Stochastic Parrots or True Virtuosos?
Digging Deeper Into the Audio-Video Understanding of AVQA Models

Sara Pernille Jensen
Department of Philosophy, Classics, History of Art and Ideas
University of Oslo
sarapje@ifikk.uio.no

Hallvard Innset Hurum
Department of Informatics
University of Oslo
hallvaih@ifi.uio.no

Anna-Maria Christodoulou
RITMO,
Department of Musicology
University of Oslo
annammc@uio.no

Abstract

Audio-video question answering (AVQA) systems for music show signs of multimodal "understanding", but it is unclear which inputs they rely on or whether their behavior reflects genuine audio-video reasoning. Existing evaluations focus on overall accuracy and rarely examine modality dependence. We address this gap by suggesting a method of using counterfactual evaluations to analyse the audio-video understanding of the models, illustrated with a case study on the audio-video spatial-temporal (AVST) architecture. This includes interventions that zero out or swap audio, video, or both, where results are benchmarked against a baseline based on linguistic patterns alone. Results show stronger reliance on audio than video, yet performance persists when either modality is removed, indicating learned cross-modal representations. The AVQA system studied thus exhibits non-trivial multimodal integration, though its "understanding" remains uneven.

1 Introduction

An increasingly popular use of machine-learning (ML) models is computational music analysis through audio, language, and video (Manco et al., 2022; Simonetta et al., 2019; Li et al., 2022; Christodoulou et al., 2025). Such analyses are believed to require a degree of *understanding* of music, motivating research into how ML models reason about it. Audio-video question–answering (AVQA) (Li et al., 2022) is a computational task that requires a model to answer questions based on both audio and video.

Various architectures have been proposed to improve music-related AVQA performance (Lin et al., 2023; Li et al., 2022; Christodoulou et al., 2025), but benchmark scores alone reveal little about how models reason from audio-video inputs. Understanding these dependencies can guide improvements in both models and data.

Counterfactual interventions offer a way to probe multimodal reasoning by systematically manipulating inputs and observing the ensuing changes in predictions, revealing dependencies, biases, and failure modes. As a case study, we apply such analyses to an AVST model (Li et al., 2022) trained on the MusiQAl dataset (Christodoulou et al., 2025). We introduce interventions that zero out modalities or create conflicting cross-modal pairs, and examine changes in predictions and confidence. We also analyse dataset patterns to distinguish multimodal reasoning from biases. Results indicate meaningful audio-video processing and possibly understanding, but uneven reliance on modalities, highlighting directions for future music-oriented QA models.

2 Related work

The AVST model (Li et al., 2022) was one of the first to explore spatial-temporal reasoning in musical performances. Tested on the MUSIC-AVQA dataset with over 45,867 QA pairs across five categories (existential, counting, location, comparative, and temporal) the model includes spatial and temporal grounding modules, enabling localization of instrument sounds and reasoning about their timing. Multimodal inputs improved performance over audio- or video-only approaches. The question templates in the MUSIC-AVQA dataset include placeholders, e.g., "Which instrument makes sounds ⟨BA⟩ the ⟨Object⟩?", with ⟨BA⟩ as "before/after" and ⟨Object⟩ as a musical object. Questions are also labeled as audio, video, or audio-video, indicating which modality is presumably needed to answer the question, though the model does not receive this information during training.

The MusiQAl dataset (Christodoulou et al., 2025) contains 310 videos and 11,793 QA pairs spanning diverse cultures and genres. Based on the MUSIC-AVQA's question templates, it extends this by introducing 47 new ones, as well as including

Proceedings of the 4th Workshop on NLP for Music and Audio (NLP4MusA 2026), pages 13–19

two new question categories, 'Causal' and 'Purpose', to probe deeper understanding of music performance. Christodoulou et al. (2025) trained both AVST (Li et al., 2022) and LAVISH (Lin et al., 2023) models on MusiQAl. A schematic of the pipeline for how the data is split up, fed into the model, and used for prediction is shown in Fig. 1. The QA task is treated as classification over the answer vocabulary, with the model output interpreted as probabilities via softmax; the highest-probability answer is taken as the prediction. Results revealed strengths in areas like performer tracking, but highlight that further development is needed in complex scenarios like style classification or source separation in ensemble performances. The overall accuracy was $\approx 71\%$.

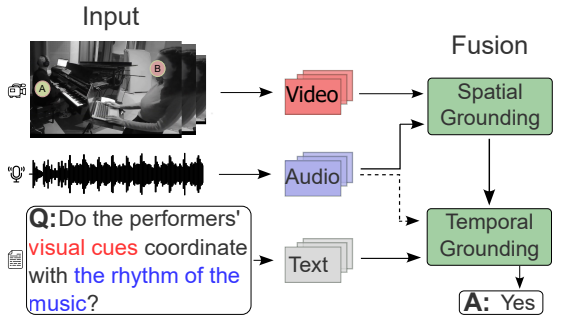

Figure 1: Illustration of the basic components of the ML model used in (Christodoulou et al., 2025) during inference, showing how the audio, video, and question are inputted to the model, and used for making the final prediction of the answer. Figure borrowed from (Christodoulou et al., 2025) with permission.

Recent work has also examined what multimodal models actually learn (Weck et al., 2024; Zang et al., 2025). Both studies analysed music QA models, investigating to what extent performance relied on processing the auditory or linguistic input data, and hence the type of understanding it is indicative of. Strong linguistic biases were found, as models often reached near-original accuracy using only the questions. This highlights the need for careful evaluation of which modalities models truly exploit.

3 Methods

To analyse how the multimodal AVST model reasons based on the different modalities, we applied a set of counterfactual interventions to the inputs and observed the resulting changes in prediction accuracy. We also examined the distribution of the model's confidence in its predicted answers to assess how its certainty shifts in response to the interventions. This constitutes a black-box interpretability approach, as we did not inspect changes in the model's internals.

The goal was to investigate the model's relative reliance on the different modalities, on which we systematically intervened, using two methods: modality ablations and modality conflicts. In addition, we derived estimates of the chance of correct predictions based on random guessing and statistical patterns in the linguistic question–answer pairs alone, and used these as a baseline for evaluating the model's relative success. All models were trained by us according to the specifications given in (Christodoulou et al., 2025). All figures are included in the accompanying appendix.

3.1 Modality Ablation

Modality ablation involved removing one or both modalities by setting the audio and/or video input tensors to zero, forcing the model to rely on the question alone or on the question and a single modality. Changes in prediction accuracy under these conditions indicate which modality the model relies on most heavily, both overall and across question categories.

To further understand differences in accuracy, test samples were also split based on whether the model answered them correctly when given the complete input. This highlights not only overall accuracy changes, but also how the modalities given affect the predictions in both directions, for better and worse. Model confidence under each ablation was also examined.

A potential limitation is that the model was trained on non-zero inputs, so zeroed tensors may not be interpreted as silence or blank images, as intended. To mitigate this, three additional models were trained with the same dataset, architecture, and hyperparameters, but with audio, video, or both modalities ablated, providing a reference to validate the ablation results for the main model.

3.2 Modality Conflict

The modality conflict intervention mixes audio and video across different question–answer pairs, providing the model with signals that do not fit together. We observe whether the model can still answer questions that depend on a single modality, and how it handles conflicting inputs when given questions that depend on both modalities.

We tested three types of conflict interventions: one in which the audio was replaced with an in-

correct sample (relative to the question), while the video was kept unchanged, one in which the video was replaced with an incorrect sample, while the audio remained the same, and one in which both audio and video were replaced with incorrect samples. Replacements were drawn randomly from the 310 unique performances in the dataset. Since performances are unevenly distributed across QA pairs, each of the 310 modalities was used to replace 3-4 samples ($\frac{1183}{310} \approx 3.8$) to reduce representation bias. For each conflict type, five independent runs were performed with different reproducible seeds, ensuring unique replacements across runs. In the "both" mode, audio and video were replaced with samples from separate performances.

3.3 Dataset Analysis

To investigate what the model had learned, we analysed the chances of success assuming no audio-video understanding. Positive evidence alone is insufficient to infer a hypothesis; plausible alternative hypotheses must also be considered and falsified (Reiss, 2015). The main hypothesis was that the model had learned to understand audio-video content across sub-categories, rather than relying on simpler strategies such as random guessing or exploiting statistical patterns in the question–answer pairs. For example, always predicting the most frequent answer for a given question template provides a baseline that requires no use of audio or video inputs, and consequently does not reflect genuine audio-video understanding.

These alternatives were formalised by computing expected accuracy from random guessing and from always predicting the most frequent answer per question template. These baselines allow for a deeper analysis of the model performance across question categories. Performance above them indicates reliance on the audio-video inputs, while performance at or below could be explained by the model's relying on linguistic patterns alone. Expected accuracy was calculated per question template, accounting for the actual number of possible answers, which varies (e.g., "yes/no?", versus "which instrument?"), and reflects the subset of answers the model could plausibly infer.

4 Results

4.1 Dataset Analysis

The expected accuracy on test data from uniform guessing is 32.14%, whereas always predicting the most frequent answer from the training data, based on the question template, gives an accuracy of 47.84%. The reason for the high uniform accuracy is that many of the questions only have a few possibly answers, with 70% of the samples having five or fewer alternatives.

Appendix Fig. 2 compares guessing strategies across question categories. Frequentist guessing generally outperforms uniform guessing, reflecting learnable linguistic patterns, such as common instrument usage. The AVST model surpasses both baselines in almost all categories. However, sample sizes differ widely, with some categories only including a few questions, which should be considered when interpreting evidence for understanding.

4.2 Modality Ablation

Overall test accuracies for the different models and modality combinations are given in Table 1. First, the standard model trained on complete data was tested with video, audio, and both modalities removed. Appendix Fig. 3 shows that removing video slightly reduces accuracy, audio more so, and both modalities cause a substantial drop. There is no clear alignment between the question label (audio, video, or both) and sensitivity to ablations, but overall, the model depends more on audio than video.

To further probe this, the subset of test questions answered incorrectly with full input was analysed under the same ablations. Removing both modalities gives the highest accuracy in this subset, audio second, and video lowest, showing which ablations lead to the greatest changes in predictions. The inverse pattern is found when looking at the questions that were answered correctly when given complete input. Both findings are consistent with the previous results in terms of relative modality significance. The resulting overall accuracy for the different interventions, on both the initially incorrectly answered questions and the complementary subset (correctly answered questions), is given in Table 1.

Separate models trained on ablated data (audio, video, or both removed) show only marginally better performance than applying ablations to the standard model, confirming the validity of the method. Hyperparameters were not tuned for these alternative models, so results are indicative rather than precise measures of information extracted from each modality.

Appendix Fig. 4 shows the distribution of pre-

Table 1: Overall test accuracy for the standard model (except last column: models trained on corresponding ablated data) under specified modality ablations. 'Success'/'Failure' indicates subsets of test data where the standard model succeeds/fails with full input.

Modality Excluded	Standard	Success	Failure	Alt. mod.
None	70.58%	100%	0%	—
Video	66.27%	87.78%	14.66%	67.79%
Audio	56.13%	70.78%	20.98%	58.50%
Both	46.49%	54.13%	28.16%	50.80%

dicted probabilities for correct and incorrect answers under each ablation. Correct predictions are most confident with full data (peak approximately 90%, mean 76%), slightly lower with audio only, lower with video only, and lowest with questions alone. Incorrect predictions show similar distributions across ablations. Differences are likely underestimated, since questions with more possible answers naturally have lower maximum predicted probabilities; models with more modalities can answer these complex questions, lowering the average predicted probability.

4.3 Modality conflict

We examine how the model responds to conflicting audio and video signals.

Table 2 shows overall accuracy under different conflict conditions compared to the baseline without conflict (approximately 71%). Video conflict reduced accuracy to 61%, audio conflict to 51%, and conflicts in both modalities to 44%, again indicating greater reliance on audio than video.

Table 2: Overall test accuracy for the standard model under modality conflicts. 'Success'/'Failure' indicates subsets of test data where the standard model succeeds/fails with full input.

Modality Conflicted	Standard	Success	Failure
None	70.58%	100%	0%
Video	61.05%	79.16%	16.72%
Audio	50.97%	65.25%	20.34%
Both	43.79%	53.84%	24.31%

The 'success' and 'failure' subsets show complementary trends. In the failure subset, video conflict allowed the model to correct 17% of previously incorrect answers, audio conflict 20%, and both conflicts 25%. This confirms that audio conflicts impact predictions more than video conflicts, consistent with overall findings.

Detailed subcategory analysis (Appendix Fig. 5) reveals that audio conflicts consistently reduce performance in audio and audio-video questions. Video-existential questions remain largely unaffected by video conflict, likely due to language bias (e.g., the correct answer to "Is the dancer in the video always dancing?" is "yes" 80% of the time). Audio-video comparative questions maintain high accuracy under all conflicts, possibly due to language bias in yes/no questions (e.g., "Do the performers' video cues coordinate with the rhythm of the music?"; 94% "yes"). Audio-video temporal questions drop sharply under audio conflict, indicating that they are primarily audio-driven. Some audio-video questions may be mislabelled, relying almost entirely on one modality.

To probe model certainty, we analysed the highest softmax probability per prediction (Appendix Fig. 6). Confidence was higher for correct answers (74–76%) than incorrect ones (57–61%). Correct predictions show small differences across conflict types, highest with no conflict (76%). Incorrect predictions show increased confidence under dual-modality conflict (61%), suggesting the model becomes overconfident when inputs conflict. Variations in accuracy do not always reflect variations in probability distributions due to differing numbers of correct/incorrect samples per conflict type.

5 Discussion and Conclusion

Across all experiments, the AVST model's performance cannot be explained solely by random guessing or question–answer statistics, but it is clearly shaped by dataset bias and asymmetric modality use. Relatively high baseline accuracies (32% for uniform guessing and 48% for frequentist guessing) reveal strong linguistic and structural biases in the dataset. Nevertheless, the model's overall test accuracy of 71% indicates that it leverages information from audio and video inputs beyond question priors.

Modality ablation and conflict experiments reveal a clear hierarchy in modality importance: audio is the dominant modality, with audio removal or conflict leading to substantially larger performance drops than corresponding video interventions. When both modalities are removed or conflicted, accuracy falls below the frequentist baseline but remains above uniform guessing, suggesting that the model integrates multimodal inputs while still relying on residual language cues. The weak

alignment between nominal question labels and ablation sensitivity further highlights limitations in the dataset's labelling scheme, particularly for questions that are answerable from either modality (Audio *or* Video, instead of Audio-Video). This distinction was not included in the labelling scheme, which it should be in future work.

Subcategory analyses show that several nominally multimodal questions are effectively unimodal or driven by answer imbalances, while temporal questions rely primarily on audio. Increased confidence under conflicting inputs suggests limited reasoning about cross-modal inconsistency and a tendency towards overconfident decisions when evidence disagrees.

Overall, the model processes audio and video in a fused representation but relies predominantly on auditory cues and linguistic biases. Performance improvements over baseline, therefore, reflect a combination of genuine multimodal processing and shortcut learning, motivating the need for improved question design and evaluation methods that more directly probe cross-modal reasoning.

As this investigation was limited to a single AVST model, it is left for future work to determine whether these findings generalise to other models trained for similar tasks. Yet, we believe the results are sufficient to motivate the need for probing deeper into the somewhat superficial accuracy scores that are usually reported, providing a richer understanding of how such models work.

Ethics and Consent

This study builds on (Christodoulou et al., 2025), using their publicly available AVST model and dataset. Models were independently trained to verify reproducibility. The dataset contains diverse musical cultures and question-answer annotations; we relied on the original annotations.

Acknowledgements

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 262762.

6 Open Access Statement

Our GitHub repository with all code and instructions to reproduce our results can be found at `https://github.com/silyeah/MusiQAl/tree/in5490`.

References

Anna-Maria Christodoulou, Kyrre Glette, Olivier Lartillot, and Alexander Refsum Jensenius. 2025. Musiqal: A dataset for music question–answering through audio–video fusion. *Transactions of the International Society for Music Information Retrieval*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. *Preprint*, arXiv:2203.14072.

Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2023. Vision transformers are parameter-efficient audio-visual learners. *Preprint*, arXiv:2212.07983.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Julian Reiss. 2015. A pragmatist theory of evidence. *Philosophy of Science*, 82(3):341–362.

Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. 2019. Multimodal music information processing and retrieval: Survey and future challenges. In *Proceedings - 2019 International Workshop on Multilayer Music Representation and Processing, MMRP 2019*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *Preprint*, arXiv:2408.01337.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-qa benchmarks. *Preprint*, arXiv:2504.00369.

Appendix

Due to page limitations, all supplementary visualizations of results are included in the appendix. These figures provide additional context and insights relevant to the analyses presented in the main manuscript.

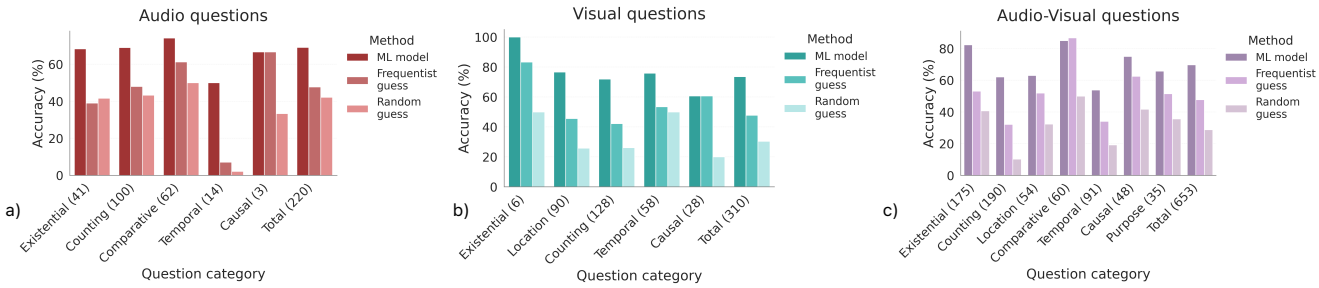

Figure 2: Comparison of AVST model performance with random and frequentist guessing across question categories. Total datapoints per category are indicated in brackets. Subgraphs: a) audio questions, b) video questions, c) audio-video questions.

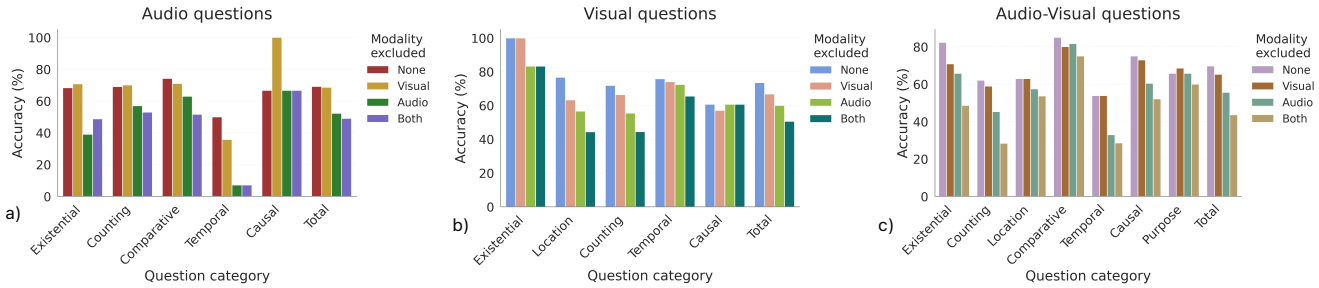

Figure 3: Test accuracy for the standard model with different modalities removed, across question categories. Subgraphs: a) audio, b) video, c) audio-video questions.

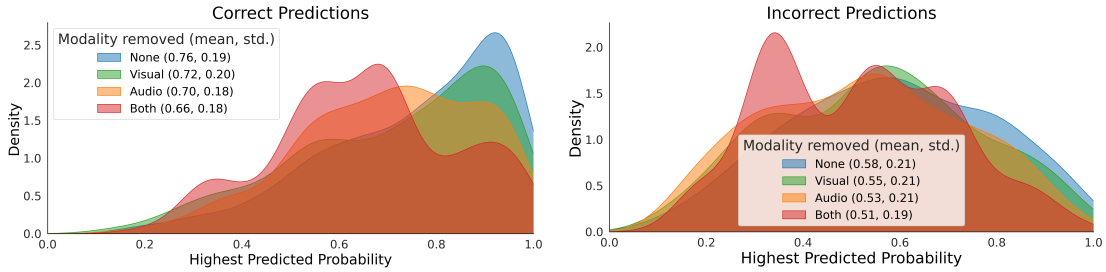

Figure 4: Distribution of the highest predicted probability for the standard model under modality ablations. Left: correct answers, right: incorrect answers. Mean and standard deviation are indicated.

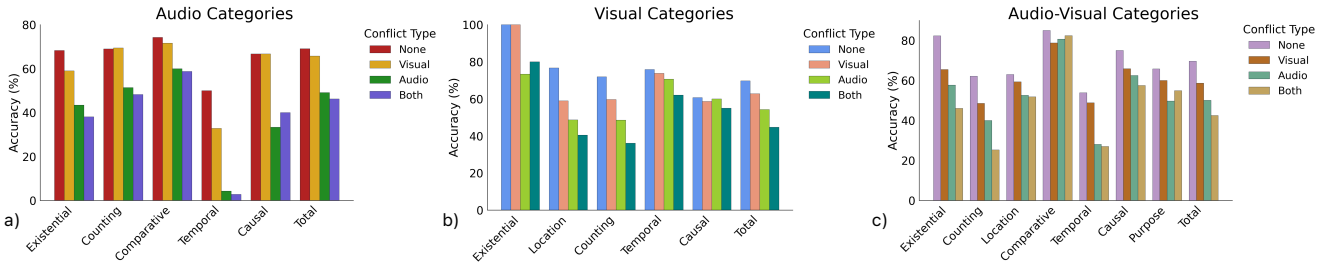

Figure 5: Overall accuracy on the test dataset with conflicts applied for a) Audio Sub-categories b) Video Sub-categories c) Audio-video Sub-categories.

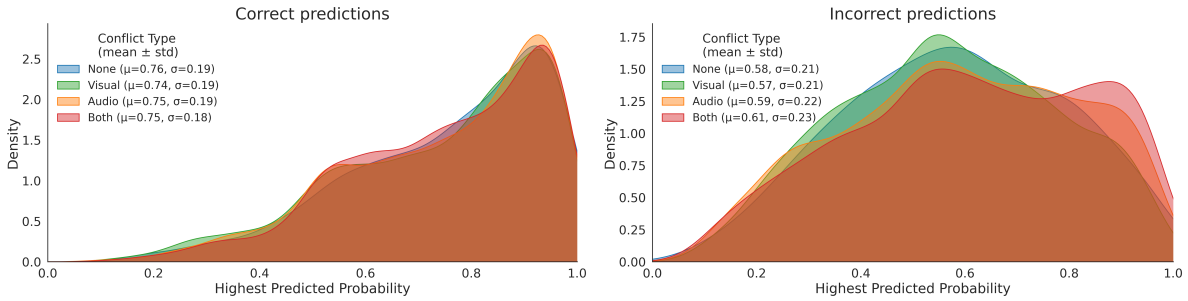

Figure 6: Distribution of highest predicted probability for the standard model on test data, for different conflict types. Mean highest predicted probability with standard deviation included. Correct answers shown to the left, incorrect to the right.