Learning Sparsity for Effective and Efficient Music Performance Question Answering

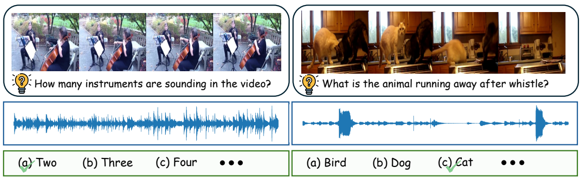
Xingjian Diao[1], **Tianzhen Yang**[2], **Chunhui Zhang**[1],
Weiyi Wu[1], **Ming Cheng**[1], **Jiang Gui**[1]
[1]Dartmouth College, [2]Yale University
xingjian.diao.gr@dartmouth.edu

Abstract

Music performances, characterized by dense and continuous audio as well as seamless audio-visual integration, present unique challenges for multimodal scene understanding and reasoning. Recent Music Performance Audio-Visual Question Answering (Music AVQA) datasets have been proposed to reflect these challenges, highlighting the continued need for more effective integration of audio-visual representations in complex question answering. However, existing Music AVQA methods often rely on dense and unoptimized representations, leading to inefficiencies in the isolation of key information, the reduction of redundancy, and the prioritization of critical samples. To address these challenges, we introduce `Sparsify`, a sparse learning framework specifically designed for Music AVQA. It integrates three sparsification strategies into an end-to-end pipeline and achieves state-of-the-art performance on the Music AVQA datasets. In addition, it reduces training time by 28.32% compared to its fully trained dense counterpart while maintaining accuracy, demonstrating clear efficiency gains. To further improve data efficiency, we propose a key-subset selection algorithm that selects and uses approximately 25% of MUSIC-AVQA v2.0 training data and retains 70–80% of full-data performance across models.

1 Introduction

Music performances, with their dense, continuous audio and seamless audio-visual integration, present both challenges and opportunities for multimodal scene understanding and reasoning (You et al., 2025; Diao et al., 2024). To address the complexities of audio-visual reasoning in music scenarios, the task of Music Performance Audio-Visual Question Answering (Music AVQA) has been proposed, and the corresponding datasets, MUSIC-AVQA (Li et al., 2022) and its extended version MUSIC-AVQA v2.0 (Liu et al., 2024), with an example shown in Figure 1(a), have been introduced to facilitate research in this emerging area.


How many instruments are sounding in the video?
What is the animal running away after whistle?
(a) Two
(b) Three
(c) Four
•••
(a) Bird
(b) Dog
(c) Cat
•••

(a) QA with **Dense Audio** from MUSIC-AVQA v2.0 (b) QA with **Sparse Audio** from VGG-Sound

Figure 1: Dense Audio QA (Liu et al., 2024) vs. Sparse Audio QA (Chen et al., 2020). Music performances contain dense and continuous audio signals with substantial inherent redundancy, much of which is irrelevant to the question being asked. **Sparse learning** has the potential to effectively filter out such redundancies, enabling more efficient and accurate reasoning.

Existing AVQA methods for music performances have evolved from early cross-modality learning approaches developed for speech recognition (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012) to more recent advancements in multimodal fusion (Yun et al., 2021; Yang et al., 2022), positive-negative pair construction (Li et al., 2022), and state-of-the-art models such as LAVisH (Lin et al., 2023), which adapts pretrained ViTs for cross-modal learning, and DG-SCT (Duan et al., 2023), which employs audio-visual prompts within frozen encoders to enhance reasoning. However, current Music AVQA methods face significant limitations in effectively modeling sparse representations, which are crucial for addressing the unique challenges posed by Music AVQA tasks. These limitations include: ① an overreliance on dense, unoptimized representations that struggle to isolate key information from dense audio-visual signals (Ye et al., 2024; Diao et al., 2024); ② a lack of effective redundancy reduction mechanisms, resulting in inefficiencies during feature extraction and model inference (Shang et al., 2024); ③ the absence of prioritization strategies to identify task-critical samples, which limits scalability and prolongs training times (Qin et al., 2024; Li et al., 2023a).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 136–146
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

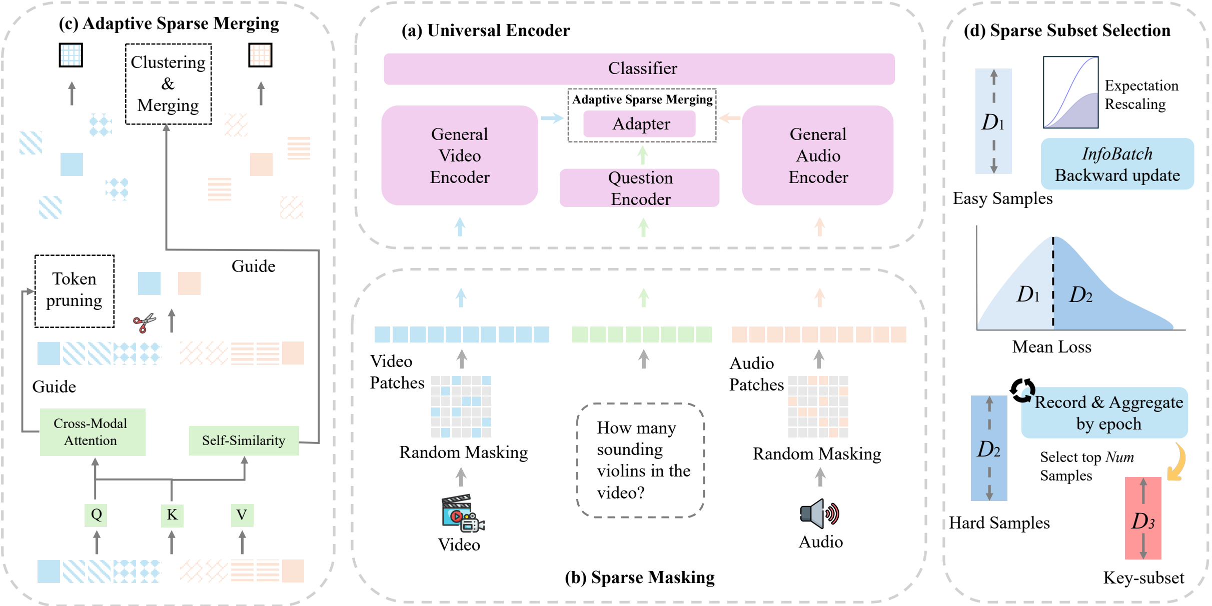


Figure 2: Overview of the Sparsify framework. Sparsify integrates a (a) Universal Encoder and three sparsification components: (b) Sparse Masking to reduce redundancy by masking audio and visual tokens; (c) Adaptive Sparse Merging to select and merge key multimodal tokens based on similarity; and (d) Sparse Subset Selection to prioritize impactful samples and reweight gradients with $InfoBatch$ (Qin et al., 2024).

To address these limitations, we propose Sparsify, a sparse learning framework designed for Music AVQA tasks. Our contributions are:

- We present an end-to-end pipeline for Music AVQA that integrates three sparsification strategies and demonstrate its effectiveness with state-of-the-art results on the MUSIC-AVQA datasets.
- Sparsify reduces training time by 28.32% while maintaining the accuracy of question answering compared to fully trained dense models, demonstrating notable efficiency improvements.
- We introduce a key-subset selection algorithm that reduces the training dataset size by approximately 75%, while retaining about 70-80% of the original performance across AVQA models.

2 Sparsify Framework

2.1 Learning Multimodal Representations

Sparsify adapts the Amuse framework (Diao et al., 2024) as its Universal Encoder, which includes a General Video Encoder built on Swin-V2 (Liu et al., 2022), a General Audio Encoder leveraging the HTS-Audio Transformer (Chen et al., 2022a), and a Question Encoder based on a standard language transformer (Vaswani et al., 2017), as shown in Figure 2(a). Cross-modal attention is applied to align features across modalities, followed by activation and linear transformation layers, resulting in unified and informative multimodal representations tailored to Music AVQA tasks.

2.2 Sparse Masking for Redundancy Reduction

Music performance data inherently contain substantial redundancies, which pose significant challenges to efficient multimodal learning. Sparse Masking addresses this issue by enforcing structured sparsity to reduce redundancy and enhance computational efficiency. As illustrated in Figure 2(b), the method draws inspiration from recent advances in random masking for multimodal models (Li et al., 2023a). This approach aligns with the objectives of sparse learning, aiming to improve efficiency while preserving model performance.

In the visual modality, Sparse Masking is applied to randomly mask 50% of the image patches, reducing input redundancy and introducing structured sparsity to encourage more efficient visual encoding. For the audio modality, we first convert raw waveforms into mel-spectrograms and apply the same masking ratio to ensure comparable sparsity levels. This unified masking design supports coherent sparsity across modalities through consistent information reduction, contributing to more effective multimodal representation learning.

2.3 Sparse Merging for Token Consolidation

In music performance AVQA tasks, dense multimodal inputs often include redundant tokens that unnecessarily increase computational overhead while diluting critical task-relevant information. To tackle this challenge, we adapt the PruMerge method proposed by Shang et al. (2024), applying it to the audio-visual setting as illustrated in Figure 2(c). This strategy dynamically prioritizes and consolidates tokens based on their significance, aligning with the objectives of sparse learning to enhance efficiency and preserve meaningful representations. Following Shang et al. (2024), our approach evaluates token importance using cross-modal attention scores $\mathbf{a} = \text{softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}^{\mathbf{T}}}{\sqrt{d}}\right)\mathbf{V}$, where query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) interactions highlight the relevance of each token. We then apply the Interquartile Range (IQR) method to these scores, dynamically identifying tokens within the top quartile of importance. IQR is particularly effective for filtering noise and ensuring robustness in token prioritization by focusing on outliers that represent highly salient features. Once the key tokens are identified, remaining tokens are clustered and adaptively merged with the closest key tokens according to their similarity, calculated as $\text{Similarity}(\mathbf{tok}_i, \mathbf{tok}_j) = \mathbf{k}_i \cdot \mathbf{k}_j^T$. This merging process retains critical tokens while integrating complementary features, preserving representational integrity. Sparse Merging ensures efficient multimodal integration with aligned sparsity across audio and visual modalities.

2.4 Sparse Subset Selection for Efficient Training

Training on dense audio-visual datasets is computationally expensive due to excessive redundancy. Sparse Subset Selection, illustrated in Figure 2 (d), addresses this by identifying and focusing on a key subset of samples that contribute the most to learning, significantly reducing training costs while preserving performance.

Following Qin et al. (2024), our method divides samples into "hard-to-learn" (D_1) and "easy-to-learn" (D_2) categories based on their loss values relative to the mean. Hard samples (D_1) are recorded and aggregated by epoch, with their importance weighted by a decay ratio r over k-epoch intervals. This ensures that difficult samples are prioritized early in training, while less critical samples are deprioritized over time. The top num samples with the highest aggregated scores are selected to form the final Key-subset (D_3). $InfoBatch$ (Qin et al., 2024) is used to rescale gradients, pruning redundant "easy-to-learn" samples (D_2) and ensuring that the reduced dataset retains the statistical properties of the original. This combination minimizes redundancy, accelerates convergence, and maintains task performance. The detailed Key-subset Selection Algorithm is presented as Algorithm 1.

Algorithm 1: Key-subset Selection Algorithm

```
Input: Model $M$, Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $N$
       samples, Loss function $L$, Number of epochs
       $E$, Merge group size $k$, Decrement ratio $r$,
       Number of key samples $n$
Output: Key-subset indices $\mathcal{K}$
# Initialization
Initialize scores vector $\mathbf{s} \leftarrow \mathbf{1} \in \mathbb{R}^N$
Initialize epochs list EpochsList $\leftarrow$ [ ]
# Compute original losses
for $i \leftarrow 1$ to $N$ do
    Compute loss $l_i \leftarrow L(M(x_i), y_i)$
    Update score $s_i \leftarrow l_i$
# InfoBatch
for epoch $e \leftarrow 1$ to $E$ do
    Initialize temporary count vector $\mathbf{t} \leftarrow \mathbf{0} \in \mathbb{R}^N$
    Compute mean loss $\mu \leftarrow \frac{1}{N}\sum_{i=1}^N s_i$
    for $i \leftarrow 1$ to $N$ do
        Compute loss $l_i \leftarrow L(M(x_i), y_i)$
        Update score $s_i \leftarrow l_i$
        if $s_i > \mu$ then
            Increment count $t_i \leftarrow t_i + 1$
    Append $\mathbf{t}$ to EpochsList
# Merge
Initialize merged scores $\mathbf{m} \leftarrow \mathbf{0} \in \mathbb{R}^N$
Compute number of groups $G \leftarrow \lceil \frac{E}{k} \rceil$
for group $g \leftarrow 1$ to $G$ do
    Compute group weight $w_g \leftarrow r^{g-1}$
    for epoch $e \leftarrow (g-1) \cdot k + 1$ to $\min(g \cdot k, E)$
     do
        Update merged scores
          $\mathbf{m} \leftarrow \mathbf{m} + w_g \cdot \text{EpochsList}[e]$
# Select the top $n$ indices as the Key-subset
$\mathcal{K} \leftarrow \text{argsort}(-\mathbf{m})[:n]$
return Key-subset indices $\mathcal{K}$
```

3 Experiments

3.1 Setup

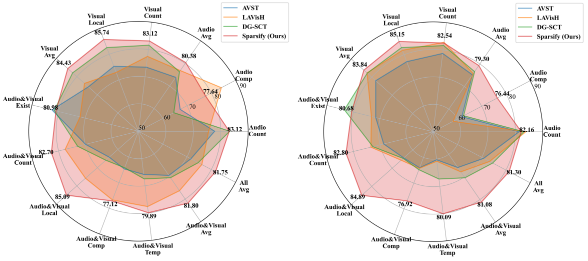
Music AVQA Datasets *(i)* MUSIC-AVQA (Li et al., 2022): Contains 9,288 videos (150 hours) spanning 22 instruments, with 45,867 QA pairs derived from 33 templates across four categories: String, Wind, Percussion, and Keyboard. Each video includes approximately five QA pairs on average. *(ii)* MUSIC-AVQA v2.0 (Liu et al., 2024): An extension addressing data bias, introducing 1,230 additional videos and 8,100 new QA pairs.

Method	Audio-related QA			Visual-related QA			Audio&Visual-related QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
MUSIC-AVQA													
AVST (Li et al., 2022)	77.78	67.17	73.87	73.52	75.27	74.40	**82.49**	69.88	64.24	64.67	65.82	69.53	71.59
LAVisH (Lin et al., 2023)	75.59	**84.13**	76.86	77.45	72.91	76.29	71.91	77.52	75.81	76.75	77.62	76.31	76.10
DG-SCT (Duan et al., 2023)	**83.27**	64.56	76.34	81.57	82.57	82.08	81.61	72.84	65.91	64.22	67.48	70.56	74.62
Sparsify (Ours)	83.12	77.64	**80.38**	**83.12**	**85.74**	**84.43**	80.98	**82.70**	**85.09**	**77.12**	**79.89**	**81.80**	**81.75**
MUSIC-AVQA v2.0													
AVST (Li et al., 2022)	81.38	61.82	75.20	78.72	77.29	78.05	71.63	68.62	64.39	64.03	60.29	65.83	70.83
LAVisH (Lin et al., 2023)	**83.82**	58.19	75.72	**82.81**	81.73	82.30	73.26	73.45	65.64	64.26	60.82	67.75	73.28
DG-SCT (Duan et al., 2023)	83.13	62.54	76.62	81.61	82.76	82.19	**83.43**	72.70	64.65	64.78	67.34	70.38	74.53
Sparsify (Ours)	82.16	**76.44**	**79.30**	82.54	**85.15**	**83.84**	80.68	**82.80**	**84.89**	**76.92**	**80.09**	**81.08**	**81.30**

Table 1: Comparison with state-of-the-art methods on the MUSIC-AVQA and MUSIC-AVQA v2.0 test set. Accuracy is reported for Audio (Counting, Comparative), Visual (Counting, Location), and Audio-Visual (Existential, Counting, Location, Comparative, Temporal) question types. Average accuracies for Audio, Visual, Audio-Visual, and Overall are also included. **Bold** marks the best results, and underlined marks the second-best.

Full Dataset Training Configuration For the experiments described in Section 3.2, Sparse Masking is applied during the first three epochs and is disabled thereafter. Adaptive Sparse Merging and $InfoBatch$ are used throughout the training. We set the masking rate to 50% for Sparse Masking. For $InfoBatch$, the ratio is set to 0.5 and the delta to 0.875, following the setup in (Qin et al., 2024).

Key-subset Selection Configuration In the key-subset selection, we apply a one-epoch warm-up phase followed by 15 epochs of training. The hyperparameters are set as follows: $N = 15$, $k = 3$, $r = 0.618$, and $Num = 10{,}819$ (i.e., the number of QA pairs). During this stage, only $InfoBatch$ (Qin et al., 2024) is employed, while Sparse Masking and Adaptive Sparse Merging are disabled.



(a) Results on MUSIC-AVQA

(b) Results on MUSIC-AVQA v2.0

Figure 3: Radar charts comparing Sparsify with state-of-the-art methods on MUSIC-AVQA and MUSIC-AVQA v2.0, across various question types.

3.2 Comparison with State-of-the-Art

Table 1 presents a detailed comparison between Sparsify and three strong baselines—AVST (Li et al., 2022), LAVisH (Lin et al., 2023), and DG-SCT (Duan et al., 2023)—on both the MUSIC-AVQA and MUSIC-AVQA v2.0 benchmarks. Sparsify achieves the highest overall average accuracy on both datasets, with consistent gains across audio, visual, and audio-visual question types. These improvements highlight the potential of sparse learning in handling the dense and continuous nature of music performance videos.

Sparse learning benefits visual question answering by promoting compact visual representations. On visual-related QA, Sparsify achieves accuracy scores of 84.43% and 83.84% on MUSIC-AVQA and MUSIC-AVQA v2.0, respectively, outperforming DG-SCT by +2.35% and +1.65%, and surpassing LAVisH by +8.14% and +1.54%. These improvements reflect the advantage of sparse inputs in retaining essential spatial and structural cues while reducing visual redundancy. In music performance QA, such representations better support reasoning over complex scenes involving performer locations, interactions, and visual composition. By limiting the influence of background clutter and redundant details, sparse visual representation enables the model to perform more robustly across diverse and fine-grained visual reasoning types.

Sparse learning supports audio question answering by reducing spectral redundancy and enabling efficient acoustic encoding. On audio-related QA, Sparsify achieves gains of +3.52% and +3.58% over LAVisH on MUSIC-AVQA and MUSIC-AVQA v2.0, respectively, and outperforms DG-SCT by +4.04% and +2.68%. These improvements suggest that sparse input representations help suppress redundant frequency patterns while retaining sufficient acoustic detail for question-relevant reasoning. By pruning less informative segments in the spectrogram, Sparsify yields more compact yet informative representations, supporting improved performance on audio-based queries involving complex musical content.

Sparse learning improves audio-visual ques-

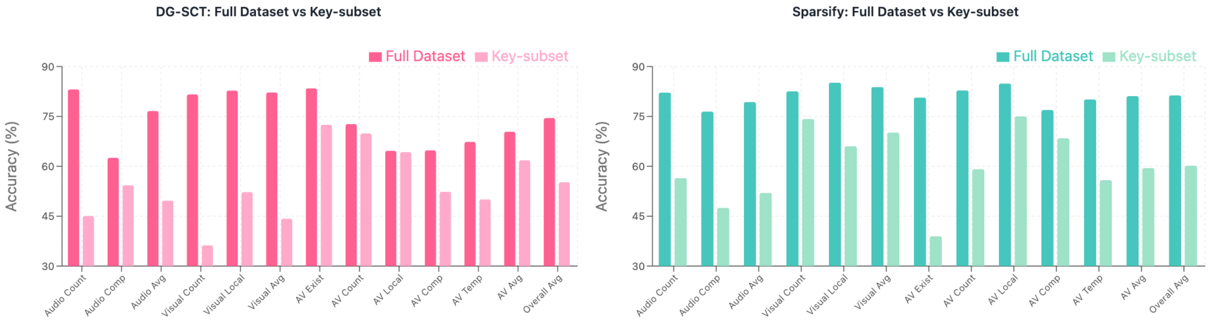


Figure 4: Accuracy comparison of DG-SCT and `Sparsify` trained on the full dataset and the key-subset (∼25% of data) (Liu et al., 2024). Training on the key-subset maintains strong performance despite substantial data reduction.

tion answering by jointly reducing modality-specific redundancies. On audio-visual QA, `Sparsify` outperforms DG-SCT by +11.24% and +10.70%, and LAVisH by +5.49% and +13.33% on MUSIC-AVQA and MUSIC-AVQA v2.0, respectively. These consistent gains suggest that sparsification across both audio and visual modalities helps eliminate less informative content and produce more streamlined multimodal representations with reduced noise. This joint reduction in redundancy allows the model to more effectively capture relevant cross-modal associations in complex performance scenarios.

3.3 Key-subset Selection and Performance Retention

We evaluate the effectiveness of our Key-subset selection algorithm on the MUSIC-AVQA v2.0 dataset (Liu et al., 2024), as illustrated in Figure 4. The selected subset comprises only ∼25% of the original training set (10,819 samples), yet enables models to retain a substantial portion of their full-data performance. When trained exclusively on this subset, `Sparsify` achieves 60.17% accuracy and DG-SCT (Duan et al., 2023) achieves 55.21%, corresponding to 74.01% and 74.08% of their respective accuracies when trained on the full dataset. These results demonstrate that our Key-subset Selection reduces training data usage while retaining much of the models' original performance, offering a data-efficient solution for Music AVQA.

3.4 Training Efficiency Gains from Sparse Learning

Figure 5 illustrates the training efficiency gains of `Sparsify`, which reduces total training time from 173 hours to 124 hours—a 28.32% improvement over its dense variant. These gains reflect the combined effect of three sparsification strategies integrated into the training pipeline. Sparse Masking reduces early-stage computational load by masking 50% of audio and visual tokens. Sparse Merging compresses intermediate representations by consolidating similar tokens, reducing token-level complexity. In parallel, using $InfoBatch$ enhances efficiency by emphasizing harder-to-learn samples, which accelerates convergence and reduces the number of required optimization steps.

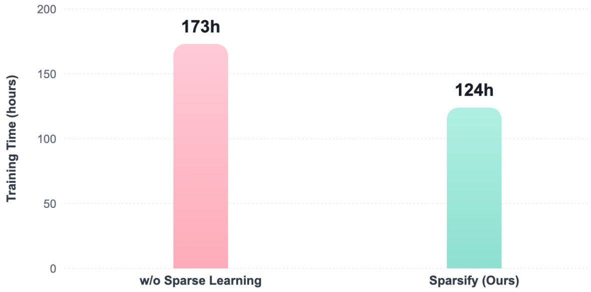


Figure 5: Comparison of the training time of `Sparsify` with a dense variant that disables all three sparsification strategies. Results are reported on the MUSIC-AVQA v2.0 dataset (Liu et al., 2024).

4 Conclusion

We present `Sparsify`, a sparse learning framework for Music AVQA that addresses the inefficiencies inherent in dense audio-visual representations. `Sparsify` achieves this by *(i)* integrating three sparsification strategies into an end-to-end pipeline and achieving state-of-the-art performance on Music AVQA datasets; *(ii)* reducing training time by 28.32% while maintaining comparable accuracy to its dense counterpart. In addition, we propose a key-subset selection algorithm that selects and uses approximately 25% of the MUSIC-AVQA v2.0 training data, while retaining 70–80% of full-data performance across models. We hope our work offers insights into efficient multimodal understanding in dense audio-visual settings.

Limitations

The effectiveness of Sparsify has been demonstrated on large-scale Music AVQA benchmarks, as it is specifically designed to address the inefficiencies of dense audio-visual representations in this domain. While it may hold promise for broader multimodal tasks, its behavior in such settings remains to be explored. Extending Sparsify to more diverse or unconstrained applications represents a valuable direction for future work.

Ethical Considerations

We examined the study describing the publicly available datasets used in this research and identified no ethical issues regarding the datasets.

Acknowledgment

This study is supported by the Department of Defense grant HT9425-23-1-0267.

References

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing & Management*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision*.

Sagar S Arya, Sofia B Dias, Herbert F Jelinek, Leontios J Hadjileontiadis, and Anna-Maria Pappa. 2023. The convergence of traditional and digital biomarkers through ai-assisted biosensing: A new era in translational diagnostics? *Biosensors and Bioelectronics*.

Alexandre Blanco-Gonzalez, Alfonso Cabezon, Alejandro Seco-Gonzalez, Daniel Conde-Torres, Paula Antelo-Riveiro, Angel Pineiro, and Rebeca Garcia-Fandino. 2023. The role of ai in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals*.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *Transactions on Audio, Speech, and Language Processing*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech and Signal Processing*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022a. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *International Conference on Acoustics, Speech and Signal Processing*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. 2022. Machine learning in drug discovery: a review. *Artificial Intelligence Review*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*.

Xingjian Diao, Ming Cheng, and Shitong Cheng. 2023. Av-maskenhancer: Enhancing video representations through audio-visual masked autoencoder. In *International Conference on Tools with Artificial Intelligence*.

Xingjian Diao, Chunhui Zhang, Tingxuan Wu, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, and Jiang Gui. 2024. Learning musical representations for music performance question answering. In *Findings of the Association for Computational Linguistics: EMNLP*.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding. *arXiv preprint arXiv:2502.06020*.

Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. 2023. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. In *Advances in Neural Information Processing Systems*.

Haytham M. Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *Transactions on Audio, Speech, and Language Processing*.

Chongyang Gao, Yiren Jian, Natalia Denisenko, Soroush Vosoughi, and VS Subrahmanian. 2024. Gem: generating engaging multimodal content. In *International Joint Conference on Artificial Intelligence*.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou

Wang, Yutong Bai, Zhuoran Yang, et al. 2024. Avodyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

Travis R Goodwin and Sanda M Harabagiu. 2016. Medical question answering for clinical decision support. In *International on Conference on Information and Knowledge Management*.

Yangfan He, Sida Li, Jianhui Wang, Kun Li, Xinyuan Song, Xinhang Yuan, Keqin Li, Kuan Lu, Menghao Huo, Jiaqi Chen, et al. 2025a. Enhancing low-cost video editing with lightweight adaptors and temporal-aware inversion. *arXiv preprint arXiv:2501.04606*.

Yangfan He, Jianhui Wang, Kun Li, Yijin Wang, Li Sun, Jun Yin, Miao Zhang, and Xueqian Wang. 2025b. Enhancing intent understanding for ambiguous prompts through human-machine co-adaptation. *arXiv preprint arXiv:2501.15167*.

Panwen Hu, Nan Xiao, Feifei Li, Yongquan Chen, and Rui Huang. 2023. A reinforcement learning-based automatic video editing method using pre-trained vision-language model. In *International Conference on Multimedia*.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2022. Sparse in space and time: Audio-visual synchronisation with trainable selectors. *arXiv preprint arXiv:2210.07055*.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient synchronization from sparse cues. In *International Conference on Acoustics, Speech and Signal Processing*.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping vision-language learning with decoupled language pre-training. In *Advances in Neural Information Processing Systems*.

Yiren Jian, Tingkai Liu, Yunzhe Tao, Chunhui Zhang, Soroush Vosoughi, and Hongxia Yang. 2024. Expedited training of visual conditioned language generation via redundancy reduction. In *Annual Meeting of the Association for Computational Linguistics*.

Balaram Yadav Kasula. 2023. Harnessing machine learning for personalized patient care. *Transactions on Latest Trends in Artificial Intelligence*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in Neural Information Processing Systems*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Bin Li and Hanjun Deng. 2023. Bilateral personalized dialogue generation with contrastive learning. *Soft Computing*.

Bin Li, Bin Sun, Shutao Li, Encheng Chen, Hongru Liu, Yixuan Weng, Yongping Bai, and Meiling Hu. 2024a. Distinct but correct: generating diversified and entity-revised medical response. *Science China Information Sciences*.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Conference on Computer Vision and Pattern Recognition*.

Shutao Li, Bin Li, Bin Sun, and Yixuan Weng. 2024b. Towards visual-prompt temporal answer grounding in instructional video. *Transactions on Pattern Analysis and Machine Intelligence*.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023a. Scaling language-image pre-training via masking. In *Conference on Computer Vision and Pattern Recognition*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *International Conference on AI in Finance*.

Jinhua Liang, Huan Zhang, Haohe Liu, Yin Cao, Qiuqiang Kong, Xubo Liu, Wenwu Wang, Mark D Plumbley, Huy Phan, and Emmanouil Benetos. 2024. Wavcraft: Audio editing and generation with large language models. *arXiv preprint arXiv:2403.09527*.

Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2023. Vision transformers are parameter-efficient audio-visual learners. In *Conference on Computer Vision and Pattern Recognition*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *European Signal Processing Conference*.

Xiulong Liu, Zhikang Dong, and Peng Zhang. 2024. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Winter Conference on Applications of Computer Vision*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Conference on Computer Vision and Pattern Recognition*.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*.

Gianluca Monaci, Friedrich T Sommer, and Pierre Vandergheynst. 2008. Learning sparse generative models of audiovisual signals. In *European Signal Processing Conference*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *International Conference on Machine Learning*.

Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. 2022. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pretraining. In *International Conference on Multimedia*.

Qi Qian, Yuanhong Xu, and Juhua Hu. 2023. Intra-modal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems*.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. 2024. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *International Conference on Learning Representations*.

Khyati Saini and Pardeep Singh. 2023. Evolution of financial question answering themes, challenges, and advances. In *International Conference on Recent Innovations in Computing*.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

Peng Shen, Satoshi Tamura, and Satoru Hayamizu. 2013. Audio-visual interaction in sparse representation features for noise robust audio-visual speech recognition. In *Auditory-Visual Speech Processing*.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*.

Teotino Gomes Soares, Azhari Azhari, Nur Rokhman, and E Wonarko. 2021. Education question answering systems: a survey. In *International MultiConference of Engineers and Computer Scientists*.

Nitish Srivastava and Russ R. Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*.

Tim Steuer, Anna Filighera, and Thomas Tregel. 2022. Investigating educational and noneducational answer selection for educational question generation. *IEEE Access*.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. 2023. Knowledge graph assisted end-to-end medical dialog generation. *Artificial Intelligence in Medicine*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Shiru Wang, Yao Chen, Lesley A Jarvis, Yucheng Tang, David J Gladstone, Kimberley S Samkoe, Brian W Pogue, Petr Bruza, and Rongxiao Zhang. 2024. Robust real-time segmentation of bio-morphological features in human cherenkov imaging during radiotherapy via deep learning. *arXiv preprint arXiv:2409.05666*.

Yanbo J Wang, Yuming Li, Hui Qin, Yuhang Guan, and Sheng Chen. 2022. A novel deberta-based model for financial question answering task. *arXiv preprint arXiv:2207.05875*.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. 2023. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*.

Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. 2022. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*.

Yuxiang Wei, Anees Abrol, and Vince D Calhoun. 2025. Hierarchical spatio-temporal state-space modeling for fmri analysis. In *International Conference on Research in Computational Molecular Biology*.

Yuxiang Wei, Yuqian Chen, Tengfei Xue, Leo Zekelman, Nikos Makris, Yogesh Rathi, Weidong Cai, Fan Zhang, and Lauren J O'Donnell. 2023. A deep network for explainable prediction of non-imaging phenotypes using anatomical multi-view data. In *International Workshop on Computational Diffusion MRI*.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. 2024. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*.

Yongchao Wu, Aron Henriksson, Martin Duneld, and Jalal Nouri. 2023. Towards improving the reliability and transparency of chatgpt for educational question answering. In *European Conference on Technology Enhanced Learning*.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *International Conference on Multimedia*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Jiawei Yao, Qi Qian, and Juhua Hu. 2024a. Customized multiple clustering via multi-modal subspace proxy learning. *arXiv preprint arXiv:2411.03978*.

Jiawei Yao, Qi Qian, and Juhua Hu. 2024b. Multi-modal proxy learning towards personalized visual multiple clustering. In *Conference on Computer Vision and Pattern Recognition*.

Qilang Ye, Zitong Yu, and Xin Liu. 2024. Answering diverse questions via text attached with key audio-visual clues. *arXiv preprint arXiv:2403.06679*.

Wenhao You, Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Zhongyu Ouyang, Chiyu Ma, Tingxuan Wu, Noah Wei, Zong Ke, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Music's multimodal complexity in avqa: Why we need more than general multimodal llms. *arXiv preprint arXiv:2505.20638*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Conference on Computer Vision and Pattern Recognition*.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *International Conference on Computer Vision*.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. 2025. Pretrained image-text models are secretly video captioners. *arXiv preprint arXiv:2502.13363*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *Transactions on Pattern Analysis and Machine Intelligence*.

Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. 2019. The sound of motions. In *International Conference on Computer Vision*.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels. In *European Conference on Computer Vision*.

Ziyi Zhou, Ming Cheng, Xingjian Diao, Yanjun Cui, and Xiangling Li. 2024. Glumarker: A novel predictive modeling of glycemic control through digital biomarkers. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Baselines

- **AVST** (Li et al., 2022): Integrates audio, visual, and question modalities for spatio-temporal reasoning in audio-visual question answering. It aligns modalities through spatial and temporal grounding, fuses features into a joint representation, and optimizes both grounding and QA objectives.
- **LAVisH** (Lin et al., 2023): Adapts frozen Vision Transformers for audio-visual tasks using lightweight adapters and latent tokens to compress and fuse audio-visual information. Cross-modal attention and adapter modules enable bidirectional interaction between modalities.
- **DG-SCT** (Duan et al., 2023): Enhances audio-visual tasks through a Dual-Guided Spatial-Channel-Temporal attention mechanism, dynamically adjusting feature extraction and facilitating bidirectional audio-visual guidance with lightweight interaction layers.

B Related Work

Audio-Visual Video Understanding Audio and visual modalities offer complementary cues that, when jointly modeled, support a more comprehensive understanding of the scene (Wei et al., 2022; Diao et al., 2023; Shu et al., 2023; Diao et al., 2025). Early work focused on joint representations for tasks like audio-visual speech recognition (Ngiam et al., 2011) and multimodal deep learning (Srivastava and Salakhutdinov, 2012). Recent methods enhance fusion techniques for sound source localization (Zhao et al., 2018) and audio-driven visual analysis (Zhao et al., 2019). Frameworks such as LAVisH (Lin et al., 2023), which proposed a latent audio-visual hybrid adapter that adapts pretrained ViTs to audio-visual tasks by injecting a small number of trainable parameters into every layer of a frozen ViT, and DG-SCT (Duan et al., 2023) which incorporates trainable cross-modal interaction layers into pre-trained audio-visual encoders, allowing adaptive extraction of crucial information from the current modality across spatial, channel, and temporal dimensions, while preserving the frozen parameters of large-scale pre-trained models. As for benchmarks, there are MUSIC-AVQA (Li et al., 2022), AVQA (Yang et al., 2022), MUSIC-AVQA v2.0 (Liu et al., 2024) and AV-Odyssey Bench (Gong et al., 2024), which focus on whether model can truly understand audio-visual information. However, existing approaches overlook the unique challenges of music performance datasets, where dense and continuous audio-visual signals lead to significant redundancy. These dense representations hinder efficient processing and dilute task-relevant features, necessitating sparsification strategies to enable efficient reasoning in this domain.

Multimodal Question Answering Multimodal question answering spans Visual QA (VQA) (Antol et al., 2015; Lei et al., 2018; Li et al., 2024b), Audio QA (AQA) (Fayek and Johnson, 2020), and Audio-Visual QA (AVQA) (Li et al., 2022), requiring the integration of modality-specific signals for complex reasoning tasks. For VQA, datasets such as MMMU (Yue et al., 2024) and MMBench (Liu et al., 2025) provide carefully curated benchmarks that evaluate vision-language models across diverse domains. For AQA, notable datasets include Clotho-AQA (Lipping et al., 2022) and AIR-Bench (Yang et al., 2024), which consist of question-answering tasks derived from environmental and event-based audio scenes. AVQA benchmarks such as MUSIC-AVQA (Li et al., 2022), AVQA (Yang et al., 2022), Pano-AVQA (Yun et al., 2021), FunQA (Xie et al., 2024), and MUSIC-AVQA v2.0 (Liu et al., 2024) emphasize spatio-temporal reasoning and multimodal fusion in complex video contexts. Among these, Music AVQA presents distinctive challenges due to its continuous and densely structured audio signals, making it valuable for multimodal reasoning (You et al., 2025).

Sparse Learning in Audio-Visual Signals Sparsity has become increasingly crucial in audio-visual signal processing due to the inherent complexity and redundancy of cross-modal data. Early approaches employ shift-invariant kernels (Monaci et al., 2008) to capture essential patterns while reducing computational overhead. This foundation leads to more sophisticated methods using group sparsity and joint dictionaries (Shen et al., 2013), which are particularly effective in handling noisy and variable signals. Current research focuses on temporal dynamics in audio-visual learning, where audio-visual relationships are often intermittent but contextually meaningful (Iashin et al., 2022). Modern transformer-based architectures with specialized selection mechanisms (Iashin et al., 2024) have shown promise in processing extended sequences efficiently. However, sparsity-based approaches remain underexplored in the context of music performance question answering, where challenges such as overlapping instruments and complex audio-visual interactions demand more efficient representations. Our work aims to bridge this gap with sparsification strategies.

C Positioning Music AVQA Among Multimodal Tasks

To contextualize Music AVQA, it is useful to distinguish it from broader multimodal tasks that also integrate information across modalities. This section contrasts Music AVQA with vision-language modeling, audio-language modeling, and other representative domain-specific question answering to highlight its unique challenges and requirements.

C.1 Vision-Language Modeling

Vision-language modeling aims to enable multimodal systems to interpret visual content—such as images and videos—in conjunction with textual descriptions (Jian et al., 2023; Bordes et al., 2024; Jian et al., 2024; Zhang et al., 2024). It has supported a wide range of applications, including text-to-image generation (Li et al., 2019; Gao et al., 2024; He et al., 2025b), video editing (Hu et al., 2023; He et al., 2025a), video captioning and grounding (Pan et al., 2022; Li et al., 2024b; Zhang et al., 2025), and proxy learning (Qian et al., 2023; Yao et al., 2024a,b). In contrast, Music AVQA requires integrated reasoning over continuous audio-visual streams, where visual understanding must be synchronized with rhythm, motion, and acoustic cues. This setting introduces challenges such as temporal alignment and redundancy reduction, which are not typically emphasized in standard vision-language tasks.

C.2 Audio-Language Modeling

Audio-language modeling (Borsos et al., 2023; Deshmukh et al., 2023; Wu et al., 2024; Su et al., 2025) builds systems that fuse audio features with text for various downstream tasks such as audio question answering (Fayek and Johnson, 2020; Lipping et al., 2022), text-to-speech generation (Min et al., 2021; Le et al., 2023), and audio editing (Wang et al., 2023; Liang et al., 2024). These tasks primarily focus on modeling relationships between acoustic signals and language, often in domains such as speech, environmental sounds, or sound events. Unlike conventional audio-language tasks that focus on modeling acoustic-linguistic associations, Music AVQA incorporates a visual modality that is intricately entangled with the audio stream in music performance contexts. This setting necessitates fine-grained multimodal reasoning, where models must jointly interpret auditory patterns, visual dynamics, and their temporal interplay to answer performance-specific questions.

C.3 Domain-Specific Question Answering

Domain-specific question answering systems are designed to operate within specialized fields by leveraging structured knowledge and domain-specific data. Examples such as educational, financial, and medical QA, as discussed below, entail distinct input modalities, reasoning demands, and representational challenges.

Educational QA Educational Question Answering (Educational QA) systems (Soares et al., 2021; Steuer et al., 2022; Wu et al., 2023) are designed to support learning processes by responding to student queries based on educational materials such as textbooks, lecture notes, and academic articles. The primary goal is to clarify concepts, explain solutions, or guide students through subject matter. In contrast, Music AVQA involves perceptual reasoning over evolving audio-visual input. Instead of extracting explicit concepts from structured curricula, models must interpret expressive cues such as gestural nuance, visual-musical alignment, and acoustic articulation in continuous video streams. This shift demands interpreting perceptual and temporal patterns, which are not typically required in conventional educational QA. The emphasis on fluid multimodal integration further distinguishes Music AVQA as a challenging reasoning setting.

Financial QA Financial Question Answering (Financial QA) (Li et al., 2023b; Huang et al., 2023; Saini and Singh, 2023) focuses on extracting insights and answering questions from a wide range of financial data (Chen et al., 2022b), such as company reports, market data, economic indicators, and financial news. These systems assist analysts, investors, and businesses in making informed decisions by providing quick access to relevant financial information and analysis. While Financial QA can involve data from multiple views (e.g., text, tables, charts) (Zhu et al., 2021), it typically does not involve the continuous, dense audio-visual streams found in music performances. The core task in Financial QA is to identify factual information, understand financial terminology, perform numerical reasoning, and interpret trends from often structured or semi-structured textual and numerical data (Wang et al., 2022). In contrast, Music AVQA centers on the temporal and semantic understanding of performance events, requiring models to interpret how visual gestures correspond to musical outcomes, such as identifying sounding instruments, tracking temporal changes, and linking expressive motion to acoustic effects, rather than extracting or reasoning over structured financial data.

Medical QA AI models are increasingly utilized across medical field, tackling a wide array of applications such as diagnostic assistance through analysis of medical images (e.g., X-rays, MRIs) (Wei et al., 2023; Wang et al., 2024; Wei et al., 2025) and dialogues (Varshney et al., 2023; Li et al., 2024a), drug discovery (Dara et al., 2022; Blanco-Gonzalez et al., 2023), digital biomarkers (Arya et al., 2023; Zhou et al., 2024), and personalized patient care (Kasula, 2023; Li and Deng, 2023). Among these, Medical Question Answering (Medical QA) (Abacha and Zweigenbaum, 2015; Goodwin and Harabagiu, 2016) is a specialized field focused on developing systems that understand and respond to health-related queries. These systems often process information from diverse sources to provide accurate medical information or support clinical decision-making. In contrast, Music AVQA centers on interpreting music-related videos, requiring models to reason over dense, continuous audio and tightly synchronized visual streams. While both involve multimodal and complex reasoning, Music AVQA uniquely demands fine-grained alignment of perceptual cues in expressive performance contexts, such as gesture, rhythm, and phrasing.