Tackling Data Bias in MUSIC-AVQA: Crafting a Balanced Dataset for Unbiased Question-Answering

Xiulong Liu *†
University of Washington
xl1995@uw.edu

Zhikang Dong *†
Stony Brook University
zhikang.dong.1@stonybrook.edu

Peng Zhang
Bytedance Inc
zhang.peng@bytedance.com

Abstract

In recent years, there has been a growing emphasis on the intersection of audio, vision, and text modalities, driving forward the advancements in multimodal research. However, strong bias that exists in any modality can lead to the model neglecting the others. Consequently, the model's ability to effectively reason across these diverse modalities is compromised, impeding further advancement.

In this paper, we meticulously review each question type from the original dataset, selecting those with pronounced answer biases. To counter these biases, we gather complementary videos and questions, ensuring that no answers have outstanding skewed distribution. In particular, for binary questions, we strive to ensure that both answers are almost uniformly spread within each question category. As a result, we construct a new dataset, named MUSIC-AVQA v2.0, which is more challenging and we believe could better foster the progress of AVQA task. Furthermore, we present a novel baseline model that delves deeper into the audio-visual-text interrelation. On MUSIC-AVQA v2.0, this model surpasses all the existing benchmarks, improving accuracy by 2% on MUSIC-AVQA v2.0, setting a new state-of-the-art performance. Dataset: https://github.com/DragonLiu1995/MUSIC-AVQA-v2.0/

1. Introduction

In audio-visual learning, the interplay between audio and visual information provides a rich avenue for understanding dynamic scenarios. One particular task that embodies this synergy is Audio-Visual Question Answering (AVQA). As Fig. 1 shows, given a musical instrument performance video and associated music audio, models are expected to answer questions that are related to them or the relationships therein. Unlike existing popular Visual Question Answering (VQA) task, which only tackles two modalities -

vision and language, AVQA is designed to bridge and reason through all three modalities - vision, language and audio, which stimulates the merge of a new dataset. MUSIC-AVQA dataset, proposed to cater for this task, acting as an important benchmark, facilitate the research progress in this field. The dataset consists of 3 major question categories by modality: Audio-Visual, Visual and Audio questions, respectively. Across all the categories, 5 question aspects are considered, including "Existential" - e.g. Is there a voiceover?, "Temporal" - e.g. Which violin makes the sound first?, "Counting" - e.g. How many sounding violins in the video?, "location" - e.g. Where is the performance?, and "Comparative" - e.g. Which object makes the sound first?. By joining the modality types with question aspects, the dataset ends up with 33 question templates, in which one can change the instrument name from one to another based on the video. Each question template contains a fix set of answers, ranging from binary answers (e.g. "yes" and "no") to counting answers (e.g. "1", "2", "3" etc.) and so on. Finally, MUSIC-AVQA dataset contains 9,290 (7,423 real, 1,867 synthetic) videos and 45,867 corresponding questions.

We notice there is a strong bias exists in the dataset, which results in undermining the reliability of this dataset as a credible benchmark. For example, in a particular question category asking about whether or not the audio track comes from the instrument played in the video, Over 90% of answer in the dataset is "yes". In audio-visual temporal question, when asking which instrument in the video sounds first, nearly 80% answer is "Simultaneously". In counting questions, answers of small number like 1 and 2 alone could take up more than 50%. These imbalance exhibit across question aspects of "Existence", "Counting", "Temporal", "Location" and "Comparative". Such outstanding bias could impact the model training negatively and thus lowering the ability of the underlying model, as model will be trained to favor towards to the most common answers in the training set and ignoring the importance of video and audio, as well as the reasoning between three of them. This problem has been explored in other tasks, e.g., VQA [3],

---

*These authors contributed equally.
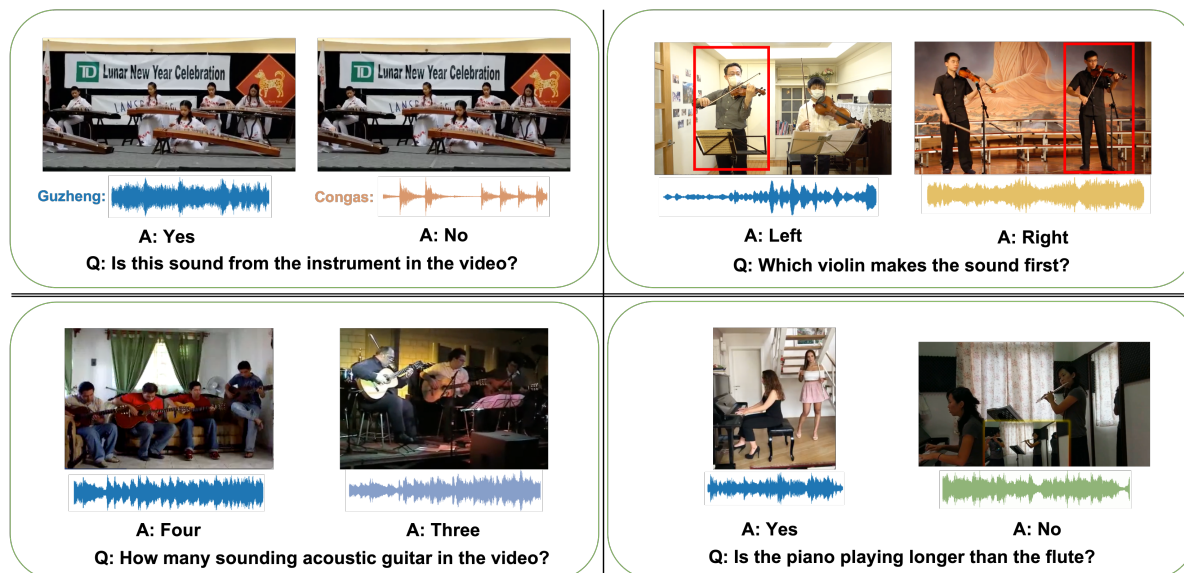†Work done while interning at Tiktok

Figure 1. An overview of *MUSIC-AVQA v2.0 QA* samples. (a) showcases two videos with the same visuals but different audio; (b) both videos display two violinists in different orders; (c) videos present acoustic guitar ensembles of different counts; (d) features a piano-flute duet where the piano plays longer in the first video. To answer accurately, models must consider these audio-visual nuances rather than just language priors.

where several datasets and models [2, 4, 18, 38] are proposed to tackle different biases. However, this problem still remains untouched in AVQA. In light of these observations, we endeavor to systematically address the issue and propose an improved version of the data, called MUSIC-AVQA v2.0.

Our initial step was to evaluate the answer distribution for each question template. Through this analysis, we identified specific templates that displayed a skewed answer distribution. After identifying these skewed templates, we proceed to select the question templates that have minority answers, designating them as our primary targets for balance. Recognizing the significance of holistic data representation, we take the additional step of manually collecting videos that corresponded with these specific question template and answer pairs. This ensures that the dataset is more representative and balanced.

Existing works have proposed models to the task of AVQA, which were trained on the public MUSIC-AVQA dataset. On the evaluation of the balanced dataset, we show that the strong data bias does mislead the model training, resulting in inferior performance. Moreover, we contribute a model that has the ability to learn the connections across all the three modalities. The model extends existing methods by: (i). adding an additional pre-trained Audio-Spectrogram-Transformer (AST) branch for audio-visual grounding, and (ii). designing a cross-modal pixel-wise attention between audio and visual spatial maps. Experiment results on balanced dataset indicate that our new baseline, even without audio-visual pretraining, sur-

passes the performance of existing state-of-the-art models by more than 2%. In summary, the primary contributions of this work are 3-folds: (i). We identify and mitigate the data imbalance issue in existing MUSIC-AVQA dataset, and establish MUSIC-AVQA v2.0, a more balanced AVQA benchmark; (ii). MUSIC-AVQA v2.0 enriches the MUSIC-AVQA dataset with an additional 1230 real videos compared to exising 7.4k real videos, leading to an additional inclusion of 8.1k QA pairs. A substantial portion of the additional videos are musical instrument ensemble involving 3 or more instruments, capturing more complex audio-visual relationship; (iii). We introduce a novel baseline model, which demonstrates enhanced capability in bridging different modalities through the use of pretrained models and an attention mechanism. Upon evaluation on both bias and balanced datasets, our model surpassed all existing baselines, establishing a new and stronger baseline.

2. Related Works

Emerging as a significant branch of multi-modal learning, audio-visual learning explores the complex relationship between audio and visual signals. Various tasks have been established under this domain, such as audio-visual source separation [5, 11–13, 30, 39, 40], audio-visual event localization [9, 31, 33–35], audio-visual generation (audio-to-visual or visual-to-audio) [8, 10, 14, 19, 25–28], audio-visual detection and segmentation.

More recently, AVQA [20, 36, 37] emerges as a novel task as a tri-modal extension of VQA. Such a QA task focuses on the aspect of understanding the video by combin-

ing both audio and visual context. Pano-AVQA [37] set up a QA benchmark that focused on 360 degree videos. Later, MUSIC-AVQA [20] proposed a new benchmark that focused on musical instrument performance videos, and the questions cover more aspects of the audio-visual associations by incorporating challenging questions like counting, temporal relation and comparative questions. Both MUSIC-AVQA and Pano-AVQA designed QA benchmark to be open-ended QA. In contrast, AVQA [36] provided a multiple-choice format, adding depth to the question types by including causal and purpose questions.

In addition to these tasks, different representation learning paradigms or model architecture designs have been proposed to effectively bridge the gap between audio and visual modality. In representation learning aspect, early works exploit the audio-visual correspondence [23, 24] by designing tasks like audio-visual matching and audio-visual synchronization. More recently, a self-supervised learning-based framework [15] that combines both contrastive learning [6, 7, 16] and masked auto-encoder technique [17] was proposed to pre-train audio-visual representations. In model architecture aspect, LAVISH [21] designed a lightweight adapter layer that bridges between vision pre-trained backbones. During training, only the adapter layer along with the task-specific layers are trained while the remaining parameters are kept frozen, which enables parameter-efficient learning without large-scale pre-training. The method achieves the state-of-art performance on 3 different tasks with benchmarks, respectively audio-visual event localization (AVE [32]), audio-visual segmentation (AVS [41]) and audio-visual question answering (MUSIC-AVQA [20]).

Biases in question answering dataset has long been a problem. In particular, for template-based open-ended question, answers imbalance, where a specific answer category takes up the majority of the question. Such biases can skew models towards certain answer categories, ignoring the actual context. , e.g., in VQA, no visual context. [38] verified and mitigated such issue by constructing a balanced synthetic VQA dataset. Such bias issue is further examined in VQA-CP [2] by crafting a dataset where answer distribution varies significantly across train and test splits such that models that overfit to specific answer category during training will be 'trapped' on the test set. In addition, different models that aim to handle the bias issue were proposed by reducing the language bias and grounding on visual modality [4, 18]. As audio-visual question answering emerged as a new task, few attentions have been focused on studying the bias issue. In particular for open-ended AVQA dataset, Pano-AVQA [37] only briefly mentioned that bias are alleviated by audio replacement or providing counter examples with different answer to same question. MUSIC-AVQA [20], as the largest and most popular open-ended AVQA benchmark, completely ignores the bias issue pre-

sented in each question category. As such, our endeavor to solve bias issue in this particular dataset becomes valuable.

3. Dataset balancing

We describe our approach to balance the dataset on the public MUSIC-AVQA dataset. This process entails two phases: pinpointing biased questions and then achieving balance. First, we scrutinize the distribution of answers for each question template. Specifically, we review questions templates under all the 9 questions types. Our criteria for identifying bias is when a single answer represents over 60% of responses for binary questions or exceeds 50% for multi-class questions (with more than two possible answers). Out of the 9 question types evaluated, 7 exhibited a skewed answer distribution for at least one of their templates. These include: audio-visual existential, audio-visual counting , audio-visual temporal, visual location, visual counting, audio counting and audio comparative questions. Within these types, multiple templates could exhibit biases, amounting to 15 out of 33 templates in total. A detailed comparison of the answer distribution before and after our balancing for each biased template is illustrated in Fig. 2. We present the balancing process below.

Audio-visual Existential Questions: Binary questions dominate this question type, where the answers are either "yes" or "no". In this case, we simply pick the most frequent answer and collect complementary pairs for it. We take two questions as an example: **Is this sound from the instrument in the video?** 90% of data samples in this question template are answered "yes". To create QA pairs whose answers are "no", we replace the audio track from the video with audio of another instrument type. To make the QA pairs non-trivial, we cluster the set of instruments into "string instrument", "woodwind instrument", "brass instrument" and "percussion instrument". When replacing the audio track, by 50% chance the audio track is replaced with a different instrument of the same cluster, while another 50% chance the audio track is replaced with instrument music belonging to other clusters. Using this method, we create 794 videos paired with non-matching audio segments. **Is there a voiceover?** Originally, this question is severely imbalanced where 79.6% of answers are "no". However, after carefully delving deep into the video data, we found that different individuals have different definitions to the concept of "voiceover", which led to inconsistencies in the labels. For example, some labelers define "voiceover" as human voice appearing on top of the instrument sound, while others define it as any general "off-screen" sound. To fix such inconsistency, we define it as **any "off-screen sounds"**. After manually checking 1,278 video-audio-question pairs from the training set, we corrected 169 mislabeled entries (13%). Despite this correction, a significant imbalance remained with 68% of labels still being no". To address this, we
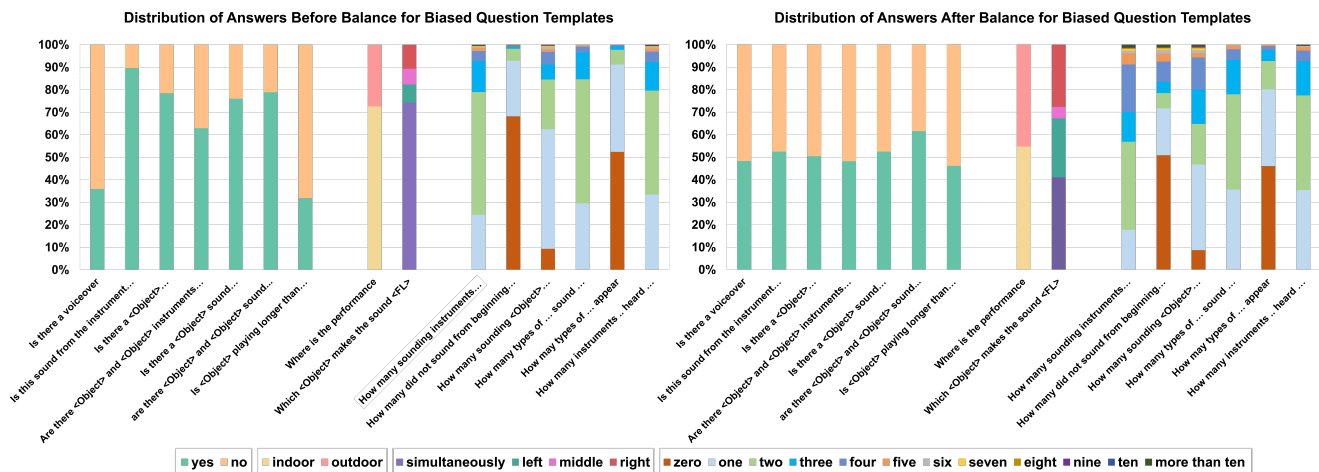
Figure 2. An overview of answer distribution in bias question templates before and after balance.

add another 456 QA pairs from our collected videos where "voiceover" presents (answered "yes"), resulting in a balanced distribution with 51.7% yes" and 48.3% no" answers.

Audio-Visual Counting Questions: This type of question asks about counting aspect of instruments in the video. The questions are structured using 4 templates that address the following aspect of counting: (i). the total number of sounding instruments (T1). (ii). the number of instrument types which are sounding (T2), (iii). the number of a specific sounding instrument type (T3), iv). and number of instruments that did not sound from the beginning to the end (T4). The answers are restricted to "0-10" and "more than ten". Upon analyzing the dataset, we observe a significant imbalance in the answers. Specifically, "0", "1" and "2" dominates the answers. For all four templates, the most frequent answer exceeded 50% of the total, with one template even reaching 60%.

To balance audio-visual counting questions, we manually collect musical ensemble performance videos. Our goal is to gather videos of musical ensemble performances where at least one answer to the aforementioned question templates exceeded 2. We sourced potential videos from the YouTube8M [1] dataset, specifically targeting those tagged with terms like "musical ensemble," "string quartet," or specific instrument names. For each instrument type, we selected videos tagged with that type and combined them with videos tagged as "musical ensemble," "string quartet," "quartet ensemble," or a specific instrument name, which helps narrow down potential candidates. From this set, we manually filtered out videos that were of very low quality, had static scenes like album covers, or had ambiguous content. For each selected video, we annotated: (i). Total number of instruments, (ii). Number of distinct instrument types, (iii). Count of each instrument type, (iv). Number of sounding instruments, (v). Number of distinct sounding

instrument types, (vi). Count of the most frequently appearing instrument that also produces sound, (vii). Number of instruments that did sound from the beginning to the end of the video. In total, we collected 591 videos for audio-visual counting questions. Using the annotations from these videos, we generated additional QA pairs for each template: 572 (+39%) for T1, 502 (+25.4%) for T2, 815 (+40.1%, 350 from originally unlabeled videos for the question template in MUSIC-AVQA dataset, 465 from our collected videos) for T3, and 413 (+30.3%) for T4. These new pairs have answers from the less frequent answer categories. After adding these pairs, imbalance issue from all 4 question templates is sufficiently mitigated, with the most frequent answer percentages decreasing by 16%, 17%, 15%, 13% for the 4 templates, respectively.

Audio-Visual Temporal Questions: This category of questions probes the order in which instruments play during a performance. The candidate answers range from 22 instrument categories, as well as positional indicators like "left", "middle", "right" that specify an instrument's location. Additionally, the term"Simultaneously" denotes that instruments play at the same time. Among 3 question templates in this question category, the question "Which <Object> makes the sound first/last?" shows a strong imbalance: 74% of answers are "simultaneously".

To address this imbalance in a multi-class setting, we labeled QA pairs with the answers "left," "right," or "middle" to diminish the dominance of the "simultaneously" category. For example, consider a video where three violinists are performing. If the violinist on the left initiates the performance, followed by the middle and right violinists, we can formulate a QA pair as: "Q: Which violin starts playing first? A: left." In addition, we augment the video by horizontally flipping it. This transforms the QA pair to: "Q: Which violin starts playing first? A: right." Follow-

ing the above procedure, we collected 203 additional targeted videos from YouTube for creating QA pairs. After augmentation, we end up creating 713 (+81.1%) additional QA pairs with answers rather than "simultaneously", reducing the most frequent answer percentage by 33% from 74% to 41%.

Visual Counting Questions: Unlike audio-visual questions, these questions rely solely on visual information. The first two templates, "Is there <Object> in the entire video" and "Are there <Object> and <Object> instruments in the video" determine the presence of specific instruments. For these templates, the majority of answers are "yes", constituting 78.4% and 62.7% respectively. To counter this bias, we generated 794 and 423 QA pairs with the answer "no" for each template. These pairs were created using labels from our collected videos.

The third template focuses on counting the types of instruments present in the video. There's a notable imbalance here, with the answers "1" and "2" making up 91% of all responses. To address this, we used labels from the 591 videos from our audio-visual counting collection. We specifically chose 204 videos where the answer to this question exceeded 2. This selection helped reduce the dominance of the top two answer categories, bringing their combined percentage down from 91% to 80%.

Visual Location Questions: This category of questions pertains to the location of the performance and the specific positioning of a performer. It seeks answers to whether the performance is indoor or outdoor and inquires about the relative position of instruments in the video. Out of the four question templates in this category, the template "Where is the performance?" with answers either "indoor" or "outdoor", exhibits an outstanding imbalance. To address this, we collect 456 QA pairs whose answer is "outdoor", from 456 videos that are not labeled for this question template from original MUSIC-AVQA dataset, resulting in reduction of QA pairs with the dominant category ("indoor") by 17.9% from 72.6% to 54.7%.

Audio Counting Questions: The first two question templates, "Is there a <Object> sound?" and "Are there <Object1> and <Object2> sounds?", focus on the audio aspect, determining the presence of specific sounds. For these templates, the majority of answers are "yes", accounting for 76.0% and 78.8% respectively. To address this imbalance, we created 794 and 423 QA pairs with the answer "no" for each template, using labels from videos where the sounding instruments were identified. This reduced the dominance of the "yes" answer to 52.5% and 61.5% for the two templates, respectively. The third template queries the total number of distinct instrument sounds heard throughout the video. Notably, the answers "1" and "2" represent over 89% of all responses, highlighting a significant imbalance. To address the issue, we labeled 572 QA pairs from

our collection of ensemble performance videos, specifically choosing videos where the answer to this question was neither "1" nor "2". After the balancing, We reduced the combined dominance of the "1" and "2" answers by 28% to 61%.

Audio Comparative Questions: This type of question compares different instruments sounding in the video in terms of loudness, duration and rhythm aspect. Among 3 question templates, a question template asking which instrument playing longer has a strong imbalance, where the answer "no" (indicating that neither instrument plays significantly longer) represents 68% of the data. To address this imbalance, we curated 182 QA pairs from previously unlabeled videos in the original dataset, ensuring that the first <Object> plays longer than the second <Object>. This reduces the dominance of "no" answer to 53.9%.

4. Approach

In addition to data balancing, we introduce a new model designed to set a robust baseline for MUSIC-AVQA v2.0. This model integrates existing audio-visual learning and QA components: (i). LAVISH [21], a 2-tower pretrained Swin-Transformer V2 [22] with "Audio-Visual Adapter" for audio-visual fusion. (ii). A spatial & temporal grounding module from AVST [20]. Our extension to the existing model include: (i). A new audio-visual fusion branch, the "AST branch". This extracts audio feature from a pretrained Audio-Spectrogram-Transformer (AST) and merges them with the visual feature branch from Swin-v2 backbone, leveraging audio-visual spatial and temporal grounding [20]. (ii). A cross-modal pixel-wise attention module, further refining audio-visual fusion at a granular level. A detailed overview of the model is illustrated in Fig. 3.

AST branch: Incorporating the pretrained audio feature as an auxiliary branch enables our model to capture richer semantic audio information compared to the vision pretrained transformer applied to the audio spectrogram in LAVISH. Specifically, we extract the final hidden output of the AST model, which is a spatial feature map $x_{as}^{(1)}$. We then apply the same operations as those used for the LAVISH [21] branch outputs, namely spatial and temporal grounding as introduced by AVST [20] (please refer to it for details). The feature map $x_{as}^{(1)}$ undergoes average pooling to produce a vector $x_{a1}^{(1)}$, which subsequently computes spatial attention in relation to the visual map, x_{vs} , output from LAVISH's vision branch. The resulting spatial grounding module output is a vector in each frame, $x_{av}^{(1)}$, capturing visual features attended by audio. Following this, temporal grounding is performed. $x_{av}^{(1)}$ is concatenated with audio features, which are then attended by the question vector produced by a LSTM encoder, x_q, along the temporal axis. The output of temporal grounding module is a question-
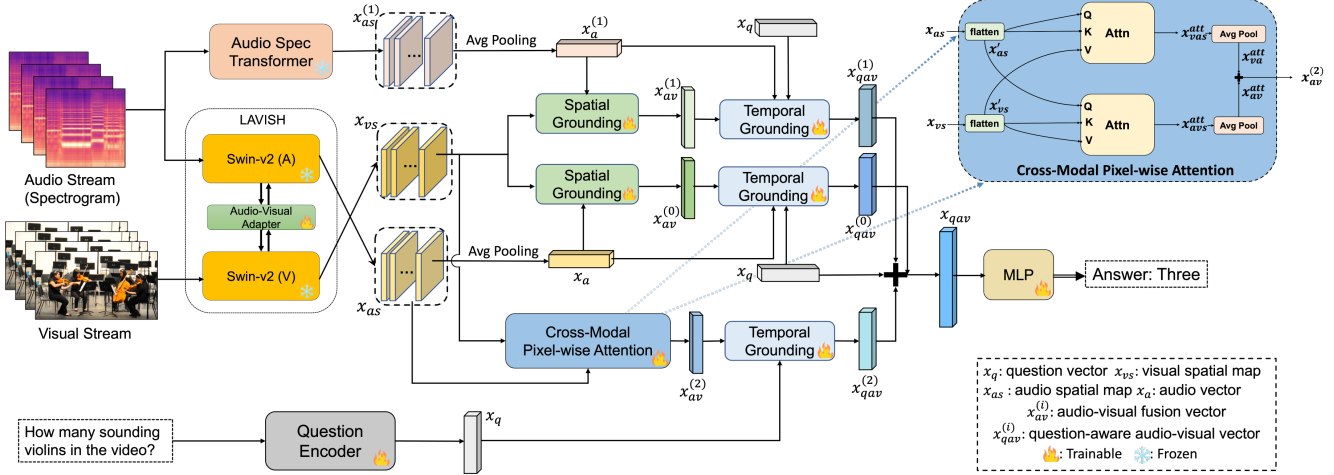
Figure 3. An overview of the new baseline model: (i). A pre-trained Audio-Spectrogram-Transformer (AST) branch is incorporated as an additional audio feature branch. (ii). A cross-modal pixel-wise attention module between audio and visual feature maps is designed to better capture the audio-visual correspondence at a granular level.

aware audio-visual fusion feature $x_{qav}^{(1)}$. This feature is concatenated with outputs from other branches and subsequntly passed to a MLP for classification.

Cross-modal pixel-wise attention module: Existing spatial grounding module [20] uses a mean-pooled audio vector to compute attention with visual spatial maps. This attention approach: i) Losing the spatial details of spectrogram features. ii) is uni-directional, where only the audio vector serves as a query to the visual maps without any reciprocal interaction. To address these limitations, we propose a refined pixel-wise cross attention between the visual and audio maps, aiming to capture the correspondence between these two modalities more effectively. Specifically, given $x_{vs} \in \mathbb{R}^{H \times W \times C}$ representing frame-level visual map output from LAVISH, and $x_{as} \in \mathbb{R}^{H \times W \times C}$ representing spectrogram frame-level feature output from LAVISH, we compute two pixel-wise audio-visual attentions between two maps. We first flatten the spatial dimension of both feature maps to be $(HW) \times C$, resulting in x'_{vs} and x'_{as}. Then we compute mutual cross-attention between these two flattened maps, where each map attends to the other :

$$x_{vas}^{att} = x'_{as} + \text{Softmax}(\frac{x'_{vs} x'^{T}_{as}}{HW})x'_{as}, \quad (1)$$

$$x_{avs}^{att} = x'_{vs} + \text{Softmax}(\frac{x'_{as} x'^{T}_{vs}}{HW})x'_{vs} \quad (2)$$

The obtained x_{vas} and x_{avs} represent the pixel-wise fusion maps for audio and vision branch. We then normalize both maps and average pool their spatial dimensions to produce two vectors, x_{av} and x_{va}. These are concatenated to yield the final module output, $x_{av}^{(2)}$. $x_{av}^{(2)}$ is same dimension as $x_{av}^{(0)}$, $x_{av}^{(1)}$ - outputs of the spatial grounding modules from the other 2 branches. As a result, subsequent opera-

tions, such as temporal grounding, remain consistent with the other two branches.

5. Experiments

5.1. Models Evaluation on Bias and Balanced Dataset

To further verify the data bias existing in the MUSIC-AVQA dataset and its negative impact to the model performance, we evaluate AVST and LAVISH models on two newly created datasets (biased and balanced datasets) here. We first create a balanced test set by sampling 20% from MUSIC-AVQA v2.0, with balanced answer distribution in every question categories using stratified sampling, which was extensively studied in works like [29]. Then within this balanced test set, we sample a bias test set by keeping the same QA distribution as MUSIC-AVQA dataset [20] (with our corrected QA pairs in "voiceover" category). Both the biased and balanced set test are used to evaluate all models.

After reserving the balanced test set from MUSIC-AVQA v2.0, the residual data remains balanced. To ensure a fair comparison, it is imperative that the training sets for both the bias and balanced datasets are of equal size. To achieve this, we first extract a bias subset from the leftover balanced data. To maximize the training samples in this bias set, any QA pair with an answer belonging to the most frequent answer category within its question template is incorporated into the biased subset.

Once the balanced test set is held out from MUSIC-AVQA v2.0, the remaining part is still a balanced dataset. To guarantee the fairness of comparison, we need to keep the size of training set same for the bias and the balanced set. To achieve this, we first sample the bias subset from the

Figure 4. Qualitative Results of 'LAST-Att' model on contrasting binary QA samples from different question categories. The model can correctly predict the pairs of QA with same question but opposite answers.

remaining balanced set. In order to maximize the number of training samples for the bias set, any QA pair whose answer is the most frequent answer category in its question template is included into the bias subset. For QA pairs in remaining answer categories in the question template, the numbers are determined by: $N_{most} \times \frac{N'_{other}}{N'_{most}}$, where N_{most} is the number of QA pairs whose answer is the most frequent answer category of the question template in our remaining balanced set, N'_{most} is the number of QA pairs whose answer is the most frequent answer category of the question template in MUSIC-AVQA training dataset, and N'_{other} is the number of QA pairs whose answer is from another less frequent category of the question template in MUSIC-AVQA training dataset. Once we create the bias subset, we sample another balanced subset by keeping the same number of samples as the bias subset. In the last step, we reserve $1/8$ for validation, with the remaining $7/8$ forming our biased and balanced training sets.

By following the outlined procedure, we create 2 datasets, a bias and a balanced set. Each dataset contains 31,513 training samples, mirroring the size of the original MUSIC-AVQA training set (approximately 31k). Additionally, both datasets share 4,502 validation samples, 10,819 balanced test samples, and 9,119 biased test samples. We then assess the performance of 2 existing open-source models, the baseline AVST model [20] and the state-of-the-art LAVISH [21] model. Both models are evaluated on the biased and balanced test sets respectively, adhering strictly to their original training guidelines without any modifications. To validate the integrity of our code and data, we trained, validated, and tested both models using the original MUSIC-AVQA dataset. The AVST model achieved a total accuracy of 71.25% on the test set (compared to the reported 71.51%), while the LAVISH model achieved 77.17% (close to the reported 77.20%). These results align closely with

the numbers reported in their publications.

We proceeded with training both models on the bias and balanced training sets, resulting in 4 distinct models. For models trained on the bias set, we use the total accuracy on the bias validation set to select the best checkpoint across epochs. Similarly, for models trained on the balanced set, we use the total accuracy on the balanced validation set for checkpoint selection. After training, we evaluated all four models on both the bias and balanced test splits. The evaluation results are presented in Table 1 for the balanced test set and in Table 2 for the bias test set. As shown in Table 1, both models trained using the balanced set achieve higher total accuracy on the balanced test set, with gains of **+1.75%** for AVST and **+2.96%** for LAVISH. Specifically, in question types where severe answer imbalance exists, such as audio-visual counting and existential questions, models trained on our balanced data consistently achieve higher accuracy on the balanced test set. For instance, the LAVISH model showed improvements of **+3.93%** and **+5.23%** respectively. Conversely, as shown in Table 2, all models trained on the bias set surpassed those trained on the balanced set when evaluated on the biased test set. These results suggest the data bias in original MUSIC-AVQA. Models trained from the data overfit to the biases, thereby undermining their ability to generalize well.

5.2. New Baseline Evaluation

To establish new baselines, we evaluate our proposed models using MUSIC-AVQA v2.0, our entire balanced dataset, comprising 36.7k QA pairs for training, 5,250 for validation, and 10,819 for testing. We compare two model variants, 'LAST' and 'LAST-Att', against existing models. The LAST variant integrates the existing LAVISH model with the AST branch, as detailed in Section 4. LAST-Att represents our full model, incorporating both the AST

Table 1. Evaluate Existing Models on Balanced Test set. Highlights are results where both models trained on balanced set consistently outperform trained on bias set. (Ext: Existential. Cnt: Counting. Temp: Temporal. Comp: Comparative.)

Model	Train set	Total	Audio-Visual					Visual		Audio	
			Ext	Temp	Cnt	Loc	Comp	Cnt	Loc	Cnt	Comp
AVST [20]	bias	69.40	70.16	61.94	62.99	63.26	63.94	75.02	78.85	77.86	63.61
	balanced	**71.15**	**71.38**	59.98	**68.85**	63.48	65.40	**77.61**	77.80	**82.13**	62.05
LAVISH [21]	bias	70.39	68.11	60.08	66.72	65.11	63.12	78.14	81.46	78.80	60.96
	balanced	**73.35**	**72.04**	63.19	**71.95**	67.07	63.40	**80.40**	82.66	**83.57**	63.14

Table 2. Evaluate Existing Models on Bias Test set (Ext: Existential. Cnt: Counting. Temp: Temporal. Comp: Comparative.)

Model	Train set	Total	Audio-Visual					Visual		Audio	
			Ext	Temp	Cnt	Loc	Comp	Cnt	Loc	Cnt	Comp
AVST [20]	bias	**73.07**	**85.68**	66.02	69.97	63.26	63.94	77.50	77.89	83.56	65.57
	balanced	72.01	75.36	64.68	70.82	63.48	65.40	77.91	76.75	84.44	62.11
LAVISH [21]	bias	**74.59**	**84.79**	67.84	73.53	65.11	63.12	80.77	81.14	84.54	63.59
	balanced	74.00	75.36	68.33	73.37	67.07	63.40	80.36	81.79	84.83	63.92

Table 3. Evaluation Results on Balanced Test Set: Our new baselines v.s existing models. (Ext: Existential. Cnt: Counting. Temp: Temporal. Comp: Comparative.)

Model	Total	Audio-Visual					Visual		Audio	
		Ext	Temp	Cnt	Loc	Comp	Cnt	Loc	Cnt	Comp
AVST [20]	71.02	72.44	59.36	68.22	65.54	63.31	77.48	77.88	82.34	60.81
LAVISH [21]	73.18	73.83	60.81	73.28	65.00	63.49	81.99	80.57	84.37	58.48
LAST	74.85	74.08	59.15	75.17	**69.02**	**66.12**	83.19	83.41	85.75	61.59
LAST-Att	**75.44**	**76.21**	60.60	75.23	68.91	65.60	**84.12**	**84.01**	**86.03**	**62.52**

Table 4. Evaluation Results on Bias Test Set: Our new baselines v.s existing models. (Ext: Existential. Cnt: Counting. Temp: Temporal. Comp: Comparative.)

Model	Total	Audio-Visual					Visual		Audio	
		Ext	Temp	Cnt	Loc	Comp	Cnt	Loc	Cnt	Comp
AVST	71.92	75.36	64.81	70.51	65.54	63.31	77.74	76.67	84.74	61.78
LAVISH [21]	73.51	74.92	66.5	75.08	65	63.49	82.08	79.59	85.32	59.14
LAST	75.24	75.58	65.78	76.16	**69.02**	**66.12**	83.63	82.36	**86.20**	61.78
LAST-Att	**75.45**	**76.47**	66.75	76.20	68.91	65.60	83.86	83.09	85.71	**63.10**

Table 5. Model Performance on Contrasting Binary audio-visual QA pairs

Model	Total Accuracy
AVST [20]	52.4
LAVISH [21]	54.47
LAST	**58.00**
LAST-Att	**58.86**

ples. Each pair contains two questions with identical phrasing but pertains to different videos that have opposite answers. For instance, the question 'Are there ukulele and violin sounds in the video?' is posed for both video A and B. While the answer for video A is "yes", it's "no" for video B. From the balanced test set, we identified 1643 such contrasting binary QA pairs. For evaluation, a model must correctly answer both questions in a pair to be deemed accurate for that pair. This ensures that the model truly comprehends the audio and visual context of each video, given that the question phrasing is the same. As shown in Table 5, in this challenging scenario, our LAST-Att model strongly outperforms the LAVISH and AVST models trained on the balanced set by margins of +4.39% and +6.46%, respectively. Some of the contrasting QA samples predicted by our LAST-Att' model are illustrated in Fig. 4. Overall, our model can correctly predict QA pairs by better reasoning the audio-visual context than relying on language priors.

6. Conclusion and Discussion

In this work, we identify the data bias that exists in the public MUSIC-AVQA dataset, which hinders the progress of multimodal research. We systematically address the data imbalance issue, and propose a new and larger balanced dataset, MUSIC-AVQA v2.0. Extensive results show that the strong bias from MUSIC-AVQA negatively affect the model ability, favoring more to dominant answers from training set. However, existing models trained on our balanced dataset do not overfit to the bias distribution and generalize well, proving that MUSIC-AVQA v2.0 is a more reliable benchmark for AVQA task. Furthermore, we propose a new model on MUSIC-AVQA v2.0. With added AST branch and an additional cross-modal pixel-wise attention, it consistently outperforms previous methods on almost every question categories, serving as a strong baseline for AVQA task.

branch and cross-modal pixel-wise attention. Further implementation specifics for both models can be found in the Supplementary Material. As shown in Table 3 and Table 4, both of our proposed baselines surpass the performance of LAVISH and AVST on both test sets. Notably, on the balanced test set, LAST-Att achieves a performance boost of **+2.26%** over LAVISH and **+4.42%** over AVST.

5.3. Contrasting Binary QA Evaluation

To assess the models' true capability in understanding audio-visual contexts and to ensure they do not merely rely on language priors, we design a contrasting binary QA evaluation. We curated a subset from the balanced test split of MUSIC-AVQA v2.0, consisting of paired binary QA sam-

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016. 4

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering, 2018. 2, 3

[3] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 1

[4] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering, 2020. 2, 3

[5] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation, 2021. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 3

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 3

[8] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045, 2021. 2

[9] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4013–4022, January 2021. 2

[10] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *ECCV*, 2020. 2

[11] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[12] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[13] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[14] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 2

[15] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder, 2023. 3

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 3

[18] Gouthaman KV and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder, 2020. 2, 3

[19] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems*, pages 3586–3596, 2019. 2

[20] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios, 2022. 2, 3, 5, 6, 7, 8

[21] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners, 2023. 3, 5, 7, 8

[22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. 5

[23] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4733–4744. Curran Associates, Inc., 2020. 3

[24] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, June 2021. 3

[25] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018. 2

[26] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[27] Kun Su, Xiulong Liu, and Eli Shlizerman. Multi-instrumentalist net: Unsupervised generation of music from body movements. *arXiv preprint arXiv:2012.03478*, 2020. 2

[28] Kun Su, Xiulong Liu, and Eli Shlizerman. How does it sound? *Advances in Neural Information Processing Systems*, 34, 2021. 2

[29] Piotr Szymański and Tomasz Kajdanowicz. A network perspective on stratification of multi-label data. In Paula Branco Luís Torgo and Nuno Moniz, editors, *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35. PMLR, 22 Sep 2017. 6

[30] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation, 2021. 2

[31] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[32] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos, 2018. 3

[33] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[34] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19989–19998, June 2022. 2

[35] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 279–286, 2020. 2

[36] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491, 2022. 2, 3

[37] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360° videos, 2021. 2, 3

[38] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016. 2, 3

[39] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019. 2

[40] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2

[41] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation, 2023. 3