Learning Sparsity for Effective and Efficient Music Performance Question Answering

Xingjian Diao, Tianzhen Yang, Chunhui Zhang, Weiyi Wu,
Ming Cheng, Jiang Gui
Dartmouth College
{xingjian.diao, chunhui.zhang, weiyi.wu, ming.cheng}.gr@dartmouth.edu
jiang.gui@dartmouth.edu

Abstract

Music performances, characterized by dense, continuous audio and seamless audio-visual integration, present unique challenges for multimodal scene understanding and reasoning. Recent advances, including the MUSIC-AVQA datasets, have highlighted the need for effective integration of audio-visual representations to tackle complex questions. However, existing AVQA methods often rely on dense and unoptimized representations, leading to inefficiencies in isolating key information, reducing redundancy, and prioritizing critical samples. To address these challenges, we propose `Sparsify`, a sparse learning framework specifically designed for music AVQA. `Sparsify` integrates three complementary sparsification strategies into an end-to-end pipeline, achieving state-of-the-art performance on the MUSIC-AVQA datasets. Additionally, it reduces training time by 28.32% while improving accuracy compared to fully trained dense model, demonstrating significant gains in both efficiency and effectiveness. To further optimize data usage, we introduce a key-subset selection algorithm that reduces the training dataset size by about 75%, retaining 70-80% of the original performance across tested models. Our code and key-subset are available at https://anonymous.4open.science/r/ARRsubmission14.

1 Introduction

Music performances, with their dense, continuous audio and seamless audio-visual integration, present challenges and opportunities in multimodal scene understanding and reasoning (Ma et al., 2024). The recently proposed MUSIC Audio-Visual Question Answering (AVQA) dataset (Li et al., 2022) and its extended version, MUSIC-AVQA v2.0 (Liu et al., 2024), exemplified by a sample in Figure 1, have garnered attention for their focus on integrating audio-visual representations to tackle questions about music performances.

(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 include dense and continuous audio signals with inherent and substantial redundancies that contribute little to answering questions. **Sparse learning** has the potential to effectively filters out such redundancies, enabling more efficient and accurate reasoning.

AVQA methods for music performances have evolved from early cross-modality learning for speech recognition (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012) to 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 in frozen encoders to enhance reasoning.

However, existing AVQA methods face significant limitations in effectively modeling sparse representations, which are critical for addressing the unique challenges posed by QA tasks for music (Yang et al., 2024b). These limitations include: *(i)* an overreliance on dense, unoptimized representations, which struggle to isolate key information from dense audio-visual signals (Ye et al., 2024; Diao et al., 2024); *(ii)* the absence of effective redundancy reduction mechanisms, leading to inefficiencies in feature extraction and model inference (Shang et al., 2024); *(iii)* the lack of prioritization strategies for identifying task-critical samples, limiting scalability and prolonging training times (Qin et al., 2024; Li et al., 2023).

To address these limitations, we propose `Sparsify`, a sparse learning framework designed

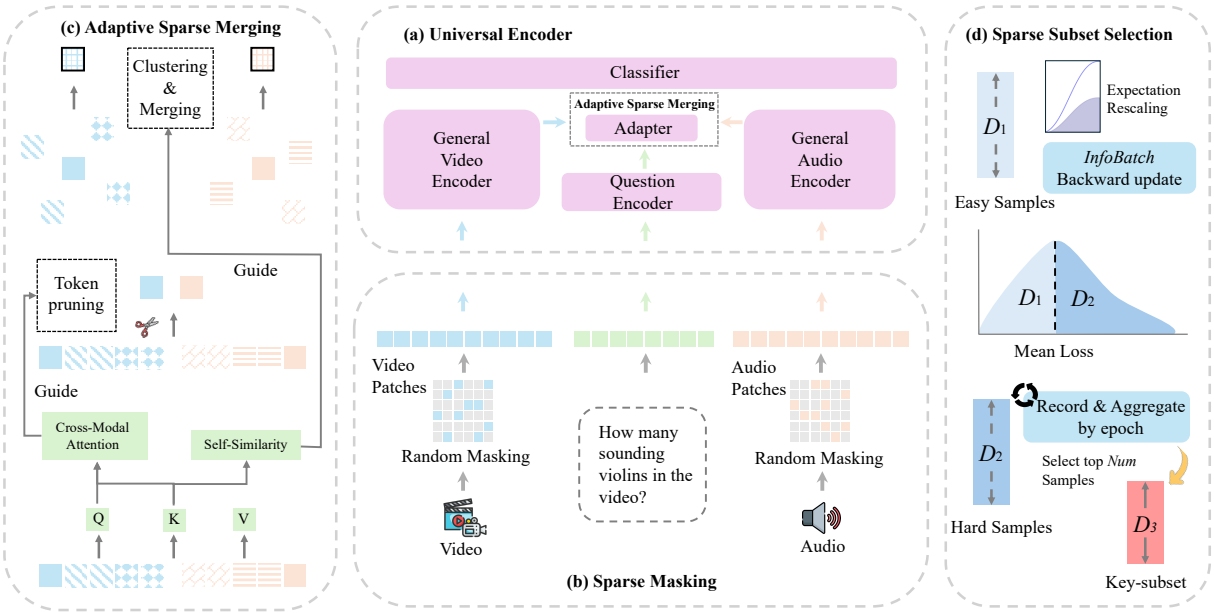


Figure 2: Sparsify framework integrates (a) a Universal Encoder with three key components: (b) Sparse Masking to reduce redundancy by sparsifying audio and visual tokens; (c) Adaptive Sparse Merging to select and merge key multimodal tokens based on similarity; and (d) Sparse Subset Selection to prioritize impactful samples and reweight gradients with InfoBatch.

for AVQA tasks. Our contributions are threefold:

- We propose an end-to-end pipeline for music AVQA, integrating three complementary sparsification strategies, and validate its effectiveness through state-of-the-art performance on the MUSIC-AVQA datasets.
- Sparsify achieves a reduction in training time of 28. 32% while improving accuracy compared to fully trained dense models, showcasing notable gains in both efficiency and effectiveness.
- 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 tested models. To advance research on efficient music performance AVQA, we release both the algorithm code and the resulting key-subset.

2 Sparsify Framework

2.1 Learning Multimodal Representations

Sparsify primarily utilizes three interactive encoders as the Universal Encoder (Diao et al., 2024): the General Video Encoder built on Swin-V2 (Liu et al., 2022), the General Audio Encoder leveraging the HTS-Audio Transformer (Chen et al., 2022), and the Question Encoder based on a language transformer (Vaswani et al., 2017). Cross-modal attention is applied to align features across these modalities, followed by activation functions and linear transformations, ensuring unified and effective multimodal representations tailored to music-AVQA tasks.

2.2 Efficient Sparse Masking for Key Token Retention

Music performance data inherently contain substantial redundancies, which pose significant challenges to efficient multimodal learning. Sparse Masking, illustrated in Figure 2 (a) and inspired by recent advances in random masking techniques for multimodal models (Li et al., 2023), addresses this issue by learning sparse representations that emphasize task-relevant information while discarding irrelevant patches. This approach aligns with the goal of sparse learning, optimizing both accuracy and computational efficiency.

In the visual modality, Sparse Masking operates on image patches by dynamically masking 50% regions, forcing the model to focus on key areas that contribute to task-specific reasoning. For the audio modality, we transform raw audio signals into mel-spectrograms and apply the same masking strategy. Treating audio spectrogram patches as visual tokens ensures consistent sparsity across modalities, enabling seamless alignment during cross-modal feature integration.

2.3 Adaptive Sparse Merging for Multimodal Pruning

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 introduce Adaptive Sparse Merging, as shown in Figure 2 (b). 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. 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. To further refine token selection, we apply the Interquartile Range (IQR) method (Shang et al., 2024) 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. Adaptive Sparse Merging ensures efficient multimodal integration with aligned sparsity across audio and visual modalities.

2.4 Sparse Subset Selection for Training Optimization

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.

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 algorithm is presented in Appendix A as Algorithm 1.

3 Experiments

3.1 Setup

Music AVQA datasets *(i)* MUSIC-AVQA (Li et al., 2022): A dataset with 9,288 videos (150 hours) covering 22 instruments, featuring 45,867 QA pairs from 33 templates in 4 categories (String, Wind, Percussion, Keyboard). Each video has 5 QA pairs on average. *(ii)* MUSIC-AVQA v2.0 (Liu et al., 2024): Expanded to address data bias, adding 1,230 new videos and 8,100 QA pairs.

Details The baseline details are in Appendix B, and the training settings are in Appendix C.

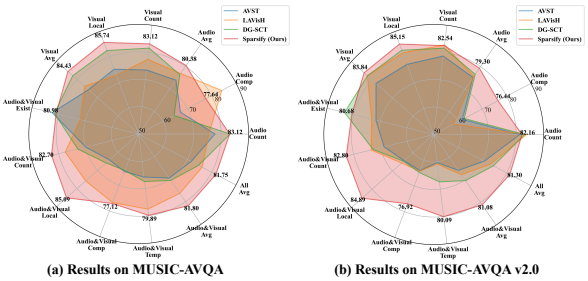


Figure 3: Radar charts comparing `Sparsify` with state-of-the-art methods on MUSIC-AVQA and MUSIC-AVQA v2.0, across various question types.

3.2 Overall Comparisons

3.2.1 Audio-Visual Question Answering

Figure 3 and Table 1 present the experimental results of methods on the MUSIC-AVQA (Li et al., 2022) and MUSIC-AVQA v2.0 (Liu et al., 2024) datasets, demonstrating that `Sparsify` outperforms existing AVQA approaches and achieves state-of-the-art performance. Specifically, `Sparsify` attains overall AVQA accuracies of 81.75% and 81.30% on MUSIC-AVQA and MUSIC-AVQA v2.0, surpassing the previous best baselines by 5.65% and 6.77%, respectively. Across individual subtasks, `Sparsify` consistently demonstrates effectiveness: on MUSIC-AVQA, it achieves 80.38% for audio QA, 84.43% for video QA, and 81.80% for audio-visual joint QA; Likewise, on MUSIC-AVQA v2.0,

Methods	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.

it achieves 79.30% for audio QA, 83.84% for video QA, and 81.30% for audio-visual joint QA. The use of Sparse Masking and Adaptive Sparse Merging strategies enables the model to dynamically select and retain critical multimodal tokens, leading to high-quality representations and superior performance in music performance AVQA tasks.

Method	Training Time (hours)	Overall Average Accuracy (%)
w/o Sparse Learning	173	77.71
Sparsify (Ours)	**124**	**81.30**
Δ	−49	+3.59

Table 2: Ablation study comparing training time and overall average accuracy of Sparsify with fully trained Sparsify model without three sparsification strategies on MUSIC-AVQA v2.0 (Liu et al., 2024).

3.2.2 Training Time Reduction

Table 2 highlights the efficiency of Sparsify, which reduces training time by 28.32% (49 hours) while achieving a 3.59% improvement in accuracy compared to Sparsify dense model without sparse learning. This efficiency stems from Sparse Masking, which reduces input size by masking 50% of audio and visual tokens to focus on task-relevant regions, Adaptive Sparse Merging, which consolidates key tokens to optimize selection and preserve modality integrity, and InfoBatch, which prioritizes hard-to-learn samples for efficient convergence. These strategies collectively reduce computational costs while enhancing performance.

3.2.3 Efficient Subset Selection

We evaluate DG-SCT and Sparsify on the MUSIC-AVQA v2.0 (Liu et al., 2024) Key-subset, as illustrated in Figure 4. The Key-subset selection algorithm effectively reduces the training dataset size to approximately 25% of the original,

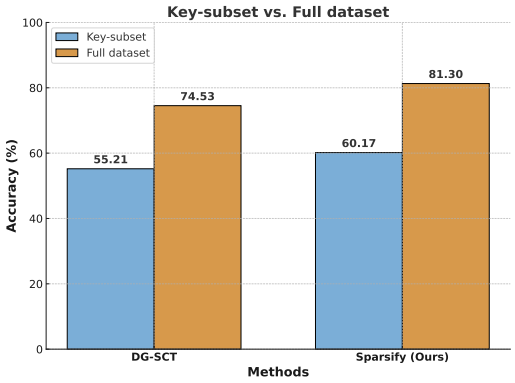


Figure 4: Accuracy comparison between DG-SCT and Sparsify trained on the Key-subset and the Full dataset. Detailed results are provided in Appendix D, Table 3.

corresponding to 10,819 samples, while maintaining a 70-80% of the performance achieved when trained on the full dataset. Specifically, DG-SCT retains 74.08% of the accuracy, while Sparsify retains 74.01%, demonstrating the Key-subset strategy's ability to reduce computational costs without severely compromising accuracy.

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 in an end-to-end pipeline, achieving state-of-the-art performance on MUSIC-AVQA datasets; *(ii)* reducing training time by 28.32% while improving accuracy compared to the dense model; and *(iii)* designing a subset selection algorithm that reduces training data by approximately 75%, retaining 70-80% of the original performance across tested models. Sparsify offers key insights to advance multimodal understanding in continuous dense audio contexts.

Limitations

The effectiveness of Sparsify has been demonstrated on large-scale music performance AVQA benchmarks, as it is specifically designed to address the inefficiencies of dense audio-visual representations in this domain. However, its generalizability to other in-the-wild multimodal tasks remains unexplored. Extending Sparsify to broader applications would be an exciting future direction for real-world use cases.

Ethical Considerations

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

A Key Subset Selection Algorithm

We outline our method for Key-subset Selection Algorithm in 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 $\mathcal{K}$

B 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.

C Training details

Full dataset training details For the experiments described in Section 3.2.1, Sparse Masking is applied during the first three epochs and is disabled for the remaining 12 epochs. Adaptive Sparse Merging and $InfoBatch$ are used throughout the training process. In our Sparse Masking approach, we apply a masking rate of 50%, while for $InfoBatch$, we set the ratio to 0.5 and the delta to 0.875 (Qin et al., 2024).

Key-subset selection training details In the key-subset selection experiment, we perform a warm-up phase for one epoch, followed by 15 epochs of training. The parameters are set as follows: $N = 15$, $k = 3$, $r = 0.618$, and Num is set to the number of pairs of QA (10,819). During this experiment, only $InfoBatch$ (Qin et al., 2024) is utilized, without applying Sparse Masking and Adaptive Sparse Merging.

D Key-subset Experiment

Table 3 presents the results of DG-SCT (Duan et al., 2023) and Sparsify evaluated on the Key-subset, which constitutes only $\sim 25\%$ (10,819 sample) of the original dataset. Despite the drastic reduction in dataset size, Sparsify achieves an overall accuracy of 60.17% and DG-SCT achieves 55.21% overall accuracy on the Key-subset, preserving 70-80% of its full dataset performance.

Method	Audio-related QA			Visual-related QA			Audio & Visual-related QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
DG-SCT (Duan et al., 2023)	45.05	54.30	49.67	36.24	52.24	44.24	72.51	69.88	64.24	52.34	50.06	61.81	55.21
Sparsify (Ours)	56.46	47.51	51.98	74.22	66.02	70.12	38.95	59.11	74.99	68.41	55.82	59.46	60.17

Table 3: Experimental results of DG-SCT (Duan et al., 2023) and Sparsify on the Key-subset of MUSIC-AVQA 2.0.

E Related Work

Audio-Visual Video Understanding Audio-visual scene understanding leverages complementary modalities for comprehensive reasoning. 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 novel 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), Audio QA (Fayek and Johnson, 2020), and Audio-Visual QA (AVQA) (Li et al., 2022), integrating cues across modalities to answer complex questions. For VQA datasets, there are MMMU (Yue et al., 2024), MMBench (Liu et al., 2025) which provides meticulously curated tasks for Vision Language Models. In regard to Audio QA, there appears to be Clotho-AQA (Lipping et al., 2022) and AIR-Bench (Yang et al., 2024a) that consist various audio tasks. Datasets like MUSIC-AVQA (Li et al., 2022) and AVQA (Yang et al., 2022), and MUSIC-AVQA v2.0 (Liu et al., 2024) emphasize spatio-temporal reasoning and multimodal fusion and ensure that no answers have outstanding skewed distribution. However, existing methods rely heavily on dense representations and computationally intensive models, limiting scalability for large-scale datasets. Our work addresses this gap by proposing a sparse learning framework that prioritizes critical data regions and dynamically reduces redundancies, improving efficiency and scalability while maintaining high accuracy.

F Results Demonstration

We showcase Sparsify's ability to handle diverse question types across the Audio-Visual domain, Visual domain, and Audio domain. Figures 5 to 12 illustrate representative question-answering examples from the MUSIC-AVQA v2.0 (Liu et al., 2024) test set. These examples demonstrate Sparsify's robustness and effectiveness in addressing complex scenarios, highlighting its comprehensive understanding and strong performance in music audio-visual question answering.

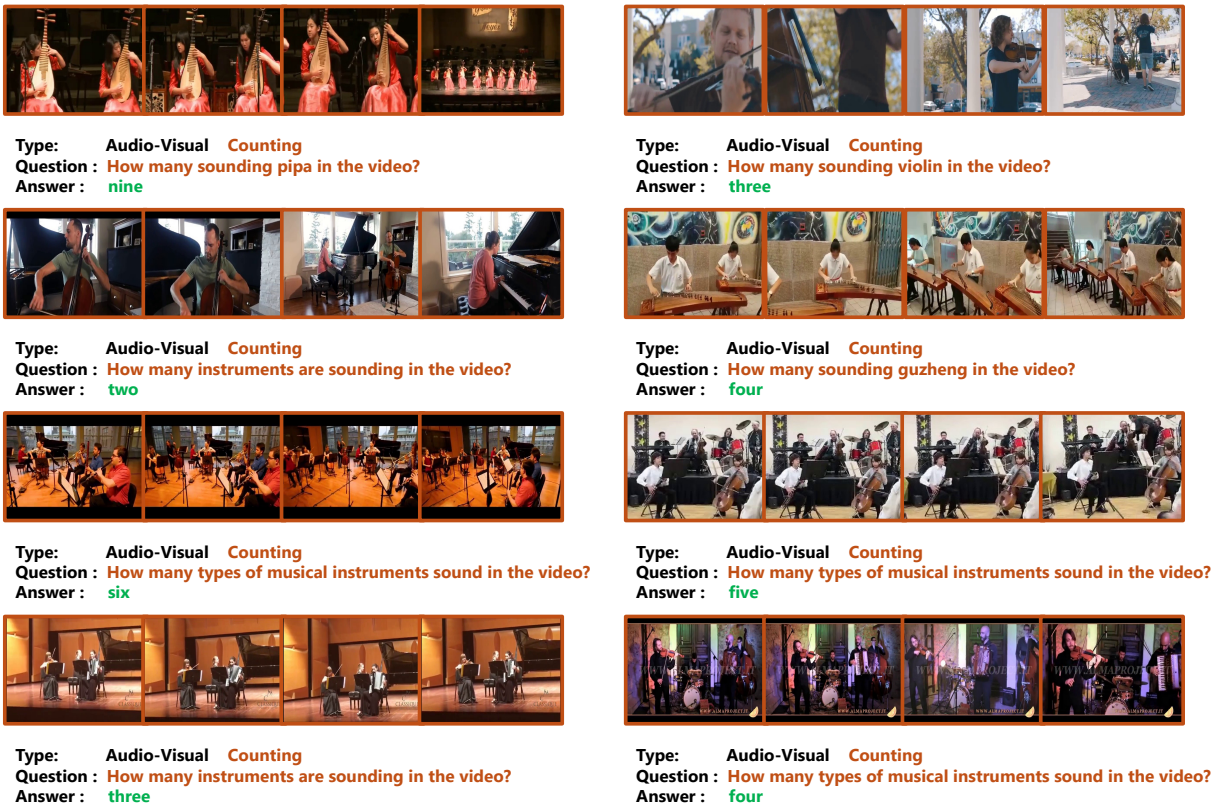


Figure 5: **Demonstration of Audio-Visual Counting Question-Answering.** Our model handles the Audio-Visual Counting questions correctly under complicated scenarios.

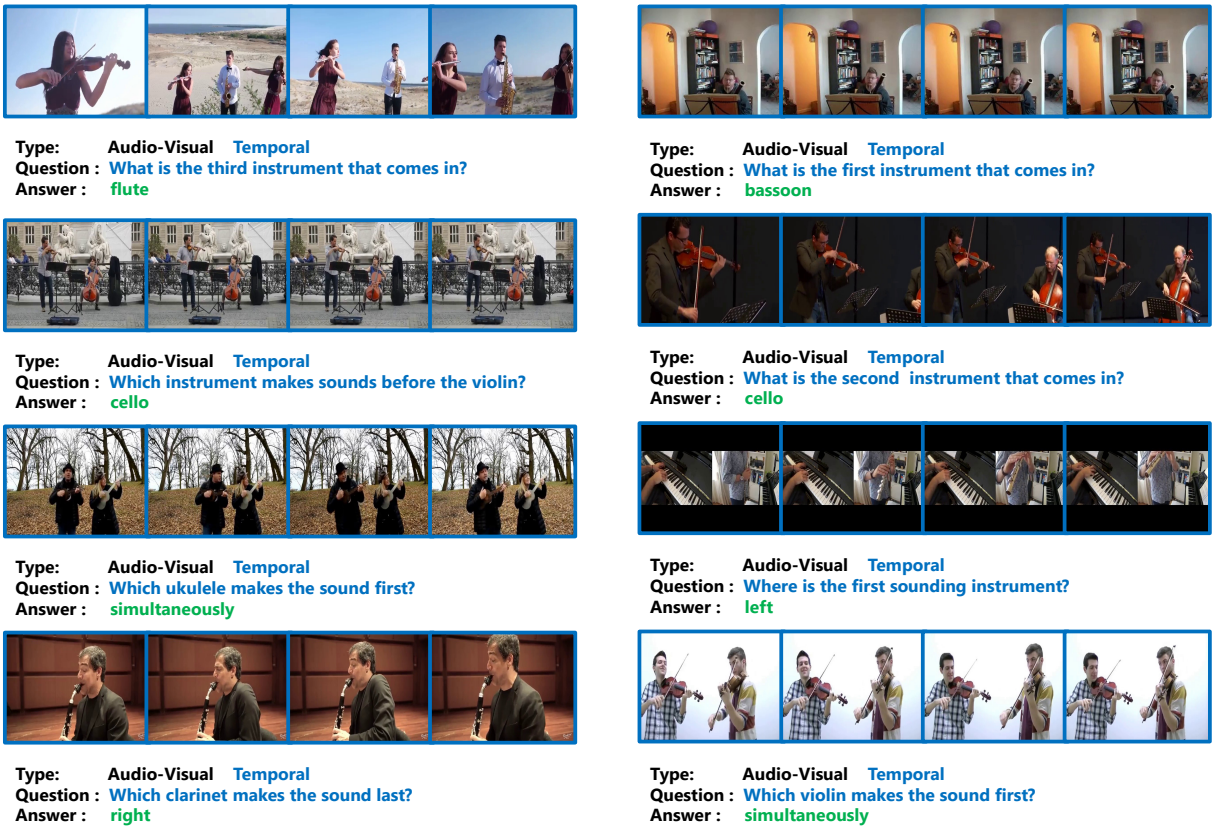
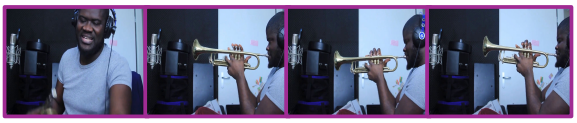


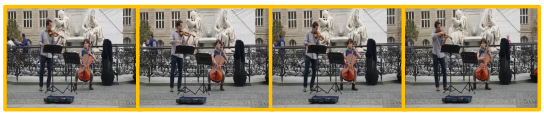
Figure 6: **Demonstration of Audio-Visual Temporal Question-Answering.** Our model handles the Audio-Visual Temporal questions correctly under complicated scenarios.

Type: Audio-Visual Existential
Question : Is there a voiceover?
Answer : no

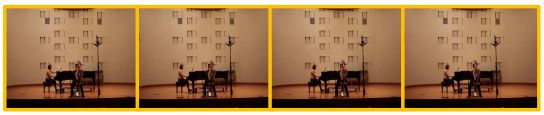


Type: Audio-Visual Existential
Question : Is there a voiceover?
Answer : yes

Figure 7: **Demonstration of Audio-Visual Existential Question-Answering.** Our model handles the Audio-Visual Existential questions correctly under complicated scenarios.



Type: Audio-Visual Comparative
Question : Is the violin on the left more rhythmic than the cello on the right?
Answer : yes



Type: Audio-Visual Comparative
Question : Is the tuba on the right more rhythmic than the piano on the left?
Answer : yes

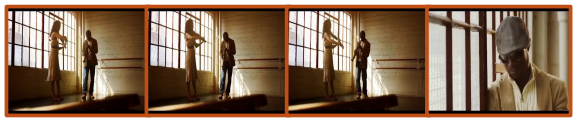
Figure 8: **Demonstration of Audio-Visual Comparative Question-Answering.** Our model handles the Audio-Visual Comparative questions correctly under complicated scenarios.

Type: Audio-Visual Location
Question : Which is the musical instrument that sounds at the same time as the pipa?
Answer : acoustic_guitar

Type: Audio-Visual Location
Question : Is the first sound coming from the middle instrument?
Answer : yes

Figure 9: **Demonstration of Audio-Visual Location Question-Answering.** Our model handles the Audio-Visual Location questions correctly under complicated scenarios.

Type: Visual Counting
Question : Are there violin and ukulele instruments in the video?
Answer : yes



Type: Visual Counting
Question : Are there pipa and electric_bass instruments in the video?
Answer : no

Figure 10: **Demonstration of Visual Counting Question-Answering.** Our model handles the Visual Counting questions correctly under complicated scenarios.

Type: Visual Location
Question : What kind of instrument is the leftest instrument?
Answer : violin

Type: Visual Location
Question : What kind of musical instrument is it?
Answer : cello

Figure 11: **Demonstration of Visual Location Question-Answering.** Our model handles the Visual Location questions correctly under complicated scenarios.

Type: Audio Counting
Question : Are there clarinet and acoustic_guitar sound?
Answer : yes

Type: Audio Counting
Question : Are there acoustic_guitar and accordion sound?
Answer : yes

Figure 12: **Demonstration of Audio Counting Question-Answering.** Our model handles the Audio Counting questions correctly under complicated scenarios.

References

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*.

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. 2022. Htsat: A hierarchical token-semantic audio transformer for sound classification and detection. In *International Conference on Acoustics, Speech and Signal Processing*.

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*.

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*.

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*.

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*.

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*.

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

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*.

Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. 2024. Look, listen, and answer: Overcoming biases for audio-visual question answering. In *Advances in Neural Information Processing Systems*.

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

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*.

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*.

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

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*.

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. 2024a. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Tianyu Yang, Yiyang Nan, Lisen Dai, Zhenwen Liang, Yapeng Tian, and Xiangliang Zhang. 2024b. SaSR-net: Source-aware semantic representation network for enhancing audio-visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

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

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*.

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*.