🎼 Music Performance Audio-Visual Question Answering Requires Specialized Multimodal Designs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances in music understanding and generation have expanded the ability of AI models to process musical content across modalities. Alongside this progress, Music Performance Audio-Visual Question Answering (Music AVQA) has emerged as a new multimodal challenge, requiring reasoning over continuous, densely layered audio-visual performances through natural language queries. This position paper argues that Music AVQA constitutes a distinct multimodal reasoning task that demands specialized input processing and architectural designs. We systematically survey existing Music AVQA datasets and methods, analyze what kinds of specialized multimodal designs are critical for accurate question answering, and propose potential music-specific design directions for advancing Music AVQA methods. We hope this work will inspire broader attention and further research in multimodal musical understanding. To support the community, we provide an anonymous GitHub repository of relevant papers that will be continuously updated at `https://anonymous.4open.science/r/Survey4MusicAVQA`.

1 Introduction

> *"Music is a moral law. It gives a soul to the Universe, wings to the mind, flight to the imagination, a charm to sadness, gaiety and life to everything. It is the essence of order, and leads to all that is good and just and beautiful."*
>
> — Plato

Music plays an integral role in human culture and expression [1, 2], and this significance has motivated extensive research on modeling musical intelligence. In the AI community, recent advances in music understanding [3, 4, 5, 6, 7, 8] and generation [9, 10, 11, 12, 13, 14] have significantly expanded the capabilities of machine learning systems to model, interpret, and produce musical content. Parallel to these developments, Music Performance Audio-Visual Question Answering (Music AVQA) has emerged as a distinctive multimodal challenge [15, 16]. Unlike common scenarios with sparse and discrete audio signals, music performances exhibit a continuous and tightly interwoven blend of audio and visual signals—offering a uniquely rich context for fine-grained audio-visual scene understanding and temporal reasoning [17, 18].

Music AVQA poses unique challenges that differentiate it from conventional Question Answering (QA) tasks, as illustrated in Figure 1. While questions are framed in natural language, answering them requires reasoning over continuous, temporally evolving, and densely layered audio-visual signals [19, 17, 20]. Unlike conventional audio QA tasks—where sound events are typically isolated and temporally distinct—music performances involve overlapping sources of instruments and complex temporal dynamics that unfold across multiple timescales. For example, questions like "*Which instrument produces the loudest sound?*" require tracking dynamic intensity across multiple

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

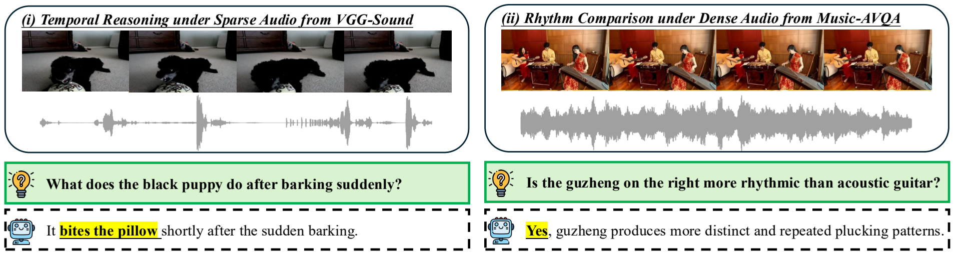


Figure 1: (i) Conventional QA with sparse audio (left) [21] vs. (ii) dense audio QA (right) [15]. (i) Involves an isolated barking sound and a closely synchronized bite action, which are relatively easy to detect. (ii) Involves overlapping instruments, rhythmic pattern modeling, and cross-modal comparison—highlighting the fine-grained temporal and spatial reasoning demands of Music AVQA, where **dense** and **continuous** audio signals pose unique challenges for multimodal understanding.

simultaneous sources. Similarly, questions like "*Is the cello on the right more rhythmic than the cello on the left?*" demand an understanding of both spatial relationships and temporal rhythmic patterns across visual and audio modalities. These examples collectively underscore the unique reasoning demands of Music AVQA and prompt a central question: what multimodal designs are well positioned to address them?

This paper argues that Music Performance Audio-Visual Question Answering (Music AVQA) constitutes a fundamentally distinct multimodal reasoning task, and that specialized multimodal designs are not only essential but empirically linked to strong model performance in this domain. To support this position, we present the first comprehensive survey of Music AVQA, focusing on how specialized multimodal design—spanning input processing and model architecture—enables effective reasoning in this uniquely challenging domain.

To advance this position, we organize the paper as follows. Section 2 provides background on Music AVQA. Section 3 reviews the evolution of benchmark datasets. Section 4 categorizes existing methods, while Section 5 focuses on how input processing pipelines are adapted for musical contexts. Section 6 analyzes existing Music AVQA methods and highlights design choices associated with strong performance. Section 7 distills insights into music-specific modeling strategies that may further advance the field. Through this work, we hope to draw more attention, elicit broader interest, and motivate additional research on multimodal understanding within rich musical environments.

2 Background

What are common music performance scene types? ① `Solo Performance` – A single musician showcasing technical skills and artistic expression on one instrument. ② `Ensemble of the Same Instrument` – Multiple musicians playing identical or related instruments, creating unified harmonies and textures. ③ `Ensemble of Different Instruments` – Musicians performing with a variety of instruments, producing diverse tonal colors and complex musical interactions. ④ `Culture-Specific Ensembles` – Traditional instrumental groups that embody the musical heritage and regional styles of specific cultures.

What are common question types in Music AVQA? ① `Existential Questions`: Determine whether a sound corresponds to a visible object in the scene (e.g., "Is this sound from the instrument in the video?"). ② `Counting Questions`: Quantify audio-visual elements that require cross-modal integration (e.g., "How many instruments are sounding in the video?"). ③ `Location Questions`: Identify the spatial position of sound sources within the visual scene (e.g., "Where is the first sounding instrument?"). ④ `Comparative Questions`: Compare properties across different audio-visual elements (e.g., "Is the instrument on the left louder than the one on the right?"). ⑤ `Temporal Questions`: Reason about the timing and sequential relationships between auditory and visual events (e.g., "Which instrument produces sound before the piano?").

What are the challenges of Music AVQA? Music AVQA presents several distinctive challenges: ① `Dense Signal Interpretation`: Unlike sparse audio events in conventional AVQA, music performances feature continuous, overlapping instrumental sources that require sophisticated separation and attribution; ② `Hierarchical Temporal Reasoning`: Musical information unfolds across multiple time scales (beats, phrases, sections), demanding models capable of reasoning across these hierarchical structures; ③ `Cross-Modal Correspondence`: Establishing reliable associations between visual instrumental actions and their acoustic outputs is complicated by temporal misalignments between physical gestures and the resulting sounds; ④ `Domain-Specific Knowledge`: Effective reasoning often depends on implicit musical knowledge, such as instrumental techniques, ensemble conventions, and acoustic properties; ⑤ `Abstract Attribute Quantification`: Questions involving subjective qualities such as "rhythmic", "melodic," or "harmonious" require computational strategies to map linguistic descriptors onto measurable signal properties; ⑥ `Data Scarcity`: The specialized nature of musical performances results in smaller and less diverse datasets compared to general AVQA tasks, limiting the generalization capabilities of trained models.

3 Evolution of MUSIC-AVQA Datasets

The development of Music AVQA research has been driven by progressively refined datasets addressing specific limitations. As summarized in Table 6 in Appendix Section C, this evolution began with the ① **MUSIC-AVQA** dataset [15], the first large-scale benchmark designed specifically for AVQA in musical contexts, comprising 9,288 performance videos and 45,867 question-answer pairs across diverse reasoning tasks. Subsequent research reveal challenges related to data bias and imbalanced answer distributions, prompting the creation of ② **MUSIC-AVQA v2.0** [16], which expands to 10,518 videos and approximately 54,000 question-answer pairs. This version balance 15 biased templates by ensuring no dominant answers exceed 60% for binary questions or 50% for multi-class questions, particularly enhancing representation in various question categories. Building on these foundations, ③ **MUSIC-AVQA-R** [18] introduce robustness evaluation through question rephrasing, expanding the test set from 9,129 to 211,572 questions. With a vocabulary five times larger than the original dataset, MUSIC-AVQA-R distinguishes between head (common) and tail (rare) samples, enabling assessment of model performance in both in-distribution and out-of-distribution scenarios. This progressive refinement of datasets has laid a solid foundation for advancing multimodal understanding and robust evaluation in music performance environments.

4 Categorization of Music AVQA Methods Based on Architecture

Music AVQA methods exhibit diverse architectural designs, particularly in how they encode and integrate textual, visual, and auditory modalities. To better organize existing approaches by their core modeling strategies, we categorize them into three groups—Transformer-based, CNN-based, and Hybrid models—as summarized in Table 1. This categorization highlights how different models are structured to handle the continuous and densely layered nature of musical performances.

Transformer-based models. Transformer-based models are characterized by the extensive use of self-attention mechanisms, which benefit in particular from their ability to handle long-range temporal dependencies and fine-grained cross-modal alignment. Methods such as Amuse utilize transformers across all modalities, combining a Swin Transformer for visual processing with an HTS-AT transformer for audio encoding, and employing cross-modal adapters to facilitate early and frequent fusion of multimodal information. Similarly, LAST-Att integrates a Swin-V2 Transformer for vision and an Audio Spectrogram Transformer (AST) for audio, emphasizing fine-grained spatial-temporal alignment through pixel-level cross-modal attention. Other methods such as LAVisH and LSTTA, adopt lightweight transformer adapters to inject multimodal cues into frozen transformer backbones, enabling efficient cross-modal reasoning while leveraging strong pre-trained representations.

CNN-based models. CNN-based methods typically utilize convolutional backbones such as ResNet or VGGish to encode modality-specific information into global or regional features, often relying on simpler late-stage fusion strategies. The AVST method exemplifies this approach, combining ResNet-18 visual embeddings and VGGish audio features through spatial attention modules to explicitly localize sound sources within visual frames. PSTP-Net extends this design by introducing a progressive refinement strategy that sequentially filters temporal segments and spatial regions, systematically narrowing down question-relevant audio-visual content prior to fusion. Although CNN-based models are computationally efficient and straightforward, their reliance on late fusion may pose challenges to capture the complex temporal dynamics characteristic of musical performances.

Table 1: Architectural summary of representative Music AVQA methods. Each method lists the text, visual, and audio encoders used, along with an indication of whether explicit spatial-temporal (S-T) modeling is incorporated. Detailed descriptions of each method are provided in Appendix D and E.

METHOD	Text Encoder	Visual Encoder	Audio Encoder	S-T
AMUSE [17]	Transformer [22]	Swin-Transformer-v2 [23]	HTS-AT [24]	✓
AUDIO FLAMINGO [25]	OPT-IML-MAX-1.3B [26]	-	ClapCap [27]	✓
AVMOE [28]	-	Swin-Transformer-v2 [23]	HTS-AT [24]	×
AVSD [29]	LSTM	LSTM	LSTM	×
AVSIAM [30]	-	ViT [31]	ViT [31]	×
AVST [15]	LSTM	ResNet-18 [32]	VGGish [33]	✓
CAT [34]	LLaMA2-7B [35]	ImageBind [36]	ImageBind [36]	×
CHATBRIDGE [37]	Vicuna-13B [38]	ViT-G [39]	BEATs [40]	×
CIGN [41]	-	ResNet-18 [32]	ResNet-18 [32]	✓
COCA [42]	Word Embedding	ResNet-18 [32]	VGGish [33]	×
CONVLSTM [43]	LSTM	-	Conv	×
CROSSMAE [44]	-	MAE [45]	AudioMAE [46]	×
DCL [47]	DeBERTa-V3-Large [48]	ViT [31]	AST [49]	✓
DG-SCT [50]	-	ViT [31]	HTS-AT [24]	✓
EEMC [51]	RoBERTa [52]	ViT [31]	VGGish [33]	✓
FCNLSTM [43]	LSTM	-	Conv	×
GPT-4O [53]	Transformer	CLIP-ViT	Transformer	×
GRU [19]	LSTM	VGGNet [54]	-	×
HCRN [55]	BiLSTM	ResNet-18 [32]	-	×
LAST-ATT [16]	LSTM	Swin-Transformer-v2 [23]	Audio-Spectrogram-Transformer	✓
LAVISH [56]	-	ViT [31]	ViT [31]	✓
LAVIT [57]	Transformer [22]	Transformer [22]	Transformer [22]	✓
LSTTA [58]	CLIP [31]	CLIP [31]	w2v-Conformer [59]	✓
MAVEN [60]	Mixtral	InternViT-300M-448px [61]	Transformer	×
MCAN [62]	GloVe [63]+LSTM	Faster R-CNN [64]	-	×
MCCD [18]	-	-	-	✓
MEERKAT [65]	LLaMA2-7B [35]	CLIP-ViT	CLAP [66]	✓
OGM [67]	-	ResNet-18 [32]	ResNet-18 [32]	×
ONELLM [68]	LLaMA2-7B [35]	CLIP-ViT	Unified Multimodal Encoder	×
OPM [67]	-	ResNet-18 [32]	ResNet-18 [32]	×
PSAC [69]	Word Embedding	CNN	-	×
PSTP-NET [70]	CLIP [31]	CLIP [31]	VGGish [33]	✓
QAP [71]	DeBERTa-V2-XLarge	CLIP [31]	CLAP [66]	×
QWEN2.5-VL [72]	MRoPE [72]	ViT [31]	-	×
REFATOMNET [73]	BERT	ViT [31]	-	✓
VALOR [74]	BERT	CLIP [31]	AST [49]	×
VAST [75]	BERT [76]	ViT [77]	BEATs [40]	×
VIDEOLLAMA-2 [78]	Transformer	CLIP [31]	BEATs [40]	✓
VITA [79]	Mixtral [80]	InternViT-300M-448px [61]	CNN	×

Hybrid models. Hybrid models combine CNNs, transformers, and large language models (LLMs) to enable unified multimodal reasoning. They typically employ pre-trained encoders from both CNN and transformer families, integrated through sophisticated cross-modal fusion mechanisms. Representative examples include ChatBridge, CAT, OneLLM, and Meerkat. ChatBridge utilizes a perceiver-based multimodal transformer to merge modalities via language-aligned latent representations, followed by a frozen LLM for reasoning. CAT introduces modality-specific clue aggregation modules on top of ImageBind encodings, enabling precise question-driven multimodal grounding before passing information to a generative LLaMA2 LLM. OneLLM further generalizes multimodal integration by introducing a universal projection mechanism that allows a single LLM to interpret diverse modality embeddings seamlessly. In contrast, Meerkat emphasizes fine-grained cross-modal alignment through an audio-visual optimal transport module that explicitly matches audio segments to corresponding visual regions, achieving strong performance on tasks requiring precise localization of sound sources, underscoring the benefit of precise local grounding for complex audio-visual interactions in musical contexts.

5 Music AVQA Requires Specialized Multimodal Input Processing

While input preparation is often treated as a fixed pipeline in general AVQA, music performance settings introduce unique challenges that make input fidelity, segmentation, and representation design especially consequential. Musical scenes are densely layered, temporally continuous, and rich in expressive detail, requiring greater care in how audio, visual, and textual inputs are captured and structured. In what follows, we examine how Music AVQA tasks motivate specialized input processing across three key fronts: maintaining high-resolution and synchronized multimodal signals, adapting tokenization to the structure of musical content, and managing the scale and diversity of music-specific data representations.

Continuous, high-fidelity, and tightly aligned inputs are foundational. Compared to event-centric AVQA tasks that typically involve short, discrete sound events and lower-resolution recordings, Music AVQA deals with continuous, polyphonic streams spanning multiple spatial and temporal scales. Audio is commonly sampled at high rates (44.1 kHz or above) and often preserved in lossless formats to retain subtle timbral and articulatory detail [81]. Visual inputs similarly tend to require higher resolution (1080p or above) and frame rates (30–60 fps) to capture nuanced performer motions such as bowing or fingering [82, 83]. Even modest temporal offsets—around 100–200ms—can affect the perceived correspondence between gesture and sound. To improve synchronization and cue isolation, some recent models adopt preprocessing strategies like beat-based segmentation [84] and harmonic-percussive separation [85], which can help surface rhythmically or acoustically meaningful content for downstream reasoning.

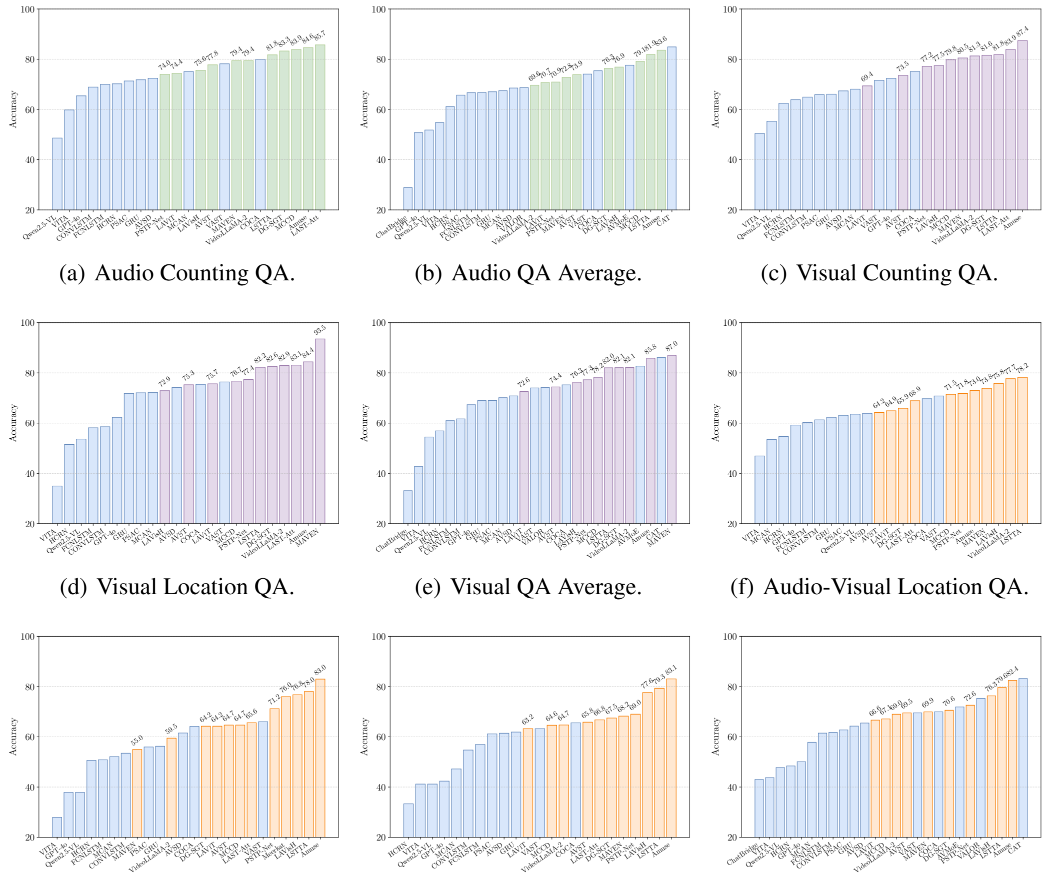
Tokenization strategies benefit from musical adaptation. Tokenization plays a central role in structuring inputs for multimodal reasoning, and recent Music AVQA models often tailor their strategies to preserve musical structure. For audio, models such as AMUSE [17], DG-SCT [50], and PSTP-NET [70] transform waveforms into Mel-spectrograms, which are then segmented via patch-based encoders like AST [49] and HTS-AT [24] or CNNs such as VGGish [33] and ResNet-18 [32]. AUDIO FLAMINGO [25], for instance, uses overlapping 7-second windows in CLAPCAP [27] to embed long-range audio context. Visual streams are frequently tokenized using ViT [31] or Swin-based [23] patch embeddings (e.g., in AVSIAM [30] and LAVISH [56]), while earlier models like AVST [15] use frame-level CNN features. Text tokenization is typically handled by subword models aligned with large language models (e.g., LLAMA2 [35], ROBERTA [52]), as seen in CHATBRIDGE [37] and ONELLM [68]. These tokenization schemes help preserve temporal granularity and modality alignment, which may be important for interpreting overlapping instruments, rhythmic changes, and localized visual cues.

Musical content introduces distinct data and representational considerations. Music AVQA tasks often involve long-form performances with overlapping sources and evolving musical dynamics, which can create challenges for segmentation, annotation, and generalization. Unlike typical AVQA datasets centered on short clips and isolated actions, music-focused benchmarks (e.g., MUSIC-AVQA [15]) include multi-instrument performances spanning several minutes. These conditions place greater demands on dataset diversity to avoid overfitting to genre-specific patterns or ensemble configurations. To broaden coverage, some models are trained on data drawn from live performances, studio recordings, and synthetic renderings. However, the absence of symbolic structure can limit the model's access to mid-level grounding. In this context, musically informed preprocessing (e.g., onset alignment, rhythmic segmentation, graph representation learning [86]) may support more interpretable and temporally aligned input representations.

6 Music AVQA Requires Specialized Spatial-Temporal Designs

We systematically analyze the models listed in Table 1 to identify architectural factors associated with strong Music AVQA performance across diverse multimodal designs. Each model is annotated based on whether it incorporates **spatial-temporal design**, defined as architectural components explicitly aimed at localizing audio-visual content in space and time—such as temporal segment selection, spatial attention, or cross-modal alignment modules. This categorization enables us to assess whether high-performing models exhibit structural traits aligned with the temporally continuous and spatially layered nature of musical performances.

To assess the empirical impact of spatial-temporal design, we evaluate Music AVQA models across representative question types grouped by modality—audio, visual, and audio-visual—as shown in Figure 2. Each subplot compares model accuracy on a specific QA type, with bars color-coded to indicate whether spatial-temporal design is applied for the relevant modality. This setup allows precise attribution of performance differences to design choices. To capture broader trends, Figure 3 summarizes average accuracy across all 13 QA categories using radar plots on two benchmarks: Music-AVQA and Music-AVQA-R. These visualizations reveal that models with spatial-temporal design consistently outperform their counterparts, particularly in tasks involving fine-grained localization or temporal sequencing. The full quantitative results supporting these figures are reported in Appendix A, Tables 2, 3, and 4. This experimental design enables systematic assessment of spatial-temporal design as a key architectural driver of multimodal reasoning in musical environments.



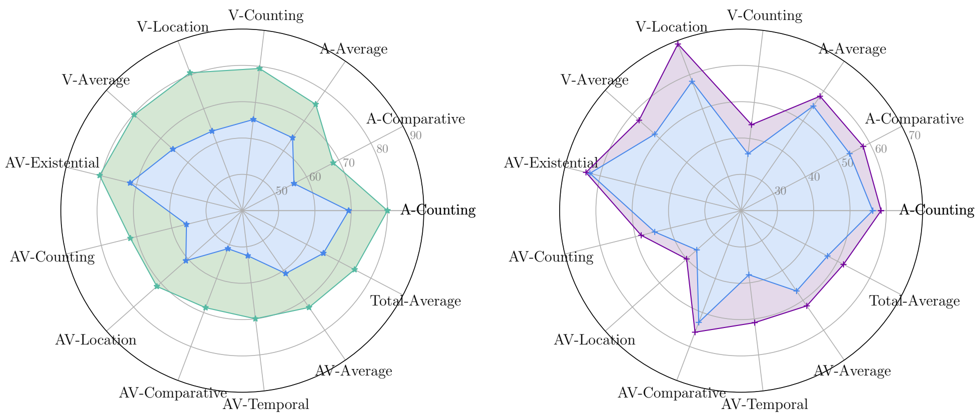
(a) Audio Counting QA. (b) Audio QA Average. (c) Visual Counting QA.

(d) Visual Location QA. (e) Visual QA Average. (f) Audio-Visual Location QA.

(g) Audio-Visual Comparative QA. (h) Audio-Visual Temporal QA. (i) Audio-Visual QA Average.

Figure 2: Accuracy comparison of Music AVQA models across representative question types, grouped by modality: (a–b) Audio, (c–e) Visual, and (f–i) Audio-Visual. Each bar corresponds to a model and is color-coded based on whether it incorporates **spatial-temporal design** for the relevant task type: bars in green, purple, and orange represent models that apply spatial-temporal modeling to Audio-related, Visual-related, and Audio-Visual-related question answering, respectively; bars in blue represent models without spatial-temporal design. Across most categories, models with spatial-temporal components tend to perform more accurately, particularly on tasks requiring temporal reasoning or spatial localization. These patterns suggest that incorporating spatial-temporal design supports more effective reasoning in musically structured multimodal environments.

Spatial-temporal design enhances audio QA by supporting fine-grained tracking of overlapping sources and temporally evolving acoustic cues. Audio-related questions in Music AVQA—such as instrument counting or loudness comparison—require models to distinguish simultaneous sound sources, localize temporal onsets, and resolve dynamic variations across time. As shown in Figures 2(a) and 2(b), models with spatial-temporal design consistently outperform others. LAST-ATT [16] achieves the highest audio counting accuracy at 85.71%, benefiting from repeated cross-attention between question-guided Swin-Transformer features and spectrogram patches from an Audio Spectrogram Transformer, which helps the model focus on musically salient moments. AMUSE [17], with 83.58% average audio QA accuracy, aligns audio-video streams using beat-synchronous features and temporally-adaptive fusion modules, allowing it to isolate relevant auditory content even under polyphonic conditions. DG-SCT [50] further introduces bidirectional attention layers across temporal, spatial, and channel dimensions, dynamically adjusting audio-visual focus based on the question's



(a) Methods on Music-AVQA [15].

(b) Methods on Music-AVQA-R [18].

Figure 3: Radar plots showing the per-type average accuracy of model groups with and without **spatial-temporal design** across 13 QA categories on (a) Music-AVQA [15] and (b) Music-AVQA-R [18]. Each axis corresponds to a QA type spanning audio, visual, and audio-visual reasoning, including the overall average (Total-Average). The filled green polygon in Figure 3(a) and purple polygon in Figure 3(b) represent the mean accuracy across QA types for models with spatial-temporal design, while the blue polygon represents the average performance of models without such design. Models with spatial-temporal design consistently achieve higher accuracy across all modality groups. These advantages persist under distribution shift in the robustness-focused Music-AVQA-R dataset.

semantics. By contrast, models lacking spatial-temporal structure—such as MCAN (67.47%) and CONVLSTM (66.73%)—often rely on global feature pooling or frame-agnostic fusion, making them vulnerable to overlap, misalignment, and temporal drift. Notably, spatial-temporal designs adopt recurring architectural motifs: temporal segment selection (PSTP-NET [70], AVST [15]), audio-guided visual attention (DG-SCT, LSTTA [58]), and fine-grained cross-modal alignment (MEERKAT [65]). These mechanisms are well-suited for modeling music's complex structure, where overlapping instruments and evolving rhythms require localized reasoning in both time and space. The strong performance of spatial-temporal models across audio QA tasks confirms their value in resolving multi-instrument scenarios and detecting temporally grounded acoustic attributes.

Spatial-temporal design improves visual QA by enhancing spatial disambiguation and capturing motion cues over time. Visual-related questions in Music AVQA—such as counting instruments or identifying positions—often involve tracking multiple performers, detecting visual cues of articulation (e.g., bowing, striking), and resolving spatial relationships within densely packed frames. As shown in Figures 2(c)–2(e), models with spatial-temporal components generally achieve stronger accuracy. For example, LSTTA [58] (82.03% visual QA average) combines short-term semantic interaction and long-term semantic filtering modules to capture both local gestures and global scene dynamics, enabling precise reasoning about when and where instruments are engaged. DG-SCT [50] (82.08%) uses cross-modal temporal attention guided by audio prompts to enhance visual token selection, focusing on visually active regions corresponding to sounding instruments. PSTP-NET [70] (77.26%) implements a region refinement module that explicitly filters visual patches within question-relevant segments, improving spatial disambiguation. While spatial-temporal modeling is effective, some models without it still perform competitively—most notably CAT [34] (86.10%), which leverages large-scale pretrained vision encoders (ImageBind) and LLaMA2 to infer structure implicitly. However, such models may rely heavily on correlation learned from pretraining, rather than explicit reasoning about visual dynamics. Spatial-temporal models, by contrast, explicitly model the temporal unfolding of gestures and the spatial focus of performer activity—important properties in musical scenes where instrument positions are static but their activation varies over time. These architectural patterns help stabilize attention and reduce confusion when multiple instruments are visually present but only some are active, contributing to more consistent visual QA performance across counting and localization tasks.

Spatial-temporal design is critical for audio-visual QA, where accurate reasoning requires precise temporal and spatial alignment between modalities. Among all Music AVQA categories, audio-visual questions impose the strongest demand on cross-modal synchronization, requiring the model to associate specific acoustic events with their visual sources over time. As shown in Figures 2(f)–2(i) and Table 2, models with spatial-temporal components consistently achieve higher accuracy across AV-Existential, AV-Counting, AV-Location, AV-Comparative, and AV-Temporal types. AMUSE [17] reaches 82.43% on overall AV questions by leveraging segment-level alignment between synchronized beat-level audio and video inputs and applying cross-modal adapters at each step. PSTP-NET [70] adopts a progressive three-stage pipeline: temporal segment selection, spatial region refinement, and audio-guided attention, resulting in 72.57% AV average. MEERKAT [65] further enhances local alignment by explicitly modeling cross-modal transport between audio patches and visual regions, and enforces bounding box constraints for grounding, yielding strong performance on AV-Comparative and AV-Location. In contrast, models without spatial-temporal design—such as MCAN (57.80%), GPT-4O (50.08%), and QWEN2.5-VL (47.75%)—struggle to resolve fine-grained multimodal relationships. While CAT [34] achieves 83.20% AV average through large-scale pre-trained encoders, its performance drops on AV-Temporal and AV-Location tasks that require precise temporal ordering or spatial binding. These results support that spatial-temporal designs—especially those involving temporally segmented reasoning, audio-guided spatial focus, and per-frame fusion—enable the model to track which instrument is sounding, when, and in which location, which is critical for answering questions such as "Did the cello on the left play after the drum on the right?". Without such structure, models tend to conflate co-occurring signals or miss temporally offset actions, leading to lower accuracy in complex cross-modal scenarios.

Spatial-temporal design provides a robust and generalizable structural advantage across diverse Music AVQA tasks. Our analysis reveals that models equipped with spatial-temporal design such as beat-synchronous segment alignment in AMUSE, progressive temporal-spatial filtering in PSTP-NET, and audio-guided token selection in DG-SCT—achieve consistently higher accuracy across audio (e.g., LAST-ATT: 85.71%), visual (e.g., LSTTA: 82.03%), and audio-visual (e.g., AMUSE: 82.43%) question types. These performance gains are particularly pronounced on tasks requiring temporal ordering or cross-modal localization, as shown in Figures 2 and 3. Despite some strong baselines using large-scale pretrained encoders, we observe that models lacking spatial-temporal design struggle with tasks requiring temporal resolution or spatial grounding. Notably, many high-performing models adopt a common architectural pattern: (1) identifying question-relevant time segments, (2) focusing on spatial regions associated with sound cues, and (3) fusing modalities with fine-grained temporal awareness. This recurring design motif underscores spatial-temporal design as not only empirically effective, but also structurally aligned with the demands of reasoning over continuous, densely layered musical performances.

7 Music AVQA Requires Specialized Musical Designs

Current Music AVQA models typically treat musical audio as generic acoustic input, operating directly on spectrograms or waveforms without incorporating structured musical attributes such as tempo, downbeats, key, or chord progressions. More fundamentally, human understanding of music relies on hierarchical temporal structure, harmonic organization, and latent causal intent—all of which are shaped by domain-specific knowledge and perceptual priors. Inspired by this observation, we argue that musical audio should not be treated as raw signal alone, but as a richly structured modality requiring specialized processing and reasoning capabilities.

To this end, we propose four concrete directions that embed musical priors and inductive structure into Music AVQA models. These ideas reflect our core insight: musical understanding demands not just better representations, but task-aware mechanisms that account for event timing, structural alignment, latent musical trajectories, and multi-step reasoning grounded in audio-visual context.

Incorporating fine-grained musical event cues. To support precise temporal reasoning over musical events—such as the entrance or exit of specific instruments—models can benefit from auxiliary timestamp supervision derived from musically meaningful proxies. For example, combining waveform peak analysis, Mel-frequency cepstral coefficients (MFCCs), and spectral change detection can help identify dynamic shifts in the audio stream. Beat-tracking algorithms (e.g., from Librosa) can segment audio by rhythm, while pitch-based estimators (e.g., Aubio's YIN) can trace changes in dominant frequency to indicate evolving instrumental activity. These mid-level cues can be used to generate pseudo-labels for training timestamp encoders, enabling models to better localize temporally anchored events. Embedding such representations into Music AVQA pipelines may improve event-level understanding and enhance the interpretability of the model's temporal predictions.

Embedding mid-level musical structure into multimodal models. Structured musical features—such as tempo, key, downbeats, and chord progressions—can provide a coherent framework for aligning audio-visual inputs across time. These symbolic or MIR-derived signals offer interpretable, temporally smooth trajectories that reflect the hierarchical organization of music, such as phrases, sections, and transitions. Crucially, they abstract away from low-level waveform fluctuations and offer a musically meaningful scaffold that persists across different genres, tempos, and instrumentation. By integrating them as auxiliary inputs or attention-guiding signals, models may improve their ability to capture long-range dependencies, maintain rhythmic continuity, and resolve ambiguous instrument interactions—especially in polyphonic or ensemble contexts. This structured conditioning can serve as a musical inductive bias, particularly helpful in complex multimodal scenes where overlapping sources challenge simple bottom-up fusion strategies, and where salient events may not be visually or acoustically distinct without temporal alignment cues.

Modeling latent musical reasoning trajectories. Many Music AVQA questions require reasoning over implicit causal or temporal relationships—for example, identifying which performer initiated a musical phrase, or determining whether an instrument's entrance shifted the ensemble's dynamic balance. These questions often lack explicit step-level supervision, making it difficult to learn reasoning paths from labels alone. To address this, models can incorporate latent reasoning trajectories: structured internal variables that represent evolving hypotheses about the musical scene. Rather than directly mapping inputs to answers, the model infers intermediate latent states—such as "which instrument is currently leading," "how the rhythmic intensity is changing," or "which performer is preparing to enter"—and updates these states over time as more multimodal evidence arrives. Architecturally, this can be implemented via hierarchical latent-variable models or recurrent variational modules, where latent states encode musical intentions, transitions, or causal flow. These hidden trajectories allow the model to simulate plausible sequences of musical events, enabling it to answer questions that require extrapolating or filling in missing links between observed signals. Crucially, this style of latent reasoning supports robust generalization by embedding inductive structure aligned with how humans infer musical cause and progression—not just surface-level audio-visual co-occurrence.

Supervising chain-of-thought reasoning in musical QA. Some musical questions—especially those involving temporal or causal dependencies—require sequential sub-decisions to reach the correct answer. For instance, the question "Which instrument enters after the piano stops?" involves: (1) detecting piano cessation, (2) identifying subsequent onsets, and (3) selecting the earliest new instrument. Rather than treating such questions as black-box classification, models can be explicitly trained to emit intermediate reasoning steps, either through supervised rationales or pseudo-labels derived from MIR-based event detection. This approach—akin to chain-of-thought (CoT) prompting in LLMs—improves transparency, encourages modular subgoal learning, and helps the model maintain alignment across modalities. Moreover, step-wise supervision can highlight failure points in temporal or semantic inference, offering clearer diagnostics for model improvement. In music contexts, CoT chains can incorporate domain-specific steps such as beat alignment, timbre matching, or onset-event attribution. These interpretable intermediate traces not only support higher accuracy on multi-stage queries but also make it easier to identify reasoning shortcuts and dataset biases.

8 Conclusion

This position paper argues that Music Performance Audio-Visual Question Answering (Music AVQA) constitutes a distinct multimodal reasoning task that demands tailored input processing and architectural strategies to meet the unique challenges of continuous, densely layered musical performances. Empirical analyses suggest that spatial-temporal designs are often associated with stronger performance across Music AVQA benchmarks. In addition, we introduce four music-specific design directions to improve musical reasoning and alignment. As the first systematic survey in this domain, we hope our position paper serves as a valuable resource for researchers interested in multimodal musical understanding and stimulates further innovation in this emerging area.

References

[1] Ian Cross. Music, cognition, culture, and evolution. *Annals of the New York Academy of sciences*, 2001. Available: https://nyaspubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/j.1749-6632.2001.tb05723.x.

[2] Stephen Malloch and Colwyn Trevarthen. The human nature of music. *Frontiers in psychology*, 2018. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.01680/full.

[3] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *ICML Machine Learning for Music Discovery Workshop*, 2019. Available: https://github.com/MTG/mtg-jamendo-dataset.

[4] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. Available: https://arxiv.org/pdf/2311.07919.

[5] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*, 2023. Available: https://arxiv.org/pdf/2309.08730.

[6] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *International Conference on Acoustics, Speech and Signal Processing*, 2024. Available: https://arxiv.org/pdf/2308.11276.

[7] Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153*, 2024. Available: https://arxiv.org/pdf/2402.16153.

[8] Mengjie Zhao, Zhi Zhong, Zhuoyuan Mao, Shiqi Yang, Wei-Hsiang Liao, Shusuke Takahashi, Hiromi Wakaki, and Yuki Mitsufuji. Openmu: Your swiss army knife for music understanding. *arXiv preprint arXiv:2410.15573*, 2024. Available: https://arxiv.org/pdf/2410.15573.

[9] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. Available: https://arxiv.org/pdf/2301.11325.

[10] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/94b472a1842cd7c56dcb125fb2765fbd-Paper-Conference.pdf.

[11] Shuqi Dai, Huiran Yu, and Roger B Dannenberg. What is missing in deep music generation? a study of repetition and structure in popular music. *arXiv preprint arXiv:2209.00182*, 2022. Available: https://arxiv.org/pdf/2209.00182.

[12] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *AAAI Conference on Artificial Intelligence*, 2024. Available: https://arxiv.org/pdf/2304.12995.

[13] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. M2ugen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*, 2023. Available: https://arxiv.org/pdf/2311.11255.

[14] Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*, 2023. Available: https://arxiv.org/pdf/2306.00110.

[15] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Conference on Computer Vision and Pattern Recognition*, 2022. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Li_Learning_To_Answer_Questions_in_Dynamic_Audio-Visual_Scenarios_CVPR_2022_paper.pdf.

[16] Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Winter Conference on Applications of Computer Vision*, 2024. Available: https://openaccess.thecvf.com/content/WACV2024/papers/Liu_Tackling_Data_Bias_in_MUSIC-AVQA_Crafting_a_Balanced_Dataset_for_WACV_2024_paper.pdf.

[17] Xingjian Diao, Chunhui Zhang, Tingxuan Wu, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, and Jiang Gui. Learning musical representations for music performance question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. Available: https://aclanthology.org/2024.findings-emnlp.159.pdf.

[18] Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. Look, listen, and answer: Overcoming biases for audio-visual question answering. In *Advances in Neural Information Processing Systems*, 2024. Available: https://openreview.net/pdf?id=twpPD9UMUN.

[19] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision*, 2015. Available: https://openaccess.thecvf.com/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf.

[20] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. Multimodal music information processing and retrieval: Survey and future challenges. In *International Workshop on Multilayer Music Representation and Processing*, 2019. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8665366.

[21] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech and Signal Processing*, 201820. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9053174.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Conference on Computer Vision and Pattern Recognition*, 2022. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.pdf.

[24] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *International Conference on Acoustics, Speech and Signal Processing*, 2022. Available: https://arxiv.org/pdf/2202.00874.

[25] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: a novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, 2024. Available: https://arxiv.org/pdf/2402.01831.

[26] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. Available: https://arxiv.org/pdf/2212.12017.

[27] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *International Conference on Acoustics, Speech and Signal Processing*, 2024. Available: https://arxiv.org/pdf/2309.05767.

[28] Ying Cheng, Yang Li, Junjie He, and Rui Feng. Mixtures of experts for audio-visual learning. In *Advances in Neural Information Processing Systems*, 2025. Available: https://openreview.net/pdf?id=SNmuKbU0am.

[29] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Conference on Computer Vision and Pattern Recognition*, 2019. Available: https://arxiv.org/pdf/1904.05876.

[30] Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *European Conference on Computer Vision*, 2024. Available: https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/02220.pdf.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. Available: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016. Available: https://arxiv.org/pdf/1512.03385.

[33] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *International Conference on Acoustics, Speech and Signal Processing*, 2017. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45857.pdf.

[34] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, 2024. Available: https://arxiv.org/pdf/2403.04640.

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. Available: https://arxiv.org/pdf/2307.09288.

[36] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Conference on Computer Vision and Pattern Recognition*, 2023. Available: https://openaccess.thecvf.com/content/CVPR2023/papers/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.pdf.

[37] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. Available: https://arxiv.org/pdf/2305.16103.

[38] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. Available: https://lmsys.org/blog/2023-03-30-vicuna/.

[39] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. Available: https://arxiv.org/pdf/2303.15389.

[40] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, 2023. Available: https://arxiv.org/pdf/2212.09058.

[41] Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *International Conference on Computer Vision*, 2023. Available: https://openaccess.thecvf.com/content/ICCV2023/papers/Mo_Class-Incremental_Grouping_Network_for_Continual_Audio-Visual_Learning_ICCV_2023_paper.pdf.

[42] Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. Coca: Collaborative causal regularization for audio-visual question answering. In *AAAI Conference on Artificial Intelligence*, 2023. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26527.

[43] Haytham M. Fayek and Justin Johnson. Temporal reasoning via audio question answering. *Transactions on Audio, Speech, and Language Processing*, 2020. Available: https://arxiv.org/pdf/1911.09655.

[44] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Conference on Computer Vision and Pattern Recognition*, 2024. Available: https://openaccess.thecvf.com/content/CVPR2024/papers/Guo_CrossMAE_Cross-Modality_Masked_Autoencoders_for_Region-Aware_Audio-Visual_Pre-Training_CVPR_2024_paper.pdf.

[45] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition*, 2022. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.pdf.

[46] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *International Conference on Neural Information Processing Systems*, 2022. Available: https://arxiv.org/pdf/2207.06405.

[47] Changsheng Lv, Shuai Zhang, Yapeng Tian, Mengshi Qi, and Huadong Ma. Disentangled counterfactual learning for physical audiovisual commonsense reasoning. In *Advances in Neural Information Processing Systems*, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/29571f8fda54fe93631c41aad4215abc-Paper-Conference.pdf.

[48] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. Available: https://arxiv.org/pdf/2111.09543.

[49] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. Available: https://arxiv.org/pdf/2104.01778.

[50] Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. In *Advances in Neural Information Processing Systems*, 2023. Available: https://openreview.net/pdf?id=9MwidIH4ea.

[51] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *European Conference on Computer Vision*, 2024. Available: https://arxiv.org/pdf/2407.10957.

[52] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Chinese National Conference on Computational Linguistics*, 2021. Available: https://aclanthology.org/2021.ccl-1.108.pdf.

[53] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. Available: https://arxiv.org/pdf/2410.21276.

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Available: https://arxiv.org/pdf/1409.1556.

[55] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Conference on Computer Vision and Pattern Recognition*, 2020. Available: https://arxiv.org/pdf/2002.10698.

[56] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Conference on Computer Vision and Pattern Recognition*, 2023. Available: https://openaccess.thecvf.com/content/CVPR2023/papers/Lin_Vision_Transformers_Are_Parameter-Efficient_Audio-Visual_Learners_CVPR_2023_paper.pdf.

[57] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *International Conference on Computer Vision*, 2021. Available: https://arxiv.org/pdf/2110.05122.

[58] Hongye Liu, Xianhai Xie, Yang Gao, and Zhou Yu. Parameter-efficient transfer learning for audio-visual-language tasks. In *International Conference on Multimedia*, 2023. Available: https://dl.acm.org/doi/pdf/10.1145/3581783.3611939.

[59] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. Available: https://arxiv.org/pdf/2005.08100.

[60] Jie Ma, Zhitao Gao, Qi Chai, Jun Liu, Pinghui Wang, Jing Tao, and Zhou Su. Fortisavqa and maven: a benchmark dataset and debiasing framework for robust multimodal reasoning. *arXiv preprint arXiv:2504.00487*, 2025. Available: https://arxiv.org/pdf/2504.00487.

[61] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Conference on Computer Vision and Pattern Recognition*, 2024. Available: https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.pdf.

[62] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition*, 2019. Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Yu_Deep_Modular_Co-Attention_Networks_for_Visual_Question_Answering_CVPR_2019_paper.pdf.

[63] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Conference on Empirical Methods in Natural Language Processing*, 2014. Available: https://aclanthology.org/D14-1162.pdf.

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2015. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[65] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, 2025. Available: https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/08071.pdf.

[66] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. In *International Conference on Acoustics, Speech and Signal Processing*, 2023. Available: https://arxiv.org/pdf/2206.04769.

[67] Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multi-modal learning. *Transactions on Pattern Analysis and Machine Intelligence*, 2024. Available: https://arxiv.org/pdf/2410.11582.

[68] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Conference on Computer Vision and Pattern Recognition*, 2024. Available: https://openaccess.thecvf.com/content/CVPR2024/papers/Han_OneLLM_One_Framework_to_Align_All_Modalities_with_Language_CVPR_2024_paper.pdf.

[69] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: positional self-attention with co-attention for video question answering. In *AAAI Conference on Artificial Intelligence*, 2019. Available: https://doi.org/10.1609/aaai.v33i01.33018658.

[70] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *International Conference on Multimedia*, 2023. Available: https://dl.acm.org/doi/pdf/10.1145/3581783.3612293.

[71] Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, and Qiang Zhu. Querying as prompt: Parameter-efficient learning for multimodal language model. In *Conference on Computer Vision and Pattern Recognition*, 2024. Available: https://openaccess.thecvf.com/content/CVPR2024/papers/Liang_Querying_as_Prompt_Parameter-Efficient_Learning_for_Multimodal_Language_Model_CVPR_2024_paper.pdf.

[72] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. Available: https://arxiv.org/pdf/2502.13923.

[73] Kunyu Peng, Jia Fu, Kailun Yang, Di Wen, Yufan Chen, Ruiping Liu, Junwei Zheng, Jiaming Zhang, M Saquib Sarfraz, Rainer Stiefelhagen, et al. Referring atomic video action recognition. In *European Conference on Computer Vision*, 2024. Available: https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/02873.pdf.

[74] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. *Transactions on Pattern Analysis and Machine Intelligence*, 2025. Available: https://arxiv.org/pdf/2304.08345.

[75] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: a vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems*, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/e6b2b48b5ed90d07c305932729927781-Paper-Conference.pdf.

[76] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. Available: https://aclanthology.org/N19-1423.pdf.

[77] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Available: https://openreview.net/pdf?id=YicbFdNTTy.

[78] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. Available: https://arxiv.org/pdf/2406.07476.

[79] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. Available: https://arxiv.org/pdf/2408.05211.

[80] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. Available: https://arxiv.org/pdf/2401.04088.

[81] Olga Slizovskaia, Gloria Haro, and Emilia Gómez. Conditioned source separation for musical instrument performances. *Transactions on Audio, Speech, and Language Processing*, 2021. Available: https://arxiv.org/pdf/2004.03873.

[82] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Conference on Computer Vision and Pattern Recognition*, 2020. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Gan_Music_Gesture_for_Visual_Sound_Separation_CVPR_2020_paper.pdf.

[83] Yitong Jin, Zhiping Qiu, Yi Shi, Shuangpeng Sun, Chongwu Wang, Donghao Pan, Jiachen Zhao, Zhenghao Liang, Yuan Wang, Xiaobing Li, et al. Audio matters too! enhancing markerless motion capture with audio signals for string performance capture. *Transactions on Graphics*, 2024. Available: https://arxiv.org/pdf/2405.04963.

[84] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. madmom: A new python audio and music signal processing library. In *International Conference on Multimedia*, 2016. Available: https://arxiv.org/pdf/1605.07008.

[85] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *International Conference on Digital Audio Effects*, 2010. Available: https://www.audiolabs-erlangen.de/media/pages/resources/aps-w23/papers/7ef50065d9-1663763115/2010_FitzGerald_HarmonicPercussiveSep_DAFx.pdf.

[86] Jincheng Huang, Yujie Mo, Ping Hu, Xiaoshuang Shi, Shangbo Yuan, Zeyu Zhang, and Xiaofeng Zhu. Exploring the role of node diversity in directed graph representation learning. In *International Joint Conference on Artificial Intelligence*, 2024. Available: https://www.ijcai.org/proceedings/2024/0229.pdf.

[87] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Annual Meeting of the Association for Computational Linguistics*, 2016. Available: https://aclanthology.org/P16-2034/.

[88] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, 2016. Available: https://proceedings.neurips.cc/paper/2016/file/9dcb88e0137649590b755372b040afad-Paper.pdf.

[89] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Conference on Computer Vision and Pattern Recognition*, 2019. Available: https://arxiv.org/pdf/1904.04357.

[90] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. Available: https://dl.acm.org/doi/pdf/10.1145/3503161.3548291.

[91] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/90ce332aff156b910b002ce4e6880dec-Paper-Datasets_and_Benchmarks.pdf.

[92] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer, 2024. Available: https://arxiv.org/pdf/2306.14899.

Contents of Appendix

A Quantitative Comparison on Music AVQA Datasets

We present comprehensive quantitative comparisons of recent state-of-the-art methods on multiple Music AVQA datasets [15, 16, 18], shown in Table 2, 3, and 4. We evaluate the models across a diverse set of question categories, spanning Audio-related, Visual-related, and Audio&Visual-related reasoning tasks. For each dataset, we report accuracy metrics for subcategories such as Counting, Comparative, Location, Existential, and Temporal reasoning, along with average scores within each modality and the overall performance.

Table 2: Comparison with state-of-the-art methods on the Music AVQA [15] test set. We report the accuracy for Audio (Counting, Comparative), Visual (Counting, Location), and Audio-Visual (Existential, Counting, Location, Comparative, Temporal) question types, along with the average accuracy for Audio, Visual, Audio-Visual, and overall.

Methods	Audio-related QA			Visual-related QA			Audio&Visual-related QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
AMUSE [17]	84.61	82.45	83.58	87.41	84.39	85.84	86.95	85.49	73.01	82.98	83.06	82.43	83.52
AUDIO FLAMINGO [25]	-	-	-	-	-	-	-	-	-	-	-	-	-
AVMOE [28]	-	-	77.60	-	-	82.70	-	-	-	-	-	71.90	75.70
AVSD [29]	72.41	61.90	68.52	67.39	74.19	70.83	81.61	58.79	63.89	61.52	61.41	65.49	67.44
AVSIAM [30]	-	-	-	-	-	-	-	-	-	-	-	-	-
AVST [15]	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
CAT [34]	-	-	84.90	-	-	86.10	-	-	-	-	-	83.20	84.30
CHATBRIDGE [37]	-	-	28.90	-	-	33.10	-	-	-	-	-	43.00	78.90
CIGN [41]	-	-	-	-	-	-	-	-	-	-	-	-	-
COCA [42]	79.94	67.68	75.42	75.10	75.43	75.23	83.50	66.63	69.72	64.12	65.57	69.96	72.33
CONVLSTM [43]	68.88	63.06	66.73	64.89	58.55	61.68	82.81	55.99	61.30	53.45	54.73	61.75	62.61
CROSSMAE [44]	-	-	-	-	-	-	-	-	-	-	-	-	-
DCL [47]	-	-	-	-	-	-	-	-	-	-	-	-	-
DG-SCT [50]	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
EEMC [51]	-	-	-	-	-	-	-	-	-	-	-	-	-
FCNLSTM [43]	69.96	61.06	66.67	63.89	58.14	60.98	83.42	56.31	60.28	50.85	56.92	61.46	62.25
GPT-4O [53]	65.42	36.07	50.75	72.36	62.30	67.33	56.12	54.84	59.23	37.84	42.35	50.08	54.06
GRU [19]	71.82	58.90	67.04	66.06	71.82	68.97	81.41	60.30	62.32	56.23	61.89	64.26	66.00
HCRN [55]	70.21	45.62	61.14	62.41	51.51	56.90	52.94	42.07	54.70	50.59	33.33	48.41	52.54
LAST-ATT [16]	85.71	63.10	-	83.86	83.09	-	76.47	76.20	68.91	65.60	66.75	-	75.45
LAVISH [56]	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
LAVIT [57]	74.36	64.56	70.73	69.39	75.65	72.56	81.21	59.33	64.91	64.22	63.23	66.64	68.93
LSTTA [58]	81.75	82.04	81.90	81.82	82.23	82.03	83.46	79.11	78.23	78.02	79.32	79.63	81.19
MAVEN [60]	79.44	54.10	72.79	80.49	93.50	86.99	87.00	66.67	73.85	54.95	68.24	69.94	74.60
MCAN [62]	75.05	54.58	67.47	68.06	72.15	70.13	81.91	54.15	53.45	52.11	47.21	57.80	62.77
MCCD [18]	83.87	71.04	79.14	79.78	76.73	78.24	80.87	51.63	71.46	64.67	64.60	67.13	72.20
MEERKAT [65]	-	-	-	-	-	-	-	85.70	-	75.98	-	-	-
ONELLM [68]	-	-	-	-	-	-	-	-	-	-	-	-	47.60
OPM [67]	-	-	-	-	-	-	-	-	-	-	-	-	70.80
PSAC [69]	71.33	56.07	65.68	65.89	72.07	69.02	78.59	54.80	63.11	55.96	61.17	62.75	64.92
PSTP-NET [70]	73.97	65.59	70.90	77.15	77.36	77.26	76.18	73.23	71.80	71.19	69.00	72.57	73.52
QAP [71]	-	-	-	-	-	-	-	-	-	-	-	-	-
QWEN2.5-VL [72]	48.60	55.00	51.80	55.28	53.66	54.47	44.00	52.17	63.57	37.84	41.18	47.75	50.14
REFATOMNET [73]	-	-	-	-	-	-	-	-	-	-	-	-	-
VALOR [74]	-	-	68.70	-	-	74.20	-	-	-	-	-	75.30	78.90
VAST [75]	78.18	67.05	74.06	71.56	76.38	74.00	81.81	64.51	70.80	66.01	63.23	69.54	71.52
VIDEOLLAMA-2 [78]	79.44	52.46	69.64	81.30	82.93	82.11	77.00	63.44	77.69	59.46	64.71	68.98	72.56
VITA [79]	59.81	45.90	54.76	50.41	34.96	42.68	54.00	49.46	46.92	27.93	41.18	43.74	45.44

Table 3: Comparison with state-of-the-art methods on the Music AVQA v2.0 [16] test set. We report the accuracy for Audio (Counting, Comparative), Visual (Counting, Location), and Audio-Visual (Existential, Counting, Location, Comparative, Temporal) question types, along with the average accuracy for Audio, Visual, Audio-Visual, and overall.

Methods	Audio-related QA			Visual-related QA			Audio&Visual-related QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
AMUSE [17]	84.76	83.88	84.34	88.15	85.16	86.74	88.30	87.47	78.77	84.41	85.38	85.51	85.16
AVST [15]	81.74	62.11	72.46	79.08	77.64	78.40	72.12	69.03	65.05	63.98	60.57	66.26	71.08
DG-SCT [50]	83.66	62.47	73.64	82.05	82.97	82.48	83.43	72.70	64.65	64.78	67.34	70.38	74.08
LAST-ATT [16]	86.03	62.52	-	84.12	84.01	-	76.21	75.23	68.91	65.60	60.60	-	75.44
LAVISH [56]	84.36	58.57	72.17	83.25	81.46	82.40	73.26	73.45	65.64	64.26	60.82	67.75	72.34

To complement the above quantitative results, Table 5 lists one representative question–answer pair for each modality–reasoning combination. These examples make the abstract metrics more concrete and illustrate the diversity of Music-AVQA tasks evaluated in this paper.

Table 4: Comparison with state-of-the-art methods on the Music-AVQA-R [18] test set. We report the accuracy for Audio (Counting, Comparative), Visual (Counting, Location), and Audio-Visual (Existential, Counting, Location, Comparative, Temporal) question types, along with the average accuracy for Audio, Visual, Audio-Visual, and overall.

Methods	Audio-related QA			Visual-related QA			Audio&Visual-related QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
ATT-BLSTM [87]	60.00	49.55	54.77	32.15	47.97	48.89	54.33	39.46	32.52	51.00	24.45	40.35	40.35
AVSD [29]	50.92	54.20	52.56	35.21	68.11	52.20	64.13	36.68	27.14	58.99	40.83	45.55	45.55
CONVLSTM [43]	55.68	60.22	57.95	35.64	51.66	52.23	72.45	53.18	32.35	57.91	43.33	51.84	51.84
FCNLSTM [43]	51.36	57.96	54.66	33.53	52.96	50.09	71.64	51.98	34.96	57.40	33.90	49.98	49.98
GRU [19]	57.78	58.95	58.36	38.08	57.67	54.17	70.53	43.33	39.70	57.29	35.85	49.34	49.34
HCATTN [88]	51.65	53.12	52.38	32.86	60.09	50.02	63.85	39.77	36.01	54.47	36.54	46.13	46.13
HCRN [55]	54.42	39.81	47.11	32.71	45.34	43.88	53.63	39.67	37.08	35.10	42.30	41.56	41.56
HME [89]	58.28	56.63	57.45	33.71	65.93	54.40	66.12	39.91	40.18	56.89	37.76	48.17	48.17
LAVISH [56]	52.86	62.72	57.79	38.33	67.47	55.83	78.65	41.48	32.38	62.18	44.05	51.75	51.75
LAVIT [57]	47.01	47.86	47.43	31.39	66.35	48.01	37.21	53.02	36.87	43.05	42.17	42.46	42.46
MCAN [62]	67.59	54.49	61.04	45.64	64.37	58.62	59.29	53.86	45.02	51.49	46.35	51.20	51.20
MCCD [18]	75.78	63.43	69.60	61.76	73.43	68.80	76.18	50.55	50.92	62.15	66.95	61.35	61.35
PSAC [69]	54.85	52.77	53.81	37.99	66.83	53.25	53.05	47.14	38.14	48.53	36.46	44.66	44.66

Table 5: Examples of questions in Music AVQA categorized by modality and reasoning type.

Modal	Reasoning Type	Question	Answer
Audio-only	Counting	How many musical instruments were heard throughout the video?	One
	Comparative	Is the drum louder than the guitar?	Yes
Visual-only	Counting	How many types of musical instruments appeared in the entire video?	Two
	Localization	Where is the guitarist standing?	On the right
Audio & Visual	Existential	Is there a conductor in the video?	No
	Counting	How many unique instruments are present in the video?	One
	Localization	Which sound is coming from the left speaker?	Piano
	Comparative	Is the violin sound more dominant than the cello?	Yes
	Temporal	Which instrument starts playing first?	The piano

B How Music AVQA Differs from Traditional Multimodal Tasks

Music AVQA diverges from typical multimodal tasks in its design and objectives, particularly because it must handle the dense, continuous signals of music performance while capturing explicit instrument-to-sound relationships. In contrast to more conventional audio-visual tasks, music-oriented QA demands specialized mechanisms for precise temporal alignment, spatial grounding, and domain-specific knowledge of musical structures.

Multimodal sentiment analysis. Multimodal sentiment analysis typically fuses textual transcripts, facial expressions, and vocal intonations to infer emotional states. These models often process short clips, emphasizing emotional indicators such as voice pitch variations, prosodic cues, or facial gestures. However, these sentiment-oriented methods rarely confront situations involving overlapping, simultaneous audio sources that need precise identification and synchronization with distinct visual objects. By contrast, music AVQA must explicitly identify discrete instrumental sources within overlapping audio signals and temporally align these audio cues with corresponding visual events, requiring sophisticated cross-modal attention and specialized alignment mechanisms to determine exactly which musician or instrument is associated with each sound.

Multimodal video captioning. Multimodal video captioning systems focus primarily on summarizing high-level actions or events, using language generation modules that selectively attend to visual and auditory inputs. These models typically handle audio as supplementary background context or speech segments rather than detailed musical signals. They seldom require detailed temporal reasoning about overlapping or continuous audio events. By contrast, music AVQA explicitly models audio using methods such as beat-based segmentation and harmonic-percussive separation to handle continuous, overlapping instrument performances. While captioning models produce holistic, narrative descriptions by attending to relevant audio-visual segments as a whole, music AVQA models must segment musical signals based on precise temporal events (e.g., beats, note onsets) and correlate them closely with visual actions to answer questions involving intricate temporal or rhythmic comparisons.

Cross-Modal retrieval. Cross-modal retrieval approaches aim to match items from different modalities by learning common latent representations. Typically, such retrieval tasks employ dual encoders, encoding modalities independently with limited interaction. The emphasis is on capturing global semantic similarities rather than explicit temporal localization or fine-grained spatial correspondences. By contrast, the majority of Music AVQA methods predominantly adopt a three-encoder architecture, separately encoding text, audio, and visual modalities while integrating them through continuous cross-attention. This design allows models to explicitly reason about when specific visual actions (e.g., finger positions or bow movements) align with corresponding audio events, ensuring fine-grained temporal synchronization. Unlike retrieval models that project entire modalities into a shared space for high-level similarity comparison, Music AVQA systems incorporate sophisticated fusion strategies that maintain modality-specific details while enabling precise alignment and localization of musical instruments and performers within dynamic audio-visual scenes.

C Details of AVQA Datasets

C.1 Music AVQA Datasets

Table 6 provides a summary of three representative datasets specifically designed for music-related Audio-Visual Question Answering (AVQA) tasks.

C.2 Other Datasets

Table 7 contrasts Music-AVQA with other widely used benchmarks, highlighting for each dataset the single factor that diverges most sharply from the music-oriented setting.

Table 6: Evolution and Characteristics of Music AVQA Datasets: A Comparative Overview of MUSIC-AVQA [15], MUSIC-AVQA v2.0 [16], and MUSIC-AVQA-R [18] Benchmarks.

Dataset	**Brief Description**
MUSIC-AVQA [15]	The MUSIC-AVQA dataset represents a significant contribution to audio-visual question answering research, comprising 9,288 videos with over 150 hours of musical performances covering 22 instruments, generating 45,867 question-answer pairs. The dataset is randomly split into training, validation, and testing sets with 32,087, 4,595, and 9,185 QA pairs respectively, spanning 33 question templates across 9 question types including existential, location, counting, comparative, and temporal questions.
MUSIC-AVQA v2.0 [16]	The MUSIC-AVQA v2.0 dataset builds upon the original MUSIC-AVQA by addressing data bias issues, comprising 10,518 videos (9,288 from the original plus 1,230 new videos) with musical performances covering 22 instruments, generating approximately 54,000 question-answer pairs. The balanced dataset splits into training, validation, and testing sets with 36,700, 5,250, and 10,819 QA pairs respectively, spanning 33 question templates across 9 question types. The authors specifically balance 15 biased templates by ensuring no dominant answers exceed 60% for binary questions or 50% for multi-class questions, particularly enhancing representation of underrepresented answers in existential, counting, temporal, location, and comparative question categories.
MUSIC-AVQA-R [18]	The MUSIC-AVQA-R dataset proposed in this paper is an extension of MUSIC-AVQA specifically designed to evaluate the robustness of audio-visual question answering models. It expands the original test set through a human-machine collaboration mechanism that rephrases each question 25 times, increasing the number of questions from 9,129 to 211,572, and introduces distribution shifts to categorize questions into head (common) and tail (rare) samples. Compared to the original dataset, MUSIC-AVQA-R features a vocabulary size of 465 (five times larger than MUSIC-AVQA), provides more diverse question formulations while preserving inherent biases in the training and validation sets, and offers three evaluation metrics—head accuracy, tail accuracy, and overall accuracy—enabling researchers to assess model performance in both in-distribution and out-of-distribution scenarios, making it the first dataset specifically designed for robustness evaluation in audio-visual question answering tasks.

Table 7: Other representative benchmarks (AVQA [90], EgoSchema [91], FunQA [92], VALOR-1M [74], and VGG-Sound [21]) and the key difference each bears with respect to Music-AVQA [15].

Dataset	**Key Difference**
AVQA [90]	Builds multiple-choice QA on everyday VGG-Sound clips; questions target generic activities and causal relations in real-life videos, so it lacks the fine-grained instrument/sound localization and music-theory knowledge required in MUSIC-AVQA.
EgoSchema [91]	Uses first-person (Ego4D) footage that is three-minutes long, stressing long-range temporal reasoning in egocentric daily tasks; audio cues are incidental and the task is 5-way multiple choice, very different from the short, professionally filmed music performances and open-ended answers in MUSIC-AVQA.
FunQA [92]	Focuses on“surprising” humour/creative/magic clips (4.3 k videos, 312 k QAs) that test commonsense violations; audio is often background and questions centre on counter-intuitive visual events, not on synchronised musical notes or instrument semantics.
VALOR-1M [74]	A pre-training corpus (1 M videos) with tri-modal captions meant for retrieval/captioning; QA supervision is extremely sparse and relies on auto-generated subtitles, so it serves as a foundation model resource rather than a targeted AVQA evaluation set like MUSIC-AVQA.
VGG-Sound [21]	It is an audio-visual correspondent dataset consisting of short clips of audio sounds from YouTuBe. And it provides raw audio–visual correspondence but no question–answer supervision or fine-grained reasoning labels.

D Details of Spatial-Temporal Music AVQA Methods

Table 8 illustrates methods incorporating explicit spatial-temporal design components in detailed. For completeness, we also list the remaining methods without such special design in the following itemized format.

Table 8: Description of Representative Methods for **Spatial-Temporal Design** for Music AVQA.

Paper/Work	Brief Description
AMUSE [17]	Focuses on music performance scenarios by aligning time segments in both audio and video streams via a cross-attention paradigm. Exploits synchronized features (such as beat-level or note-level alignment) to capture subtle temporal dependencies among instruments in dense music passages. By integrating rhythmic cues and cross-modal interactions, it is particularly suited for questions that involve multiple instruments playing simultaneously or changing their patterns over time.
AVST [15]	Proposes a spatio-temporal grounded audio-visual approach that explicitly localizes sounding objects in each frame while applying a question-guided temporal attention mechanism. The model grounds audio-visual events and emphasizes which frames (visual) and which segments (audio) are most relevant for question answering. By combining localized visual features and temporal cues, it captures object interactions over time and can better handle questions involving spatial and temporal reasoning.
CIGN [41]	Learns audio-visual class tokens and an Audio-Visual Continual Grouping module that, at every time-step, pulls together frame-level spectrogram features and region features into category-aware clusters. A token-distillation schedule preserves past knowledge while the regrouping logic tracks objects and sounds through the video's timeline, giving the model temporally consistent, cross-modal semantics for spatial-temporal reasoning.
DCL [47]	Introduces a Disentangled Counterfactual Learning framework to handle physical audio-visual commonsense reasoning tasks. Decomposes video signals into static (time-invariant) and dynamic (time-varying) factors using a VAE-based encoder, enabling clearer separation of constant background features from changing events. Additionally employs a counterfactual intervention module on the dynamic factors to perform causal reasoning, helping the model answer "what if" questions related to temporal and event relationships.
DG-SCT [50]	Introduces a Dual-Guided Spatial-Channel-Temporal (DG-SCT) attention layer that is injected in every frozen audio and visual transformer block. Audio prompts steer visual tokens (and vice-versa) via bidirectional attention that highlights salient spatial regions, discriminative feature channels, and pivotal temporal segments, producing fine-grained spatio-temporal alignments that boost related tasks.
EEMC [51]	Divides audio/video into 1-s slices and fuses them with text through a Temporal Bi-modal Transformer backed by a cached-memory mechanism that magnifies sudden changes across time. The resulting multimodal cue stream then serves as a cross-attention prompt for the segmentation decoder, enabling precise localisation of objects and events as their spatial footprints and temporal order evolve.

Continued on next page

Paper / Work	Brief Description
LAST-ATT [16]	Tackles audio-visual question answering with a repeated cross-attention strategy. Uses Swin-Transformer-v2 for visual frame features and a specialized Audio Spectrogram Transformer for audio, then merges them based on the question. By repeatedly "attending" to the most relevant frames and spectrogram patches, it effectively localizes musical actions (e.g., a pianist's keystrokes) over time. This design is well suited for intricate temporal queries and locating key audio events in dense musical content.
LAVISH [56]	Adds a lightweight Latent Audio-Visual HYbrid adapter to every layer of a frozen ViT. A compact pool of latent tokens acts as a cross-attention bottleneck, letting audio frames gate visual tokens (and vice-versa) as the video unfolds, so spatial patches and framewise dynamics are fused early while keeping the backbone frozen.
LAVIT [57]	Targets 360° videos with a transformer that augments each patch by a quaternion-based spherical coordinate and aligns it with audio via joint contrastive objectives. The spherical embedding plus an auxiliary audio-skewness prediction head lets the model reason about where (on the sphere) and when a sound arises, delivering fine-grained spatial-temporal grounding beyond normal FOV clips.
LSTTA [58]	A parameter-efficient transfer learning approach for audio-visual-language tasks that adds dedicated adapter modules while freezing large pretrained backbones. Splits temporal modeling into two scales: a short-term semantic interaction module (for capturing local correlations such as brief instrumental flourishes) and a long-term semantic filtering module (for broader progressions over many frames). This structure helps the model identify when, how, and for how long different instruments contribute, achieving a refined spatio-temporal representation.
MAVEN [60]	Employs a Multimodal Audio-Visual Epistemic Network that cycles between audio, video and text logits, using debiasing constraints to keep modality-specific and fused predictions consistent over time. The cycle guidance implicitly anchors each question to the correct temporal segments while suppressing spurious correlations.
MCCD [18]	Introduces a Multifaceted Cycle-Collaborative Debiasing objective: KL penalties enlarge the gap between uni-modal and tri-modal logits at every timestep, then force the three unimodal paths to agree with each other. This temporal-cycle training steers attention toward frames (and sounds) that all modalities truly share, yielding stabler spatial-temporal grounding under distribution shift.
MEERKAT [65]	Employs a two-stage mechanism for fine-grained audio-visual grounding in space and time. First uses an Audio-Visual Optimal Transport (AVOpT) module for fine-grained local alignment between audio features and specific image patches. Next, the Audio-Visual Attention Consistency Enforcement (AVACE) module refines cross-modal attention maps to precisely locate audio sources within bounding boxes, enforcing spatial constraints and ensuring attention is focused on the correct visual objects that correspond to the audio signal.
PSTP-NET [70]	Proposes a Progressive Spatio-Temporal Perception framework for audio-visual QA. Divides the selection of relevant information into three modules: (1) the Temporal Segment Selection Module (TSSM) for picking key time segments pertinent to the question; (2) the Spatial Region Selection Module (SRSM) to identify essential visual patches within those segments; and (3) the Audio-guided Visual Attention Module (AVAM) to align selected visual patches with the audio signals. This stepwise process helps isolate question-relevant data and reduce interference.

Continued on next page

Paper / Work	Brief Description
REFATOMNET [73]	For referring atomic actions, it runs three streams—visual, text and location-semantic tokens— and merges them through novel cross-stream agent-attention blocks. The location-semantic stream provides per-person bounding-box hints over time, letting the network lock onto the described individual before classifying frame-level atomic actions, thus tightly coupling spatial localisation with temporal action cues.
VIDEOLLAMA-2 [78]	Builds a video-LLM around a Spatial-Temporal Convolution (STC) connector that first performs per-frame spatial mixing and then downsamples temporally, giving the language model a compact yet order-aware token sequence. A jointly-trained audio branch injects synchronized spectrogram tokens, enabling the model to answer audio-visual questions that hinge on both where events happen on screen and when they unfold.

E Details of Existing Music AVQA Methods

- AVMOE [28]: The paper proposes a parameter-efficient transfer learning framework for audio-visual tasks by dynamically integrating intra-modal and inter-modal information through a mixture of experts. The approach introduces unimodal adapters to capture within-modality details and cross-modal adapters to model interactions between audio-visual streams, while a lightweight modality-agnostic router dynamically allocates expert weights based on input characteristics. By combining these components, AVMoE adaptively balances modality-specific and cross-modal features, addressing challenges like missing modalities or noisy inputs, thereby enhancing robustness and performance across diverse audio-visual tasks such as AV localization, segmentation, and question answering without requiring full model fine-tuning.
- AVSD [29]: The paper proposes an end-to-end baseline for audio-visual scene-aware dialog to enhance virtual assistants by integrating multimodal signals. The method employs an attention mechanism to differentiate useful signals from distractions, while maintaining spatial features from video frames (VGG19/I3D-Kinetics) to preserve contextual details and temporally subsampling frames to improve efficiency. By fusing attended vectors across audio, video, and text modalities, the approach dynamically focuses on relevant cues during answer generation. This integrated framework addresses challenges in holistic dialog management, leveraging cross-modal interactions to outperform prior methods without relying on rigid pipelines, as demonstrated on the audio-visual scene-aware dataset.
- AVST [15]: The paper proposes a novel approach to Audio-Visual Question Answering (AVQA) by integrating multimodal understanding and spatio-temporal reasoning in dynamic audio-visual scenarios. It introduces the MUSIC-AVQA dataset with 45K QA pairs to benchmark the task, while addressing spatial associations through an attention-based sound source localization module (AV-Loc) to link sounds with visual sources. Temporal grounding (Q-Temp) is achieved by using question features to highlight key timestamps, enabling effective encoding of question-aware audio-visual embeddings. These components are fused to jointly represent spatial and temporal cues, overcoming challenges in cross-modal reasoning and enhancing performance in complex audio-visual scenes without relying on single-modality methods. The integrated framework demonstrates superior scene understanding by leveraging multisensory perception and fine-grained spatio-temporal analysis.
- AVSIAM [30]: The paper proposes an efficient and scalable audio-visual learning framework using a shared vision transformer backbone to unify audio and visual processing. The AVSiam model employs a contrastive audio-visual matching objective with a multi-ratio random masking scheme to enhance representation robustness while enabling larger batch sizes for effective contrastive learning. By sharing parameters across modalities, the approach reduces GPU memory footprint and computational costs compared to dual-backbone methods, while maintaining competitive performance on classification and retrieval tasks. This integrated design addresses scalability challenges and modality-handling flexibility without compromising accuracy.
- AMUSE [17]: The paper proposes a framework for music audio-visual question answering that addresses the unique challenges of dense, continuous audio-visual signals in musical performances.

To exploit multimodal interconnectivity, it employs cross-modal adapters to facilitate early-stage token interactions between Swin-V2 (video), HTS-Audio (audio), and language transformers, while annotating rhythm and music sources in datasets to explicitly model musical characteristics. For temporal alignment, it designs specialized encoders to link musical signals with time dimensions. This integrated approach overcomes limitations of general-purpose AVQA methods by capturing intricate audio-visual relationships in performances, enhancing accuracy for music-specific questions through rhythm-aware and temporally grounded representations.

- ATT-BLSTM [87]: The paper proposes an attention-based bidirectional LSTM network (Att-BLSTM) for relation classification to capture decisive semantic information without relying on lexical resources or NLP systems. The model processes raw text through an embedding layer to generate word vectors, while bidirectional LSTM (BLSTM) layers learn high-level features by incorporating both past and future context. An attention mechanism then assigns weights to key words, merging word-level features into a sentence-level vector for classification. By integrating these components, the approach overcomes limitations of manual feature engineering and dependency on external tools, effectively identifying critical semantic cues across sentence positions to improve relation classification performance.
- AUDIO FLAMINGO [25]: The paper proposes Audio Flamingo, a novel audio language model designed to enhance large language models' (LLMs) understanding of non-speech sounds and non-verbal speech through three key innovations. It employs a sliding-window audio feature extractor to preserve temporal information in variable-length audio, while cross-attention mechanisms efficiently fuse audio inputs into the LM to reduce computational overhead. The model leverages a curated heterogeneous dataset and a two-stage training approach (pre-training and supervised fine-tuning) to balance close-ended and open-ended tasks. Additionally, it integrates in-context learning (ICL) and retrieval-augmented generation (RAG) through tailored templates and cross-attention masks, enabling few-shot adaptation without fine-tuning. To support multi-turn dialogues, the model is fine-tuned on GPT-4-generated datasets with correlated context. By combining these techniques, Audio Flamingo addresses challenges in audio feature extraction, heterogeneous data training, task adaptation, and dialogue coherence, achieving state-of-the-art performance across.
- CAT [34]: The paper proposes an enhanced Multimodal Large Language Model (MLLM) to improve question answering in dynamic audio-visual scenarios by addressing ambiguity and localization challenges. Key components include a clue aggregator to dynamically capture question-aware audio-visual features for fine-grained grounding, a mixed training strategy combining video-text and audio-text pairs with a novel AVinstruct dataset to strengthen cross-modal awareness, and an AI-assisted Ambiguity-aware Direct Preference Optimization (ADPO) to retrain the model for precise responses. By integrating these innovations, CAT effectively mitigates ambiguous outputs and enhances audio-visual reasoning, outperforming existing methods in Audio-Visual Question Answering (AVQA) tasks.
- CIGN [41]: The paper proposes a novel framework for continual audio-visual learning by disentangling class-aware cross-modal representations to mitigate catastrophic forgetting. It introduces learnable audio-visual class tokens to continually aggregate category-wise features through the Audio-Visual Continual Grouping module, while the Audio-Visual Class Tokens Distillation module preserves knowledge from previous tasks by aligning old and new token distributions. By integrating these components, the approach effectively addresses the challenge of mixed audio semantics and forgetting in sequential tasks, enhancing discriminative feature learning across modalities without relying on single-modality or rehearsal-based methods. The CIGN framework demonstrates superior performance in class-incremental audio-visual scenarios through its ability to maintain compact and disentangled representations.
- COCA [42]: The paper proposes a collaborative causal regularization framework (COCA) to address multi-shortcut biases in Audio-Visual Question Answering (AVQA) by integrating causal intervention and dynamic debiasing. The Bias-centered Causal Regularization (BCR) mitigates specific shortcut biases (Q→G, V&Q→G, A&Q→G) through counterfactual interventions to disrupt bias-irrelevant causal effects and factual regularization to maintain semantic consistency, while the Multi-shortcut Collaborative Debiasing (MCD) dynamically adjusts debiasing focus per sample using an entropy-driven metric to balance bias contributions. By jointly addressing unimodal and joint-modal biases through causal introspection and instance-aware adaptation, COCA enhances multimodal reasoning robustness without over-correcting, achieving state-of-the-art performance on MUSIC-AVQA.

- CONVLSTM [43]: The paper proposes a novel approach to enhance temporal reasoning in Audio Question Answering (AQA) by introducing the Diagnostic Audio Question Answering (DAQA) dataset, which comprises natural sound events and programmatically generated questions to probe temporal reasoning skills, while adapting visual question answering methods to AQA reveals their limitations. To address this, the authors develop Multiple Auxiliary Controllers for Linear Modulation (MALiMo), which extends Feature-wise Linear Modulation (FiLM) by incorporating an additional auxiliary controller to process subsampled audio features, thereby enabling dynamic modulation of convolutional network processing based on both input modalities. This integrated approach improves relational and temporal reasoning by jointly leveraging audio and question inputs, overcoming the shortcomings of existing methods in handling complex temporal dependencies within sound sequences.

- CHATBRIDGE [37]: The paper proposes a multimodal language model that leverages large language models (LLMs) as a universal interface to bridge diverse modalities through language-paired data. ChatBridge integrates modality-specific encoders and perceiver modules to project embeddings into the LLM's semantic space, enabling cross-modal correlation without requiring all paired data combinations. The model undergoes two-stage training: first aligning modalities with language to emergent multimodal abilities, then instruction-finetuning on the MULTIS dataset to enhance zero-shot task generalization. By using language as a catalyst, ChatBridge addresses the challenge of limited multimodal paired data while achieving strong performance across text, image, video, and audio tasks through unified multimodal reasoning and user intent alignment.

- CROSSMAE [44]: The paper proposes a region-aware audio-visual pre-training framework to enhance cross-modality interaction and fine-grained alignment by extending masked autoencoders. It introduces Cross-Conditioned Reconstruction to reconstruct input pixels conditioned on cross-modal Attentive Tokens, while Cross-Embedding Reconstruction leverages Learnable Queries with positional cues to guide feature reconstruction between modalities, supplemented by contrastive loss for global alignment. By integrating these components, CrossMAE addresses the limitations of prior global feature-based methods, enabling effective region-level understanding and improving performance in both classification and dense prediction tasks without task-specific fine-tuning.

- DCL [47]: The paper proposes a disentangled counterfactual learning approach for physical audio-visual commonsense reasoning to infer objects' physics properties from video and audio inputs. The method decouples videos into static (time-invariant) and dynamic (time-varying) factors through a disentangled sequential encoder (DSE) using a variational autoencoder and contrastive loss to maximize mutual information while minimizing cross-factor interference. It further introduces a counterfactual learning module (CLM) to model physical knowledge relationships among objects by applying counterfactual interventions as confounders to enhance causal reasoning. By integrating DSE's disentangled representations with CLM's causal learning, the approach effectively addresses challenges in extracting implicit physical knowledge from multi-modal data, improving reasoning explainability and performance without relying on mixed feature representations.

- DG-SCT [50]: The paper proposes a novel Dual-Guided Spatial-Channel-Temporal (DG-SCT) attention mechanism to enhance large pre-trained models for audio-visual tasks by dynamically adjusting feature extraction through cross-modal guidance. The DG-SCT mechanism leverages audio and visual modalities as soft prompts to adaptively refine features across spatial, channel, and temporal dimensions, while preserving frozen pre-trained parameters. By integrating trainable cross-modal interaction layers into encoders, the approach emphasizes task-relevant information in each modality, addressing limitations of single-modality pre-training. This bidirectional prompting enables fine-grained feature fusion, improving performance on downstream tasks like AVE, AVVP, AVS, and AVQA without full retraining, while also excelling in few-shot and zero-shot scenarios.

- EEMC [51]: The paper proposes a novel task called Reference Audio-Visual Segmentation (Ref-AVS) to segment visual objects using expressions enriched with multimodal audio-visual cues, addressing the limitations of unimodal approaches. It introduces the Ref-AVS benchmark with pixel-level annotations and diverse multimodal-cue expressions to enable training and evaluation, while an end-to-end framework leverages a crossmodal transformer to process and integrate multimodal cues for precise segmentation. By simultaneously utilizing audio and visual descriptions in natural language, the approach overcomes challenges in locating objects in dynamic audio-visual scenes, enhancing segmentation accuracy in complex real-world scenarios without relying on manual mask annotations or single-modality references.

- FCNLSTM [43]: The paper proposes a novel approach to enhance temporal reasoning in Audio Question Answering (AQA) by introducing the Diagnostic Audio Question Answering (DAQA) dataset, which comprises natural sound events and programmatically generated questions to probe temporal reasoning skills. While adapting existing visual question answering methods to AQA reveals their limitations in temporal reasoning, the authors develop Multiple Auxiliary Controllers for Linear Modulation (MALiMo) to extend Feature-wise Linear Modulation (FiLM) by incorporating an additional auxiliary controller to process subsampled audio features, thereby enabling dynamic modulation of convolutional network processing based on both principal and supplementary inputs. This integrated approach addresses the challenge of in-depth temporal reasoning by facilitating relational and temporal analysis, leading to improved performance on DAQA without relying on spatial reasoning or static inputs.

- GPT-4o [53]: The paper proposes GPT-4o, an autoregressive omni model designed to process any combination of text, audio, image, and video inputs while generating text, audio, or image outputs through end-to-end training across modalities. By integrating Web Data, Code and Math, and Multimodal Data during pre-training, the model learns diverse reasoning skills and multimodal interpretation, while post-training alignment and red-teaming mitigate risks such as bias and harmful content. This unified approach enhances real-time responsiveness, multilingual performance, and multimodal understanding while addressing safety concerns through layered mitigations and external evaluations.

- GRU [19]: The paper proposes a free-form, open-ended Visual Question Answering (VQA) task to generate natural language answers from images and questions, mirroring real-world scenarios like assisting the visually impaired. The approach leverages a large dataset (0.25M images, 0.76M questions, 10M answers) combining real images from MS COCO and abstract scenes to enable both low-level vision and high-level reasoning. By supporting diverse question types (e.g., fine-grained recognition, commonsense reasoning) and offering automatic evaluation through open-ended or multiple-choice formats, the framework addresses the need for detailed image understanding and multi-modal knowledge integration, advancing AI-complete challenges beyond generic captioning.

- HCAttn [88]: The paper proposes a hierarchical co-attention model for Visual Question Answering (VQA) that jointly reasons about image and question attention to improve answer accuracy. It introduces a co-attention mechanism to simultaneously perform question-guided visual attention (to identify relevant image regions) and image-guided question attention (to focus on key words), while employing a hierarchical question representation through word-level embeddings, phrase-level 1D CNNs (to capture n-gram features), and question-level LSTMs (to encode contextual meaning). By recursively combining co-attended features across these levels, the model addresses challenges like linguistic variation and multi-modal alignment, enhancing robustness and fine-grained understanding for VQA tasks.

- HCRN [55]: The paper proposes a general-purpose neural unit for video question answering that enables hierarchical relational reasoning and multimodal fusion. The Conditional Relation Network (CRN) processes input object arrays through sparse high-order relations while modulating encodings with contextual features, allowing flexible replication and stacking into Hierarchical CRNs (HCRN). The architecture integrates appearance features with clip motion as initial context, then progressively incorporates linguistic context and video-level motion through layered CRNs to enable multi-step reasoning. By hierarchically combining localized clip relations with global video and question contexts, HCRN addresses challenges of modeling distant temporal dependencies and heterogeneous modalities in VideoQA, demonstrating robust performance across diverse question types requiring appearance, motion, and temporal reasoning.

- HME [89]: The paper proposes a novel VideoQA framework that integrates heterogeneous memory and multimodal attention to enhance video-question reasoning. It introduces a heterogeneous memory module to jointly learn global context from appearance and motion features through synchronized attention, while a redesigned question memory captures complex semantics and highlights queried subjects by storing global contexts. These components interact through a multimodal fusion layer that aligns visual and textual hints via self-updated attention, enabling multi-step reasoning. By unifying feature integration with attention learning and maintaining global context throughout, the approach addresses challenges of spatiotemporal alignment and complex question semantics, improving VideoQA performance without separating feature and attention steps.

- LAST-ATT [16]: The paper proposes a method to address data bias in audio-visual question answering (AVQA) by constructing a balanced dataset and introducing an enhanced multimodal model. It identifies skewed answer distributions in the MUSIC-AVQA dataset and rectifies them by collecting complementary videos and questions to ensure uniform answer spread, particularly for binary questions, resulting in the MUSIC-AVQA v2.0 benchmark. The baseline model strengthens audio-visual-text interrelations through a pretrained Audio-Spectrogram-Transformer (AST) branch for audio grounding and cross-modal pixel-wise attention to align audio and visual spatial maps. By integrating these components, the approach mitigates modality neglect and improves reasoning across vision, audio, and language, establishing a robust foundation for unbiased AVQA evaluation.

- LAVIT [57]: The paper proposes a novel benchmark for grounded audio-visual question answering on 360° videos to address spherical spatial reasoning and audio-visual relationships. It introduces the Pano-AVQA dataset with 51.7K QA pairs, featuring bounding-box grounding for two task types: spherical spatial relation QAs to assess relative object positioning on a sphere, and audio-visual relation QAs to link sounds with visual sources. Through quaternion-based spatial embeddings and multimodal training objectives, the framework integrates panoramic audio-visual cues while addressing challenges like spherical distortion and diverse sound localization. This holistic approach enhances semantic understanding of omnidirectional environments without relying on predefined fields of view.

- LAVISH [56]: The paper proposes adapting frozen vision transformers (ViTs) pretrained on visual data to audio-visual tasks without finetuning their original parameters. This is achieved through a latent audio-visual hybrid (LAVISH) adapter, which injects trainable parameters into each ViT layer to enable audio specialization and cross-modal fusion. The LAVISH adapter employs latent tokens to compress modality-specific information, reducing the quadratic cost of standard cross-attention while facilitating bidirectional audio-visual interaction. By integrating these components, the approach addresses the inefficiency of modality-specific models and costly audio pretraining, enabling frozen ViTs to leverage shared representations for enhanced audio-visual understanding without external encoders or extensive parameter updates.

- LSTTA [58]: The paper proposes a parameter-efficient transfer learning approach for audio-visual-language tasks by introducing the Long Short-Term Trimodal Adapter (LSTTA), which integrates pre-trained unimodal/bimodal models without full fine-tuning. LSTTA employs a long-term semantic filtering module to suppress redundant video frames by characterizing temporal importance, while the short-term semantic interaction module models local cross-modal alignments through two sub-modules (AL2V and VL2A) to facilitate fine-grained information transfer. By combining these complementary mechanisms, LSTTA addresses the challenges of uneven global semantics and unannotated local correspondences in trimodal learning, enhancing performance on tasks like Music-AVQA and CMU-MOSEI without requiring large-scale trimodal pretraining.

- MAVEN [60]: The paper proposes a robust multimodal reasoning framework for Audio-Visual Question Answering (AVQA) to address dataset biases and enhance model robustness. It introduces FortisAVQA, a novel dataset constructed by rephrasing test questions to diversify linguistic forms and introducing distribution shifts to evaluate performance across frequent and rare question types. The Multimodal Audio-Visual Epistemic Network (MAVEN) employs a Multifaceted Cycle Collaborative Debias (MCCD) strategy to mitigate bias learning by enlarging distribution differences between unimodal and multimodal logits through KL divergence optimization while using cycle guidance to align unimodal logit distributions. This integrated approach reduces reliance on spurious correlations in individual modalities, improving generalization across in-distribution and out-of-distribution scenarios without requiring balanced training data.

- MCAN [62]: The paper proposes a deep Modular Co-Attention Network (MCAN) to enhance visual question answering (VQA) by jointly modeling intra- and inter-modal interactions through a modular architecture. The framework integrates Self-Attention (SA) units to capture dense word-to-word and region-to-region relationships within questions and images, while Guided-Attention (GA) units model word-to-region cross-modal dependencies. By cascading Modular Co-Attention (MCA) layers composed of SA and GA units, MCAN enables deep reasoning while addressing the limitations of shallow co-attention models. This integrated approach improves fine-grained semantic understanding by simultaneously refining self-attention within modalities and guided-attention across modalities, leading to more accurate visual-textual alignment and robust performance on complex VQA tasks.

- MCCD [18]: The paper proposes a robust framework for Audio-Visual Question Answering (AVQA) to address dataset biases and enhance model robustness. It introduces MUSIC-AVQA-R, a novel dataset crafted by rephrasing test questions and introducing distribution shifts to evaluate performance on both frequent and rare samples, while the Multifaceted Cycle Collaborative Debiasing (MCCD) strategy mitigates bias learning by enlarging distribution differences between uni-modal and multi-modal logits and employing cycle guidance to align uni-modal distributions. This integrated approach ensures diverse question evaluation and reduces bias dependency, improving generalization across in- and out-of-distribution scenarios without relying on balanced training data.
- MEERKAT [65]: The paper proposes an audio-visual LLM for fine-grained spatio-temporal grounding in images and audio, addressing the limitations of existing MLLMs in handling fine-grained tasks. It introduces a modality alignment module based on optimal transport to learn cross-modal patch alignment in a weakly-supervised manner, while a cross-attention module enforces audio-visual consistency to improve joint representation learning. These components are integrated through the AVFIT dataset (3M instruction samples) and MeerkatBench, a unified benchmark for five tasks, enabling the model to tackle challenges like disparate task formats and lack of large-scale training data. The approach enhances performance by unifying spatial and temporal grounding capabilities, achieving state-of-the-art results across diverse audio-visual tasks.
- OPM [67]: The paper proposes an adaptive modulation approach to address imbalanced multimodal learning by dynamically balancing uni-modal optimization during joint training. It introduces On-the-fly Prediction Modulation (OPM) to weaken dominant modality influence in the feed-forward stage by probabilistically dropping its features, while On-the-fly Gradient Modulation (OGM) mitigates gradient dominance in back-propagation through adaptive noise injection. By monitoring inter-modal discriminative discrepancies, these strategies jointly alleviate under-optimization of weaker modalities while preserving dominant modality contributions. The integrated framework enhances multimodal representation learning across diverse tasks by ensuring balanced feature optimization without additional training overhead, as validated through extensive experiments on audio-visual benchmarks.
- ONELLM [68]: The paper proposes a unified framework to align multiple modalities with language using a shared architecture, eliminating the need for modality-specific encoders. It introduces lightweight modality tokenizers to convert input signals into tokens, while a universal encoder (CLIP-ViT) extracts cross-modal features and a universal projection module (UPM) dynamically routes mixed projection experts to map diverse modalities into the LLM's embedding space. Through progressive alignment and a curated multimodal instruction dataset spanning eight modalities, the integrated approach overcomes scalability limitations of prior MLLMs by unifying encoding and projection, enabling flexible modality expansion and enhanced multimodal understanding without architectural redundancy.
- PSAC [69]: The paper proposes a novel self-attention-based architecture for video question answering (VQA) to overcome the limitations of RNNs in modeling long-range dependencies and parallel processing. It introduces Positional Self-Attention (PSA) to capture global dependencies in video and question sequences by attending to all positions while incorporating absolute positional encodings to preserve temporal/spatial information. Through Video-based PSA (VPSA) and Question-based PSA (QPSA), the model encodes video frames and textual questions in parallel. A Video-Question Co-Attention (VQ-Co) block then simultaneously attends to relevant visual and textual features via bidirectional attention, enhancing cross-modal alignment. By integrating PSA with co-attention, the framework efficiently models complex video-question interactions, addressing challenges in sequential data processing and multimodal fusion while improving accuracy and computational efficiency.
- PSTP-NET [70]: The paper proposes a progressive spatio-temporal perception framework for audio-visual question answering (AVQA) to address challenges in complex multi-modal video understanding. The Temporal Segment Selection Module (TSSM) identifies relevant video segments to reduce redundancy, while the Spatial Region Selection Module (SRSM) locates question-aware visual patches within selected segments to enhance spatial reasoning. The Audio-guided Visual Attention Module (AVAM) models audio-visual associations by aligning sound features with visual patches. By progressively integrating these components, the approach effectively filters irrelevant content, localizes key spatio-temporal regions, and strengthens cross-modal interactions, leading to improved scene understanding and question answering performance.

- QAP [71]: The paper proposes a parameter-efficient multimodal language model learning strategy that bridges modalities through query-based prompts and lightweight resampling. The core innovation involves Querying Prompts (QP) to simultaneously extract modality information and interact with text, while Text-Conditioned Resamplers (TCR) adaptively inject text-relevant multimodal features into frozen language model layers. By integrating QP and TCR, the approach efficiently compresses modality inputs and leverages the model's inherent fusion capabilities, addressing computational inefficiency and redundancy in traditional projection-based methods while outperforming existing techniques across multiple multimodal tasks with minimal trainable parameters.
- QWEN2.5-VL [72]: The paper proposes Qwen2.5-VL, a vision-language model advancing multimodal understanding through enhanced visual recognition, object localization, and document parsing while addressing computational and contextual challenges. Key innovations include dynamic resolution processing to handle varying image sizes and video durations, absolute time encoding to improve temporal dynamics perception, and a native dynamic-resolution ViT with Window Attention to reduce overhead while preserving resolution. By integrating these components, the model achieves robust performance in fine-grained visual tasks, long-video comprehension, and real-world agentic applications without task-specific fine-tuning, while maintaining strong linguistic capabilities inherited from Qwen2.5 LLM. The approach overcomes bottlenecks in computational complexity and inconsistent sequence-length performance, enabling precise spatial-temporal reasoning and cross-domain generalization.
- REFATOMNET [73]: The paper proposes a novel approach for Referring Atomic Video Action Recognition (RAVAR) to identify atomic actions of a specific person guided by textual descriptions and video data, addressing limitations in traditional action recognition. Key components include RefAtomNet, which employs a multi-stream architecture connecting video, text, and location-semantic streams to interpret referring expressions and localize target individuals, while cross-stream agent attention and token fusion enhance relevance filtering across modalities. This integrated approach tackles challenges like irrelevant visual distractions and enables end-to-end action recognition for referred individuals, outperforming existing methods in RAVAR without requiring manual pre-processing. The RefAVA dataset with 36,630 annotated instances supports this task.
- VALOR [74]: The paper proposes a Vision-Audio-Language Omni-Perception pretraining model (VALOR) to jointly model tri-modality interactions for understanding and generation tasks. It employs three single-modality encoders to process vision, audio, and language separately, while a multimodal decoder enables conditional text generation through two pretext tasks: Multimodal Grouping Alignment (MGA) projects modalities into a shared space to align vision-language, audio-language, and audiovisual-language groups via contrastive learning, and Multimodal Grouping Captioning (MGC) reconstructs masked text tokens conditioned on visual, auditory, or combined inputs to enhance generative capabilities. By integrating these components with a large-scale human-annotated dataset (VALOR-1M), the approach addresses the limitations of existing bimodal systems, enabling comprehensive cross-modal alignment and flexible text generation across diverse modality combinations for downstream tasks like retrieval, captioning, and question answering.
- VAST [75]: The paper proposes an omni-modality foundation model to enhance video-text cross-modality learning by integrating vision, audio, and subtitle information. It introduces VAST-27M, a large-scale dataset automatically generated through a two-stage pipeline: first training separate vision and audio captioners to produce single-modality descriptions, then employing an LLM to synthesize these with subtitles into omni-modality captions. The VAST model leverages three modality encoders and cross-attention-based text fusion, trained with objectives (OM-VCC/VCM/VCG) to unify multi-modal understanding. This approach addresses the lack of comprehensive video-text corpora by automating caption generation, enabling joint modeling of complementary modalities to improve performance on diverse downstream tasks like retrieval, captioning, and QA without manual annotation costs.
- VITA [79]: The paper proposes VITA, an open-source Multimodal Large Language Model (MLLM) capable of simultaneous processing and interactive analysis across video, image, text, and audio modalities. Starting with Mixtral 8×7B as a language foundation, it expands Chinese vocabulary through bilingual instruction tuning to enhance multilingual proficiency, while endowing visual and audio capabilities via two-stage multi-task learning for multimodal alignment and instruction tuning. To improve interaction, VITA introduces state tokens to distinguish input queries for non-awakening interaction and employs a duplex pipeline deployment scheme, where one

model generates responses while another monitors environmental inputs, enabling audio interrupt interaction. This integrated approach addresses the lack of open-source models with unified multimodal processing and natural interaction, advancing seamless multimodal understanding and human-computer engagement without relying on wake-up words or sequential query handling.

- VIDEOLLAMA-2 [78]: The paper proposes VideoLLaMA 2, a Video Large Language Model designed to enhance spatial-temporal modeling and audio understanding in multimodal video tasks. It introduces a Spatial-Temporal Convolution (STC) connector to capture intricate spatial and temporal dynamics in video data, while integrating an Audio Branch through joint training to incorporate audio cues for richer multimodal understanding. By combining these components, the model addresses challenges in processing temporal dynamics and audio-visual synchronization, improving performance in video question answering and captioning tasks without compromising contextual integrity or processing efficiency.

F Computational Resources and Reproducibility

To support reproducibility, we detail the compute environment used for all experiments we re-implemented in this study. Our local experiments were run on a server equipped with two NVIDIA H100 GPUs. All experimental settings were consistent with those described in the corresponding papers, and we re-implemented key components of related work (properly cited in the main paper) using the original hyperparameters whenever available.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction appropriately represent the paper's scope as a position-driven paper. Our work focus on highlighting the distinct challenges of Music AVQA and align with the content presented throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: This is a position paper that does not propose new methods or conduct original experiments; thus, direct limitations of author-performed work do not apply. However, in Section 6, we compare existing methods and highlight potential limitations based on observed performance trends.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, the paper fully discloses all necessary details to reproduce the main experimental results to the extent that they substantiate the primary claims and conclusions. The descriptions of the experimental settings, model architectures, datasets, evaluation metrics, and hyperparameters are sufficiently detailed, even if actual code and data are not provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide an anonymous GitHub repository of relevant papers and brief introduction that will be continuously updated at https://anonymous.4open.science/r/Survey4MusicAVQA to support the community.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: As a position paper, our work provides the necessary settings and datasets used to contextualize and support its claims. However, full training and testing details are all followed by the selected papers.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our paper includes experiments but not consist of error bars or other statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details can be find in Appendix F. The remaining experiments in this paper are partly derived from previously published works, which are clearly cited.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We respect and follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not release data or models with a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All used assets are properly cited, and the licenses are mentioned.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide an anonymous GitHub repository of relevant papers that will be continuously updated at https://anonymous.4open.science/r/Survey4MusicAVQA.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in our paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.