysis of predictions of training split of the TVQA dataset.

The task is :

•Watching a short video from a TV show (contains video, audio and subtitle)

• Create a multiple choice question (5 options) which requires 2 or more modalities in order to answer it (here modalities = {audio, video frames, subtitles})

•It's supposed to take between 3-5 min per question. So the payment 0.6\$ per example.

For example:

Q: What does Phoebe do after saying she has only had six drinks?

- puts food in her mouth
 smiles a big smile
 takes a drink from a wine glass
- 4. touches her necklace
- 5. rubs her hair

So of course video frames are essential in order to answer. but audio\text is also since, in the video we see Phoebe does all the things which are described above, but **she puts food in her mouth only** after she said she has only had six drinks. As a result, the use of both text/audio and image is necessary to answer this revised question.

Figure 9: The instructions for annotating the multi modal questions

each worker is provided with a set of existing videos from the TVQA dataset and instructed to create 10 questions that require the integration of at least two modalities for answering. Only those workers who successfully produce a minimum of 6 multimodal questions pass the qualification test.

Data Collection. Within the group of qualified workers, some are responsible for creating the multimodal questions, while others are assigned the task of validating them. Any questions that were not deemed multimodal or answerable were discarded during this validation process

Payment. We hire 6 workers, each worker is compensated at a rate of \$14-16 per hour for their participation in the project.

A.5.2 Extra Examples



Question: What is the character who says 'What was that about?' wearing?

- 1: A blue shirt
- 2: A striped button-down shirt
- 3: A blue jacket
- 4: A green T-shirt
- 5: Nobody says it

Figure 10: Example no.1 of multi modal questions from our created test set. The video frames appear in chronological order from left to right. This example is multimodal because it requires audio or text to determine that Penny said it, and then the image modality is needed to identify what she is wearing (other optional answers are clothing worn by others in the scene).



Question: How many people are in the scene with Sheldon who don't talk?

- 1:0
- 2:1
- 3: 2
- 4: 3
- 5:4

Figure 11: Example no.2 of multi modal questions from our created test set. The video frames appear in chronological order from left to right. This example is multimodal because it requires audio or text to determine how many people speak, and then the image modality is needed to see how many people are in the scene.



Question: What does the guy who is wrapped with bubble wrap say?

1: Nothing

- 2: From?
- 3: We're taking a break
- 4: No one is wrapped with a bubble wrap
- 5: I'm sorry

Figure 12: Example no.3 of multi modal questions from our created test set. The video frames appear in chronological order from left to right. This example is multimodal because it requires the image modality to identify Joey is wrapped with bubble wrap, and then audio or text are needed to determine he says nothing.