Listening or Reading? An Empirical Study of Modality Importance Analysis Across AQA Question Types

Zeyu Yin[1], *Yiqiang Cai*[1], *Pingsong Deng*[1], *Xinyang Lyu*[1], *Shengchen Li*[1],

[1] Xi'an Jiaotong-Liverpool University, School of Advanced Technology, Suzhou, China,
{zeyu.yin22, yiqiang.cai21, pingsong.deng23, xinyang.lyu23}@student.xjtlu.edu.cn,
shengchen.li@xjtlu.edu.cn

***Abstract*—Audio Question Answering (AQA) challenges a system to integrate acoustic perception with natural–language reasoning, yet how much each modality actually matters remains unclear. We propose a controlled modality–weight study on the DCASE 2025 Task5 benchmark to quantify this balance. Building on a dual-tower BEATs+BERT architecture, we introduce a scalar fusion hyper-parameter that linearly mixes audio and text embeddings. We evaluate model performance across six distinct question types and use statistical analysis to characterize how accuracy shifts as modality weights change. Our results reveal a clear asymmetry: while text alone supports strong performance on many questions, audio contributes significantly only to tasks that require perceptual grounding. Some tasks benefit most from a balanced fusion of both modalities, whereas for others, increased audio weight can even reduce accuracy. This protocol yields a practical guidance of which question categories depend primarily on audio, on text, or on a certain balanced fusion, providing guidance for future AQA model design.**

***Index Terms*—Audio Question Answering, modality weighting, Acoustic Reasoning, DCASE 2025**

1. INTRODUCTION

Multimodal learning—integrating information across text, vision, audio and other sensory streams—has become a central theme in contemporary AI research. Groundbreaking systems such as CLIP (image–text) [1], Flamingo (vision–language) [2], and Pengi (audio–language) [3] demonstrate how cross-modal pre-training can unlock zero-shot reasoning capabilities that are unattainable with unimodal models. Audio Question Answering (AQA) emerges as a genuinely multimodal task. It requires models to not only listen to the waveform and classify sounds but also reason about them in natural language, offering a rich playground to probe cross-modal alignment, fusion strategies and modality biases [4], [5]. The DCASE 2025 Challenge Task 5 amplifies this by introducing a multi-domain AQA benchmark consisting of three subsets – Bioacoustics QA, Temporal Soundscapes QA, and Complex QA(MMAU) – each designed to evaluate distinct reasoning skills grounded in acoustic perception [4].

Early work in audio QA was limited to narrow domains or synthetic data [6]. Datasets like CLEAR [7] and DAQA [8] programmatically generated QA pairs for musical notes or generic sound events, while Clotho-AQA [6] was the first crowdsourced AQA dataset built on general environmental sound clips. With the advent of large audio-language models (LALMs) [9]–[11], AQA has gained broader traction. Recent foundation models combine powerful audio encoders and text decoders – e.g., Pengi [3] and Qwen2-Audio [12] integrate a pretrained audio front-end with a language model to handle audio-based queries. These models leverage massive training sets (including synthetic data like OpenAQA-5M [13]), and achieve impressive general audio understanding. Pretrained audio transformers such as BEATs (Bidirectional Encoder from Audio Transformers) [14] have set state-of-the-art results on AudioSet [15], providing rich acoustic representations, while text models like BERT remain strong backbones for language understanding [16]. This progress has led to notable performance gains on benchmarks and new multi-task evaluations (e.g. MMAU for audio reasoning). However, most prior works focus on pushing overall accuracy, with less emphasis on how each modality (audio vs. text) contributes to answering different question types.

In multimodal QA it is well-known that models can exploit dataset biases by over-relying on one modality. For instance, a question-only model can answer some Visual QA questions correctly without even "looking" at the image [17]–[20]. A similar concern arises in AQA: certain questions might be answerable through textual cues or common knowledge while others truly require listening to the audio. Which question types rely more on the audio content, and which can be answered mainly via text? And if our model is simply reading or listening. Answering this is crucial for understanding the modality grounding of AQA tasks and guiding the design of a more robust Audio Question Answering model.

In this work, we systematically investigate the modality grounding of Audio Question Answering (AQA) by analyzing how different question types depend on audio, text, or a combination of both. To address this research gap, we design a controlled experimental framework in which a fusion coefficient governs the relative influence of audio and text inputs for each question. We evaluate model accuracy across six distinct AQA question types, ranging from sound counting and temporal detection to knowledge recall and contextual reasoning. Using both accuracy trends and one-way ANOVA statistical tests, we provide a nuanced picture of how modality balance impacts performance. The results not only highlight where current models excel or fall short, but also expose which question types are genuinely grounded in the audio, and those vulnerable to textual bias.

2. METHOD

In this section, we introduce our approach to analyzing the sensitivity of modality in Audio Question Answering (AQA) performance. We present EchoTwin-QA [21], a dual-tower AQA model designed to systematically control and observe the influence of audio and textual modalities. EchoTwin-QA integrates a state-of-the-art BEATs audio encoder as its audio tower and a BERT model as its text tower. To precisely investigate cross-modal interactions, we incorporate a scalar fusion parameter λ that linearly mixes the audio and textual feature embeddings prior to answer classification, allowing us to sweep across a spectrum of modality balances. Performance is then evaluated across eleven λ values for each question type, and one-way ANOVA tests are applied to determine whether changes in the balance of modality lead to statistically significant differences in accuracy.

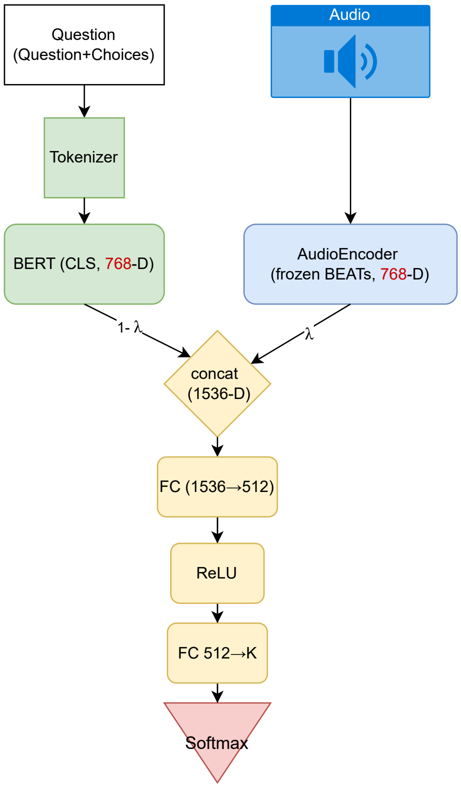

Fig. 1: Dual-tower architecture: a BEATs audio encoder and a BERT text encoder produce modality-specific embeddings, concatenated and fed to a lightweight MLP classifier with a scalar fusion parameter.

2.1. Model Architecture

Our system *EchoTwin–QA* follows a dual–tower paradigm as shown in Fig.1 in which an **audio tower** (BEATs) and a **text tower** (BERT) are fused by a single scalar, $\lambda \in [0, 1]$, that controls the relative weight of each modality.
A waveform $\mathbf{x} \in \mathbb{R}^N$ is fed to BEATs, yielding a sequence of hidden states $\mathbf{A} = \big[\mathbf{a}_1, \dots, \mathbf{a}_{T_a}\big] \in \mathbb{R}^{T_a \times d_a}$. We mean–pool these states to obtain a fixed-length embedding

$$\mathbf{h}_a = \frac{1}{T_a} \sum_{t=1}^{T_a} \mathbf{a}_t \in \mathbb{R}^{d_a}. \tag{1}$$

The question concatenated with its answer choices is tokenised and processed by BERT, producing hidden states $\mathbf{H} = \big[\mathbf{h}_1, \dots, \mathbf{h}_T\big]$. The standard `[CLS]` vector serves as the text embedding

$$\mathbf{h}_t = \mathbf{h}_{\mathrm{CLS}} \in \mathbb{R}^{d_t}. \tag{2}$$

A single scalar λ modulates the contribution of each tower:

$$\tilde{\mathbf{h}}_a = \lambda\, \mathbf{h}_a, \tag{3}$$

$$\tilde{\mathbf{h}}_t = (1 - \lambda)\, \mathbf{h}_t. \tag{4}$$

The weighted embeddings are concatenated

$$\mathbf{z} = \big[\tilde{\mathbf{h}}_a \parallel \tilde{\mathbf{h}}_t\big] \in \mathbb{R}^{d_a + d_t}, \tag{5}$$

and passed to a lightweight two–layer multilayer perceptron

$$\mathbf{u} = \mathrm{ReLU}\big(W_1 \mathbf{z} + b_1\big), \tag{6}$$

$$\hat{\mathbf{y}} = \mathrm{softmax}\big(W_2 \mathbf{u} + b_2\big) \in \mathbb{R}^K, \tag{7}$$

where K is the number of answer choices.
We train with label–smoothed cross–entropy:

$$\mathcal{L} = (1 - \varepsilon)\, \mathcal{L}_{\mathrm{CE}}\big(\hat{\mathbf{y}}, \mathbf{y}\big) + \varepsilon\, \frac{1}{K}. \tag{8}$$

Only the parameters of the two-layer MLP (and optionally the top L layers of BEATs) are trainable; both encoders are otherwise frozen.

2.2. Statistical Analysis

For each fusion coefficient λ we record, for every QA pair i, a binary correctness flag

$$c_i^{(\lambda)} = \begin{cases} 1, & \text{if the predicted answer matches ground truth,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Flags are stratified according to the organiser-supplied question types ($|\mathcal{Q}| = 11$). For a fixed type $q \in \mathcal{Q}$ we obtain $G = 11$ groups

$$\mathcal{G}_q(\lambda) = \big\{\, c_i^{(\lambda)} \,\big|\, q_i = q \big\}, \qquad \lambda \in \{0.0, 0.1, \dots, 1.0\},$$

each containing N_λ Bernoulli observations. We test the null hypothesis $H_0 \ : \ \mathbb{E}\big[\mathcal{G}_q(\lambda)\big]$ are equal for all λ by a fixed-effect one-way ANOVA [22]. Let $\bar{c}_\lambda$ denote the group mean for a given λ and $\bar{c}$ the global mean for type q. The F-statistic is

$$F_q = \frac{\frac{1}{G-1} \sum_\lambda N_\lambda\, (\bar{c}_\lambda - \bar{c})^2}{\frac{1}{N_q - G} \sum_\lambda \sum_{c \in \mathcal{G}_q(\lambda)} (c - \bar{c}_\lambda)^2}, \tag{10}$$

where $N_q = \sum_\lambda N_\lambda$.

We report the pair $\big(F_q, p_q\big)$ for every question type and declare the effect of λ *significant* when $p_q < 0.05$.

We measured the top 1 accuracy $\mathrm{Acc}(\lambda, s)$ of EchoTwin–QA on the official development set while linearly mixing the two modality embeddings via weighted concatenation. Fig. 2 plots the mean accuracy

$$\mathrm{A\bar{c}c}(\lambda) = \tfrac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathrm{Acc}(\lambda, s) \tag{11}$$

F-statistic measures how much the means vary between λ groups relative to the variance within groups; a larger F indicates stronger dependence on λ.
p-value is the probability of observing such an F under the null hypothesis that accuracy is constant across λ. A small p (< 0.05) means λ has a significant effect.

3. EXPERIMENTS

3.1. Dataset

All experiments are conducted exclusively on the official **DCASE 2025 Task 5** AQA corpus [4]. The training portion comprises 0.7 k *Bioacoustics QA*, 1.0 k *Temporal Soundscapes QA* and 6.4 k *Complex QA (MMAU)* items, while the development set provides 0.2 k, 0.6 k and 1.6 k items, respectively. No external audio or text resources are used. The official development split contains 11 annotated question types, but we discarded some type with too few instances. To obtain reliable modality–sensitivity curves we restrict our study to the six most populous and conceptually distinct categories listed in Section 3.2: *Sound Counting, Remember, Both, Sound Detection, Apply Frequency,* and *Understand Acoustics*. Their sample sizes on the development set accounted for over 97 % of the available validation data.

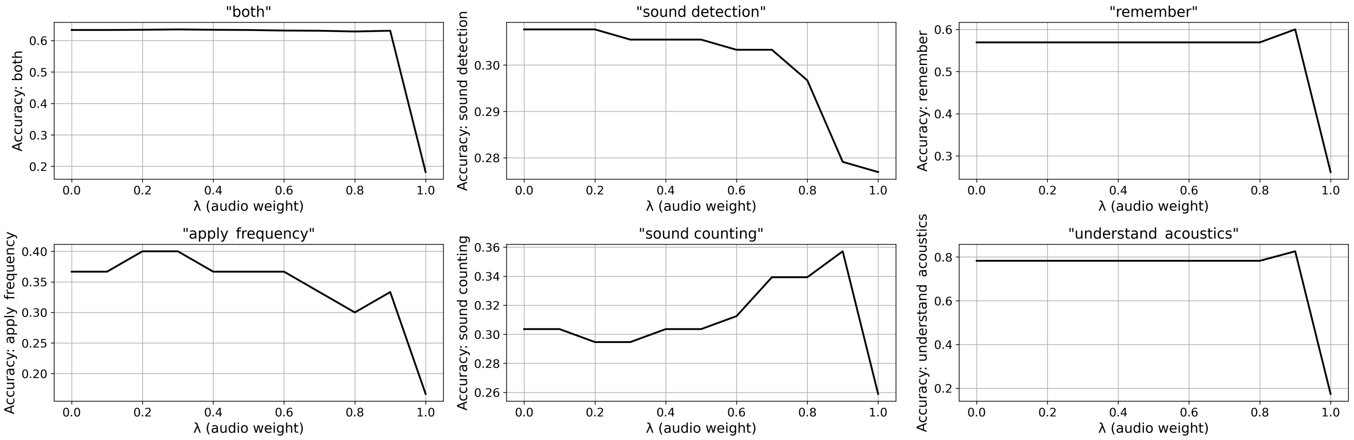

Fig. 2: Accuracy variation across question types as a function of the audio feature weight λ. Each subplot corresponds to a specific question category: *Understand Acoustics*, *Remember*, *Sound Counting*, *Both*, *Apply Frequency*, and *Sound Detection*. The plots demonstrate how model accuracy varies as λ increases from 0 (purely text-based reasoning) to 1 (purely audio-based reasoning).

3.2. Question type

Sound Counting Asks how many distinct sound events or types are present or how many times a specific sound (e.g., crash, phone, scream) occurs in an audio clip (e.g. “How many different sounds are present in the audio clip?”, “How many times does the crash sound/ female scream occur?”)
Remember (Knowledge Recall) A factual statement about the animal species that produced the recording—e.g., a unique trait, behavior, historical fact, or conservation concern—chosen from several candidate statements. (e.g., “Which of the following is a defining characteristic of the species that produced the sound?”, “What unique behavior distinguishes the species?”)
Both (Contextual Reasoning) They ask about contextual or situational information inferred from the audio, including time, place, emotional tone, environmental setting, and background interactions. (e.g., “What time of day is the audio likely set?”, “What might indicate that the speaker is not alone?”)
Sound Detection (Temporal Reasoning) A temporal detail about specific sound events: start time, end time, duration, order, or which sound occurs immediately before/after another. (e.g., “What is the start time of the drawer sound?”, “Which sound occurs immediately after the cough sound?”)
Apply Frequency (Spectral Reasoning) Asks for a comparison of pitch or frequency range between multiple sound events, often in a specified order relative to silences. (e.g.,“Which sound is most dominant at higher frequencies?”)
Understand Acoustics (Feature-Based Reasoning) The choice of the text description that best matches the overall acoustic characteristics or main feature/pattern of the recording. (e.g. “Which of the following best describes the main feature of the recording?”, “What acoustic characteristics describe the signal?”)

3.3. Training Configuration

We fine-tune the model once and then freeze all weights thereafter for the λ-sweep evaluation, with Optimiser: AdamW ($\eta_{\text{text}} = 1 \times 10^{-5}$, $\eta_{\text{audio}} = 1 \times 10^{-6}$, weight-decay = 0), a batch size of 32; gradient clip: $\|g\|_2 \leq 1.0$; AMP enabled. For Loss we use label-smoothed cross-entropy ($\varepsilon = 0.05$). After fine-tuning, we *only* vary the scalar fusion coefficient $\lambda \in \{0.0, 0.1, \ldots, 1.0\}$ while keeping all network weights fixed.

Table 1: Accuracy (%) under audio only ($\lambda = 1.0$), text only ($\lambda = 0.0$) and the **best-performing** audio + text mixed for each question type.

Question type	Audio only	Question only	Audio+Question
Apply Frequency	16.7	36.7	**40.0**
Both	18.2	63.4	**63.5**
Remember	26.2	56.9	**60.0**
Sound Counting	25.9	30.4	**35.7**
Sound Detection	27.7	**30.8**	**30.8**
Understand Acoustics	17.4	78.3	**82.6**

Table 2: One-way ANOVA results for the six question types.

Question type	F	p
Apply Frequency	0.56	8.49×10^{-1}
Both	136.80	$\mathbf{2.27 \times 10^{-277}}$
Remember	2.36	$\mathbf{9.42 \times 10^{-3}}$
Sound Counting	0.37	9.58×10^{-1}
Sound Detection	0.28	9.87×10^{-1}
Understand Acoustics	4.57	$\mathbf{6.02 \times 10^{-6}}$

4. RESULTS AND ANALYSIS

4.1. Modality Ablation in AQA

As we can observe in Table 1, the model performs poorly when relying solely on audio input. In contrast, when provided only with the textual question, the model achieves relatively high accuracy on several question types, especially for *Sound Detection* and *Both*, sometimes approaching or even surpassing performance with modality fusion. This asymmetry in ablation performance highlights a key limitation: while the task is designed to encourage multimodal reasoning, many instances can be answered using text alone, the soundscape is often ambiguous without the semantic guidance.

4.2. Impact of Modality Across Question Types

Sound Counting questions, which involve identifying discrete acoustic events or enumerating occurrences of specific sounds, show a clear dependence on audio information. Their accuracy improved monotonically as audio weight increased—from under 30% to 35.7%. The upward slope is visually apparent in Fig. 2. This pattern intuitively aligns with the fundamental nature of these questions, which inherently require detailed acoustic perception rather than textual reasoning, yet the between λ variance is still dwarfed by the within-λ noise. To perform better, a model must leverage explicit acoustic cues rather than text-based or common-knowledge shortcuts.

Understand Acoustics and *Remember* types show a similar dependence, but with a distinct delayed response to increased audio weighting. This indicates their substantial reliance on fine-grained acoustic detail. The between-level variance is almost five times the within-level variance, so the jump could be explained, suggesting a threshold phenomenon: the option texts already contain descriptors, which loosely match many recordings. Only when audio dominates can the model compute fine-grained spectral summaries, and factual recall in bioacoustics similarly requires identifying species-specific acoustic signatures. Thus, despite the possible availability of textual cues, these question types ultimately require a robust grounding in acoustic features.

Apply Frequency questions show a different and illuminating trend: accuracy peaks significantly at a low audio weight and decreases as λ increases. The rapid drop-off implies that frequency-based comparative judgments can be partially resolved via semantic or pragmatic cues embedded within the textual question and answer choices. It suggests that these tasks do not demand detailed acoustic resolution as much as contextual reasoning about the frequency order and association with different sounds. A high weighting on acoustic details might obscure semantic subtleties vital for these spectral reasoning tasks.

Both achieve maximum accuracy at a moderately light audio weight (λ = 0.3). Then follows a slight decline, strongly indicates that contextual reasoning tasks benefit most from balanced multimodal integration, slightly favoring textual modalities. The ANOVA test corroborates this behavior, indicating that over 90% of the performance variance is attributable to the choice of modality weight. These questions inherently depend on interpreting contextual scenarios—such as time, emotional tone, or background interactions—often inferred through textual knowledge or semantic framing rather than precise acoustic features alone. Hence, multimodal systems should prioritize a balanced and nuanced fusion strategy for optimal performance on these questions.

Sound Detection category demonstrates an intriguing and counter-intuitive trend, achieving its highest accuracy without audio feature. Accuracy notably decreases as audio contribution increases. This result reveals that temporal reasoning, as framed by the dataset, might heavily rely on textual patterns or conventional temporal relationships implicitly embedded in the questions themselves. The deterioration with increased audio input indicates a possible misalignment between the provided acoustic evidence and the inferred temporal logic from the textual query. we argue that current models lack the appropriate inductive biases or representational mechanisms to extract and align temporal acoustic patterns in a way that meaningfully supports semantic inference.

4.3. Discussions

From section 4.1 we found a fundamental limitation: the audio branch in current AQA models remains underutilized and often poorly integrated. Instead of serving as a robust perceptual complement to language, audio provides at most a subtle and inconsistent benefit.

In Section 4.2 we find that questions that require more audio are those whose answers depend on direct, perceptual properties of the sound. For instance, *Sound Counting* tasks demand temporal parsing and event segmentation that are simply not recoverable from the question text or choices. The model must "listen" for actual events.

Conversely, question types that are answerable through text alone typically ask for information that is either (1) explicitly provided in the question and choices, or (2) easily inferred through world knowledge or dataset regularities. They often leak clues via textual templates—like unique timestamps or the phrasing of alternatives ("highest frequency", "starts first")—which the model can learn to exploit without referencing the audio at all. Our ANOVA tests confirm that, for these categories, changing the audio–text balance has no systematic effect: performance is governed by the text modality, and extra audio either adds noise or actively impedes accuracy.

The most intriguing are those tasks that require a nuanced combination of both modalities. Contextual Reasoning (Both) and Understand Acoustics show peak performance at intermediate λ values, with ANOVA statistics confirming overwhelming sensitivity to the modality balance. These tasks demand both situational inference (text: e.g., "Is it day or night?" / "Is someone alone?") and perceptual validation (audio: background noises, emotional tone, spectral features). Only when models can integrate and align semantic cues with acoustic evidence do they achieve optimal grounding and robust performance. The most valuable are those requiring authentic cross-modal grounding, pushing models to **"listen, read and understand"** rather than to do text-only reading-comprehension guesswork of likely answers.

4.4. Limitation and Future Work

Our analysis is currently limited to the EchoTwin-QA model and official DCASE 2025 Task 5 Dataset, and results may not generalize to other AQA benchmarks with different domain coverage or question design. The total number of questions for each type is unbalanced, which may reduce statistical power and obscure subtle modality effects.

A comprehensive study should be conducted to better understand which question types most effectively test the range of abilities required for robust audio-language grounding. This includes investigating the cognitive demands, linguistic features, and perceptual complexity associated with each type. Moreover, a rigorous taxonomy of question types should be established. Categories might be based on the modality reliance, the reasoning skill required , or the nature of the answer. Such a framework would not only improve the interpretability and fairness of model evaluation, but also guide dataset design to ensure coverage of all critical skills.

The mere existence of strong, type-specific modality biases suggests that a single static fusion weight is sub-optimal. The evidence argues for dynamic fusion—either a learned gating scalar conditioned on the question embedding or a transformer that modulates cross-attention weights per token. An alternative is a Mixture-of-Experts architecture where separate audio-heavy and text-heavy experts compete, and a router network selects the appropriate one per question.

5. CONCLUSION

This work set out to uncover how different AQA question types depend on audio versus question, motivated by concerns over modality bias known in multimodal QA. Through experiments varying the balance between audio and text, we found that despite the multimodal nature of AQA, current models rely more on textual information. Audio input, while essential for certain perceptual tasks such, provides only limited and often inconsistent improvements—and can even impede accuracy for text-dominated question types. The effectiveness of audio remains constrained by poor integration and the dominance of language priors. Our findings emphasize the value of adaptive modality fusion and more precise question-type taxonomies to unlock the full potential of AQA systems. To advance AQA, future work should focus on developing adaptive fusion strategies and establishing robust, modality-aware question taxonomies, ensuring models are grounded in both sound and language.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millicah, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[3] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: an audio language model for audio tasks," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[4] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar, O. Nieto, R. Duraiswami, D. Manocha, G. Kim, J. Du, R. Valle, and B. Catanzaro, "Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge," 2025. [Online]. Available: https://arxiv.org/abs/2505.07365

[5] A. K. Sridhar, Y. Guo, and E. Visser, "Enhancing temporal understanding in audio question answering for large audio language models," 2024. [Online]. Available: https://arxiv.org/abs/2409.06223

[6] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," 2022. [Online]. Available: https://arxiv.org/abs/2204.09634

[7] J. Abdelnour, G. Salvi, and J. Rouat, "Clear: A dataset for compositional language and elementary acoustic reasoning," 2019. [Online]. Available: https://dx.doi.org/10.21227/7x26-a025

[8] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, p. 2283–2294, Aug. 2020. [Online]. Available: https://doi.org/10.1109/TASLP.2020.3010650

[9] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," 2024. [Online]. Available: https://arxiv.org/abs/2406.11768

[10] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, "Audiopalm: A large language model that can speak and listen," 2023. [Online]. Available: https://arxiv.org/abs/2306.12925

[11] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," 2023. [Online]. Available: https://arxiv.org/abs/2311.07919

[12] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," 2024. [Online]. Available: https://arxiv.org/abs/2407.10759

[13] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," 2024. [Online]. Available: https://arxiv.org/abs/2305.10790

[14] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[17] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring nearest neighbor approaches for image captioning," 2015. [Online]. Available: https://arxiv.org/abs/1505.04467

[18] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, *RUBi: reducing unimodal biases for visual question answering*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[19] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5014–5022.

[20] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," 2015. [Online]. Available: https://arxiv.org/abs/1512.02167

[21] Z. Yin, Z. Zhou, Y. Cai, S. Li, and X. Shao, "Echotwin-qa: A dual-tower beatsbert system for dcase 2025 task 5 audio question answering," DCASE2025 Challenge, Tech. Rep., June 2025.

[22] A. Edwards, "R.a. fischer, statistical methods for research workers, first edition (1925)," *Landmark Writings in Western Mathematics 1640-1940*, pp. 856–870, 12 2005.