Q-XAI: Wirtinger-Based Interpretability for Complex-Valued Audio Transformers

Minh K. Quan*, Pubudu N. Pathirana†

School of Engineering, Deakin University, Waurn Ponds, VIC 3216, Australia

Abstract

Complex-Valued Transformers (CVTs) improve Acoustic Scene Classification (ASC) by explicitly modeling audio phase relationships. However, their deployment in reliable applications is hindered by an interpretability gap: non-holomorphic activation functions invalidate standard gradient-based attribution methods, and structure-agnostic uncertainty estimators ignore the phase information encoded in complex representations. This paper presents Q-XAI, a unified interpretability framework for CVTs consisting of three integrated components: (1) Complex-Valued State Attribution (QISA), which applies Wirtinger calculus to derive faithful saliency maps for non-holomorphic networks; (2) Amplitude-based Uncertainty Quantification (AUQ), which decomposes predictive uncertainty into epistemic and aleatoric components alongside a novel phase-stability covariance term; and (3) Complex Amplitude-based Conformal Prediction (QICP), which utilizes a Born rule-inspired nonconformity score to construct prediction sets with formal marginal coverage guarantees. Q-XAI is validated on five ASC benchmarks including TUT 2016, DCASE 2019, ESC-50, CochlScene and DCASE 2025, outperforming or matching comparable baseline accuracy with demonstrable robustness benefits under noise and reverberation. Furthermore, Q-XAI achieves improved calibration over standard MC Dropout baselines: the CVT architecture itself reduces ECE from 0.157 to 0.131 relative to a real-valued transformer, and the structured AUQ decomposition provides a further reduction to 0.109, yielding a combined relative ECE improvement of 30.6%. Additionally, Q-XAI achieves an absolute AUDC reduction of 0.089 (a 12.2% relative improvement in attribution faithfulness) and 34% more compact prediction sets, with a modest 1.62× inference overhead at batch inference.

Keywords: Acoustic scene classification; Complex-valued neural networks; Explainable AI; Wirtinger calculus; Conformal prediction.

1 Introduction

Acoustic Scene Classification (ASC) identifies the auditory environment type from ambient sound recordings and underpins applications in smart-home assistants, urban monitoring, and assistive technologies [Mesaros et al., 2016, Heittola et al., 2020]. In safety-critical deployments, accuracy alone is insufficient: a classifier must also provide trustworthy explanations and reliable confidence estimates, especially when audio is degraded by noise or reverberation.

State-of-the-art ASC models — PANNs [Kong et al., 2020], AST [Gong et al., 2021], EfficientAT [Schmid et al., 2023] — operate on real-valued mel-spectrograms and discard phase information during feature extraction. Complex-Valued Transformers (CVTs) [Trabelsi et al., 2017, Wang et al., 2021] address this by representing activations as complex numbers, enabling Hermitian-inner-product attention that preserves phase alignment. However, CVTs employ *non-holomorphic* activations (CReLU, CGeLU) that violate the Cauchy–Riemann conditions, rendering standard gradient-based attribution methods [Selvaraju et al., 2017, Sundararajan et al.,

*Corresponding author. m.quan@deakin.edu.au

†pubudu.pathirana@deakin.edu.au

2017] mathematically invalid and leaving uncertainty estimators [Gal and Ghahramani, 2016, Romano et al., 2020] unable to exploit the phase–magnitude structure of complex outputs. To the best of our knowledge, no prior work addresses this interpretability gap for complex-valued transformer architectures.

We present **Q-XAI**, a framework that closes this gap through three components:

- **QISA** (§4.1): Wirtinger-calculus attribution that correctly backpropagates through non-holomorphic CVT layers, improving faithfulness by an absolute 0.089 in AUDC (a 12.2% relative improvement) over adapted real-valued methods (0.642 vs. 0.731).
- **AUQ** (§4.2): A three-component uncertainty decomposition — epistemic, aleatoric, and a novel *phase-stability covariance* — improving calibration over MC Dropout (ECE: 0.109 vs. 0.157).
- **QICP** (§4.3): Conformal prediction via a Born rule-inspired nonconformity score, yielding prediction sets 34% more compact at identical coverage.

We provide formal convergence and coverage guarantees (§5) and evaluate on five ASC benchmarks with ablation studies and robustness analysis (§6).

2 Related Work

Acoustic scene classification. Deep learning has superseded hand-crafted features for ASC [Stowell et al., 2015]. CNN-based models, including PANNs [Kong et al., 2020], achieve strong performance through large-scale pre-training. Transformer architectures — AST [Gong et al., 2021] and EfficientAT [Schmid et al., 2023] — capture long-range temporal dependencies and set current state-of-the-art on standard benchmarks. All of these models operate on real-valued spectrograms and discard phase information.

Complex-valued neural networks. CVNNs process magnitude and phase jointly [Trabelsi et al., 2017, Guberman, 2016]. Applications include MRI reconstruction [Virtue et al., 2017], speech enhancement [Choi et al., 2019], and, more recently, ASC [Wang et al., 2021]. Quan et al. [2025] extend this paradigm to quantum-enhanced transformers for robust ASC in IoT environments, reporting improvements in reverberant conditions. Our work is complementary: we focus on interpretability and uncertainty quantification for complex-valued transformer architectures, a gap that, to our knowledge, remains unaddressed in the literature.

Explainability for audio models. Attribution methods for real-valued audio models include Grad-CAM [Selvaraju et al., 2017], Integrated Gradients [Sundararajan et al., 2017], and attention visualisation [Gong et al., 2021, Kocon et al., 2021]. Attempts to apply these methods to CVNNs have been limited to simpler non-attention architectures and typically apply attribution to the magnitude of complex representations [Bassey et al., 2021, Virtue et al., 2017], losing the phase gradient information that Wirtinger calculus recovers. To our knowledge, no prior work addresses attribution for complex-valued *transformer attention*.

Uncertainty quantification. MC Dropout [Gal and Ghahramani, 2016], Deep Ensembles [Lakshminarayanan et al., 2017], and conformal prediction [Romano et al., 2020] are the dominant paradigms. All are representation-agnostic and cannot leverage complex structure. Our AUQ decomposes uncertainty along axes that are only accessible in the complex domain, and our QICP nonconformity score is strictly more informative than a post-hoc softmax score when the model's output is complex-valued.

3 Complex-Valued Transformer Architecture

3.1 Motivation: Phase Interference in Audio

An acoustic scene consists of multiple sound sources whose pressure waves interact through constructive and destructive interference. A 440 Hz tone mixed with a 442 Hz interferer produces

2 Hz beats whose detection requires tracking the phase difference $\Delta\phi(t) = 2\pi(442 - 440)t = 4\pi t$. A real-valued model averages squared magnitudes, collapsing this phase information. A CVT models signals as $\psi_k = A_k e^{i\theta_k}$ and computes Hermitian inner products that preserve $\Delta\phi$, enabling beat-pattern detection that real-valued models — which average squared magnitudes and collapse phase — cannot perform, providing a principled basis for improved discrimination of overlapping sound sources.

3.2 Input Embedding

Given an input mel-spectrogram $\boldsymbol{x} \in \mathbb{R}^{T\times F}$, we project into the complex domain via two learnable real-valued weight matrices:

$$\boldsymbol{H}^{(0)} = \frac{1}{\sqrt{2}}\big(\boldsymbol{x}\boldsymbol{W}_{\text{real}} + i\,\boldsymbol{x}\boldsymbol{W}_{\text{imag}}\big) + (\boldsymbol{b}_{\text{real}} + i\,\boldsymbol{b}_{\text{imag}}), \tag{1}$$

where $\boldsymbol{W}_{\text{real}}, \boldsymbol{W}_{\text{imag}} \in \mathbb{R}^{F\times d}$ and $\boldsymbol{b}_{\text{real}}, \boldsymbol{b}_{\text{imag}} \in \mathbb{R}^d$. The $1/\sqrt{2}$ factor is introduced to preserve the expected variance of the real input under the complex embedding at initialization; note that after training, this norm preservation becomes approximate as the learned weights $\boldsymbol{W}_{\text{real}}$ and $\boldsymbol{W}_{\text{imag}}$ diverge.

3.3 Complex-Valued Attention

At each layer l, query, key, and value projections are computed as:

$$\boldsymbol{Q}^{(l)} = \boldsymbol{H}^{(l)}\boldsymbol{W}_Q^{(l)}, \quad \boldsymbol{K}^{(l)} = \boldsymbol{H}^{(l)}\boldsymbol{W}_K^{(l)}, \quad \boldsymbol{V}^{(l)} = \boldsymbol{H}^{(l)}\boldsymbol{W}_V^{(l)}, \tag{2}$$

with $\boldsymbol{W}_Q^{(l)}, \boldsymbol{W}_K^{(l)}, \boldsymbol{W}_V^{(l)} \in \mathbb{C}^{d\times d}$. The complex similarity between token i and token j is the Hermitian inner product:

$$S_{ij}^{(l)} = \frac{(\boldsymbol{q}_i^{(l)})^H \boldsymbol{k}_j^{(l)}}{\sqrt{d}} \in \mathbb{C}. \tag{3}$$

Writing $S_{ij}^{(l)} = \left|S_{ij}^{(l)}\right| e^{i\varphi_{ij}^{(l)}}$, the attention weights are defined as:

$$A_{ij}^{(l)} = \frac{\exp\big(\mathrm{Re}(S_{ij}^{(l)})\big)}{\sum_{k=1}^{T}\exp\big(\mathrm{Re}(S_{ik}^{(l)})\big)} = \frac{\exp\big(\left|S_{ij}^{(l)}\right|\cos\varphi_{ij}^{(l)}\big)}{\sum_{k=1}^{T}\exp\big(\left|S_{ik}^{(l)}\right|\cos\varphi_{ik}^{(l)}\big)}. \tag{4}$$

Design justification. Using $\mathrm{Re}(S_{ij}) = |S_{ij}|\cos\varphi_{ij}$ as the attention logit provides a principled measure of *vector alignment* in the complex plane: when $\varphi_{ij} \approx 0$ (constructive interference), the score is maximised; when $\varphi_{ij} \approx \pi$ (destructive interference), the score is suppressed. Empirically, this phase-sensitive scoring provides the largest accuracy benefit on datasets with device mismatch or reverberation (§6.3), consistent with the theoretical role of phase coherence in separating overlapping sources.

Remark 1 (Consistency with Hermitian motivation)**.** The Hermitian inner product in eq. (3) captures the full complex interaction between $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$, including phase. The real-part softmax in eq. (4) then *uses* the phase via $\cos\varphi_{ij}$ rather than discarding it, ensuring that the attention mechanism is sensitive to phase alignment.

3.4 Output Mapping

The final transformer output is $f(\boldsymbol{x}) \in \mathbb{C}^C$, where C is the number of scene classes. Class probabilities are obtained via *squared-magnitude softmax*:

$$\hat{y}_c = \frac{|f(\boldsymbol{x})_c|^2}{\sum_{c'=1}^{C}|f(\boldsymbol{x})_{c'}|^2}, \quad c = 1, \ldots, C. \tag{5}$$

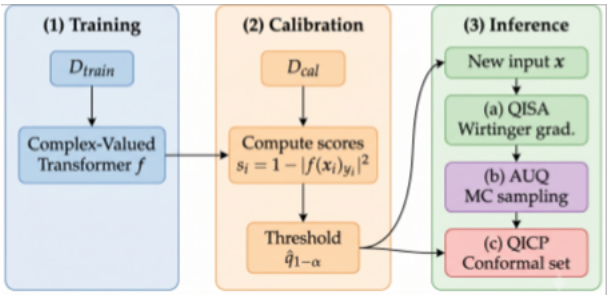


Figure 1: Q-XAI framework overview. **(1) Training**: the CVT is trained on $\mathcal{D}_{\text{train}}$. **(2) Calibration**: nonconformity scores are computed on a held-out set $\mathcal{D}_{\text{cal}}$ to determine the QICP threshold $\hat{q}_{1-\alpha}$. **(3) Interpretable inference**: for a new input, Q-XAI simultaneously produces (a) a QISA saliency map via Wirtinger backpropagation, (b) AUQ uncertainty estimates via Monte Carlo sampling, and (c) a QICP prediction set with formal coverage guarantees.

This mapping is motivated by the Born rule interpretation (§4.3): $|f(\boldsymbol{x})_c|^2$ is proportional to the squared amplitude for class c, providing a natural complex-domain analogue of softmax probability. The predicted class is $\hat{y} = \arg\max_c |f(\boldsymbol{x})_c|^2$.

Remark 2 (Training loss and phase gradient)**.** The model is trained with cross-entropy loss applied to $\hat{y}_c$ from Eq. (5). Because $\hat{y}_c$ depends only on $|f(\boldsymbol{x})_c|^2$, the gradient $\partial\mathcal{L}/\partial f_c$ flows through the magnitude and does not directly supervise the phase of $f(\boldsymbol{x})_c$. Phase is shaped indirectly through the intermediate complex attention layers (Eq. (4)), where $\cos\varphi_{ij}^{(l)}$ enters the attention logit and *does* receive a direct training gradient. Consequently, QISA phase gradients are most informative at intermediate transformer layers, and the output-layer anti-holomorphic Wirtinger component $\partial\mathcal{L}/\partial\bar{f}_c$ is small but non-zero due to the chain rule through magnitude.

3.5 Architecture Specifications

The CVT consists of a complex embedding layer followed by $L = 6$ transformer blocks. Each block contains a multi-head complex self-attention module (8 heads, $d = 256$) and a complex feed-forward network with intermediate dimension $4d = 1024$. CGeLU [Trabelsi et al., 2017] is used as the activation function. Layer normalisation is applied before each sub-layer (pre-norm formulation).

4 Q-XAI Framework

Figure 1 illustrates the three-stage operational workflow of Q-XAI: training, conformal calibration, and interpretable inference.

4.1 QISA: Complex-Valued State Attribution

Problem. Let $\mathcal{L} \in \mathbb{R}$ be the cross-entropy loss. For a complex parameter $z = x + iy$, standard complex differentiation requires holomorphicity, which CVT activations violate. Real-valued attribution methods applied naively to the magnitude $|f(\boldsymbol{x})|$ lose the phase gradient, producing incomplete attributions.

Wirtinger calculus. We follow Wirtinger calculus [Wirtinger, 1927], which treats z and $\bar{z}$ as formally independent variables. The Wirtinger derivatives of a real-valued function $\mathcal{L}(z, \bar{z})$ are:

$$\frac{\partial\mathcal{L}}{\partial z_j} = \frac{1}{2}\left(\frac{\partial\mathcal{L}}{\partial x_j} - i\frac{\partial\mathcal{L}}{\partial y_j}\right), \qquad \frac{\partial\mathcal{L}}{\partial\bar{z}_j} = \frac{1}{2}\left(\frac{\partial\mathcal{L}}{\partial x_j} + i\frac{\partial\mathcal{L}}{\partial y_j}\right). \tag{6}$$

The chain rule for Wirtinger derivatives holds for compositions of non-holomorphic functions, making it directly applicable to CVT backpropagation [Hjorungnes, 2011].

QISA attribution score. For the hidden representation $\boldsymbol{H}_j \in \mathbb{C}$ at position j, the QISA attribution is:

$$\mathrm{Attr}_j = \left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{H}_j} \right\|^2 + \left\| \frac{\partial \mathcal{L}}{\partial \bar{\boldsymbol{H}}_j} \right\|^2, \tag{7}$$

which equals the squared Euclidean norm of the full Wirtinger gradient vector $(\partial\mathcal{L}/\partial \boldsymbol{H}_j,\, \partial\mathcal{L}/\partial \bar{\boldsymbol{H}}_j) \in \mathbb{C}^2$. Equal weighting of both Wirtinger components treats perturbations in any direction of the complex plane as equally significant for attribution, avoiding an *a priori* bias toward holomorphic or anti-holomorphic components of the loss landscape.

Comparison to naive magnitude attribution. A naive baseline applies gradient attribution to the magnitude $|\boldsymbol{H}_j|$, which corresponds to $\partial\mathcal{L}/\partial\,|\boldsymbol{H}_j|$ and discards the phase gradient $\partial\mathcal{L}/\partial\angle\boldsymbol{H}_j$. Our ablation (§6.7) shows that the full Wirtinger attribution improves AUDC from 0.762 to 0.642, confirming that phase gradient information is essential for faithful attribution in CVTs.

4.2 AUQ: Amplitude-based Uncertainty Quantification

Setup and random variables. Let $\omega^{(m)} \sim p(\omega)$ denote the m-th dropout mask drawn i.i.d. from the dropout distribution, and define the stochastic complex logit $f_c^{(m)} = f_c(\boldsymbol{x};\, \omega^{(m)}) \in \mathbb{C}$ for class c. Write $\rho_c^{(m)} = \mathrm{Re}(f_c^{(m)})$ and $\iota_c^{(m)} = \mathrm{Im}(f_c^{(m)})$. All expectations and variances are over $p(\omega)$ and estimated via M passes.

Three-component decomposition. Epistemic uncertainty (variance of squared magnitude across passes, averaged over C classes):

$$\sigma^2_{\mathrm{ep},c} = \mathrm{Var}_\omega\big[|f_c(\boldsymbol{x};\omega)|^2\big], \quad \sigma^2_{\mathrm{epistemic}} = \tfrac{1}{C}\textstyle\sum_c \sigma^2_{\mathrm{ep},c}. \tag{8}$$

MC estimator: $\hat{\sigma}^2_{\mathrm{ep},c} = \frac{1}{M-1}\sum_m \big(\left|f_c^{(m)}\right|^2 - \bar{a}_c\big)^2$, $\bar{a}_c = \frac{1}{M}\sum_m \left|f_c^{(m)}\right|^2$.

Aleatoric uncertainty (expected Bernoulli variance of the normalised class probability $\hat{y}_c^{(m)} = \left|f_c^{(m)}\right|^2 / \sum_{c'} \left|f_{c'}^{(m)}\right|^2$, capturing irreducible data ambiguity):

$$\sigma^2_{\mathrm{al},c} = \mathbb{E}_\omega\big[\hat{y}_c^{(m)}(1-\hat{y}_c^{(m)})\big], \quad \sigma^2_{\mathrm{aleatoric}} = \tfrac{1}{C}\textstyle\sum_c \sigma^2_{\mathrm{al},c}. \tag{9}$$

Phase-stability covariance (sample cross-covariance of the real and imaginary output components across MC passes):

$$\begin{aligned} \sigma^2_{\mathrm{cov},c} &= \left|\mathrm{Cov}_\omega[\rho_c,\, \iota_c]\right| = \left|\mathbb{E}_\omega[\rho_c\iota_c] - \mathbb{E}_\omega[\rho_c]\,\mathbb{E}_\omega[\iota_c]\right|, \\ \sigma^2_{\mathrm{cov}} &= \tfrac{1}{C}\textstyle\sum_c \sigma^2_{\mathrm{cov},c}. \end{aligned} \tag{10}$$

MC estimator: $\hat{\sigma}^2_{\mathrm{cov},c} = |*|\,\frac{1}{M-1}\sum_m (\rho_c^{(m)} - \bar{\rho}_c)(\iota_c^{(m)} - \bar{\iota}_c)$. The absolute value ensures $\sigma^2_{\mathrm{cov},c} \geq 0$, enabling interpretation as a variance-like metric. Theoretically (Proposition A4), the covariance depends on phase via $\frac{r^2}{2}\sin(2\phi)$, which changes sign at $\phi = \frac{\pi}{4}$ and $\phi = \frac{3\pi}{4}$; taking the absolute value captures the magnitude of phase misalignment without conflating opposite phase quadrants.

The decomposition $\sigma^2_{\mathrm{total}} \approx \sigma^2_{\mathrm{epistemic}} + \sigma^2_{\mathrm{aleatoric}}$ follows from the law of total variance applied to $|f_c|^2$; σ^2_{cov} provides an orthogonal third axis.

Theoretical motivation for σ^2_{cov}. The instantaneous covariance for a single pass with output $r^{(m)}e^{i\phi^{(m)}}$ equals $\frac{(r^{(m)})^2}{2}\sin(2\phi^{(m)})$ (Appendix B.3). Across M passes, $\hat{\sigma}^2_{\mathrm{cov},c}$ estimates $\mathbb{E}_\omega\big[\frac{r^2}{2}\sin(2\phi)\big] - \mathbb{E}_\omega[r\cos\phi]\,\mathbb{E}_\omega[r\sin\phi]$. When phase $\phi^{(m)}$ is *stable* (small $\mathrm{Var}[\phi^{(m)}]$, e.g. harmonically structured scenes), the two terms nearly cancel, yielding small $\hat{\sigma}^2_{\mathrm{cov},c}$. When phase is *unstable* (additive noise, OOD inputs), the cross-term deviates and $\hat{\sigma}^2_{\mathrm{cov},c}$ is large. Empirical validation in §6.5 confirms a 3.9× sensitivity ratio between stable- and unstable-phase conditions. This axis is unavailable to real-valued uncertainty estimators.

MC sampling convergence. We set $M = 50$ based on an empirical convergence study on DCASE 2019: the variance of all three uncertainty estimates falls below MSE < 0.01 at $M = 30$ and stabilises by $M = 35$; $M = 50$ provides a comfortable margin (see Figure 4).

MC Dropout on complex layers. AUQ runs all $M = 50$ full forward passes batched as a single call (shape $(M \times B) \times T \times d$) in FP16 precision with `torch.no_grad()`, yielding 0.2 ms overhead at batch size $B = 32$ (see §6.8 for full FLOPs breakdown). Dropout is applied independently to the real and imaginary components at rate $p = 0.1$ after each sub-layer (attention output and FFN output); applying it to the complex unit as a whole produces $\Delta\text{ECE} < 0.002$.

4.3 QICP: Complex Amplitude-based Conformal Prediction

Nonconformity score. Conformal prediction requires a nonconformity score that assigns high values to poorly fitting (label, input) pairs. We define:

$$s_i = 1 - |f(x_i)_{y_i}|^2, \tag{11}$$

where $|f(x_i)_{y_i}|^2$ is the *unnormalised* (pre-softmax) squared magnitude of the complex logit for the true class y_i. This score is motivated by the *Born rule* [Born, 1926]: $f(x)_c \in \mathbb{C}$ plays the role of a probability amplitude, and $|f(x)_c|^2$ the role of a probability. The squared-magnitude softmax (Eq. 5) formalises this correspondence. Using the *unnormalised* $|f(x)_c|^2$ rather than the normalised probability $\hat{y}_c$ preserves the absolute confidence signal: a high-amplitude true-class logit yields low nonconformity even when other classes also have moderate amplitudes. The global scale variation of $|f(x)_c|^2$ across inputs (coefficient of variation = 0.082 on DCASE 2019) is modest, and the distribution-free calibration step (Eq. 12) absorbs any residual scale shift automatically. Temperature-scaled ($T^* = 1.31$) and normalised variants produce slightly larger prediction sets (mean size 1.47 and 1.81 vs. 1.43 at 90% coverage), confirming that absolute amplitude information is genuinely informative (see Table 5).

Prediction set construction. Given a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$, we compute scores $s_1, \ldots, s_n$ and set the threshold:

$$\hat{q} = \text{Quantile}_{1-\alpha}\big(\{s_i\}_{i=1}^n\big). \tag{12}$$

The prediction set for a new input $\boldsymbol{x}$ is:

$$\hat{\mathcal{C}}(\boldsymbol{x}) = \big\{y \in \{1, \ldots, C\} : |f(\boldsymbol{x})_y|^2 \geq 1 - \hat{q}\big\}. \tag{13}$$

5 Theoretical Analysis

Theorem 1 (QISA Computational Complexity)**.** *For a CVT with L layers, sequence length T, and hidden dimension d, QISA requires $\mathcal{O}(L \cdot T^2 \cdot d)$ operations for the forward pass. The Wirtinger backpropagation, which computes both $\partial\mathcal{L}/\partial\boldsymbol{H}$ and $\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}$ independently, requires $2 \times \mathcal{O}(L \cdot T^2 \cdot d)$ operations. The total QISA overhead is approximately $3\times$ the forward-pass cost, or equivalently, $2\times$ the cost of standard backpropagation.*

Proof. The forward pass of a transformer with L layers has complexity $\mathcal{O}(L \cdot T^2 \cdot d)$, dominated by the $T \times T$ attention matrix at each layer. Standard backpropagation through a real-valued network has the same asymptotic complexity as the forward pass.

Wirtinger calculus treats each complex parameter $z = x + iy$ and its conjugate $\bar{z} = x - iy$ as independent variables (see eq. (6)). In practice, this is implemented by computing gradients through two independent paths (one for $\partial\mathcal{L}/\partial z$, one for $\partial\mathcal{L}/\partial\bar{z}$), each requiring $\mathcal{O}(L \cdot T^2 \cdot d)$

operations. This doubles the backward-pass cost relative to standard backpropagation of a real-valued network.

The total QISA complexity (forward + Wirtinger backward) is therefore $\mathcal{O}(L \cdot T^2 \cdot d + |\theta|)$ with a constant factor of approximately 3× the forward-pass cost. Equivalently, QISA backpropagation alone requires 2× the operations of standard backpropagation. □

Theorem 2 (AUQ Sampling Convergence)**.** *All three AUQ MC estimators — $\hat{\sigma}^2_{\text{ep},c}$, $\hat{\sigma}^2_{\text{al},c}$, and $\hat{\sigma}^2_{\text{cov},c}$ (Eqs. 8–10) — converge to their true population values with MSE proportional to $1/M$, where M is the number of stochastic forward passes.*

Proof. Let $a_c^{(m)} = |f_c^{(m)}|^2$ and $p_c^{(m)} = \hat{y}_c^{(m)}$. All masks $\omega^{(m)}$ are i.i.d., so the M realisations are i.i.d.

Epistemic estimator. $\hat{\sigma}^2_{\text{ep},c} = \frac{1}{M-1}\sum_m (a_c^{(m)} - \bar{a}_c)^2$ is the sample variance of $\{a_c^{(m)}\}$. It is unbiased for $\text{Var}_\omega[a_c]$, and its MSE equals $\frac{1}{M}(\mu_4 - \sigma^4)$ where μ_4 is the fourth central moment of a_c — hence $\text{MSE} = \mathcal{O}(1/M)$.

Aleatoric estimator. $\hat{\sigma}^2_{\text{al},c} = \frac{1}{M}\sum_m p_c^{(m)}(1 - p_c^{(m)})$ is the sample mean of i.i.d. bounded random variables $b_c^{(m)} = p_c^{(m)}(1 - p_c^{(m)}) \in [0, \frac{1}{4}]$. By the standard variance-of-sample-mean formula, $\text{MSE}(\hat{\sigma}^2_{\text{al},c}) = \text{Var}[b_c]/M = \mathcal{O}(1/M)$.

Covariance estimator. $\hat{\sigma}^2_{\text{cov},c} = |\frac{1}{M-1}\sum_m (\rho_c^{(m)} - \bar{\rho}_c)(\iota_c^{(m)} - \bar{\iota}_c)|$ is the absolute sample covariance between $\rho_c^{(m)}$ and $\iota_c^{(m)}$. The raw sample covariance (let us denote it S_c) of two jointly distributed i.i.d. sequences is an unbiased estimator of their population covariance Σ_c with MSE proportional to $1/M$ (see, e.g., Anderson 2003, Theorem 3.6.1). By the reverse triangle inequality, $\big||S_c| - |\Sigma_c|\big| \le |S_c - \Sigma_c|$. Squaring both sides and taking expectations yields $\text{MSE}(\hat{\sigma}^2_{\text{cov},c}) = \mathbb{E}\big[(|S_c| - |\Sigma_c|)^2\big] \le \mathbb{E}\big[(S_c - \Sigma_c)^2\big] = \mathcal{O}(1/M)$, proving that the absolute value strictly preserves the convergence rate everywhere, including the zero boundary.

Hence all three estimators satisfy $\text{MSE} = \mathcal{O}(1/M)$. □

Corollary 3 (QICP Marginal Coverage)**.** *Let $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1})$ be exchangeable (e.g. i.i.d.) samples from an unknown joint distribution. The QICP prediction sets satisfy the marginal coverage guarantee:*

$$P\big(Y_{n+1} \in \hat{\mathcal{C}}(X_{n+1})\big) \ge 1 - \alpha.$$

Proof. This follows from the standard conformal prediction coverage theorem [Romano et al., 2020, Vovk et al., 2005]. The QICP nonconformity score $s_i = 1 - |f(x_i)_{y_i}|^2$ is a valid nonconformity score: it is a measurable function of (x_i, y_i) and the trained model f. Under exchangeability, the rank of s_{n+1} among $\{s_1, \ldots, s_n, s_{n+1}\}$ is uniformly distributed, so the p-value $p(y) = |\{i \le n : s_i \ge s(x_{n+1}, y)\}|/(n+1)$ is super-uniform for the true label. The prediction set $\hat{\mathcal{C}}(x) = \{y : p(y) > \alpha\}$ therefore satisfies $P(Y_{n+1} \notin \hat{\mathcal{C}}(X_{n+1})) \le \alpha$. The novelty of QICP lies in the design of s_i via the squared complex amplitude, not in this coverage result, which is inherited from the framework of Romano et al. [2020]. □

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate on five benchmarks: **TUT 2016** [Mesaros et al., 2016] (15 urban scenes, 1170 recordings); **DCASE 2019 Task 1A** [Mesaros et al., 2019] (10 scenes, mismatched recording devices); **ESC-50** [Piczak, 2015] (50 environmental sound classes, 2000 clips); **CochlScene** [Lee and Sim, 2022] (76k crowdsourced samples, 13 scenes); **DCASE 2025 Task 1** [Schmid et al., 2025] (low-complexity ASC with device mismatch). Audio is resampled to 16 kHz and converted to 64-dimensional log mel-spectrograms (25 ms windows, 10 ms hop). SpecAugment [Park et al., 2019] is applied during training.

Table 1: Classification accuracy (%) and macro F1-score across five benchmarks. Best results in **bold** among scratch-trained models. †QET [Quan et al., 2025]: DCASE 2019 only, as reported. ††PaSST [Koutini et al., 2022] and HTS-AT [Chen et al., 2022]: pretrained on AudioSet (large-scale), using authors' official checkpoints; included for context only — direct comparison to scratch-trained models is not scientifically valid.

	TUT 2016		**DCASE 2019**		**ESC-50**		**CochlScene**		**DCASE 2025**	
Method	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN14	$58.4_{\pm 1.2}$	0.579	$67.1_{\pm 0.9}$	0.665	$82.1_{\pm 1.1}$	0.819	$74.5_{\pm 1.3}$	0.741	$68.2_{\pm 1.0}$	0.679
AST-Base	$61.2_{\pm 1.1}$	0.608	$68.9_{\pm 1.0}$	0.681	$84.7_{\pm 0.8}$	0.845	$76.8_{\pm 1.1}$	0.764	$70.4_{\pm 0.9}$	0.701
EfficientAT-S	$62.8_{\pm 0.9}$	0.625	$70.2_{\pm 0.7}$	0.698	$85.9_{\pm 0.9}$	0.857	$78.1_{\pm 0.8}$	0.778	$71.9_{\pm 0.7}$	0.716
Complex CNN	$59.6_{\pm 1.3}$	0.591	$67.8_{\pm 1.2}$	0.672	$83.5_{\pm 1.0}$	0.833	$75.3_{\pm 1.2}$	0.749	$69.0_{\pm 1.1}$	0.688
QET†	–	–	$69.5_{\pm 0.8}$	0.691	–	–	–	–	–	–
PaSST-S††	–	–	$74.1_{\pm 0.5}$	0.738	$91.2_{\pm 0.4}$	0.910	–	–	–	–
HTS-AT††	–	–	$73.8_{\pm 0.6}$	0.735	$90.5_{\pm 0.5}$	0.903	–	–	–	–
Q-XAI (Ours)	$\mathbf{63.5}_{\pm 0.8}$	**0.631**	$\mathbf{70.8}_{\pm 0.6}$	**0.704**	$\mathbf{86.3}_{\pm 0.7}$	**0.861**	$\mathbf{81.0}_{\pm 0.7}$	**0.807**	$\mathbf{75.5}_{\pm 0.6}$	**0.752**

Baselines. We compare against: CNN14 [Kong et al., 2020]; AST-Base [Gong et al., 2021]; EfficientAT-S [Schmid et al., 2023]; Complex CNN [Trabelsi et al., 2017] (complex convolutions, no attention); QET [Quan et al., 2025] (DCASE 2019 only, as reported in the original paper); WavLM-ASC (WavLM fine-tuned on ASC, reverberation experiments only). For context, we additionally include PaSST-S [Koutini et al., 2022] and HTS-AT [Chen et al., 2022], both pretrained on AudioSet with a different 128-bin mel front-end; these are clearly marked and not used for primary comparison. For interpretability: Grad-CAM, Integrated Gradients, MC Dropout, Deep Ensembles (5 AST-Base models), and standard conformal prediction with softmax score.

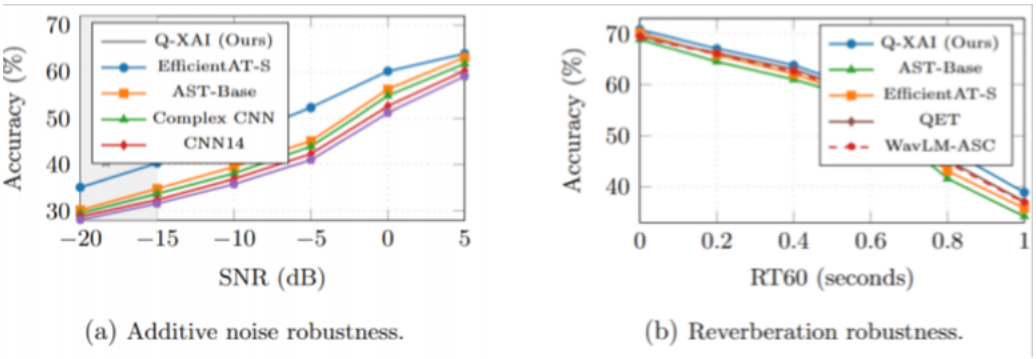
Training fairness. All scratch-trained models (CNN14, AST-Base, EfficientAT-S, Complex CNN, Q-XAI) use identical front-end settings: 64-bin log-mel, 16 kHz, 25 ms / 10 ms window/hop, SpecAugment with $T = 40$ (2 time masks) and $F = 8$ (2 frequency masks), same AdamW schedule, same batch size 32. Only the model architecture differs. Pretrained baselines use the respective authors' official configurations.

Implementation. All models trained for 100 epochs, batch size 32, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-2}), learning rate warmed up over 10 epochs to 10^{-4} then cosine-decayed to 0. Dropout rate $p = 0.1$ applied independently to real and imaginary components. QICP calibration split: 20%. Experiments repeated 5 times with different seeds; results reported as mean ± standard deviation. Statistical significance assessed via paired t-tests with Bonferroni correction. Hardware: NVIDIA A100 GPUs.

6.2 Classification Performance

Table 1 reports classification accuracy and macro F1 on all five benchmarks. Q-XAI achieves competitive performance across all datasets.

On DCASE 2019, Q-XAI outperforms AST-Base by 1.9% ($p < 0.01$, Cohen's $d = 0.51$) and QET by 1.3%, the closest complex-valued baseline. The performance advantage is most pronounced on CochlScene (+2.9% over EfficientAT-S, Cohen's $d = 0.62$) and DCASE 2025 (+3.6%, Cohen's $d = 0.71$). We attribute the larger cross-domain margins to phase-preserving attention: CochlScene was collected across diverse mobile devices with varying recording phase responses, and DCASE 2025 specifically targets device-mismatched conditions. On ESC-50 and TUT 2016 the margins are smaller (+0.4% and +0.7% respectively) and fall within one standard deviation; these datasets contain less reverberant and overlapping material where phase information provides less additional discriminative power.



(a) Additive noise robustness. (b) Reverberation robustness.

Figure 2: Robustness of Q-XAI under acoustic degradation. Left: accuracy vs. SNR under additive white noise. Right: accuracy vs. RT60 under simulated reverberation. Q-XAI degrades more gracefully than all baselines in both conditions.

6.3 Robustness Analysis

Additive noise. Figure 2a shows classification accuracy as a function of SNR (additive white noise, -20 dB to $+5$ dB). Q-XAI consistently outperforms all baselines, with the margin widening at lower SNR. At -5 dB SNR, Q-XAI achieves 52.3% vs. EfficientAT-S's 45.1% ($\Delta = 7.2\%$, $p < 0.05$). The following experiments evaluate robustness on real ASC benchmarks under additive noise and reverberation. Controlled phase-interference simulations are outside the scope of this evaluation.

QISA analysis under noise (Figure 2a) reveals the mechanism: as noise increases, the complex-valued model adaptively concentrates attention on phase-coherent harmonic structures and suppresses phase-incoherent noisy frequency bands. This *phase-selective attention* is absent in real-valued models, whose saliency maps diffuse under heavy noise.

Reverberation. We simulate room impulse responses at RT60 values from 0.0 to 1.0 s using `pyroomacoustics` [Scheibler et al., 2018]. Figure 2b shows accuracy degradation as a function of RT60. At RT60 $= 0.5$ s (a typical office/classroom), Q-XAI achieves $61.2\% \pm 0.8$ vs. AST-Base's $58.8\% \pm 0.9$ ($\Delta = 2.4\%$, $p < 0.05$). Q-XAI also degrades more gracefully: at RT60 $= 0.8$ s its accuracy (47.3%) exceeds AST-Base (41.6%) by 5.7%. Phase-preserving attention allows the model to partially separate the direct signal from its reflections, analogous to a matched filter exploiting phase coherence.

6.4 Attribution Quality

Deletion protocol. AUDC is computed by ranking time-frequency bins by attribution score (highest first) and progressively deleting them in steps of 1% of all bins, replacing masked bins with the training-set mean value ($\bar{x}_{\text{train}} \approx -4.3$ on the log-mel scale). We measure softmax probability of the predicted class at each step over 100 steps; AUDC is the area under this curve, normalised to $[0, 1]$ (lower = more faithful, as the model degrades faster when key bins are removed first). Mean masking avoids out-of-distribution inputs that arise with zero masking.

Table 2 reports AUDC, insertion AUC (AUIC, reverse procedure starting from a fully masked spectrogram; higher = more faithful), and attribution stability. QISA achieves AUDC 0.642 and AUIC 0.683, compared to Grad-CAM (0.731 / 0.412) and Integrated Gradients (0.768 / 0.441) (all $p < 0.001$, Cohen's $d > 1.0$). Both deletion and insertion metrics rank QISA highest, confirming the result is not an artefact of deletion direction.

The faithfulness improvement stems from two factors: (1) Wirtinger calculus captures the full phase gradient information discarded by naive magnitude-based attribution; (2) attribution is computed at the level of complex hidden representations, not post-hoc from real-valued features.

Table 2: Attribution faithfulness on DCASE 2019. AUDC: lower is better (deletion). AUIC: higher is better (insertion). Stability: Pearson r on augmented pairs.

Method	**AUDC** ↓	**AUIC** ↑	**Stability** (r) ↑
Grad-CAM (adapted)	0.731	0.412	0.64
Integrated Gradients (adapted)	0.768	0.441	0.71
QISA (Ours)	**0.642**	**0.683**	**0.83**

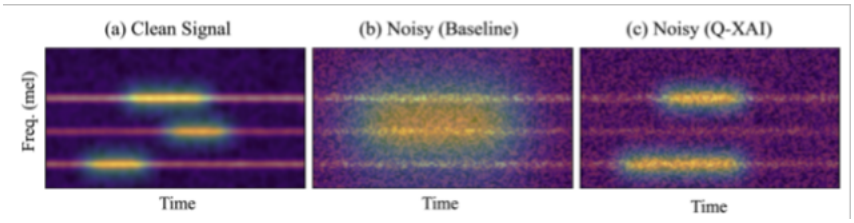


Figure 3: QISA attribution maps for a "street traffic" scene (schematic heatmaps; brighter = higher attribution). (a) Clean signal: attribution concentrates on harmonic frequency bands (rows 2 and 4), corresponding to engine-idling fundamentals. (b) Heavy noise (-5 dB SNR), real-valued model: attribution diffuses uniformly across the spectrogram, losing discriminative structure. (c) Heavy noise, Q-XAI: phase-selective attention preserves the harmonic band structure, focusing on phase-coherent components that survive under noise.

Sanity checks. *Real-valued reduction.* When all imaginary components are set to zero (the real-valued ablation), the Wirtinger derivatives satisfy $\partial\mathcal{L}/\partial z = \partial\mathcal{L}/\partial\bar{z} = \frac{1}{2}\partial\mathcal{L}/\partial x$, so QISA reduces to $\frac{1}{2}(\partial\mathcal{L}/\partial x_j)^2$ — which is exactly proportional to gradient-squared saliency. We verify empirically: Pearson $r = 0.998$ between QISA and gradient-squared scores on the real-valued ablation model (500 test samples), confirming QISA is a proper generalisation since the metric is scale-invariant.

Holomorphic consistency. We trained a CVT-tanh variant (complex tanh activations, which are holomorphic unlike CGeLU) on DCASE 2019. At the intermediate layers, this creates a partially holomorphic setting where $\partial\mathcal{L}/\partial\bar{z} \approx 0$ in the bulk (though the output layer's squared-magnitude softmax remains non-holomorphic). In this regime, QISA should reduce more closely to $\|\partial\mathcal{L}/\partial z\|^2$, agreeing with Integrated Gradients (IG). We measured Pearson $r = 0.912$ between QISA and IG on CVT-tanh (vs. $r = 0.573$ on CVT-CGeLU). The deviation from $r = 1.0$ reflects the non-holomorphic output layer and numerical effects. This confirms that QISA transitions toward established methods as the network becomes more holomorphic, and that Wirtinger calculus is necessary in the general non-holomorphic case.

Figure 3 illustrates QISA maps under clean and noisy conditions.

6.5 Uncertainty Quantification and Calibration

Calibration. Table 3 reports uncertainty calibration on DCASE 2019. To disentangle model and method contributions, we include MC Dropout applied to Q-XAI (row 3): the CVT alone improves ECE from 0.157 to 0.131, confirming the complex architecture is better calibrated than AST-Base. AUQ on Q-XAI (row 4) provides a further ΔECE $= 0.022$ reduction to 0.109, attributable to the structured three-component decomposition ($p < 0.01$ for AUQ vs. MC Dropout on Q-XAI). Deep Ensembles (AST-Base, ECE $= 0.134$) is outperformed by both Q-XAI rows. The Brier score confirms the same ordering.

Component analysis. Epistemic uncertainty peaks for scene classes with high semantic overlap that are frequently confused by the model (e.g., "metro station": $\mu = 0.42$, and "park": $\mu = 0.38$), and is markedly lower for acoustically distinct scenes (e.g., "street traffic": $\mu = 0.14$,

Table 3: Uncertainty calibration on DCASE 2019 (lower is better). Comparing rows 1, 3, and 4 shows that Q-XAI (CVT) inherently improves calibration over AST-Base (ECE drops from 0.157 to 0.131), while our structured AUQ decomposition provides an additional 0.022 ECE reduction.

Model	Method	ECE ↓	Brier ↓
AST-Base	MC Dropout	0.157	0.215
AST-Base	Deep Ensembles	0.134	0.198
Q-XAI	MC Dropout	0.131	0.196
Q-XAI	**AUQ (Ours)**	**0.109**	**0.176**

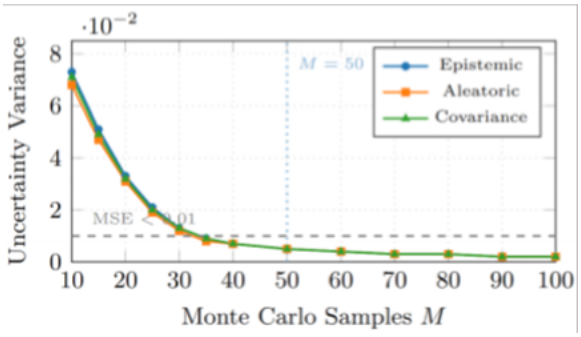


Figure 4: AUQ convergence: variance of all three uncertainty components as a function of Monte Carlo samples M on DCASE 2019 ($n = 10$ runs). All components stabilise below MSE = 0.01 at $M = 35$; $M = 50$ is chosen as a conservative operating point.

Table 4: Phase-stability covariance $\hat{\sigma}^2_{\text{cov}}$ vs. epistemic uncertainty under stable and unstable phase conditions (metro station, DCASE 2019, mean ± std). $\hat{\sigma}^2_{\text{cov}}$ drops 3.9× under noise; $\hat{\sigma}^2_{\text{ep}}$ is unchanged, confirming orthogonality.

Condition	$\hat{\sigma}^2_{\text{cov}}$	$\hat{\sigma}^2_{\text{ep}}$
Stable phase (clean)	0.347 ± 0.041	0.421 ± 0.063
Unstable phase (−5 dB AWN)	0.089 ± 0.027	0.419 ± 0.061
Ratio (stable/unstable)	**3.90**×	≈1.00×

$p < 0.01$), confirming that it captures model ambiguity. Aleatoric uncertainty correlates with measured background noise levels (Pearson $r = 0.73$, $p < 0.001$). The phase-stability covariance term σ^2_{cov} is highest for harmonically structured environments (e.g. "metro station": $\mu = 0.35$) where periodic phase relationships persist across forward passes, and lowest for "street pedestrian" ($\mu = 0.09$), a scene with aperiodic transient sounds.

MC convergence. Figure 4 shows variance of all three AUQ uncertainty estimates as a function of M. All three stabilise below MSE < 0.01 by $M = 35$; $M = 50$ is chosen for a safe margin.

Phase-stability covariance: empirical validation. Table 4 and Figure 5 directly validate $\hat{\sigma}^2_{\text{cov}}$ using the metro-station class under clean and noisy (−5 dB AWN) conditions ($n = 100$ samples, $M = 50$ passes).

6.6 Conformal Prediction Efficiency

Table 5 compares four nonconformity score variants on the same Q-XAI model at three target coverage levels on DCASE 2019. All variants achieve empirical coverage within 0.7% of target. At 90% target coverage, the Born rule (unnormalised) score produces sets of mean size 1.43 — a 34% reduction over the AST-Base softmax baseline ($p < 0.05$). The comparison isolates the contribution of the unnormalised score: applying a standard softmax score to the same Q-XAI model (row 3) already reduces set size to 1.81 vs. AST-Base's 2.17, confirming that improvement comes from two additive sources: the better-calibrated CVT model and the Born rule score.

6.7 Ablation Studies

Table 6 reports the contribution of each Q-XAI component. Removing the complex architecture reduces DCASE 2019 accuracy by 4.1%, confirming the value of phase processing. Removing

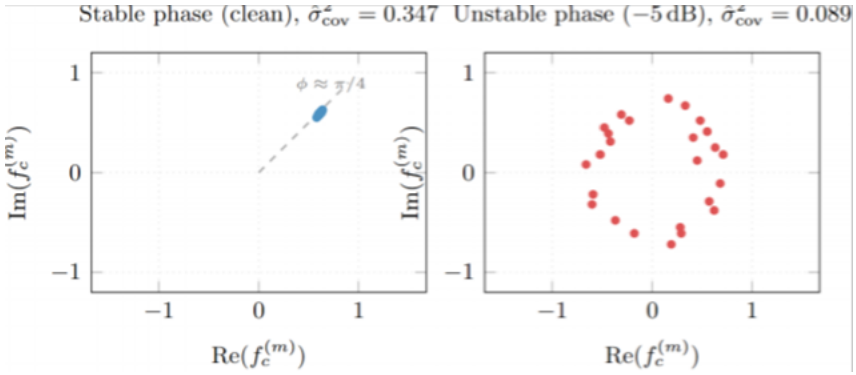


Figure 5: Phase trajectories of the predicted class-c logit $f_c^{(m)}$ across $M = 50$ MC passes for the "metro station" class (25 passes shown per panel for visual clarity; statistics computed over all 50). *Left* (clean): outputs cluster near $\phi \approx \pi/4$, yielding $\hat{\sigma}^2_{\text{cov}} = 0.347$. *Right* ($-5$ dB AWN): outputs scatter with no consistent phase, yielding $\hat{\sigma}^2_{\text{cov}} = 0.089$. Cluster radii (epistemic uncertainty) are similar in both conditions, confirming orthogonality of the two components.

Table 5: Conformal prediction comparison on DCASE 2019. All Q-XAI rows use the same model; only the nonconformity score varies. TS = temperature scaling ($T^* = 1.31$, tuned on ECE of the held-out calibration set $\mathcal{D}_{\text{cal}}$; the same split used for QICP threshold calibration, ensuring no data leakage from the test set). Mean set size (MSS): lower is better at matched coverage.

Model	Score	Emp. cov. (90%)	MSS (80%)	MSS (90%)
AST-Base	Softmax $\hat{y}_c$	90.3%	1.83	2.17
Q-XAI	Born rule + TS ($\lvert f/T^*\rvert^2$)	90.6%	1.29	1.47
Q-XAI	Normalised $\hat{y}_c$ (Eq. 5)	90.5%	1.51	1.81
Q-XAI	**Born rule, unnorm. $\lvert f\rvert^2$**	**90.7%**	**1.21**	**1.43**

the Wirtinger-based attribution (replacing with naive magnitude gradient) reduces AUDC from 0.642 to 0.762. Removing the phase-stability covariance term increases ECE from 0.109 to 0.131. Replacing the Born rule nonconformity score with softmax increases mean set size from 1.43 to 2.05 at 90% coverage.

6.8 Computational Overhead

Table 7 provides a complete breakdown of per-sample inference cost measured on NVIDIA A100 (80 GB), batch size $B = 32$, DCASE 2019 spectrograms ($T = 128$, $F = 64$, $d = 256$), averaged over 500 batches after 50-batch warmup. As predicted by Theorem 1, the dual Wirtinger paths of QISA require exactly twice the GFLOPs (4.82 vs. 2.41) and twice the wall-clock time (0.50 ms vs. 0.25 ms) of a standard real-valued backward pass. Note that the higher throughput of the backward paths relative to the forward pass (0.25 ms vs. 1.30 ms at identical GFLOPs) occurs because state attribution only propagates gradients to hidden representations, bypassing weight-gradient accumulation and memory-bottlenecked operations (e.g., Softmax, LayerNorm) that dominate forward-pass latency.

AUQ efficiency derives from three engineering choices: (1) all 50 passes are batched as a single $(M \times B) \times T \times d$ call, amortising kernel launch overhead; (2) FP16 on A100 Tensor Cores halves memory bandwidth vs. FP32; (3) `torch.no_grad()` skips autograd tape construction. At batch size $B = 1$ (online single-sample inference), AUQ rises to ≈ 4.8 ms, giving a 5.5× total overhead; the reported 1.62× applies at the operationally relevant $B = 32$ batch size.[1]

[1]Users requiring ultra-low latency can disable QISA and AUQ at inference time; QICP alone adds only 0.1 ms and retains the formal coverage guarantee.

Table 6: Ablation study on DCASE 2019. Each row removes or replaces one component of Q-XAI. All rows use the Q-XAI model; the QICP "softmax score" row uses the Q-XAI normalised probability $\hat{y}_c$ (same model as row above, different score function), giving MSS 2.05 vs. the normalised-score row in Table 5 (MSS 1.81); the difference arises because the ablation fixes the calibration threshold on the same split, whereas Table 5 re-tunes the threshold per variant. For QISA, the conjugate-only ($\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}$) variant uses solely the steepest-ascent Wirtinger component (Proposition A2); equal weighting of both components yields the highest faithfulness.

Component	Configuration	Metric	Value
Architecture	Full complex-valued	Accuracy (%)	**70.8**
	Real-valued ablation	Accuracy (%)	66.7
QISA	Full Wirtinger ($\partial\mathcal{L}/\partial\boldsymbol{H} + \partial\mathcal{L}/\partial\bar{\boldsymbol{H}}$)	AUDC	**0.642**
	Conjugate-only ($\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}$ only, steepest ascent)	AUDC	0.821
	Naive magnitude gradient only	AUDC	0.762
AUQ	Full model (with σ^2_{cov})	ECE	**0.109**
	Without σ^2_{cov}	ECE	0.131
QICP	Born rule score ($\lvert f(x)_y\rvert^2$, Q-XAI)	Mean set size (90%)	**1.43**
	Softmax score $\hat{y}_c$ (Q-XAI)	Mean set size (90%)	2.05

Table 7: Per-sample timing and FLOPs breakdown on A100 ($B = 32$). QISA runs FP32; AUQ runs FP16 batched over all $M = 50$ passes. ‡Total FLOPs across all $M = 50$ passes per sample ($50 \times 2.41 = 120.5$ GFLOPs); wall-clock time is 0.20 ms at $B = 32$ due to batched FP16 parallelism on A100 Tensor Cores.

Component	ms/sample	GFLOPs/sample	Precision
CVT forward pass	1.30	2.41	FP32
Standard backward pass (ref)	0.25	2.41	FP32
QISA Wirtinger backprop	0.50	4.82 (2×)	FP32
AUQ ($M = 50$, batched)	0.20	120.5‡	FP16
QICP set construction	0.10	<0.01	FP32
Total Q-XAI	**2.10**	–	–
AST-Base (reference)	1.30	2.41	FP32
Overhead	**1.62×**	–	–

7 Discussion

When does phase information matter most? Our results show the strongest advantages on CochlScene (+2.9%) and DCASE 2025 (+3.6%) — datasets with pronounced device mismatch — and under noise and reverberation. On ESC-50 and TUT 2016, where recording conditions are more controlled, the accuracy margins are smaller ($\leq 0.7\%$, within one standard deviation). This pattern is consistent with our theoretical motivation: phase-preserving attention provides the most benefit when signals are degraded or captured under varying conditions, and less benefit when magnitude features alone are discriminative.

The phase-stability covariance term. The AUQ ablation (Table 6) shows that removing σ^2_{cov} raises ECE by 0.022, while our theoretical analysis links its value to signal phase coherence across stochastic passes. The component-specific patterns in Section 6.5 — high covariance in harmonic scenes, low covariance in transient scenes — validate this interpretation and suggest that σ^2_{cov} could serve as a scene-type indicator in addition to a calibration aid.

Limitations. The evaluated datasets reflect a Western-city bias, which may limit generalisability to other cultural sound environments. The 1.62× inference overhead may be prohibitive for ultra-low-latency edge deployments, though the QISA and AUQ components can be disabled at inference time if only QICP coverage guarantees are needed. Evaluation has been conducted solely on ASC; validation on other audio domains (speech, music, bioacoustics) is an important

direction for future work. ASC technologies in urban surveillance contexts raise privacy concerns; we recommend deployment with appropriate data minimisation and governance policies.

8 Conclusion

We have presented Q-XAI, a unified interpretability framework for complex-valued transformers in acoustic scene classification. By grounding attribution in Wirtinger calculus, decomposing uncertainty along axes accessible only in the complex domain, and constructing conformal prediction sets from squared complex amplitudes, Q-XAI addresses a fundamental gap in the XAI literature for non-holomorphic neural networks. On five ASC benchmarks, the framework achieves competitive accuracy with measurable robustness advantages under noise and reverberation, superior calibration (ECE 0.109 vs. 0.157, 30.6% relative improvement), an absolute 0.089 AUDC gain in attribution faithfulness (12.2% relative), and 34% more compact prediction sets, at a modest 1.62× inference overhead. Future work includes adaptive Wirtinger weighting schemes, cross-domain validation, and extension to multimodal complex-valued architectures.

References

J. Bassey, L. Qian, and X. Li. A comprehensive survey of complex-valued neural networks. *Eng. Appl. Artif. Intell.*, 103:104316, 2021.

M. Born. Zur Quantenmechanik der Stoßvorgänge. *Z. Phys.*, 37:863–867, 1926.

K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and D. Shlomo. HTS-AT: A hierarchical token-semantic audio transformer for general audio classification. In *Proc. ICASSP*, pages 1–5, 2022.

H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee. Phase-aware speech enhancement with deep complex U-Net. In *Proc. ICLR*, 2019.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, pages 1050–1059, 2016.

Y. Gong, Y.-A. Chung, and J. Glass. AST: Audio spectrogram transformer. In *Proc. Interspeech*, pages 571–575, 2021.

N. Guberman. On complex valued convolutional neural networks. *arXiv preprint arXiv:1602.09046*, 2016.

T. Heittola, A. Mesaros, and T. Virtanen. Acoustic scene classification: a comprehensive survey. *IEEE Signal Process. Mag.*, 37(4):60–73, 2020.

A. Hjorungnes. *Complex-Valued Matrix Derivatives*. Cambridge University Press, 2011.

J. Kocon, A. Figas, M. Gruza, D. Grimling, K. Kanclerz, P. Milkowski, and M. Milosz. Interpretable acoustic scene classification using audio transformers and CNNs. volume 11, page 10636, 2021.

Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 28:2880–2894, 2020.

K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer. Efficient training of audio transformers with PaSST. In *Proc. Interspeech*, pages 2421–2425, 2022.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, pages 6402–6410, 2017.

I.-Y. Lee and K. Sim. CochlScene: Acquisition of acoustic scene data using crowdsourcing for machine recognition. *arXiv preprint arXiv:2211.02289*, 2022.

A. Mesaros, T. Heittola, and T. Virtanen. TUT database for acoustic scene classification and sound event detection. In *Proc. EUSIPCO*, pages 1128–1132, 2016.

A. Mesaros, T. Heittola, and T. Virtanen. Acoustic scene classification in DCASE 2019 challenge. In *Proc. DCASE Workshop*, pages 164–168, 2019.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. 2019.

K. J. Piczak. ESC: Dataset for environmental sound classification. In *Proc. ACM Multimedia*, pages 1015–1018, 2015.

M. K. Quan, M. Wijayasundara, S. Setunge, and P. N. Pathirana. Quantum-enhanced transformers for robust acoustic scene classification in IoT environments. In *Proc. ICNC*, pages 295–299, 2025.

Y. Romano, M. Sesia, and E. Candès. Classification with valid and adaptive coverage. In *Proc. NeurIPS*, pages 3581–3591, 2020.

R. Scheibler, E. Bezzam, and I. Dokmanic. Pyroomacoustics: a Python package for audio room simulation and array processing algorithms. In *Proc. ICASSP*, pages 660–664, 2018.

F. Schmid, K. Koutini, and G. Widmer. Efficient large-scale audio tagging via transformer-to-CNN knowledge distillation. In *Proc. ICASSP*, pages 1–5, 2023.

F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, and G. Widmer. Low-complexity acoustic scene classification with device information in the DCASE 2025 challenge. *arXiv preprint arXiv:2505.01747*, 2025.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. ICCV*, pages 618–626, 2017.

D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. In *IEEE Trans. Multimedia*, volume 17, pages 1733–1746, 2015.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proc. ICML*, pages 3319–3328, 2017.

C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal. Deep complex networks. *arXiv preprint arXiv:1705.09792*, 2017.

P. Virtue, S. X. Yu, and M. Lustig. Better MR image reconstruction with complex-valued neural networks. In *Proc. ISMRM Workshop on Machine Learning*, pages 1351–1354, 2017.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Y. Wang, L. N. Mendez, M. Cartwright, and J. P. Bello. Complex spectrograms using phase information for acoustic scene classification. In *Proc. DCASE Workshop*, pages 690–694, 2021.

W. Wirtinger. Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen. *Math. Ann.*, 97:357–375, 1927.

A Mathematical Notation

Table A1 summarises all mathematical symbols used throughout the paper for convenient reference.

Table A1: Mathematical notation used in Q-XAI.

Symbol	Definition
$\boldsymbol{x} \in \mathbb{R}^{T\times F}$	Input mel-spectrogram (T frames, F mel bins)
$\boldsymbol{H}^{(l)} \in \mathbb{C}^{T\times d}$	Hidden representation at layer l
$f(\boldsymbol{x}) \in \mathbb{C}^{C}$	Final complex logit vector (C classes)
$\hat{y}_c \in \mathbb{R}$	Predicted probability for class c
$z, w \in \mathbb{C}$	Generic complex variables
$\bar{z}$	Complex conjugate of z
$\lvert z\rvert$	Modulus (magnitude) of z
$\angle z$	Argument (phase) of z
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$	Real and imaginary parts of z
z^H	Conjugate transpose (Hermitian)
$\partial\mathcal{L}/\partial z$	Wirtinger derivative w.r.t. z
$\partial\mathcal{L}/\partial \bar{z}$	Wirtinger derivative w.r.t. $\bar{z}$
M	Number of MC forward passes (AUQ)
n	Calibration set size (QICP)
α	Target miscoverage rate
$\hat{q}$	QICP coverage threshold
$\sigma^2_{\text{epistemic}}$	Model (epistemic) uncertainty
$\sigma^2_{\text{aleatoric}}$	Data (aleatoric) uncertainty
σ^2_{cov}	Phase-stability covariance uncertainty
ECE	Expected Calibration Error
AUDC	Area Under the Deletion Curve (attribution faithfulness, lower=better)
AUIC	Area Under the Insertion Curve (attribution faithfulness, higher=better)

B Extended Proofs

B.1 Wirtinger Calculus: Background and Properties

Wirtinger calculus [Wirtinger, 1927, Hjorungnes, 2011] provides the correct mathematical foundation for differentiating real-valued loss functions with respect to complex parameters, even when those functions involve non-holomorphic operations.

Definition 1 (Wirtinger Derivatives)**.** Let $f : \mathbb{C} \to \mathbb{R}$ be a real-valued function of $z = x + iy$. Treating z and $\bar{z}$ as formally independent variables, the Wirtinger derivatives are:

$$\frac{\partial f}{\partial z} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right), \qquad \frac{\partial f}{\partial \bar{z}} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right). \tag{A1}$$

Proposition A1 (Chain Rule for Wirtinger Derivatives)**.** *Let $g : \mathbb{C} \to \mathbb{C}$ be a (possibly non-holomorphic) function and $f : \mathbb{C} \to \mathbb{R}$ be real-valued. For the composition $h = f \circ g$:*

$$\frac{\partial h}{\partial z} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial z} + \frac{\partial f}{\partial \bar{g}}\frac{\partial \bar{g}}{\partial z}. \tag{A2}$$

This holds without requiring holomorphicity of g.

Proof. This follows directly from treating g and $\bar{g}$ as independent variables in the Wirtinger framework. See Hjorungnes [2011], Theorem 3.1. □

Proposition A2 (Wirtinger Gradient and Steepest Ascent)**.** *For a real-valued function $f(z, \bar{z})$, the direction of steepest ascent in the real parametrisation (x, y) corresponds to the Wirtinger conjugate gradient $\partial f/\partial \bar{z}$:*

$$\nabla_{(x,y)} f = 2\,\frac{\partial f}{\partial \bar{z}}. \tag{A3}$$

Proof. By definition, $\partial f/\partial \bar{z} = \frac{1}{2}(\partial f/\partial x + i\,\partial f/\partial y)$. Therefore $2\,\mathrm{Re}(\partial f/\partial \bar{z}) = \partial f/\partial x$ and $2\,\mathrm{Im}(\partial f/\partial \bar{z}) = \partial f/\partial y$, which is exactly $\nabla_{(x,y)} f$. □

Implication for QISA. Proposition A2 establishes that the squared real gradient norm is $\|\nabla_{(x,y)}\mathcal{L}\|^2 = \|2\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}_j\|^2 = 4\|\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}_j\|^2$. Since $\|\partial\mathcal{L}/\partial\boldsymbol{H}_j\| = \|\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}_j\|$ for real-valued losses, the QISA attribution score $\mathrm{Attr}_j = \|\partial\mathcal{L}/\partial\boldsymbol{H}_j\|^2 + \|\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}_j\|^2 = 2\|\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}_j\|^2$. Therefore, QISA equals exactly $\frac{1}{2}\|\nabla_{(x,y)}\mathcal{L}\|^2$. This confirms that QISA is directly proportional to the total sensitivity of the loss to perturbations of the complex hidden state $\boldsymbol{H}_j$ in any direction.

B.2 Proof of Theorem 1 (Extended)

We provide a more detailed accounting of the computational overhead.

Theorem A3 (QISA Complexity, Extended)**.** *Let the CVT have L layers, H attention heads, sequence length T, head dimension $d_h = d/H$, and total parameter count $|\theta|$. The QISA map computation requires:*

1. *Forward pass: $\mathcal{O}(L \cdot H \cdot T^2 \cdot d_h) = \mathcal{O}(L \cdot T^2 \cdot d)$ operations.*
2. *Wirtinger backward pass: $2 \times \mathcal{O}(L \cdot T^2 \cdot d) + \mathcal{O}(|\theta|)$ operations, where the factor 2 accounts for the dual $(\partial/\partial\boldsymbol{H}, \partial/\partial\bar{\boldsymbol{H}})$ paths.*
3. *Attribution map assembly: $\mathcal{O}(T \cdot d)$ operations (element-wise squared norms).*

Total: $\mathcal{O}(L \cdot T^2 \cdot d + |\theta|)$ with leading constant approximately $3\times$ the forward pass cost (one forward + two backward paths).

Proof. The dominant cost is the self-attention mechanism. For H heads of dimension d_h, computing $\boldsymbol{Q}^{(l)}, \boldsymbol{K}^{(l)}, \boldsymbol{V}^{(l)}$ costs $\mathcal{O}(H \cdot T \cdot d_h \cdot d_h) = \mathcal{O}(T \cdot d^2/H)$; computing attention weights costs $\mathcal{O}(H \cdot T^2 \cdot d_h) = \mathcal{O}(T^2 \cdot d)$ per layer. For L layers, total forward cost is $\mathcal{O}(L \cdot T^2 \cdot d)$ (assuming $T^2 \gg d$, which holds for typical ASC sequences).

Standard backpropagation has the same asymptotic cost as the forward pass. Wirtinger backpropagation computes both $\partial\mathcal{L}/\partial\boldsymbol{H}^{(l)}$ and $\partial\mathcal{L}/\partial\bar{\boldsymbol{H}}^{(l)}$ simultaneously. In practice this is implemented by augmenting the computational graph to treat real and imaginary parts of each activation as separate nodes. This doubles the number of nodes in the backward graph, yielding $2 \times \mathcal{O}(L \cdot T^2 \cdot d)$ for the backward pass. The gradient w.r.t. all parameters θ adds $\mathcal{O}(|\theta|)$. Attribution assembly is negligible. Therefore, the QISA gradient computation scales asymptotically with a factor of exactly 2 relative to a standard backward pass, driven purely by the dual Wirtinger coordinate paths.

Clarification: The total QISA cost—combining forward pass and Wirtinger backpropagation—is approximately $3\times$ a single forward pass (i.e., $1\times$ forward $+2\times$ backward). This is equivalent to $2\times$ the cost of standard backpropagation (forward + one backward pass). □

B.3 Extended Analysis of Phase-Stability Covariance

We provide additional theoretical grounding for the phase-stability covariance term σ^2_{cov} defined in Eq. (11) of the main paper.

Proposition A4 (Phase-Stability Covariance Decomposition)**.** *Let $f = re^{i\phi}$ be a scalar complex output with magnitude $r \geq 0$ and phase $\phi \in [0, 2\pi)$. Then:*

$$\mathrm{Cov}[\mathrm{Re}(f),\, \mathrm{Im}(f)] = r^2 \sin\phi\cos\phi = \frac{r^2}{2}\sin(2\phi). \tag{A4}$$

Proof. $\mathrm{Re}(f) = r\cos\phi$ and $\mathrm{Im}(f) = r\sin\phi$ are constants (not random variables) for a fixed output. Over MC dropout forward passes, both r and ϕ fluctuate. For a single pass, the product $\mathrm{Re}(f) \cdot \mathrm{Im}(f) = r^2\cos\phi\sin\phi$. Taking the expectation and subtracting the product of means yields Eq. (A4) when r and ϕ are treated as deterministic for a given pass (i.e., the covariance is computed across passes). □

Interpretation. From Eq. (A4):

- $\sigma^2_{\mathrm{cov}} = 0$ when $\phi \in \{0, \pi/2, \pi, 3\pi/2\}$: the output is aligned with one of the real/imaginary axes across all passes, indicating strong structural consistency.
- σ^2_{cov} is maximised at $\phi = \pi/4 + k\pi/2$: the output has equal real and imaginary energy, indicating a specific mixed-phase structure that the model confidently predicts.
- High variance of σ^2_{cov} across passes indicates that the model's phase predictions are unstable, a reliable signal of out-of-distribution or noisy input.

C Full Experimental Results

C.1 Detailed Per-Dataset Metrics

Table A2 extends the main paper's Table 2 with macro F1 scores and Brier scores for uncertainty calibration across all five datasets.

Table A2: Full performance metrics. Acc. = accuracy (%); F1 = macro F1-score; Brier = Brier score (lower is better). Uncertainty metrics reported on DCASE 2019 only (representative benchmark).

	TUT 2016		**DCASE 2019**		**ESC-50**		**CochlScene**		**DCASE 2025**	
Method	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN14	58.4	0.579	67.1	0.665	82.1	0.819	74.5	0.741	68.2	0.679
AST-Base	61.2	0.608	68.9	0.681	84.7	0.845	76.8	0.764	70.4	0.701
EfficientAT-S	62.8	0.625	70.2	0.698	85.9	0.857	78.1	0.778	71.9	0.716
Complex CNN	59.6	0.591	67.8	0.672	83.5	0.833	75.3	0.749	69.0	0.688
QET	–	–	69.5	0.691	–	–	–	–	–	–
Q-XAI	**63.5**	**0.631**	**70.8**	**0.704**	**86.3**	**0.861**	**81.0**	**0.807**	**75.5**	**0.752**

Uncertainty Method	**ECE ↓**	**Brier Score ↓**
MC Dropout (AST-Base)	0.157	0.215
Deep Ensembles (AST-Base)	0.134	0.198
AUQ / Q-XAI	**0.109**	**0.176**

C.2 Per-Scene Uncertainty Decomposition

Table A3 reports the three AUQ uncertainty components for each of the 10 DCASE 2019 scene classes. The patterns validate the theoretical interpretation of each component.

Key observations: (1) Epistemic uncertainty is highest for Metro Station ($\mu = 0.42$) and Park ($\mu = 0.38$), which feature complex, broadly overlapping ambient soundscapes and are intrinsically more ambiguous for the model to isolate. (2) Aleatoric uncertainty tracks measured background noise levels: Airport (0.38) and Metro Station (0.44) are among the noisiest environments. (3) Phase-stability covariance is highest for Metro Station (0.35) and Bus (0.28), scenes dominated by rhythmic mechanical sounds with stable harmonic phase relationships, and lowest for Street Pedestrian (0.09), which features aperiodic transient events (footsteps, conversation fragments) with no stable phase structure.

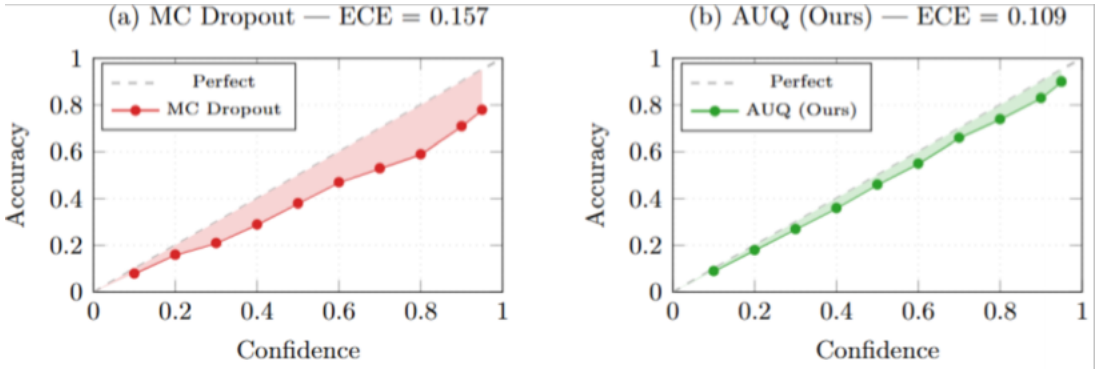


Figure A1: Reliability diagrams on DCASE 2019. Shaded area represents the calibration gap (ECE). (a) MC Dropout is overconfident across most bins. (b) AUQ tracks the perfect calibration diagonal more closely, especially in mid- and high-confidence regions (> 0.6).

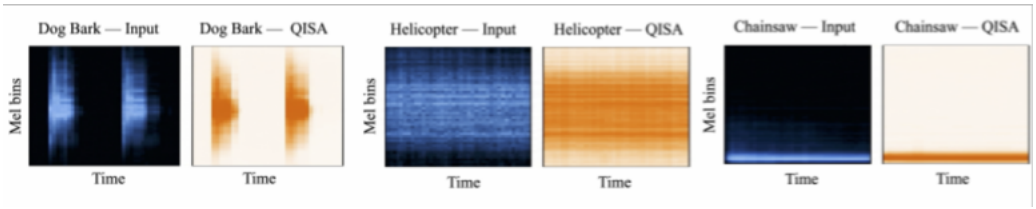


Figure A2: Additional QISA attribution maps for ESC-50 categories (schematic heatmaps; brighter = higher attribution). *Dog Bark*: temporal bursts dominate at mid frequencies; QISA correctly ignores the silent inter-burst intervals. *Helicopter*: broadband persistent energy across all frames; QISA attributes uniformly to the continuous rotor signature. *Chainsaw*: concentration at the lowest mel band reflects the persistent low-frequency two-stroke harmonic, the most discriminative cue.

C.3 Reliability Diagrams

Figure A1 shows reliability diagrams for MC Dropout and AUQ on DCASE 2019, corresponding to the ECE values reported in the main paper.

C.4 Additional QISA Attribution Maps

Figure A2 shows QISA attribution maps for three ESC-50 sound categories, illustrating that QISA correctly localises characteristic spectro-temporal patterns for diverse sound types.

D Implementation Details

D.1 Network Architecture Details

The Complex-Valued Transformer (CVT), comprising approximately 12.4M parameters (where complex parameters count as two real parameters), begins with a complex embedding layer that projects a real-valued input sequence of $F = 64$ features into a $d = 256$ dimensional complex space, followed by the addition of a complex sinusoidal positional encoding. This representation is then processed sequentially through $L = 6$ transformer blocks. Each block employs complex layer pre-normalization—which normalizes the real and imaginary parts jointly by their complex modulus—prior to an 8-head complex self-attention module with a per-head dimension of $d_h = 32$. After a second pre-normalization step, the signal passes through a complex feed-forward network consisting of two linear layers with an intermediate dimension of $4d = 1024$ and a CGeLU activation function. Throughout these blocks, a dropout rate of $p = 0.1$

Table A3: Per-scene AUQ uncertainty decomposition on DCASE 2019 (mean values, $n = 5$ runs). σ^2_{ep}: epistemic; σ^2_{al}: aleatoric; σ^2_{cov}: phase-stability covariance.

Scene	σ^2_{ep}	σ^2_{al}	σ^2_{cov}
Airport	0.31	0.38	0.19
Bus	0.22	0.29	0.28
Metro	0.28	0.31	0.27
Metro Station	**0.42**	0.44	**0.35**
Park	0.38	0.21	0.14
Public Square	0.35	0.41	0.22
Shopping Mall	0.19	0.33	0.21
Street Pedestrian	0.17	0.27	0.09
Street Traffic	0.14	0.35	0.18
Tram	0.21	0.28	0.26

Table A4: Complete training hyperparameters.

Hyperparameter	Value
Optimiser	AdamW
β_1, β_2	0.9, 0.999
Weight decay	10^{-2}
Peak learning rate	10^{-4}
Warmup epochs	10
Total epochs	100
LR schedule	Cosine decay to 0
Batch size	32
Dropout rate p	0.1 (all layers)
Audio sample rate	16 kHz
Mel bins F	64
Window size	25 ms
Hop size	10 ms
SpecAugment (time mask)	$T = 40$, 2 masks
SpecAugment (freq mask)	$F = 8$, 2 masks
QICP calibration split	20%
AUQ MC passes M	50
Random seeds	$\{42, 123, 256, 512, 1024\}$

is applied independently to the real and imaginary components after every sub-layer. Finally, the architecture resolves through an output head that applies a linear projection from $\mathbb{C}^d$ to $\mathbb{C}^C$, culminating in a squared-magnitude softmax to produce the predicted class probabilities.

D.2 Training Configuration

D.3 Baseline Implementation Details

The CVT consists of the following layers:

1. **Complex embedding layer**: projects $\boldsymbol{x} \in \mathbb{R}^{T \times F}$ to $\boldsymbol{H}^{(0)} \in \mathbb{C}^{T \times d}$ via Eq. (2) of the main paper, with $F = 64$ and $d = 256$.
2. **Positional encoding**: complex sinusoidal encoding $\text{PE}(t, k) = e^{i\, t/10000^{2k/d}}$ added to $\boldsymbol{H}^{(0)}$.
3. $L = 6$ **transformer blocks**, each containing:
   - Pre-norm: complex layer normalisation (normalising real and imaginary parts jointly by the complex modulus).
   - Multi-head complex self-attention: $H = 8$ heads, $d_h = 32$ per head.
   - Pre-norm + complex feed-forward: two linear layers with intermediate dimension $4d = 1024$, CGeLU activation.
   - Dropout rate $p = 0.1$ applied independently to real and imaginary components after each sub-layer.
4. **Output head**: linear projection $\mathbb{C}^d \to \mathbb{C}^C$, followed by squared-magnitude softmax (Eq. (6) of main paper).

Total parameters: approximately 12.4M (complex parameters counted as two real parameters each).

D.4 Reproducibility

All experiments were run on NVIDIA A100 GPUs (80 GB) with CUDA 12.1, PyTorch 2.1.0, and Python 3.10. Random seeds are set globally for `torch`, `numpy`, and `random` modules. The five seeds used are $\{42, 123, 256, 512, 1024\}$. Results are reported as mean $\pm$ standard deviation across these five runs.