ECHOTWIN-QA: A DUAL-TOWER BEATSBERT SYSTEM FOR DCASE 2025 TASK 5 AUDIO QUESTION ANSWERING

Technical Report

Zeyu Yin[1], *Ziyang Zhou*[1], *Yiqiang Cai*[1], *Shengchen Li*[1], *Xi Shao*[2]

[1] Xi'an Jiaotong-Liverpool University, School of Advanced Technology, Suzhou, China,
{zeyu.yin22, ziyang.zhou22, yiqiang.cai21}@student.xjtlu.edu.cn, shengchen.li@xjtlu.edu.cn
[2] Nanjing University of Posts and Telecommunications,
College of Telecommunications and Information Engineering, Nanjing, China,
shaoxi@njupt.edu.cn

ABSTRACT

Task 5 of the DCASE 2025 Challenge frames *Audio Question Answering* (AQA) as a multi-choice test of acoustic reasoning across marine bio-acoustics, temporal soundscapes and everyday recordings. We present a light-weight dual-tower system that couples a BEATs-Base audio encoder with a BERT-Base text encoder; a two-layer MLP, amounting to $\sim$132M trainable parameters, maps the concatenated embeddings to answer logits. On the official development set our best submission achieves 54.46 % accuracy, surpassing the strongest baseline (Gemini-2.0-Flash, 52.5%)[1].

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

1. INTRODUCTION

Audio Question Answering (AQA) has recently emerged as a unified benchmark for agents that must both perceive and reason over real-world soundscapes [1]. The DCASE 2025 Task 5 Audio Question Answering (AQA) formalises this by requiring systems to answer open–domain questions on marine-mammal calls, temporal soundscapes, and complex real-world recordings. [2].

AQA is conceptually close to Automated Audio Captioning (AAC), which translates an audio clip into free-text descriptions. The 2024 DCASE AAC task demonstrated that pairing strong audio encoders with large language models (LLMs) can achieve state-of-the-art FENSE scores [3]. These trends are highly relevant to AQA because they show how audio perception and language reasoning can be combined.

End-to-end audio-LLMs excel at zero-shot generalisation but require billions of trainable parameters. In contrast, contrastive models such as CLAP align audio and text in a shared space with far fewer weights [4]. We therefore adopt a dual-tower design in our submission: a BEATs audio encoder [5] and a BERT-base text encoder [6]. Our dual-tower architecture resembles the two-branch baseline introduced for the Clotho-AQA dataset, where independent audio and question encoders are concatenated before downstream classification [7].

[1] https://github.com/HuffmanJoey/dcase25_t5_EchoTwin-QA

2. METHOD

2.1. Overview

Our system follows a dual-tower paradigm: a frozen audio encoder and a frozen text encoder transform their respective modalities into fixed-dimensional embeddings, which are concatenated and passed to a lightweight multilayer perceptron (MLP) classifier. All learnable parameters therefore reside only in the MLP and (optionally) the N highest audio layers we unfreeze for domain adaptation. Figure 1 in the final paper will depict this pipeline.

2.2. Data Preparation

Corpus structure Each JSON file supplies a question, a list of choices, an answer letter, the audio_url, and a question_type tag.
Waveform loading Clips are resampled to 16 kHz, converted to mono, peak-normalised, clipped to 10 s (160 000 samples) and zero-padded if shorter. Files with $<$ 25 ms of content are discarded.

$$x_{\text{pad}}[n] \;=\; \begin{cases} \tilde{x}[n], & 0 \le n < L, \\ \varepsilon, & L \le n < N, \end{cases} \qquad L = \text{len}(\tilde{x}),\; N = 160\,000.$$

Data Augmentation

- **Time shift** random ±1 s circular roll with zero padding.

$$x_{\text{shift}}[n] \;=\; x\big[(n-s) \bmod N\big].$$

- **SpecAug** 1 × frequency mask (F=20 bins) + 1 × time mask (T=80 frames) on a magnitude spectrogram (400-pt FFT, 25 ms hop) [8]. A 16-step Griffin–Lim vocoder reconstructs the waveform [9].

$$S_{\text{mask}}(k,t) \;=\; \begin{cases} 0, & f_0 \le k < f_0 + F, \\ S(k,t), & \text{otherwise.} \end{cases}$$

$$S_{\text{mask}}(k,t) \;=\; \begin{cases} 0, & t_0 \le t < t_0 + T, \\ S(k,t), & \text{otherwise.} \end{cases}$$

- **Random gain** uniform [-6 dB, +6 dB] (p = 0.5).

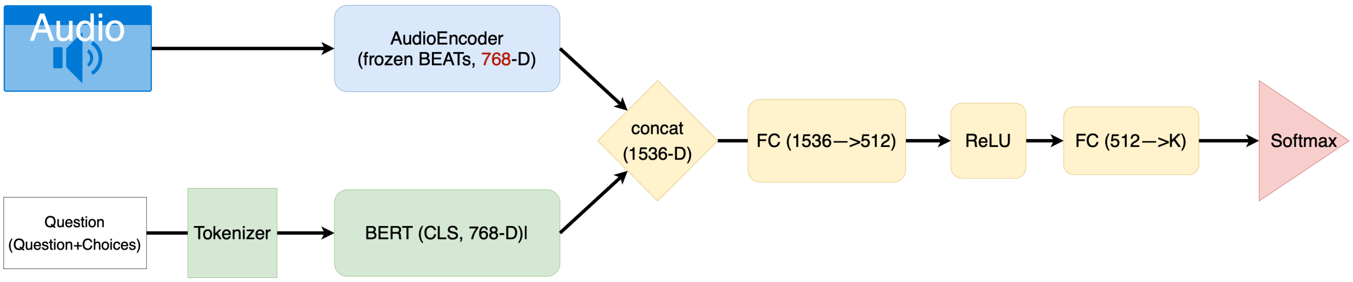

Figure 1: Dual-tower architecture: a frozen BEATs audio encoder and a frozen BERT text encoder produce modality-specific embeddings, which are concatenated and fed to a lightweight MLP classifier. Only the MLP (and optionally the top L BEATs layers) are fine-tuned.

- **Additive noise** Gaussian noise mixed at a random 10–30 dB SNR (p = 0.3).

$$y[n] = x[n] + \sigma\,\eta[n], \tag{1}$$

$$\sigma = \sqrt{\frac{P_x}{10^{\mathrm{SNR}/10}}}, \tag{2}$$

$$P_x = \frac{1}{N}\sum_{n=0}^{N-1} x[n]^2, \qquad \eta[n] \sim \mathcal{N}(0,1). \tag{3}$$

Tokenizer The question string is concatenated with all answer choices, lower-cased, and tokenised by **BERT-Base-Uncased**; sequences are padded or truncated to 128 sub-words. (For padding we reuse the `[PAD]` token so that attention masks remain binary.)

2.3. Model Architecture

Our system is a *dual-tower* network (Fig. 1). Given an audio clip $x \in R^N$ and a token sequence $\mathbf{q} = (q_1, \ldots, q_T)$, the forward path is

$$\mathbf{a} = f_{\mathrm{BEATs}}(x) \xrightarrow{\text{mean-pool}} \bar{\mathbf{a}} \in R^{d_a}, \tag{4}$$

$$\mathbf{t} = f_{\mathrm{BERT}}(\mathbf{q}) = \mathbf{h}_{[\mathrm{CLS}]} \in R^{d_t}, \tag{5}$$

$$\mathbf{z} = \left[\bar{\mathbf{a}} \parallel \mathbf{t}\right] \in R^{d_a+d_t}, \tag{6}$$

$$\hat{\mathbf{y}} = \mathrm{MLP}(\mathbf{z}) \xrightarrow{\text{softmax}} \mathbf{p} \in [0,1]^C, \tag{7}$$

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

- **Fusion & classification.** We concatenate the two embeddings and feed them to a two-layer multilayer perceptron

$$\mathbf{z} = \mathrm{ReLU}\big(\mathbf{W}_1[\mathbf{a};\mathbf{t}]+\mathbf{b}_1\big) \in R^{512}, \qquad \hat{\mathbf{y}} = \mathbf{W}_2\mathbf{z}+\mathbf{b}_2 \in R^K,$$

where K is the number of answer choices. The model is trained with label-smoothed cross-entropy.

2.4. Training Strategy

We train the model using a purely supervised objective without any contrastive loss. The training configuration is as follows:

- **Loss function:** We use *label-smoothed cross-entropy* with $\varepsilon = 0.05$ over the multiple-choice logits:

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{smooth}} = (1-\varepsilon)\cdot\mathcal{L}_{\mathrm{CE}} + \varepsilon\cdot\mathcal{L}_{\mathrm{uniform}}.$$

- **Optimiser:** *AdamW* with two learning rates:
  - 1×10^{-5} for all trainable text and fusion MLP parameters,
  - 1×10^{-6} for optionally unfrozen BEATs layers.

  Weight decay is applied in our training and we experimented with $\lambda_{\mathrm{wd}} \in \{0.01,\ 0.001\}$ and a no-decay run ($\lambda_{\mathrm{wd}} = 0$).
- **Scheduler:** *Cosine annealing* schedule with warmup:

$$\eta_t = \eta_{\min} + \tfrac{1}{2}(\eta_0 - \eta_{\min})\left(1+\cos\left(\pi\cdot\frac{t}{T_{\max}}\right)\right).$$

- **Gradient handling:** Gradients are clipped to ℓ_2 norm $\le$ 1.0. *Automatic mixed precision (AMP)* is enabled via `torch.cuda.amp`.
- **Data augmentation:** During training we apply a *class-conditional MixUp* at the waveform level. At every optimisation step we sample $\lambda \sim \mathrm{Beta}(\alpha,\alpha)$ with $\alpha = 0.4$ and blend two clips *only if they share the same ground-truth label*:

$$\tilde{x} \;=\; \lambda\,x_i \;+\; (1-\lambda)\,x_j, \qquad y_i = y_j.$$

- **Early stopping:** Validation accuracy is tracked at the end of each epoch. The model with the highest dev-set accuracy is saved as the final checkpoint.

2.5. Inference and Evaluation

At test time both encoders remain frozen; the system performs a single forward pass and chooses the answer with the highest softmax probability. We report:

Overall accuracy on the development sets.

Per-question-type accuracy, enabling fine-grained error analysis (audio_tagging, vocalization, counting . . .).

During inference on the evaluation set, the model predicts only the choice letter (e.g., A, B, C). A post-processing script then maps this letter to its corresponding answer text and writes the full string to the final CSV file submitted to DCASE.

Table 1: Dev-set accuracy (%) and key hyper-parameters of our submissions.

ID	Unfreeze layer	Data Aug.	WD	Params (M)	Dev (%)
Sub 1	3	No	0	131.5	54.46
Sub 2	3	No	0	131.5	54.26
Sub 3	4	Yes	0.001	138.6	54.01
Sub 4	—	Ens.	—	— (reuse)	**55.07**

Table 2: Dev-set accuracy (%) by question type. Best value is **bold**.

Type	Sub1	Sub2	Sub3	Sub4
overall	54.46	54.26	54.01	**55.07**
both	63.36	63.54	**63.78**	**63.78**
sound counting	29.46	30.36	**33.93**	**33.93**
sound detection	**30.55**	29.89	27.91	30.42
audio tagging	50.00	52.50	**55.00**	**55.00**
remember	**56.92**	55.38	40.00	**56.92**
vocalization	**48.65**	45.95	45.95	**48.65**
apply_frequency	**36.67**	30.00	33.33	**36.67**
apply_duration	**38.10**	**38.10**	28.57	**38.10**
understand_acoustics	78.26	69.57	**82.61**	**82.61**
species	**25.00**	18.75	**25.00**	**25.00**
audio detection	0.00	0.00	**50.00**	**50.00**

3. SUBMISSIONS AND RESULTS

3.1. System Variants

We submitted three single-model runs and one router–ensemble (Table 1):

- **Sub 1** — best checkpoint of a run with *three* unfrozen BEATs layers and *no* data augmentation.
- **Sub 2** — final checkpoint of the same configuration as Sub 1, illustrating the effect of continued training.
- **Sub 3** — best checkpoint with *four* unfrozen BEATs layers, full waveform augmentation, and weight decay 10^{-3}.
- **Sub 4** — a lightweight router that sends each `question_type` to the model (Sub 1 or Sub 3) that performs best for that type; no additional parameters are trained.

3.2. Comparison with Baseline Systems

Table 3 contrasts our best run with the official baseline systems released for DCASE 2025 Task 5. The ensemble exceeds the strongest baseline (Gemini-2.0-Flash) by **2.57 pp** on the development set while being two orders of magnitude smaller.

Table 3: Dev-set accuracy of baseline models versus our ensemble.

System	Dev Acc. (%)
Qwen2-Audio-7B	45.0
AudioFlamingo 2	45.7
Gemini-2.0-Flash	52.5
Ours (Sub 4)	**55.07**

3.3. Discussion

Layer unfreezing. Moving from three to four trainable BEATs layers (Sub 1 $\rightarrow$ Sub 3) adds $\approx$ 7 M parameters that adapt low-level acoustic filters. The gain is clearest on categories that depend on *fine temporal resolution* or *event boundaries*: *sound counting* (+4.5pp) and *audio tagging* (+5pp) in Table 2. In contrast, tasks that rely on *stable semantic alignment between audio and text—remember*, *vocalization*— lose 8–17pp, suggesting that too much feature drift in the audio tower can mis-align with the frozen BERT embeddings.

We attribute the pattern in Table 2 to three interacting factors:

Spectro-temporal focus of higher BEATs blocks. Layers 10–12 of BEATs attend to short onsets and energy envelopes that mark individual events. Unfreezing them lets the model re-shape these detectors toward the mixed marine–urban domain of DCASE, which boosts *sound counting*, *audio tagging* and the rare *audio detection* queries that hinge on clear event boundaries. However, the same re-tuning distorts mid-level abstractions (phonetic identity, harmonicity) that BERT relies on for semantic grounding, reducing accuracy on narrative or memory-style tasks.

Invariance introduced by SpecAug + MixUp. Frequency masking and MixUp drive the network to ignore localised spectral content and focus on *presence* rather than *exact position*. This is ideal for binary decisions such as "is there a whistle in the clip?" but detrimental when the start/end positions or fine durational structure matter (*apply_duration*, *remember*). Noise and gain jitter further damp amplitude cues that BERT's CLS embedding might use to align words like "first" or "louder" with specific acoustic segments.

Sample-size imbalance across types. Bio-acoustic and counting questions are the sparsest categories in the training set ($<$ 5% of clips). Augmentation effectively enlarges these sub-corpora, allowing the four-layer model to generalise better on them even at the cost of minor degradation elsewhere. For plentiful types such as *both* and *remember*, the benefit saturates, so any drift in the representation becomes a net loss.

These factors explain why Sub 1 (fewer trainable weights, no augmentation) preserves global semantic alignment, whereas Sub 3 (four trainable layers, heavy augmentation) excels on event-centric, data-sparse tasks. The router-ensemble in Sub 4 simply harvests whichever inductive bias is more appropriate for each question type, yielding the best overall score without additional parameters.

Waveform augmentation. The SpecAug+MixUp regime in Sub 3 injects frequency masks, time masks, gain jitter and noise. These perturbations mimic the variability of real-world sound events, helping categories that hinge on short bursts of energy (*audio detection*, which climbs from 0% to 50%). Conversely, the same distortions can blur long-context cues needed for duration or narrative questions, explaining the drop on *apply_duration* (–10pp) and the partial loss on *remember*.

Complementarity and ensemble. Because Sub 1 excels at semantics-heavy and holistic queries while Sub 3 excels at event-centric ones, routing each `question_type` to its stronger expert (Sub 4) combines the best of both worlds. The ensemble preserves Sub 3's improvements on *sound counting*, *audio tagging*, and *understand_acoustics* while restoring Sub 1's advantages on *remember*, *vocalization* and *apply_duration*. The net result is a further **+0.61pp** in overall accuracy,

4. SUMMARY AND FUTURE WORK

We introduced **EchoTwin-QA**, a lightweight dual-tower system that couples frozen BEATs and BERT encoders with a shallow classification head for DCASE 2025 Task 5 Audio Question Answering. A three-layer variant without augmentation already surpasses the strongest baseline by 1.96 pp on the development set, and a simple router–ensemble that combines this model with a four-layer, heavily augmented counterpart raises overall accuracy to **55.07 %**.

Our next steps centre on addressing the clear domain gaps revealed by the per-type evaluation.

We will pursue an **ensemble strategy** in which small specialist encoders—one trained on bio-acoustic corpora and another on short-event sound detection—are combined with the current BEATs + BERT tower. A lightweight "router" module will analyse each (question, audio) pair and dispatch it to the most suitable expert before logit fusion, allowing us to exploit strengths that a single generic model cannot capture. We believe these directions will push our audio QA system closer to human-level acoustic reasoning in forthcoming challenges.

5. REFERENCES

[1] "Audio question answering – dcase challenge 2025 task 5," https://dcase.community/challenge2025/task-audio-question-answering, accessed 5 May 2025.

[2] C.-H. H. Yang, S. Ghosh, and Q. W. et al., "Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge," 2025.

[3] "Automated audio captioning – dcase challenge 2024 task 6," https://dcase.community/challenge2024/task-automated-audio-captioning-results, accessed 14 Apr 2025.

[4] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[5] 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.

[6] 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

[7] 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

[8] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," 09 2019, pp. 2613–2617.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.