Cross-Attentive Bayesian Low-Rank Adaptation for Multimodal Uncertainty Estimation

SPECTRA: Stochastic Posterior Estimation with Cross-modal Token-level Rank Adaptation

Habibeh Naderi[1] **Behrouz Haji Soleimani**[1,2] **Stan Matwin**[1]

[1] Faculty of Computer Science, Dalhousie University, Halifax, Canada [2] Kinaxis, Ottawa, Canada

NeurIPS 2026 · Main Track

Why multimodal uncertainty for PEFT?

Human-centred, low-resource settings (e.g., clinical interviews): scarce labels, costly confident mistakes

PEFT (e.g., LoRA)

✓ Cheap adaptation of a frozen LLM

✗ Deterministic: **overconfident** under noise, shift and ambiguity

Bayesian LoRA

BLoB, C-LoRA

✓ Uncertainty in the adapter space

✗ Conditioned **only on internal text features**

Why multimodal uncertainty for PEFT?

Human-centred, low-resource settings (e.g., clinical interviews): scarce labels, costly confident mistakes

PEFT (e.g., LoRA)

✓ Cheap adaptation of a frozen LLM

✗ Deterministic: **overconfident** under noise, shift and ambiguity

Bayesian LoRA

BLoB, C-LoRA

✓ Uncertainty in the adapter space

✗ Conditioned **only on internal text features**

Speech: reliability is acoustic



Noise, channel, prosody, speaker state: **invisible in the transcript**

Naive audio–text feature fusion → higher dimensionality, over-parameterization in low-resource tasks, and harder calibration when the modalities disagree.

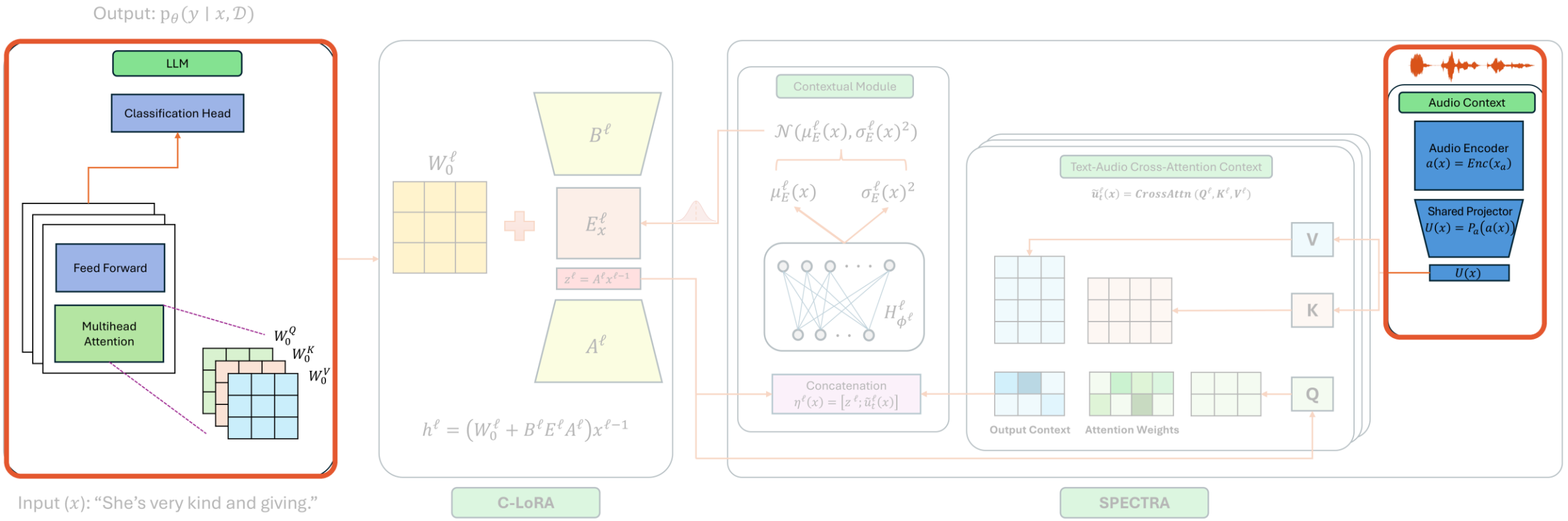
Key idea: audio as a reliability signal

Let token-level audio context decide how much the adapter should trust itself: it sets the **mean and variance** of a Bayesian low-rank adapter posterior.

SPECTRA: Stochastic Posterior Estimation with Cross-modal Token-level Rank Adaptation

Method	Bayesian	Input-dependent posterior	Audio context	Token-level	Frozen backbones
LoRA	✗	✗	✗	✗	✓
BLoB	✓	✗	✗	✗	✓
C-LoRA	✓	✓	✗	✗	✓
Transfer (feature fusion)	✗	✗	pooled	✗	✓
SPECTRA-G (ours)	✓	✓	global	✗	✓
SPECTRA-X (ours)	✓	✓	token→frame	✓	✓

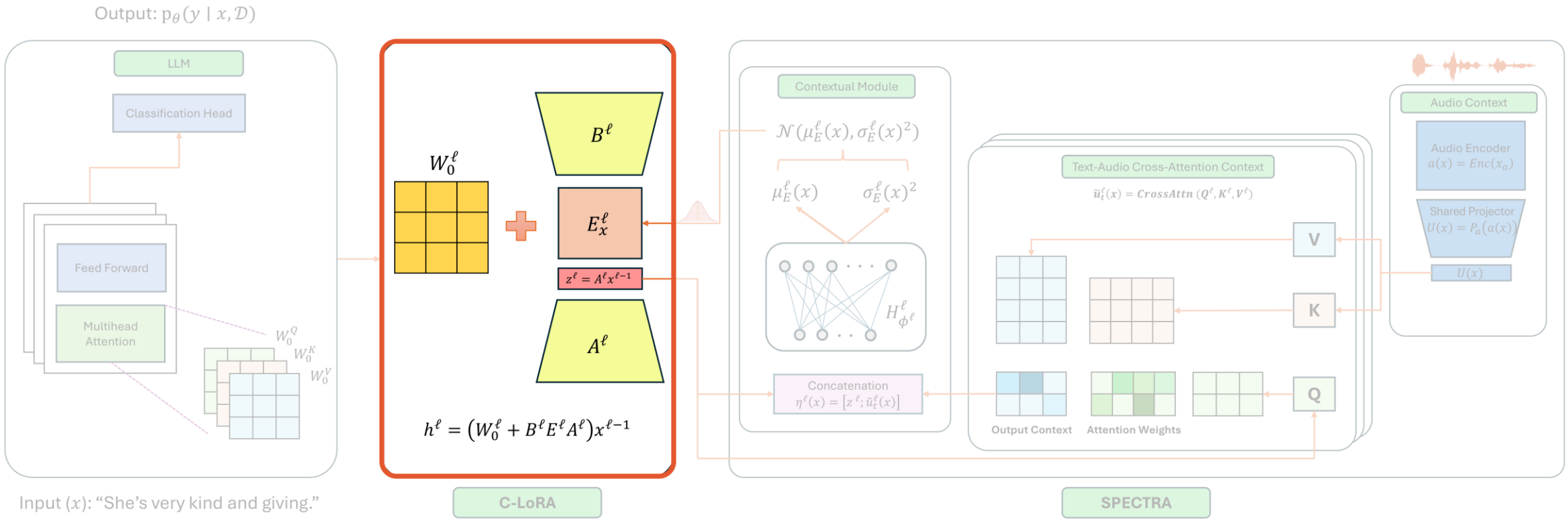
SPECTRA: cross-attentive Bayesian low-rank adaptation



① Frozen backbones

A frozen **text LLM** yields token hidden states; a frozen **audio encoder** yields frame-level embeddings $a_s(x)$, projected to a small shared context space.

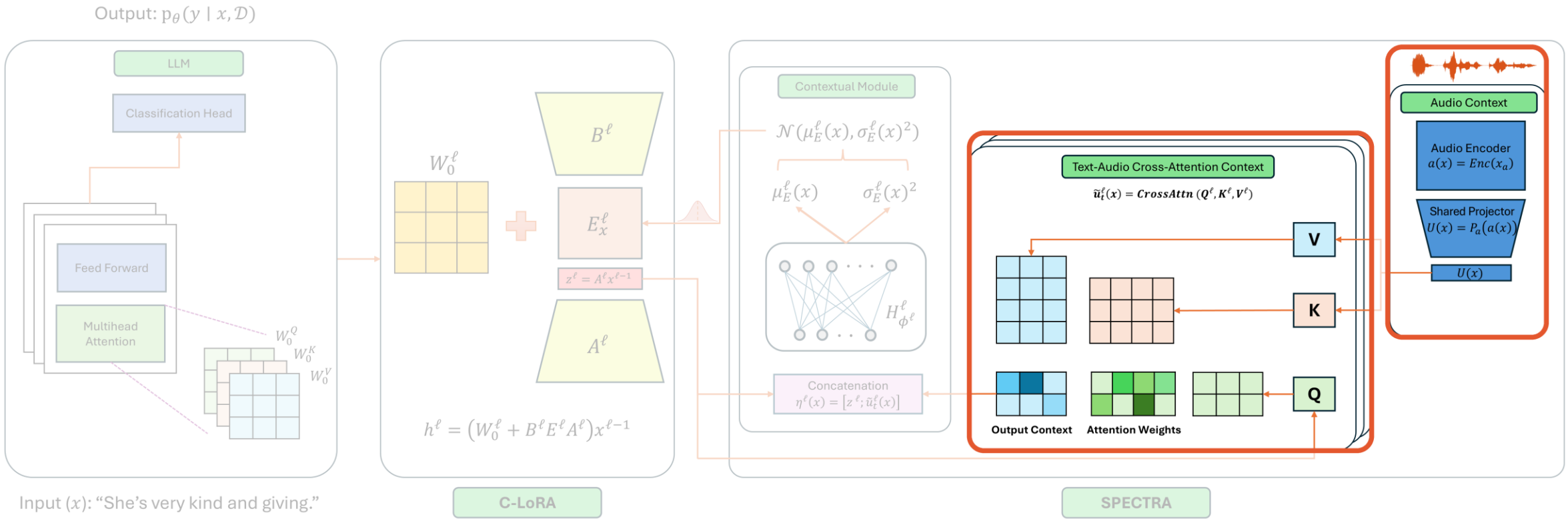
SPECTRA: cross-attentive Bayesian low-rank adaptation



② **Stochastic low-rank adapter: A, B deterministic, only an r × r latent E is random**

$$h_t^\ell = W_0^\ell x_t^{\ell-1} + B^\ell E_{x,t}^\ell A^\ell x_t^{\ell-1}, \qquad E_{x,t}^\ell \in \mathbb{R}^{r \times r}$$

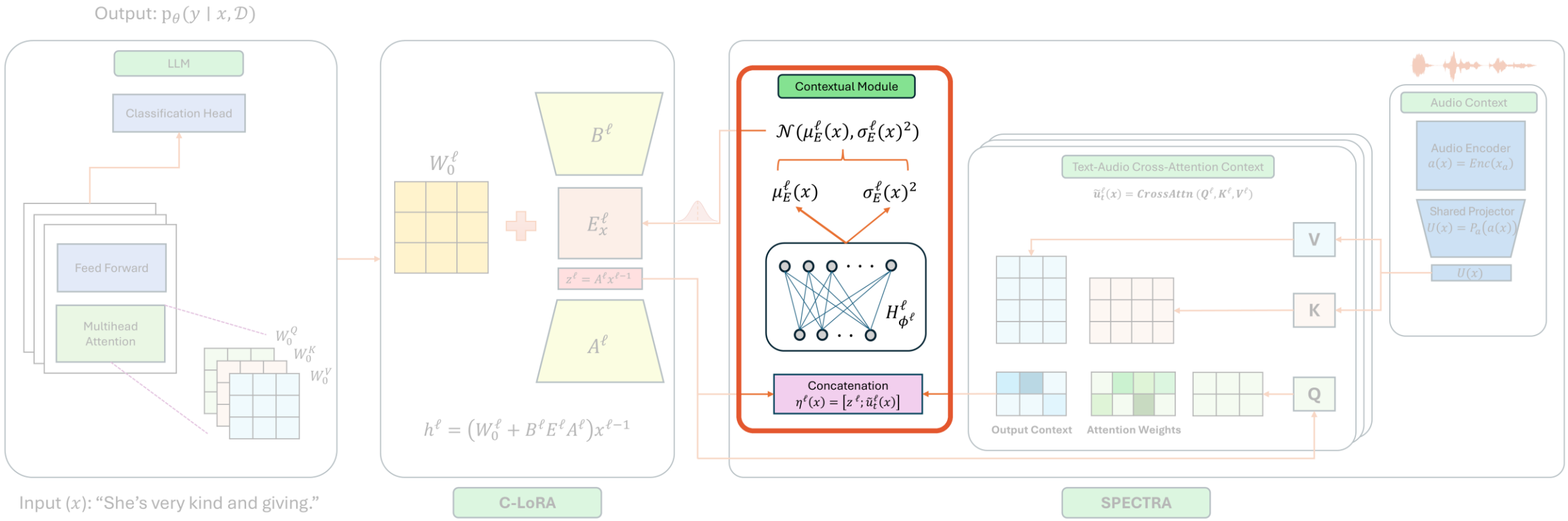
SPECTRA: cross-attentive Bayesian low-rank adaptation



③ Per-layer, token-level text → audio cross-attention

$$\tilde{u}_t^\ell(x) \;=\; W_O^\ell \,\mathrm{softmax}\!\left(\frac{q_t^\ell\, K^\ell(x)^\top}{\sqrt{d_c}}\right) V^\ell(x), \qquad q_t^\ell = W_Q^\ell z_t^\ell \;(\text{text}), \quad K^\ell, V^\ell \text{ from audio frames}$$

SPECTRA: cross-attentive Bayesian low-rank adaptation



④ Context-conditioned variational posterior → heteroscedastic uncertainty

$$q_{\phi^\ell}\left(\mathrm{vec}(E_{x,t}^\ell) \mid x\right) = \mathcal{N}\left(\mu_{E,t}^\ell(\eta_t^\ell),\ \mathrm{diag}\,\sigma_{E,t}^\ell(\eta_t^\ell)^2\right), \qquad \eta_t^\ell = [\, z_t^\ell\ ;\ \tilde{u}_t^\ell(x)\,]$$

Posterior, training objective and uncertainty

Posterior heads

$$(\mu^{\ell}_{E,t},\ \log v^{\ell}_{E,t}) \;=\; H_{\phi^{\ell}}\big([\, z^{\ell}_{t}\,;\, \tilde{u}^{\ell}_{t}(x)\,]\big), \qquad \sigma^{\ell}_{E,t} = \varepsilon\,\mathrm{softplus}(\log v^{\ell}_{E,t}) + \delta$$

same words → different uncertainty, depending on how they sound

Training (ELBO)

$$\mathcal{L}(\theta,\phi) \;=\; \sum_{i}\Big[\,\mathbb{E}_{q_\phi}\log p_\theta(y_i \mid x_i, E_{x_i}) \;-\; \gamma \sum_{\ell}\frac{1}{T_{x_i}}\sum_{t}\mathrm{KL}\big(q_{\phi^{\ell}}(E^{\ell}_{x_i,t} \mid x_i)\,\big\|\,\mathcal{N}(0,\beta^2 I)\big)\Big]$$

token-averaged KL · prior $\mathcal{N}(0, \beta^2 I)$ · one MC draw per example

Prediction

$$p(y^* \mid x^*, \mathcal{D}) \;\approx\; \frac{1}{M}\sum_{m=1}^{M} p_\theta\big(y^* \mid x^*, E^{(m)}_{x^*}\big), \qquad E^{(m)}_{x^*} \sim q_\phi(E \mid x^*)$$

Monte Carlo averaging over M = 10 adapter samples

Uncertainty

$$U_{\text{total}} = H[\bar{\mathbf{p}}], \qquad U_{\text{alea}} = \frac{1}{M}\sum_{m} H[\mathbf{p}^{(m)}], \qquad U_{\text{epi}} = U_{\text{total}} - U_{\text{alea}}$$

epistemic part = adapter-space uncertainty · all scores normalized by log C

Why it scales: only $r^2 = 64$ stochastic latents per token and layer (r = 8): $O(r^2)$ Bayesian overhead, independent of the hidden size d.

Experimental setup

IEMOCAP · emotion recognition

4 classes: neutral, happy (+ excited), sad, angry

5 sessions, 10 speakers; standard 5-fold splits

FORBOW · clinical interviews

369 parents (MDD, bipolar, schizophrenia, controls) and their children; ≈ 17 words / 6.5 s per segment

12 parent + 17 offspring segment-level tasks, **up to 99% class imbalance**

Frozen encoders

6 text: nliRoBERTa, nliDistilRoBERTa, allRoBERTaLarge, allDistilRoBERTa, allMiniLM12, paraTinyBERT

3 audio: whisperMedium, wav2vec2LargeFT, hubertLargeFT

Baselines

Text-only PEFT: LoRA, C-LoRA

Transfer: pooled audio ⊕ text → MLP head

Ours

SPECTRA-G: global audio context

SPECTRA-X: token→frame cross-attention (+ Shared-KV variant)

r = 8, α = 32 · AdamW, 50 epochs · β = 0.2, ε = 0.05, γ = 0.008 · M = 10 MC samples · speaker-level 5-fold CV · metric: AUC

Results on IEMOCAP (AUC %)

93.0 vs 88.4

audio-only vs. best text-only

9 / 12

pairs: best SPECTRA > Transfer

11 / 12

pairs: SPECTRA-X > SPECTRA-G

94.68

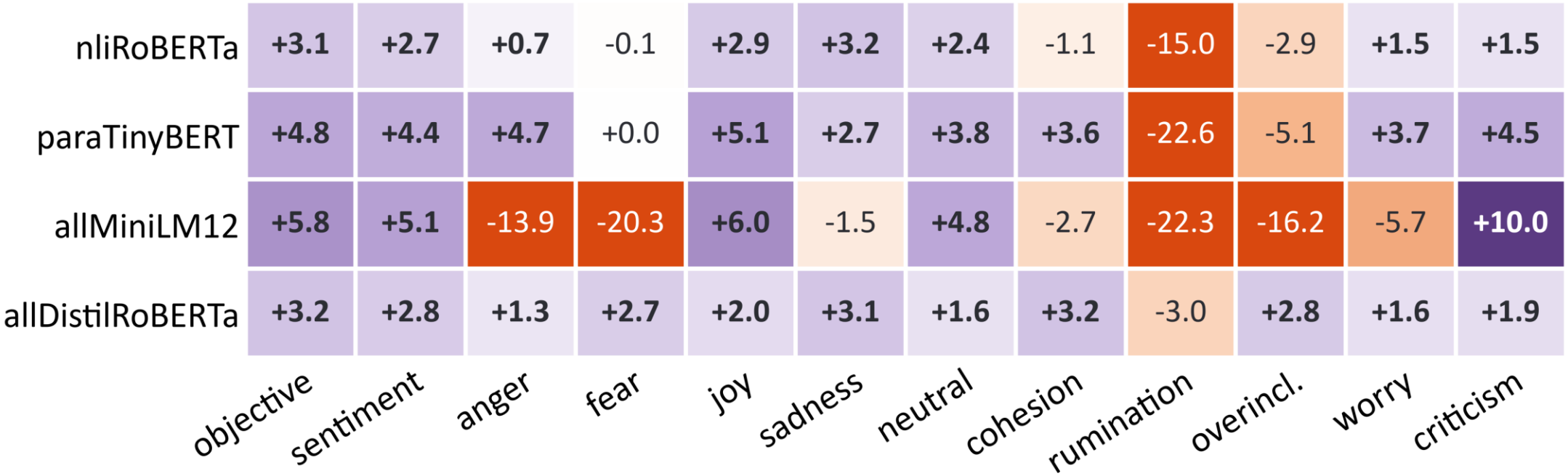
best multimodal AUC (SPECTRA-X)

Text encoder	Audio	Transfer	SPECTRA-G	X: Shared-KV	X: Cross-Attn
nliRoBERTa	whisperM	93.40	93.27	94.01	93.05
	hubertLFT	89.93	90.40	90.20	91.06
nliDistilRoBERTa	whisperM	93.39	92.91	93.15	93.99
	hubertLFT	**89.76**	89.21	89.46	89.51
allRoBERTaLarge	whisperM	**94.64**	92.48	92.59	94.05
	hubertLFT	88.06	89.21	88.49	88.81
allDistilRoBERTa	whisperM	94.48	92.56	93.14	94.68
	hubertLFT	90.28	90.00	89.62	91.09
allMiniLM12	whisperM	**93.32**	92.93	92.84	92.99
	hubertLFT	89.75	89.37	89.78	89.62
paraTinyBERT	whisperM	92.51	91.62	92.27	93.36
	hubertLFT	89.13	88.32	89.62	88.97

Mean over 5 folds (std. in the paper); best per row in bold.

Results on FORBOW clinical interviews (parent data)

ΔAUC = best SPECTRA variant − best baseline (LoRA / C-LoRA / Transfer), per text encoder and task



33 / 48

encoder–task pairs where SPECTRA beats all baselines

+10.0

max gain over the best baseline

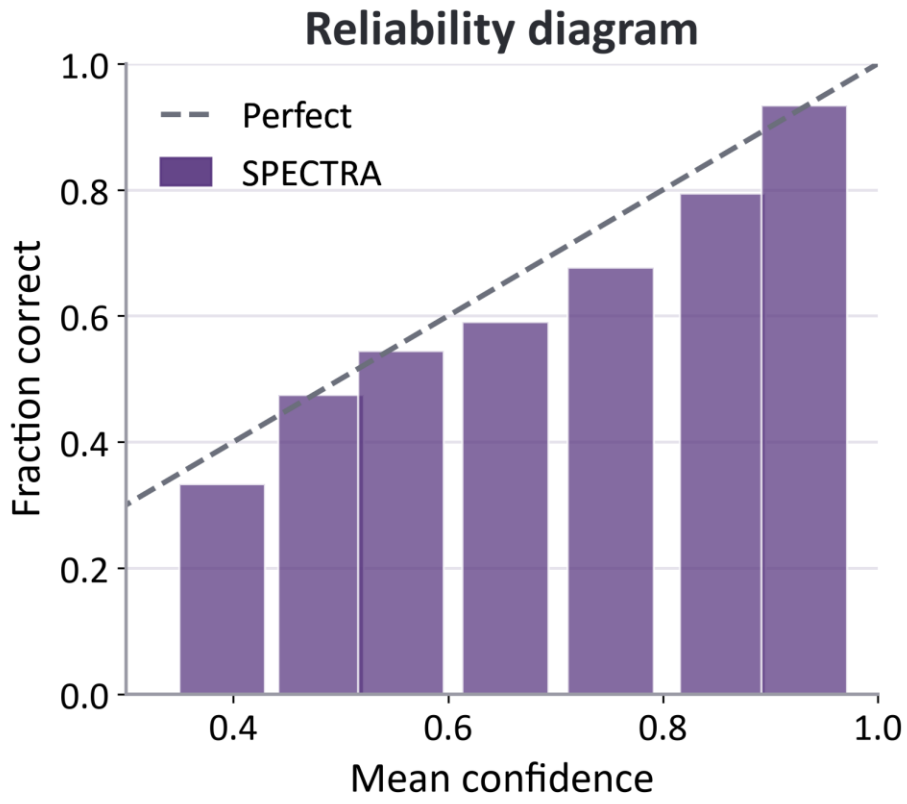
+25.0

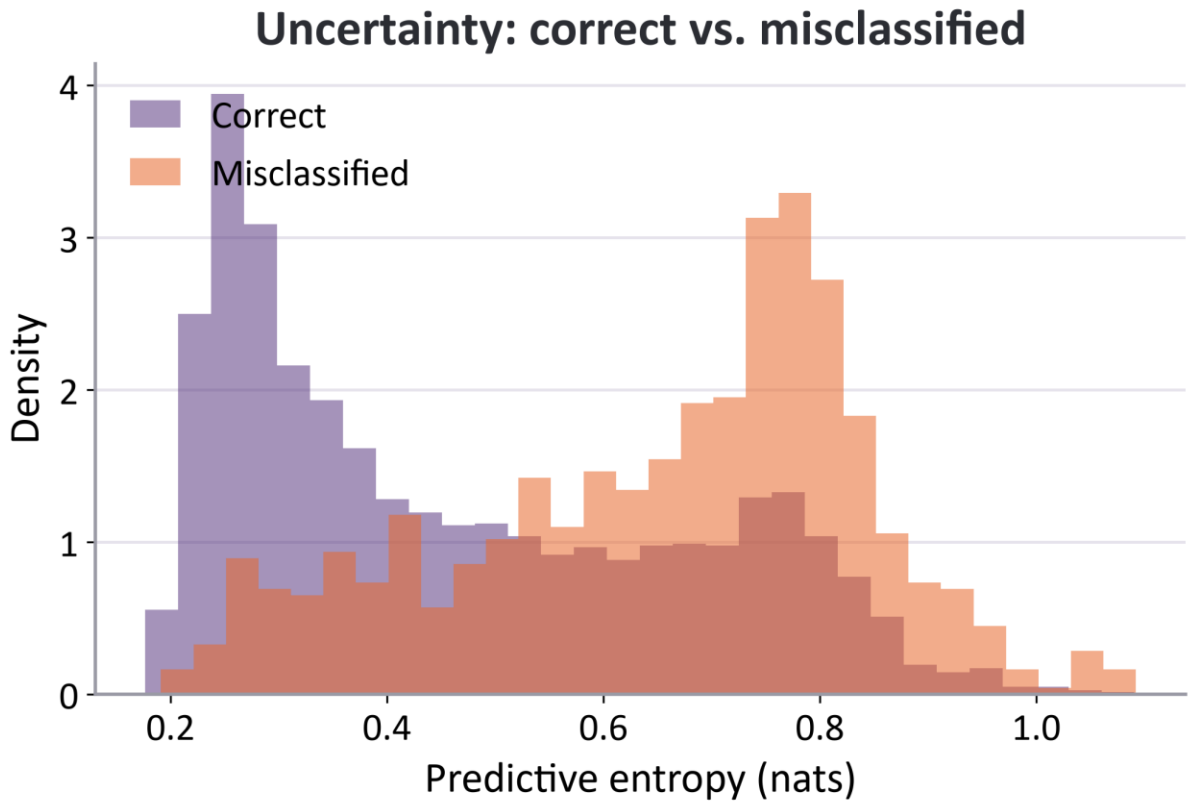
max gain over text-only PEFT (allMiniLM12)

Losses: ultra-rare labels (rumination ≈ 1 % positives) and weak encoders.

Children's data: more mixed (appendix).

Calibrated and informative uncertainty





Confidence closely tracks accuracy (near-diagonal reliability).

Errors carry higher predictive entropy → uncertainty **flags likely mistakes**.

Modality-disagreement stress test

Same trained model · same held-out FORBOW examples (N = 5,596) · labels fixed

Condition	AUC ↑	NLL ↓	ECE ↓	Brier ↓
Clean text + matched audio	**0.835**	0.419	0.042	0.260
Clean text + noisy audio	0.812	0.435	0.021	0.278
Clean text + mismatched audio	0.716	0.560	0.087	0.346
Corrupted text + clean audio	0.795	0.458	0.049	0.290

Noisy audio

Small AUC drop; **lowest ECE**: the model turns cautious

Mismatched audio

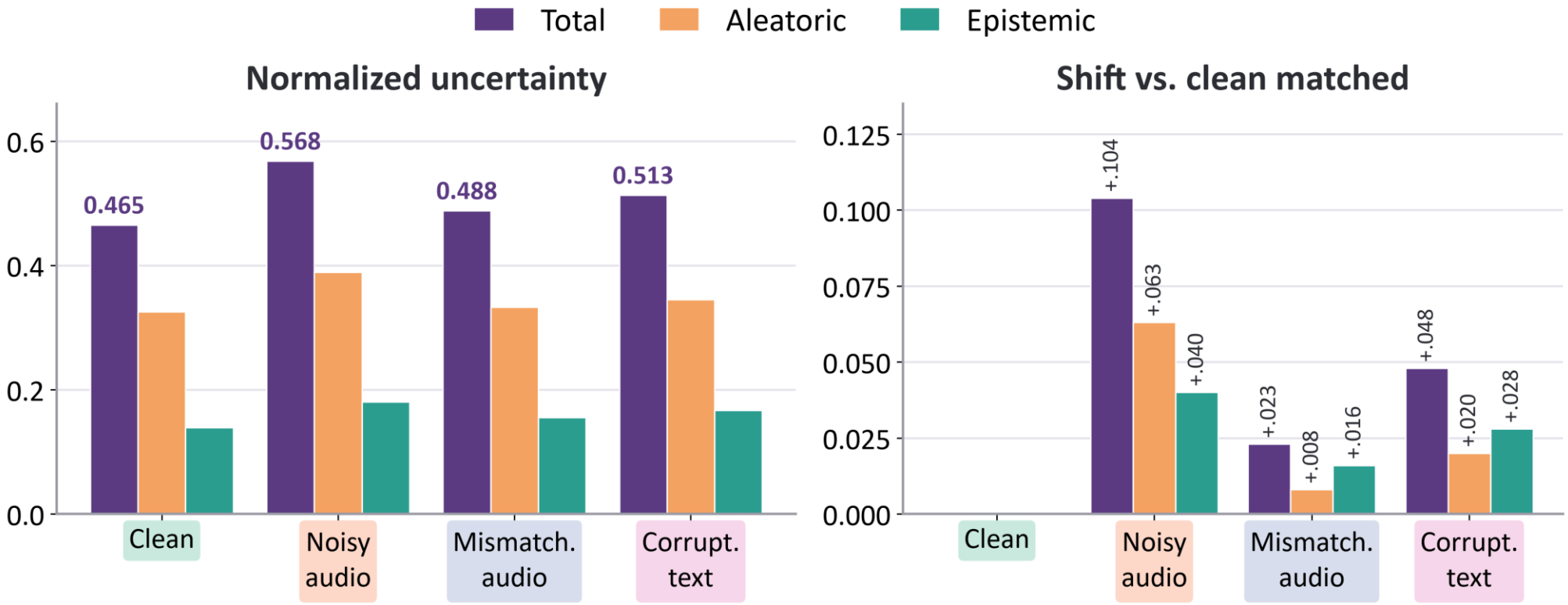
Hurts most: AUC 0.716, ECE doubles, NLL and Brier rise

Corrupted text

Clean audio partially compensates (AUC 0.795)

Where does the uncertainty go?

Normalized total = aleatoric + adapter-space epistemic uncertainty, per stress-test condition



Noisy audio: total +0.104 (aleatoric +0.063, epistemic +0.040)

Mismatched audio: AUC −0.12 but uncertainty only +0.023 → residual overconfidence

Take-home messages

1. **Audio as a reliability signal:** token-level acoustic context sets the mean and variance of a tiny r × r Bayesian adapter latent.

2. **Token-level cross-attention** gives the most reliable gains, at PEFT cost (frozen backbones, $O(r^2)$ stochastic overhead).

3. **Modality-agnostic core:** video, image patches or physiological signals can replace audio frames as keys / values.

→ **Next:** make the posterior variance itself reliability-aware, to close the mismatched-audio overconfidence gap.

Paper + code

openreview.net/forum?id=szn6IUxvw6

Contact

habibeh.naderi@dal.ca

Thank you!