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

Habibeh Naderi
Faculty of Computer Science
Dalhousie University
habibeh.naderi@dal.ca

Behrouz Haji Soleimani
Kinaxis
Ottawa, ON
bhajisoleimani@kinaxis.com

Stan Matwin
Faculty of Computer Science
Dalhousie University
stan@cs.dal.ca

Abstract

Parameter-efficient fine-tuning (PEFT) enables efficient adaptation of frozen language models, but most PEFT methods remain deterministic or unimodal, limiting their reliability in low-resource audio-text settings where uncertainty depends on both linguistic evidence and acoustic conditions. We introduce SPECTRA (Stochastic Posterior Estimation with Cross-modal Token-level Rank Adaptation), a multimodal Bayesian low-rank adaptation framework for uncertainty-aware audio-text learning. SPECTRA keeps the text and audio backbones frozen and confines stochasticity to a compact rank-r latent matrix inside each LoRA adapter. At each transformer layer, text-derived low-rank token features query frame-level audio embeddings through lightweight cross-attention; the resulting token-specific acoustic context parameterizes the mean and variance of an amortized variational posterior over the adapter latent. This design treats audio not merely as an additional feature stream, but as a localized reliability signal that modulates both adaptation and confidence while preserving the scalability of PEFT. Posterior prediction is performed with Monte Carlo adapter samples, enabling a Bayesian uncertainty analysis that decomposes normalized predictive uncertainty into total, aleatoric, and adapter-space epistemic components and evaluates whether uncertainty distinguishes correct from misclassified predictions. Across IEMOCAP and clinical interview prediction tasks, SPECTRA is consistently competitive with or improves upon text-only Bayesian PEFT and conventional multimodal transfer-learning baselines, with token-level cross-attention yielding the most reliable gains. Modality-disagreement stress tests further show that mismatched audio causes substantially larger degradation in AUC, likelihood, calibration, and Brier score far more than noisy but matched audio, highlighting the importance of modeling cross-modal reliability. These results suggest that Bayesian cross-modal conditioning in low-rank adapter space provides an efficient and principled mechanism for calibrated multimodal adaptation.

1 Introduction

Pre-trained Large Language Models (LLMs) have become the dominant backbone for a wide range of prediction tasks Brown et al. [2020], Minaee et al. [2024]. However, adapting these large models to specialized domains remains challenging when labeled data are scarce and predictions must be reliable. This challenge is particularly pronounced in human-centered applications such as behavioral and clinical analysis, where downstream targets may include affect, cognitive state, and interaction

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

quality Naderi et al. [2026b,a, 2025]. In such settings, models must simultaneously satisfy three requirements: (i) strong predictive performance under limited supervision, (ii) computationally efficient adaptation of large pretrained backbones, and (iii) reliable uncertainty estimates that reflect when model predictions should be trusted.

Parameter-efficient fine-tuning (PEFT) methods Fu et al. [2023] address the first two requirements by adapting large models using small trainable modules while keeping the backbone weights frozen. Among these methods, Low-Rank Adaptation (LoRA) and related approaches introduce low-dimensional parameter updates that allow efficient adaptation with a small number of additional parameters Hu et al. [2022], Zhang et al. [2023], Liu et al. [2022], Shi et al. [2025]. PEFT techniques significantly reduce memory and training costs while often maintaining performance comparable to full fine-tuning. However, most PEFT approaches are deterministic and optimized primarily for predictive accuracy, which can lead to overconfident predictions in the presence of noisy inputs, domain shifts, or ambiguous examples Leng et al. [2025], Xiong et al. [2024]. This limitation motivates uncertainty-aware PEFT methods that incorporate Bayesian principles to produce calibrated predictions Wang et al. [2024], Rahmati et al. [2025], Yang et al. [2024], Pandey et al. [2024], Samplawski et al. [2025].

Bayesian variants of LoRA introduce uncertainty by placing probabilistic distributions over adapter parameters while keeping the backbone deterministic. Bayesian LoRA by Backpropagation (BLoB) Wang et al. [2024] models epistemic uncertainty in LoRA weights using variational inference. More recently, Contextual LoRA (C-LoRA) Rahmati et al. [2025] introduced input-dependent Bayesian adapters by conditioning the variational posterior on intermediate low-rank features. This contextualization produces heteroscedastic uncertainty, allowing the model to express varying confidence across different inputs while maintaining the computational efficiency of PEFT. Despite these advances, existing uncertainty-aware PEFT methods are largely unimodal and primarily condition uncertainty on internal language-model representations.

A second challenge arises when the predictive signal is inherently multimodal. Many real-world prediction tasks involve both lexical content and acoustic information. In speech-centered applications, text often provides strong semantic cues while audio carries complementary signals related to prosody, speaker state, and environmental conditions Borsos et al. [2023], Thimonier et al. [2025]. Multimodal transfer learning studies consistently show that these modalities can provide complementary evidence, especially when one modality becomes unreliable. However, incorporating multimodal information introduces additional modeling complexity. Naive feature fusion strategies may increase representational dimensionality, amplify over-parameterization in low-resource settings, and complicate calibration when modalities disagree. These issues become particularly critical when the reliability of one modality varies across time, such as when audio segments contain noise, silence, or variable recording conditions.

In many speech-based tasks, uncertainty is driven not only by the linguistic content but also by external acoustic conditions. For example, background noise, channel variability, or speaker-specific factors may reduce the reliability of audio segments that correspond to otherwise informative text tokens. Existing uncertainty-aware PEFT approaches do not explicitly incorporate such cross-modal reliability signals. Consequently, they may fail to adjust predictive confidence when acoustic evidence contradicts or weakens the textual signal.

In this work, we introduce Stochastic Posterior Estimation with Cross-modal Token-level Rank Adaptation (SPECTRA), a multimodal uncertainty-aware PEFT framework that integrates token-level cross-modal context into Bayesian low-rank adaptation (Figure 1). SPECTRA extends contextual Bayesian adapters by conditioning the variational posterior over the low-rank latent space on cross-attention between text tokens and frame-level audio representations. Each transformer layer forms a token-specific view of the audio sequence through lightweight cross-attention modules. The resulting cross-modal context acts as a reliability signal that modulates both the mean and variance of the adapter posterior.

Conceptually, SPECTRA combines three complementary research directions. From the perspective of PEFT, the method preserves the scalability advantages of LoRA by restricting trainable parameters to lightweight adapters and keeping the backbone frozen. From the uncertainty modeling perspective, SPECTRA extends contextual Bayesian LoRA approaches by conditioning the posterior distribution on multimodal evidence, enabling uncertainty estimates that respond to cross-modal interactions. From the multimodal learning perspective, SPECTRA introduces a lightweight alternative to high-

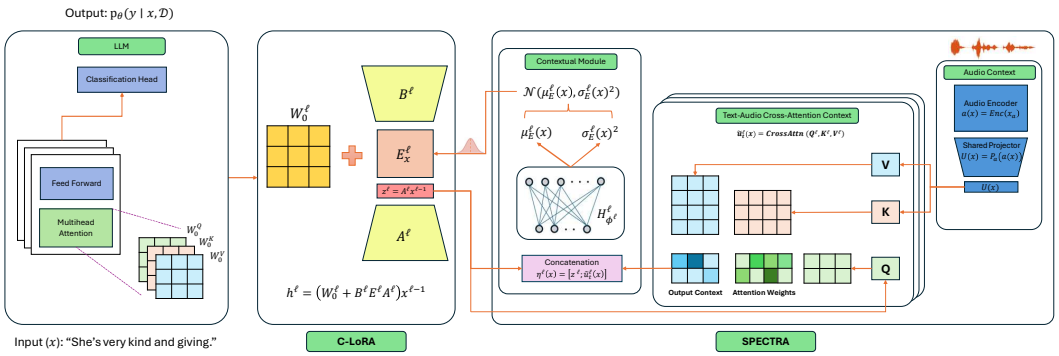


Figure 1: Overview of the proposed SPECTRA architecture. Per-layer text-audio cross-attention conditions the variational adapter distribution, enabling uncertainty-aware multimodal low-rank adaptation. A frozen text backbone produces token-level hidden states that enter LoRA adapters at each transformer layer, while a frozen audio encoder produces frame-level acoustic embeddings. The audio frames are first projected into a shared context space and then consumed by lightweight per-layer cross-attention modules, where LoRA text features act as queries and audio frames serve as keys and values. The resulting token-conditioned cross-modal context is concatenated with the local low-rank feature to parameterize a variational posterior over the compact latent matrix $E_{x,t}^\ell$. Samples from this posterior modulate the low-rank update $B^\ell E_{x,t}^\ell A^\ell$, enabling context-aware Bayesian adaptation with calibrated uncertainty while preserving the efficiency of PEFT.

dimensional feature fusion: instead of merging modalities directly in the representation space, audio information influences the distribution over adapter behavior in the low-rank space. This design allows the model to dynamically adjust its confidence based on temporally localized acoustic cues while maintaining parameter efficiency.

Another key design principle of SPECTRA is to treat audio as a contextual reliability signal rather than a primary feature stream. Each transformer layer computes token-conditioned cross-attention over the audio frames, enabling depth-specific alignment between text tokens and acoustic segments. These cross-attended features are then used to parameterize the variational posterior of a compact latent matrix within the low-rank adapter space. Because stochasticity is confined to this low-dimensional latent component, the Bayesian complexity scales with the adapter rank rather than the full backbone dimension, making the approach practical for large transformer models.

Overall, SPECTRA aims to address a fundamental problem in multimodal low-resource prediction: how to adapt large language models efficiently while producing uncertainty estimates that respond to temporally localized cross-modal reliability cues. Our contributions can be summarized as follows:

- We propose SPECTRA, a multimodal, uncertainty-aware PEFT framework that conditions adapter uncertainty on external audio evidence in addition to internal low-rank text features, yielding cross-modal, heteroscedastic uncertainty.
- We introduce a per-layer, token-level text-audio cross-attention mechanism that allows the variational posterior to depend on temporally localized acoustic information, enabling depth-specific and token-specific cross-modal conditioning without heavy multimodal fusion.
- We preserve lightweight Bayesian complexity by restricting stochasticity to a compact latent component in the rank-r space, maintaining the scalability advantages of contextual LoRA-style Bayesianization.
- We provide a unified perspective connecting PEFT, uncertainty-aware LoRA (e.g., BLoB and C-LoRA), and multimodal transfer learning, showing how cross-modal context can be incorporated as a principled uncertainty signal rather than only as a feature-fusion signal.

2 Proposed Method

2.1 Preliminaries: Low-Rank and Bayesian Adapter Tuning

Low-rank adaptation. Low-Rank Adaptation (LoRA) adapts a frozen linear transformation $W_0 \in \mathbb{R}^{d\times k}$ by learning a low-rank update rather than modifying all entries of W_0 Hu et al. [2022]. For an input $x \in \mathbb{R}^k$, the adapted forward pass is

$$h = (W_0 + \Delta W)x = (W_0 + BA)x, \qquad B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k). \tag{1}$$

Only A and B are trained, reducing the number of trainable parameters from dk to $r(d+k)$ while keeping the pretrained backbone fixed. However, standard LoRA is deterministic and therefore provides only point estimates of the adapter parameters. Bayesian LoRA methods address this limitation by placing a variational posterior over adapter parameters and optimizing an evidence lower bound (ELBO) Wang et al. [2024]. Contextual LoRA (C-LoRA) further amortizes this posterior from input-dependent low-rank features, yielding heteroscedastic uncertainty while keeping Bayesian inference in a compact adapter subspace Rahmati et al. [2025]. SPECTRA builds on this idea but conditions the adapter posterior on both text-derived low-rank features and token-level audio context.

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

SPECTRA is a multimodal uncertainty-aware PEFT framework for audio-text prediction. The text transformer and the audio encoder are kept frozen. Instead of fusing a high-dimensional audio representation with the final text embedding, SPECTRA uses audio as a contextual reliability signal that shapes the distribution over low-rank adapter behavior at each layer. Figure 1 illustrates the architecture, and pseudocode is provided in Appendix 5.

Setup and stochastic low-rank adapter. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be the training set, where each text input x_i is paired with an audio sequence $a(x_i)$. Consider a frozen transformer with L layers. In a linear sub-layer at layer ℓ, let $W_0^\ell \in \mathbb{R}^{d\times k}$ be the frozen weight, $x_t^{\ell-1} \in \mathbb{R}^k$ the hidden state at token position $t \in [T_x]$, and $h_t^\ell \in \mathbb{R}^d$ the output. SPECTRA uses deterministic LoRA factors $A^\ell \in \mathbb{R}^{r\times k}$ and $B^\ell \in \mathbb{R}^{d\times r}$, but inserts a stochastic latent matrix $E_{x,t}^\ell \in \mathbb{R}^{r\times r}$ in the rank-r space:

$$z_t^\ell = A^\ell x_t^{\ell-1}, \qquad h_t^\ell = W_0^\ell x_t^{\ell-1} + B^\ell E_{x,t}^\ell z_t^\ell, \qquad \Delta W_t^\ell(x) = B^\ell E_{x,t}^\ell A^\ell. \tag{2}$$

Thus, stochasticity is confined to $E_{x,t}^\ell$, whose size is r^2 per token and layer, while A^ℓ and B^ℓ remain deterministic trainable parameters. This design preserves the efficiency of LoRA and avoids placing a posterior over the full backbone or over all entries of the LoRA factors.

Per-layer text-audio cross-attention. For each sample, a frozen audio encoder produces frame-level embeddings $a(x) = [a_1(x), \ldots, a_{T_a}(x)]$, with $a_s(x) \in \mathbb{R}^{d_a}$. We first project each frame into a low-dimensional context space,

$$u_s(x) = P_a(a_s(x)) \in \mathbb{R}^c, \qquad U(x) = [u_1(x), \ldots, u_{T_a}(x)] \in \mathbb{R}^{T_a\times c}. \tag{3}$$

At layer ℓ, the local low-rank text feature z_t^ℓ queries the projected audio frames to obtain a token-specific acoustic context: $\tilde{u}_t^\ell(x) = \mathrm{CrossAttn}^\ell(q_t^\ell, K^\ell(x), V^\ell(x)) \in \mathbb{R}^r$. Let $q_t^\ell = W_Q^\ell z_t^\ell$, $K^\ell(x) = U(x)W_K^\ell$, and $V^\ell(x) = U(x)W_V^\ell$. Then: $\tilde{u}_t^\ell(x) = W_O^\ell \left[\mathrm{softmax}\Big(\frac{q_t^\ell (K^\ell(x))^\top}{\sqrt{d_c}}\Big) V^\ell(x)\right]$

This per-layer cross-attention lets different transformer layers attend to different temporal regions of the audio. The resulting context is used only to parameterize the adapter posterior, so SPECTRA does not increase the dimensionality of the downstream classifier.

Context-conditioned variational posterior. SPECTRA models $E_{x,t}^\ell$ with an input-dependent diagonal Gaussian posterior. The posterior is conditioned on both the local text adapter feature and

the token-specific audio context:

$$\eta_t^\ell(x) = [z_t^\ell; \tilde{u}_t^\ell(x)] \in \mathbb{R}^{2r}, \tag{4}$$

$$(\mu_{E,t}^\ell(x), \log v_{E,t}^\ell(x)) = H_{\phi^\ell}^\ell(\eta_t^\ell(x)), \qquad H_{\phi^\ell}^\ell : \mathbb{R}^{2r} \to \mathbb{R}^{2r^2},$$

$$q_{\phi^\ell}(\mathrm{vec}(E_{x,t}^\ell) \mid x) = \mathcal{N}\Big(\mu_{E,t}^\ell(x), \mathrm{diag}\big(\sigma_{E,t}^\ell(x)^2\big)\Big), \tag{5}$$

$$\sigma_{E,t}^\ell(x) = \varepsilon\, \mathrm{softplus}\Big(\log v_{E,t}^\ell(x)\Big) + \delta, \qquad \varepsilon > 0,\ \delta \approx 10^{-6}. \tag{6}$$

Because both the posterior mean and variance depend on $\eta_t^\ell(x)$, SPECTRA induces heteroscedastic uncertainty: the same text activation can lead to different adapter uncertainty depending on the acoustic frames attended by that token.

Reparameterized adapter sampling. During training and Monte Carlo inference, we sample $E_{x,t}^\ell$ using the reparameterization trick:

$$\mathrm{vec}(E_{x,t}^\ell) = \mu_{E,t}^\ell(x) + \sigma_{E,t}^\ell(x) \odot \xi_t^\ell, \qquad \xi_t^\ell \sim \mathcal{N}(0, I_{r^2}), \tag{7}$$

$$\Delta h_t^\ell(x) = B^\ell E_{x,t}^\ell z_t^\ell = B^\ell E_{x,t}^\ell A^\ell x_t^{\ell-1}. \tag{8}$$

Thus, each stochastic draw changes the effective low-rank update while keeping the pretrained backbone fixed. The Bayesian overhead scales with $O(r^2)$ in the latent adapter space rather than with the full hidden dimension of the transformer.

Prior and training objective. We use a factorized Gaussian prior over the compact latent matrix,

$$p(\mathrm{vec}(E^\ell)) = \mathcal{N}(0, \beta^2 I_{r^2}), \tag{9}$$

where β controls the prior scale. Let θ denote the deterministic trainable parameters, including $\{A^\ell, B^\ell\}_{\ell=1}^L$, P_a, the cross-attention projections, and the task head, and let $\phi = \{\phi^\ell\}_{\ell=1}^L$ denote the variational inference heads. SPECTRA maximizes

$$\mathcal{L}(\theta, \phi) = \sum_{i=1}^N \left[\mathbb{E}_{q_\phi(E_{x_i} \mid x_i)} \log p_\theta(y_i \mid x_i, E_{x_i}) - \gamma \sum_{\ell=1}^L \frac{1}{T_{x_i}} \sum_{t=1}^{T_{x_i}} \mathrm{KL}\Big(q_{\phi^\ell}(E_{x_i,t}^\ell \mid x_i) \,\|\, p(E^\ell)\Big) \right]. \tag{10}$$

The KL term is averaged over token positions so that its magnitude does not grow trivially with sequence length. In practice, we minimize the negative ELBO and use one Monte Carlo draw per training example,

$$\mathbb{E}_{q_\phi(E_{x_i} \mid x_i)} \log p_\theta(y_i \mid x_i, E_{x_i}) \approx \log p_\theta(y_i \mid x_i, E_{x_i}^{(1)}), \qquad E_{x_i}^{(1)} \sim q_\phi(E_{x_i} \mid x_i). \tag{11}$$

Only the adapters, cross-attention modules, audio projection, inference heads, and classifier are updated; the text and audio backbones remain frozen.

Posterior prediction and uncertainty scores. At test time, SPECTRA can use either posterior means or Monte Carlo model averaging. For a test input x^*, the posterior predictive distribution is approximated as

$$p(y^* \mid x^*, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^M p_\theta(y^* \mid x^*, E_{x^*}^{(m)}), \qquad E_{x^*}^{(m)} \sim q_\phi(E_{x^*} \mid x^*). \tag{12}$$

For classification with C classes, let $\mathbf{p}_i^{(m)} = p_\theta(y_i \mid x_i, E_i^{(m)})$ and $\bar{\mathbf{p}}_i = M^{-1} \sum_{m=1}^M \mathbf{p}_i^{(m)}$. We report total predictive uncertainty, aleatoric uncertainty, and adapter-space epistemic uncertainty as

$$U_{\mathrm{total}}(x_i) = H[\bar{\mathbf{p}}_i], \qquad U_{\mathrm{alea}}(x_i) = \frac{1}{M} \sum_{m=1}^M H[\mathbf{p}_i^{(m)}], \qquad U_{\mathrm{epi}}(x_i) = U_{\mathrm{total}}(x_i) - U_{\mathrm{alea}}(x_i), \tag{13}$$

where $H[\mathbf{p}] = -\sum_{c=1}^C p_c \log(p_c + \epsilon_h)$. Since our tasks include both binary and multi-class labels, we use normalized scores $\widetilde{U}_b(x_i) = U_b(x_i)/\log C$ for $b \in \{\mathrm{total}, \mathrm{alea}, \mathrm{epi}\}$. In SPECTRA, U_{epi} measures disagreement over the stochastic adapter latent rather than uncertainty over all backbone parameters; it is therefore best interpreted as adapter-space epistemic uncertainty. High epistemic uncertainty indicates that different context-conditioned adapter samples induce different predictions, while high aleatoric uncertainty indicates that individual sampled adapters remain uncertain.

Table 1: Dataset information for the parent and offspring data.

	Task	# Samples	# Classes	Imbalanced (%)	Label 0	Label 1	Label 2	Label 3
Parent	objective	23693	2	19.08	13096	10597	-	-
Parent	sentiment	23479	3	57.39	4725	11088	7666	-
Parent	anger	23694	2	92.22	21984	1710	-	-
Parent	fear	23694	2	95.29	22628	1066	-	-
Parent	joy	23690	2	66.58	17756	5934	-	-
Parent	sadness	23694	2	92.70	22081	1613	-	-
Parent	neutral	23693	2	16.28	10797	12896	-	-
Parent	cohesion	23694	2	81.32	19964	3730	-	-
Parent	rumination	23694	2	98.95	23447	247	-	-
Parent	overinclusive	23694	2	97.20	23049	645	-	-
Parent	worry	23694	2	91.76	21890	1804	-	-
Parent	criticism	23694	2	88.42	21235	2459	-	-
Offspring	objective	10319	2	1.90	5209	5110	-	-
Offspring	sentiment	10222	3	52.51	2404	5062	2756	-
Offspring	richness	10273	3	97.08	1366	8654	253	-
Offspring	reference	10328	4	77.89	2072	4970	2187	1099
Offspring	irrelevance	10301	2	87.98	9196	1105	-	-
Offspring	anger	10329	2	95.55	9889	440	-	-
Offspring	fear	10329	2	99.04	10231	98	-	-
Offspring	joy	10329	2	94.49	9790	539	-	-
Offspring	sadness	10329	2	94.62	9802	527	-	-
Offspring	neutral	10328	2	80.92	1655	8673	-	-
Offspring	coherence	10329	2	68.35	7846	2483	-	-
Offspring	rumination	10329	2	99.01	10228	101	-	-
Offspring	worry	10329	2	99.13	10240	89	-	-
Offspring	anxiousness	10317	2	96.51	9969	348	-	-
Offspring	aggression	2182	2	99.49	2171	11	-	-
Offspring	criticism	8090	2	95.04	7708	382	-	-
Offspring	self-criticism	10329	2	96.77	10006	323	-	-

Relation to prior work. SPECTRA differs from BLoB by using an amortized input-dependent posterior instead of a fixed posterior over adapter weights. It differs from C-LoRA by conditioning this posterior not only on internal low-rank text features but also on per-layer, token-level audio context. Consequently, SPECTRA preserves the lightweight Bayesian structure of contextual LoRA while allowing uncertainty and adaptation to respond to temporally localized cross-modal evidence.

3 Dataset

FORBOW. The data used in this work consists of audio speech samples from 369 subjects participating in the Families Overcoming Risks and Building Opportunities for Well Being (FORBOW) research project Uher et al. [2014]. Participants are parents, 266 mothers and 103 fathers, in the age range of 28-51 years. In these clinical interviews, parents were asked to talk about their children for five minutes without interruption. Out of these subjects, 149 were diagnosed with Major Depressive Disorder (MDD), 66 were diagnosed with Bipolarity Disorder (BD), 19 were diagnosed with Schizophrenia, and 129 were the control group with no major mood disorders. In addition to the parents' interview files, FORBOW research project collected interviews with the children themselves. The audio interviews of children consists of 3 parts: 1) a three minute interview where children talk about themselves, 2) a two minute interview talking about a positive experience they had, and 3) a two minute interview where they talk about a negative experience they had. All three interviews are uninterrupted with a total of 7 minutes of speech from each child. We transcribed and broke down each sample into multiple segments based on changes in emotion, sentiment, objectivity/subjectivity, etc. Average word count in a segment is 17 and average audio length for a segment is 6.47 seconds.Table 1 summarizes the dataset statistics, including the set of prediction tasks at different levels of granularity (segment-level, document-level, and psychological and cognitive tasks), as well as the label distributions and the degree of class imbalance.

IEMOCAP. In addition to our defined set of prediction tasks, we evaluate SPECTRA on IEMOCAP as the most widely used public dataset for emotion recognition Busso et al. [2008]. The IEMOCAP corpus contains a total of 5 sessions and 10 different speakers, with each session being a conversation of two exclusive speakers. We follow the conventional evaluation protocol: merge 'excited' with 'happy' to better balance the size of each emotion class and drop the unbalanced emotion classes to leave the final four classes (neutral, happy, sad, angry) with a similar amount of data points and cross-validate on five folds of the standard splits.

Table 2: Unimodal emotion recognition results on IEMOCAP (AUC %).

Modality	Model	Transfer	LoRA	C-LoRA
Text	nliRoBERTa	86.58±0.46	87.67±0.62	**88.08±0.50**
	nliDistilRoBERTa	86.35±0.63	87.33±0.71	**87.38±1.12**
	allRoBERTaLarge	86.88±0.64	84.12±0.54	**88.40±1.07**
	allDistilRoBERTa	85.64±0.57	87.28±0.49	**87.89±0.60**
	allMiniLM12	83.05±0.88	75.36±0.83	**87.19±0.71**
	paraTinyBERT	84.69±0.36	**86.24±0.76**	86.13±1.06
Audio	whisperMedium	93.00±0.44	-	-
	hubertLargeFT	87.86±0.87	-	-

Table 3: Multimodal emotion recognition results on IEMOCAP (AUC %).

Text Model	**Audio Model**	**Transfer**	**SPECTRA-G**	**SPECTRA-X**	
			Audio Ctx	**Shared-KV**	**Cross-Attn**
nliRoBERTa	whisperMedium	93.40±0.63	93.27±0.60	**94.01±0.53**	93.05±0.99
	hubertLargeFT	89.93±0.46	90.40±0.54	90.20±0.69	**91.06±0.70**
nliDistilRoBERTa	whisperMedium	93.39±0.75	92.91±0.64	93.15±0.19	**93.99±0.95**
	hubertLargeFT	**89.76±0.54**	89.21±0.72	89.46±0.59	89.51±0.97
allRoBERTaLarge	whisperMedium	**94.64±0.49**	92.48±0.76	92.59±0.95	94.05±0.33
	hubertLargeFT	88.06±0.14	**89.21±1.67**	88.49±2.18	88.81±1.39
allDistilRoBERTa	whisperMedium	94.48±0.38	92.56±0.81	93.14±0.76	**94.68±0.58**
	hubertLargeFT	90.28±0.52	90.00±0.43	89.62±0.82	**91.09±0.55**
allMiniLM12	whisperMedium	**93.32±0.57**	92.93±0.47	92.84±0.64	92.99±0.67
	hubertLargeFT	89.75±1.00	89.37±0.60	**89.78±0.83**	89.62±1.17
paraTinyBERT	whisperMedium	92.51±0.49	91.62±0.61	92.27±0.70	**93.36±0.52**
	hubertLargeFT	89.13±0.74	88.32±0.51	**89.62±0.42**	88.97±0.77

4 Experiments

We evaluate SPECTRA on emotion recognition and clinical interview prediction to test whether audio-conditioned Bayesian low-rank adaptation improves parameter-efficient multimodal learning. The main paper reports compact summaries of the experimental grid; full per-backbone, per-audio-encoder, per-task results are provided in Appendix B.

Protocol. We use six frozen text encoders, nliRoBERTa, nliDistilRoBERTa, allRoBERTaLarge, allDistilRoBERTa, allMiniLM12, and paraTinyBERT, and three frozen audio encoders, whisperMedium, wav2vec2LargeFT, and hubertLargeFT. We compare three method families: (i) text-only PEFT using LoRA and C-LoRA, (ii) multimodal transfer learning, which fuses a pooled audio embedding with the text representation and trains a lightweight MLP head, and (iii) SPECTRA, which uses audio as a context signal for Bayesian low-rank adaptation. SPECTRA-G conditions the adapters on a global audio embedding, while SPECTRA-X uses token-to-frame cross-modal conditioning. On IEMOCAP we also evaluate a Shared-KV version of SPECTRA-X, where audio key/value projections are shared across layers to reduce overhead.

All PEFT methods use rank $r = 8$ and scaling $\alpha = 32$. Text and audio backbones are frozen; only adapter and task-head parameters are trained. Models are optimized for 50 epochs with AdamW using learning rate 10^{-3} and weight decay 10^{-3}. For the Bayesian adapter, we use $\beta = 0.2$, $\varepsilon = 0.05$, and $\gamma = 0.008$, and estimate predictive uncertainty with $M = 10$ Monte Carlo forward passes. All clinical experiments use 5-fold cross-validation with speaker-level separation. We report AUC (%) for predictive performance because the clinical labels are imbalanced.

4.1 Main Predictive Results

On IEMOCAP, the results indicate that the task is already strongly driven by the acoustic modality, with whisperMedium achieving the best unimodal performance overall at 93% AUC, substantially higher than all text-only models (Table 2). This suggests that emotional content in IEMOCAP is heavily reflected in prosodic and paralinguistic cues, which is consistent with the dataset's acted conversational setting. Nevertheless, the text-only comparison remains informative: C-LoRA is consistently competitive and usually outperforms deterministic LoRA, yielding the best text-only

Table 4: AUC (%) on a 5 fold cross-validation on segment-level predictions on parent data.

Method	LLM	Audio Encoder	**objective**	**sentiment**	**anger**	**fear**	**joy**	**sadness**	**neutral**	**cohesion**	**rumination**	**overinclude**	**worry**	**criticism**
LoRA	nRoBERTa	-	88.02	92.38	90.87	89.81	89.77	89.92	79.49	78.30	59.01	85.76	89.99	91.93
C-LoRA	nRoBERTa	-	86.82	91.42	88.76	75.32	87.74	87.95	77.39	66.40	55.68	65.06	86.94	88.91
Transfer	nRoBERTa	whisperM	85.70	90.39	89.56	89.10	89.13	89.51	78.28	79.63	82.40	83.55	89.03	90.15
		wav2vec2LFT	85.15	91.07	89.20	88.43	88.69	88.76	77.36	76.89	79.16	82.25	88.47	90.15
		hubertLFT	85.08	90.94	88.31	87.87	88.56	88.95	76.92	77.06	79.70	82.31	88.26	89.67
SPECTRA-G	nRoBERTa	whisperM	90.03	94.86	89.84	88.52	91.65	91.52	80.53	78.15	59.43	82.45	91.09	92.34
		wav2vec2LFT	90.11	94.58	88.54	88.64	91.64	91.09	76.32	74.71	66.43	76.74	90.52	**93.46**
		hubertLFT	89.91	94.57	90.92	85.92	91.69	87.55	80.93	76.15	65.35	74.84	90.78	93.43
SPECTRA-X	nRoBERTa	whisperM	91.08	95.09	90.34	89.71	92.08	**93.17**	**81.92**	78.51	60.41	82.86	91.09	93.17
		wav2vec2LFT	90.74	94.89	89.11	88.71	92.65	91.4	77.05	75.53	67.42	78.01	91.36	93.23
		hubertLFT	90.68	94.71	**91.55**	86.26	91.94	88.34	81.71	76.71	66.12	75.21	**91.5**	92.95
LoRA	pTinyBERT	-	86.91	90.23	74.97	53.94	87.82	75.71	77.12	75.47	50.00	52.10	75.31	84.09
C-LoRA	pTinyBERT	-	85.91	89.98	78.39	65.44	87.24	82.61	76.00	72.43	54.66	62.36	80.94	86.21
Transfer	pTinyBERT	whisperM	84.07	88.75	86.82	86.52	86.95	87.00	71.37	79.08	**83.61**	83.07	87.11	87.77
		wav2vec2LFT	83.04	88.42	86.18	84.52	86.21	86.69	73.06	76.84	80.53	79.95	85.88	86.73
		hubertLFT	83.11	88.10	85.11	85.01	86.07	86.06	72.77	76.57	78.33	80.44	85.93	86.92
SPECTRA-G	pTinyBERT	whisperM	89.86	94.09	90.14	85.94	91.51	88.70	80.27	79.08	57.04	66.70	79.42	91.54
		wav2vec2LFT	90.54	94.23	90.60	78.13	91.31	89.00	80.03	81.70	58.44	77.19	84.65	87.78
		hubertLFT	90.51	93.78	89.11	75.13	91.59	86.15	80.09	81.87	60.90	75.91	90.02	91.07
SPECTRA-X	pTinyBERT	whisperM	90.15	94.63	90.28	86.52	92.88	89.67	80.36	79.71	57.55	67.4	81.05	92.14
		wav2vec2LFT	91.73	94.31	91.49	78.5	91.58	89.34	80.95	**82.64**	58.6	77.97	84.98	88.59
		hubertLFT	90.96	93.87	90.14	75.49	92.54	87.13	80.59	82.0	61.01	76.97	90.77	92.29
LoRA	aMiniLM12	-	86.91	90.16	58.83	51.26	87.16	50.74	77.14	71.31	50.00	50.00	68.33	72.70
C-LoRA	aMiniLM12	-	86.38	90.10	60.29	53.76	79.97	55.29	75.86	62.41	51.39	52.36	54.23	60.67
Transfer	aMiniLM12	whisperM	82.68	83.32	80.24	81.81	84.06	81.81	72.29	78.50	78.86	80.40	82.44	80.56
		wav2vec2LFT	81.28	84.12	80.56	78.22	83.01	80.08	69.64	75.65	72.78	75.07	80.36	82.11
		hubertLFT	80.88	83.60	77.45	78.13	81.97	78.77	68.50	74.87	69.31	74.70	79.53	79.78
SPECTRA-G	aMiniLM12	whisperM	91.94	**95.30**	64.82	57.31	91.38	79.54	81.35	75.48	55.00	58.86	66.89	91.72
		wav2vec2LFT	91.11	94.92	65.59	60.84	91.88	56.38	81.26	71.78	55.70	61.57	76.15	81.24
		hubertLFT	91.59	95.13	55.46	56.76	92.17	75.05	79.96	70.35	55.14	63.50	72.13	89.06
SPECTRA-X	aMiniLM12	whisperM	**92.69**	95.12	65.26	57.99	92.52	80.33	81.9	75.77	55.2	59.76	67.2	92.15
		wav2vec2LFT	91.71	94.7	66.66	61.55	**93.15**	56.99	81.68	72.43	56.57	61.97	76.76	81.82
		hubertLFT	91.98	94.86	55.51	57.61	92.94	76.05	80.98	71.06	55.84	64.25	72.88	89.84
LoRA	aDRoBERTa	-	87.39	91.31	89.70	88.26	88.94	88.95	78.30	77.73	78.73	84.76	89.29	90.40
C-LoRA	aDRoBERTa	-	86.40	90.70	88.09	86.44	88.21	87.70	76.74	76.75	68.47	80.21	87.64	89.57
Transfer	aDRoBERTa	whisperM	82.99	82.35	82.86	82.70	84.83	83.49	72.47	78.90	81.43	80.39	84.48	83.41
		wav2vec2LFT	81.77	85.53	81.74	80.99	84.01	82.14	70.76	75.81	75.23	76.46	82.26	83.41
		hubertLFT	81.73	84.80	80.38	79.58	83.01	81.52	68.97	75.77	68.52	75.51	81.40	82.23
SPECTRA-G	aDRoBERTa	whisperM	88.31	92.69	89.91	88.49	89.54	90.51	78.98	79.73	76.13	83.36	89.19	91.38
		wav2vec2LFT	88.30	92.99	90.34	89.94	90.39	89.49	79.08	79.23	75.62	84.56	89.51	91.08
		hubertLFT	88.93	92.89	88.74	82.09	90.08	90.25	78.43	81.11	77.65	85.93	89.39	90.46
SPECTRA-X	aDRoBERTa	whisperM	88.81	93.27	90.77	88.88	90.01	92.01	79.89	79.99	76.55	85.05	89.83	92.26
		wav2vec2LFT	88.84	94.1	91.03	**90.95**	90.75	90.81	79.39	79.6	76.81	85.62	90.94	91.68
		hubertLFT	90.63	93.71	89.12	83.23	90.89	90.97	79.28	82.12	78.42	**87.59**	89.98	91.27

Table 5: Modality-disagreement stress test. Higher AUC is better; lower NLL, ECE, and Brier score are better. AUC is reported on the $[0, 1]$ scale.

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

result for five of the six language backbones. This trend supports the value of contextual Bayesian low-rank adaptation even in the unimodal setting, where input-dependent uncertainty appears to improve robustness relative to deterministic PEFT. At the same time, the gap between the best text-only result (88.40% for allRoBERTaLarge) and the best audio-only result highlights the importance of incorporating acoustic evidence when the target is emotion recognition.

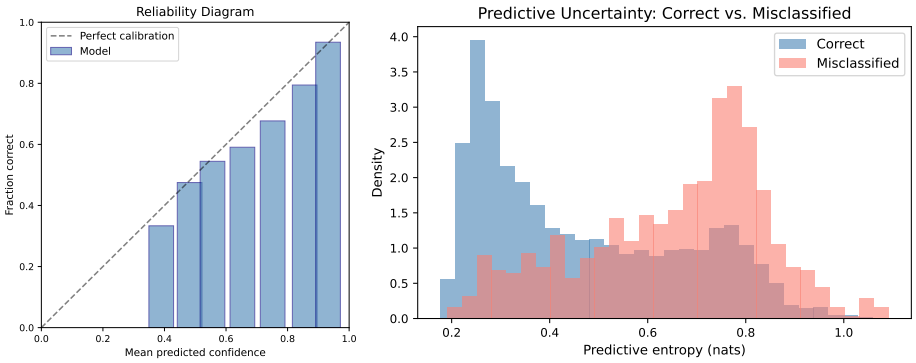
The multimodal results in Table 3 show that SPECTRA is most beneficial when the text backbone and audio backbone are reasonably complementary, rather than when one modality already dominates the prediction. Across the 12 multimodal model combinations, the best SPECTRA variant outperforms the transfer-learning baseline in 9 cases, with the gains most often achieved by the token-level SPECTRA-X variants. In particular, the full cross-attention version performs best for several pairings, suggesting that fine-grained token-to-frame conditioning can provide useful localized acoustic context beyond simple global fusion. The shared-KV variant also performs strongly, achieving the best overall multimodal result of 94.01% for nliRoBERTa combinations, which indicates that much of the benefit of SPECTRA can be retained with reduced parameter overhead. However, improvements are not universal: for combinations already dominated by a very strong audio backbone, such as allRoBERTaLarge+whisperMedium, the transfer baseline remains best. This pattern suggests that SPECTRA is especially effective when cross-modal conditioning helps resolve uncertainty or complement weaker textual representations, whereas its advantage is naturally smaller when a single modality already provides near-saturated predictive performance.

Table 4 shows a clear and consistent advantage of multimodal contextualized adaptation over both text-only PEFT and conventional transfer-based fusion on the parent segment-level tasks. Overall, the strongest results are typically achieved by the SPECTRA variants, with SPECTRA-X most often providing the best or near-best performance, which suggests that token-level text-audio conditioning is more effective than relying only on a global audio summary. This trend is especially visible on relatively well-structured affective and discourse-related tasks such as objective, sentiment, joy, sadness, neutral, and criticism, where SPECTRA substantially improves over both LoRA/C-LoRA and the transfer baselines across multiple text encoders. These gains indicate that, in this clinical interview setting, acoustic cues provide useful localized evidence that can refine the adapter behavior beyond what deterministic text-only adaptation or pooled multimodal fusion can capture. Importantly, the improvements are not limited to the strongest text models: even weaker backbones such as paraTinyBERT and allMiniLM12 benefit substantially from SPECTRA, suggesting that cross-modal Bayesian conditioning can compensate for limited textual representational capacity. The results on offspring data is provided in the Appendix, Table 7.

4.2 Modality-Disagreement Stress Test

Standard multimodal evaluation uses correctly paired text and audio, but this does not test whether a model reacts appropriately when one modality becomes unreliable. We therefore evaluate the same trained SPECTRA model under four held-out conditions: clean matched text-audio pairs, clean text with noisy audio, clean text with mismatched audio from another example, and corrupted text with clean audio. The label is kept fixed in all conditions, so changes in performance reflect sensitivity to modality reliability rather than changes in the task. We report AUC, negative log-likelihood (NLL), expected calibration error (ECE), and Brier score in Table 5. Detailed perturbation definitions, calibration equations, reliability diagrams, and uncertainty decomposition results are provided in Appendix B.1.

The clean matched condition performs best overall. Noisy but correctly paired audio causes only a moderate AUC drop, from 0.835 to 0.812, and the model remains well calibrated. In contrast, mismatched audio is substantially more harmful: AUC drops to 0.716, while NLL, ECE, and Brier



(a) Calibration diagram (b) Predictive uncertainty

Figure 2: Calibration and predictive uncertainty analysis.

score all increase sharply. This shows that semantically inconsistent audio is more damaging than degraded but still matched audio, and it exposes a residual overconfidence failure mode under cross-modal contradiction. Corrupted text with clean audio leads to an intermediate degradation, suggesting that valid acoustic evidence can partially compensate when lexical evidence is degraded. These stress-test results support the central motivation of SPECTRA: audio is not merely an additional feature stream, but a contextual signal that can shape both adaptation and predictive confidence.

Figure 2 provides complementary evidence about the uncertainty behavior of the model. The reliability diagram in Figure 2a shows that the predicted confidence and empirical accuracy follow the diagonal trend very well, indicating that the model is well calibrated overall. Figure 2b further shows that predictive entropy is informative about error likelihood: correctly classified samples are concentrated at lower uncertainty values, while misclassified samples tend to shift toward higher entropy. This separation suggests that the model's uncertainty estimates are meaningful, as larger predictive uncertainty is generally associated with a greater chance of incorrect predictions.

5 Conclusion

We introduced SPECTRA, a multimodal uncertainty-aware PEFT framework that integrates token-level audio-text context into Bayesian low-rank adaptation. By conditioning a variational posterior in the rank-r adapter space on per-layer cross-attended acoustic context, SPECTRA provides a lightweight mechanism for incorporating multimodal reliability cues without modifying the frozen backbone or relying on heavy feature-fusion architectures. This design preserves the scalability advantages of PEFT while enabling heteroscedastic, input-dependent uncertainty that reflects localized cross-modal interactions.

Experimental results on multiple datasets show that SPECTRA is often superior to text-only PEFT baselines and conventional multimodal transfer-learning approaches. In particular, the token-level cross-attention variant yields the most consistent improvements, suggesting that fine-grained cross-modal contextualization is more effective than global audio conditioning in many low-resource prediction settings. Architecturally, SPECTRA's Bayesian core is not audio-specific. It consumes a text query sequence and a masked sequence of context tokens projected to a small common dimension. Image patches, video clips, physiological measurements, or other temporal sensor features can replace audio frames as keys/values while leaving $q_\phi\big(E^\ell_{x,t} \mid [z^\ell_t; \tilde{u}^\ell_t]\big)$ and $B^\ell E^\ell_{x,t} A^\ell$ unchanged. The implementation separates encoder interfaces, shared projection, cross-attention, and the modality-agnostic posterior head. Multiple context modalities could be concatenated as a typed K/V memory or attended separately and gated before the posterior head.

Overall, our findings highlight that multimodal information can play a dual role: not only as a source of predictive features, but also as a signal for modulating model confidence and adaptation. This provides a principled and parameter-efficient direction for uncertainty-aware multimodal learning with large frozen backbones. Future work may extend SPECTRA to additional modalities, longer temporal contexts, and other high-stakes low-resource applications where reliable uncertainty estimation is essential.

References

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *AAAI*, volume 37, pages 12799–12807, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *ICLR*, 2025.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *NeurIPS*, 35:1950–1965, 2022.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Reinhold Müller and Petra Büttner. A critical discussion of intraclass correlation coefficients. *Statistics in medicine*, 13(23-24):2465–2476, 1994.

Habibeh Naderi, Behrouz Haji Soleimani, and Stan Matwin. Mac: Multimodal Attentive Contrastive Learning Framework. *Proceedings of the Canadian Conference on Artificial Intelligence*, may 19 2025. https://caiac.pubpub.org/pub/zpm3p8jv.

Habibeh Naderi, Behrouz Haji Soleimani, and Stan Matwin. Joint-centric dual contrastive alignment with structure-preserving and information-balanced regularization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 38–55. Springer, 2026a.

Habibeh Naderi, Behrouz Haji Soleimani, and Stan Matwin. From token imbalance to balanced routing: An elbo-regularized probabilistic framework for contrastive multimodal learning. In *29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026b.

Deep S Pandey, Spandan Pyakurel, and Qi Yu. Be confident in what you know: Bayesian parameter efficient fine-tuning of vision foundation models. *Advances in Neural Information Processing Systems*, 37:44814–44844, 2024.

Amir Hossein Rahmati, Sanket Jantre, Weifeng Zhang, Yucheng Wang, Byung-Jun Yoon, Nathan Urban, and Xiaoning Qian. C-loRA: Contextual low-rank adaptation for uncertainty estimation in large language models. In *NeurIPS*, 2025.

Colin Samplawski, Adam D Cobb, Manoj Acharya, Ramneet Kaur, and Susmit Jha. Scalable bayesian low-rank adaptation of large language models via stochastic variational subspace inference. *arXiv preprint arXiv:2506.21408*, 2025.

Haizhou Shi, Yibin Wang, Ligong Han, Huan Zhang, and Hao Wang. Training-free bayesianization for low-rank adapters of large language models. *Advances in Neural Information Processing Systems*, 38:41663–41700, 2025. arXiv:2412.05723.

Hugo Thimonier, Antony Perzo, and Renaud Seguier. Emosllm: Parameter-efficient adaptation of llms for speech emotion recognition. *arXiv preprint arXiv:2508.14130*, 2025.

Rudolf Uher, Jill Cumby, Lynn E MacKenzie, Jessica Morash-Conway, Jacqueline M Glover, Alice Aylott, Lukas Propper, Sabina Abidi, Alexa Bagnell, Barbara Pavlova, et al. A familial risk enriched cohort as a platform for testing early interventions to prevent severe mental illness. *BMC psychiatry*, 14(1):344, 2014.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *NeurIPS*, 37:67758–67794, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *ICLR*, 2024.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *ICLR*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *ICLR*, 2023.

Technical appendices and supplementary material

In this appendix, we present further details and additional experimental results that are not included in the main text due to space constraints.

A Proposed Method: SPECTRA Algorithm

The pseudo-code of SPECTRA is provided in Algorithm 1. A frozen text backbone produces token-level hidden states that enter LoRA adapters at each transformer layer, while a frozen audio encoder produces frame-level acoustic embeddings. The audio frames are first projected into a shared context space and then consumed by lightweight per-layer cross-attention modules, where LoRA text features act as queries and audio frames serve as keys and values. The resulting token-conditioned cross-modal context is concatenated with the local low-rank feature to parameterize a variational posterior over the compact latent matrix $E^{\ell}_{x,t}$. Samples from this posterior modulate the low-rank update $B^{\ell}E^{\ell}_{x,t}A^{\ell}$, enabling context-aware Bayesian adaptation with calibrated uncertainty while preserving the efficiency of PEFT.

Algorithm 1 Training SPECTRA

Input: Training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, frozen transformer backbone, frozen audio encoder
Initialize low-rank parameters $\{A^{\ell}, B^{\ell}\}_{\ell=1}^{L}$
Initialize audio projection P_a, per-layer cross-attention parameters $\{W_Q^{\ell}, W_K^{\ell}, W_V^{\ell}, W_O^{\ell}\}_{\ell=1}^{L}$, and inference heads $\{H_{\phi^{\ell}}^{\ell}\}_{\ell=1}^{L}$
for each minibatch (x, y) **do**
  Encode the text sequence with the frozen transformer and obtain hidden states $\{x_t^{\ell-1}\}$
  Encode the accompanying audio and obtain frame-level embeddings $\{a_s(x)\}_{s=1}^{T_a}$
  Project audio frames: $u_s(x) = P_a(a_s(x))$
  for each layer $\ell = 1, \ldots, L$ **do**
    for each token position $t = 1, \ldots, T_x$ **do**
      Compute local LoRA feature: $z_t^{\ell} = A^{\ell}x_t^{\ell-1}$
      Form token-conditioned audio context $\tilde{u}_t^{\ell}(x)$ using Eqs. (17)-(19)
      Form contextual summary $\eta_t^{\ell}(x) = [z_t^{\ell}; \tilde{u}_t^{\ell}(x)]$
      Compute posterior parameters $(\mu_{E,t}^{\ell}(x), \sigma_{E,t}^{\ell}(x))$ using Eqs. (21)-(22)
      Sample $E_{x,t}^{\ell}$ via Eq. (23)
      Compute stochastic adapter update:

$$\Delta h_t^{\ell}(x) = B^{\ell}E_{x,t}^{\ell}A^{\ell}x_t^{\ell-1}$$

    end for
  end for
  Compute task likelihood and ELBO objective in Eq. (26)
  Update only trainable parameters θ, ϕ by gradient ascent on the ELBO (or equivalently gradient descent on the negative ELBO)
end for

A.1 Preliminaries

Low-Rank Adaptation (LoRA) Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method designed to adapt large pre-trained models to downstream tasks while keeping the backbone weights frozen Hu et al. [2022]. The core assumption behind LoRA is that task-specific weight updates lie in a low-dimensional subspace. Instead of updating the full weight matrix, LoRA learns a low-rank decomposition of the weight update.

Consider a linear transformation with frozen pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$. The adapted forward pass is defined as:

$$h \;=\; (W_0 + \Delta W)x \;=\; (W_0 + BA)x, \tag{14}$$

where $x \in \mathbb{R}^k$ is the input, $h \in \mathbb{R}^d$ is the output, and $\Delta W = BA$ is the low-rank update. Here, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable matrices with rank $r \ll \min(d, k)$. This factorization reduces the number of trainable parameters from $d \times k$ to $r \times (d + k)$, yielding substantial savings in memory and computation while often matching full fine-tuning performance.

Bayesian Uncertainty-Aware Low-Rank Adaptation While LoRA provides efficient deterministic adaptation, it does not model uncertainty in the learned parameters. Full Bayesian inference over all backbone weights is computationally infeasible for modern transformer models. A scalable alternative is to restrict uncertainty modeling to the low-rank adapters.

Bayesian LoRA by Backpropagation (BLoB) introduces uncertainty into the LoRA parameters while keeping the frozen backbone intact. Given the LoRA decomposition $\Delta W = BA$, BLoB maintains B as deterministic and places a variational posterior over A. Specifically, a mean-field variational inference approximation is used: $q(A) = \mathcal{N}\big(A \mid \mu_A, \Omega_A^2\big)$, where μ_A and Ω_A^2 denote the variational mean and variance parameters. The variational parameters are learned by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}' = \mathbb{E}_q\big[\log p(\mathcal{D} \mid A, B)\big] - \mathrm{KL}\big[q(A) \,\|\, p(A)\big].$$

The first term corresponds to the expected log-likelihood under the variational posterior, while the second term regularizes the solution via the Kullback-Leibler divergence to the prior. Optimization is performed using the reparameterization trick, enabling gradient-based training of both mean and variance parameters.

By confining Bayesian inference to the low-rank adapters, BLoB preserves the scalability benefits of LoRA while providing principled uncertainty estimation. This makes it an effective compromise between deterministic fine-tuning and fully Bayesian neural networks.

A.2 SPECTRA: Stochastic Posterior Estimation with Cross-modal Token-level Rank Adaptation

The preliminaries above highlight a practical gap in low-resource, high-stakes settings. Deterministic PEFT methods such as LoRA are efficient, but they can be overconfident when supervision is scarce or inputs are ambiguous. Bayesianizing adapters, as in BLoB, introduces epistemic uncertainty, yet its posterior is typically input-independent, limiting its ability to respond to sample-specific noise. C-LoRA addresses this by amortizing a contextual variational posterior in the low-rank space, producing heteroscedastic (input-dependent) uncertainty, but it conditions that uncertainty primarily on internal language-model features. In multimodal speech-centered applications, however, predictive uncertainty is often driven by external factors such as background noise, channel variability, prosodic variation, or speaker state, which are observable in audio but not necessarily recoverable from text alone.

We therefore propose SPECTRA (Stochastic Posterior Estimation with Cross-modal Token-level Rank Adaptation), a multimodal, uncertainty-aware PEFT framework that extends contextual Bayesian adapters by conditioning the variational posterior over the low-rank latent on per-layer text-audio cross-attention. A visual representation of the framework is shown in Figure 1. The key design principle is to use audio as a contextual reliability signal rather than as a heavy feature-fusion stream. Concretely, we keep the transformer backbone frozen and avoid expanding representation dimensionality at the classifier level. Instead, we allow each layer to form a lightweight token-conditioned view of the audio sequence and use that cross-modal context to shape uncertainty and adaptation within the rank-r space.

Setup and notation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a dataset of N i.i.d. samples. We consider a pre-trained transformer with L layers. In a linear sub-layer at layer $\ell \in [L]$, let $W_0^\ell \in \mathbb{R}^{d\times k}$ be the frozen pre-trained weight. For sequence inputs, let $x_t^{\ell-1} \in \mathbb{R}^k$ denote the hidden representation at token position $t \in [T_x]$, and let $h_t^\ell \in \mathbb{R}^d$ denote the corresponding output. LoRA parameterizes a low-rank update ΔW^ℓ via matrices $B^\ell \in \mathbb{R}^{d\times r}$ and $A^\ell \in \mathbb{R}^{r\times k}$ with rank $r \ll d$. Following the lightweight factorization introduced in C-LoRA Rahmati et al. [2025], we insert an additional matrix $E_{x,t}^\ell$ in the low-dimensional space:

$$h_t^\ell \;=\; \big(W_0^\ell + \Delta W_t^\ell(x)\big)x_t^{\ell-1} \;=\; W_0^\ell x_t^{\ell-1} + B^\ell E_{x,t}^\ell A^\ell x_t^{\ell-1}, \qquad \Delta W_t^\ell(x) \;=\; B^\ell E_{x,t}^\ell A^\ell. \tag{15}$$

The LoRA intermediate feature is defined token-wise as: $z_t^\ell = A^\ell x_t^{\ell-1} \in \mathbb{R}^r$.

Where uncertainty lives and what cross-modal context controls. SPECTRA restricts stochasticity to the compact latent $E^{\ell}_{x,t} \in \mathbb{R}^{r\times r}$, keeping A^{ℓ} and B^{ℓ} deterministic. This preserves the scalability of PEFT while enabling uncertainty estimation through variational inference. The key distinction from prior work is what the posterior conditions on. Instead of relying only on internal text-derived features, SPECTRA conditions the posterior on a token-specific cross-modal context obtained by letting text query the accompanying audio frames. As a result, uncertainty can increase when the audio suggests unreliability (e.g., noisy or weakly informative regions) and decrease when the acoustic signal is temporally aligned and consistent with the text.

Per-layer text-audio cross-attention context. Each input sample x is accompanied by a sequence of frame-level audio embeddings

$$a(x) = \big[a_1(x), \dots, a_{T_a}(x)\big], \qquad a_s(x) \in \mathbb{R}^{d_a},$$

obtained from a frozen or separately trained audio encoder, where T_a denotes the number of audio frames. We first project each audio frame into a shared context space of dimension $c \ll d_a$:

$$u_s(x) = P_a\big(a_s(x)\big) \in \mathbb{R}^{c}, \qquad s \in [T_a], \tag{16}$$

and collect the projected audio sequence as $U(x) = \big[u_1(x), \dots, u_{T_a}(x)\big] \in \mathbb{R}^{T_a\times c}$. At layer ℓ, SPECTRA forms a token-conditioned audio context by cross-attending the local LoRA text features $\{z^{\ell}_t\}_{t=1}^{T_x}$ to the projected audio frames. Specifically, for each token position t, we define

$$\tilde{u}^{\ell}_t(x) = \mathrm{CrossAttn}^{\ell}\big(q^{\ell}_t,\, K^{\ell}(x),\, V^{\ell}(x)\big) \in \mathbb{R}^{r}, \tag{17}$$

where the query is derived from the local text feature, $q^{\ell}_t = W^{\ell}_Q z^{\ell}_t \in \mathbb{R}^{d_c}$, and the keys and values are derived from the projected audio sequence,

$$K^{\ell}(x) = U(x)W^{\ell}_K \in \mathbb{R}^{T_a\times d_c}, \qquad V^{\ell}(x) = U(x)W^{\ell}_V \in \mathbb{R}^{T_a\times d_c}. \tag{18}$$

Using scaled dot-product attention, the token-conditioned context is

$$\tilde{u}^{\ell}_t(x) = W^{\ell}_O \left[\mathrm{softmax}\!\left(\frac{q^{\ell}_t (K^{\ell}(x))^{\top}}{\sqrt{d_c}}\right) V^{\ell}(x)\right], \tag{19}$$

where W^{ℓ}_O maps the attention output back to $\mathbb{R}^r$. This per-layer design allows different transformer layers to attend to different temporal regions and acoustic cues, yielding a depth-specific cross-modal reliability signal without high-dimensional multimodal fusion.

Context-conditioned variational posterior over $\mathbf{E^{\ell}_{x,t}}$. We model $E^{\ell}_{x,t} \in \mathbb{R}^{r\times r}$ as a latent random matrix with an input-dependent Gaussian variational posterior. The posterior is conditioned on (i) the local LoRA feature z^{ℓ}_t, which captures how the current text activation enters the low-rank adapter, and (ii) the cross-attended audio context $\tilde{u}^{\ell}_t(x)$, which captures token-specific external reliability information. We concatenate these to form the contextual summary

$$\eta^{\ell}_t(x) = [\, z^{\ell}_t; \tilde{u}^{\ell}_t(x)\,] \in \mathbb{R}^{2r}. \tag{20}$$

Let $H^{\ell}_{\phi^{\ell}} : \mathbb{R}^{2r} \to \mathbb{R}^{2r^2}$ output a mean vector $\mu^{\ell}_{E,t}(x) \in \mathbb{R}^{r^2}$ and a log-variance vector $\log v^{\ell}_{E,t}(x) \in \mathbb{R}^{r^2}$. Then

$$q_{\phi^{\ell}}\Big(\mathrm{vec}(E^{\ell}_{x,t}) \mid x\Big) \;=\; \mathcal{N}\Big(\mu^{\ell}_{E,t}(x),\, \mathrm{diag}\big(\sigma^{\ell}_{E,t}(x)^2\big)\Big), \tag{21}$$

where $\sigma^{\ell}_{E,t}(x) \in \mathbb{R}^{r^2}_{>0}$ is parameterized using a scaled softplus:

$$\sigma^{\ell}_{E,t}(x) = \varepsilon\,\mathrm{softplus}\big(\log v^{\ell}_{E,t}(x)\big) + \delta, \qquad \varepsilon > 0,\ \delta \approx 10^{-6}. \tag{22}$$

Equation (21) induces heteroscedastic uncertainty because both $\mu^{\ell}_{E,t}(x)$ and $\sigma^{\ell}_{E,t}(x)$ depend on the local text feature and the token-specific cross-modal context. Consequently, the same lexical pattern can yield different posterior variance depending on which acoustic frames it attends to and how informative or reliable those frames appear.

Reparameterized sampling and stochastic low-rank update. We sample $E^{\ell}_{x,t}$ via the reparameterization trick. Let $\xi^{\ell}_t \sim \mathcal{N}(0, I_{r^2})$:

$$\mathrm{vec}(E^{\ell}_{x,t}) = \mu^{\ell}_{E,t}(x) + \sigma^{\ell}_{E,t}(x) \odot \xi^{\ell}_t, \tag{23}$$

and reshape to $E^{\ell}_{x,t} \in \mathbb{R}^{r\times r}$. The resulting stochastic adapter contribution is

$$\Delta h^{\ell}_t(x) = B^{\ell} E^{\ell}_{x,t} z^{\ell}_t = B^{\ell} E^{\ell}_{x,t} A^{\ell} x^{\ell-1}_t. \tag{24}$$

Importantly, stochasticity remains confined to $r \times r$ variables, keeping the Bayesian complexity per token and per layer at $O(r^2)$ rather than scaling with the backbone dimension d.

Prior and ELBO objective. We place a factorized Gaussian prior on $\mathrm{vec}(E^\ell)$:

$$p\Big(\mathrm{vec}(E^\ell)\Big) = \mathcal{N}\big(0, \beta^2 I_{r^2}\big)\,, \tag{25}$$

where $\beta > 0$ controls the prior scale. Let θ collect all deterministic parameters, including $\{A^\ell, B^\ell\}_{\ell=1}^L$, the audio projection P_a, and the per-layer cross-attention parameters $\{W_Q^\ell, W_K^\ell, W_V^\ell, W_O^\ell\}_{\ell=1}^L$. Let $\phi = \{\phi^\ell\}_{\ell=1}^L$ denote the variational parameters of the inference heads. SPECTRA maximizes the ELBO:

$$\mathcal{L}(\theta,\phi) = \sum_{i=1}^{N}\left[\mathbb{E}_{q_\phi(E_{x_i}|x_i)}\log p_\theta(y_i \mid x_i, E_{x_i}) - \gamma \sum_{\ell=1}^{L}\frac{1}{T_{x_i}}\sum_{t=1}^{T_{x_i}}\mathrm{KL}\Big(q_{\phi^\ell}\Big(E^\ell_{x_i,t} \mid x_i\Big) \,\|\, p(E^\ell)\Big)\right], \tag{26}$$

with $\gamma > 0$ a KL reweighting coefficient. The complete learning objective is expressed as a summation over all samples: $\mathcal{L} = \sum_{(x,y)\in\mathcal{D}} \mathcal{L}(x,y)$. We optimize the model by minimizing the negative ELBO with respect to all trainable parameters, where the deterministic adapter parameters $A^\ell, B^\ell{}_{\ell=1}^L$ are included in θ. Equivalently, training minimizes the expected negative log-likelihood (NLL) with the KL regularization term. Averaging the KL across token positions prevents the regularization magnitude from growing trivially with sequence length, while still encouraging the contextual posterior to remain close to the prior unless the data support deviation from it. The conditioning on the cross-attended context $\tilde{u}_t^\ell(x)$ means that this deviation can systematically reflect the reliability of the accompanying audio signal.

Stochastic optimization. Using one Monte Carlo draw per sample (as in C-LoRA), we approximate the expected log-likelihood term as

$$\mathbb{E}_{q_\phi(E_{x_i}|x_i)}\log p_\theta(y_i \mid x_i, E_{x_i}) \;\approx\; \log p_\theta\Big(y_i \mid x_i, E^{(1)}_{x_i}\Big)\,, \qquad E^{(1)}_{x_i} \sim q_\phi(E_{x_i} \mid x_i). \tag{27}$$

Optimization is performed end-to-end using the reparameterization trick, updating only the low-rank adapter parameters, cross-attention projections, and variational inference heads, while keeping the backbone frozen. Monte Carlo sampling provides uncertainty estimates that reflect variability induced by both the contextual conditioning and the stochastic adapter representation.

Posterior predictive inference and cross-modal uncertainty. At test time, SPECTRA supports either a deterministic approximation using posterior means or Monte Carlo model averaging. For a test input x^* with accompanying frame-level audio sequence $a(x^*)$:

$$p(y^* \mid x^*, \mathcal{D}) \;\approx\; \frac{1}{M}\sum_{m=1}^{M} p_\theta\Big(y^* \mid x^*, E^{(m)}_{x^*}\Big)\,, \qquad E^{(m)}_{x^*} \sim q_\phi(E_{x^*} \mid x^*). \tag{28}$$

Because $q_\phi(E_{x^*} \mid x^*)$ is conditioned on both the local text features $\{z_t^\ell\}$ and the per-layer cross-attended audio context $\{\tilde{u}_t^\ell(x^*)\}$, predictive uncertainty becomes explicitly multimodal, token-sensitive, and heteroscedastic. Practically, this provides a principled mechanism for modulating confidence based on temporally localized acoustic evidence without changing the frozen backbone or introducing heavy multimodal fusion blocks.

Predictive uncertainty decomposition. SPECTRA produces predictive uncertainty through Monte Carlo marginalization over the context-conditioned stochastic adapter latent variables. For a held-out input x_i, let

$$E_i^{(m)} = \Big\{E_{i,t}^{\ell,(m)}\Big\}_{\ell=1,t=1}^{L,T_i} \sim q_\phi(E_i \mid x_i), \qquad m = 1,\dots,M,$$

denote the m-th draw from the variational posterior over the collection of stochastic low-rank adapter latents across layers and token positions. Each draw induces a categorical predictive distribution

$$\mathbf{p}_i^{(m)} = p_\theta(y_i \mid x_i, E_i^{(m)}) = \mathrm{softmax}\left(f_\theta(x_i, E_i^{(m)})\right) \in \Delta^{C-1},$$

where C is the number of classes and Δ^{C-1} denotes the probability simplex. The posterior predictive distribution is approximated by Monte Carlo averaging:

$$\bar{\mathbf{p}}_i = p(y_i \mid x_i, \mathcal{D}) \approx \frac{1}{M}\sum_{m=1}^{M}\mathbf{p}_i^{(m)}.$$

The final prediction is: $\hat{y}_i = \arg\max_{c \in \{1,\ldots,C\}} \bar{p}_{ic}$. The predictive entropy of the averaged distribution measures the total predictive uncertainty:

$$U_{\text{total}}(x_i) = H[\bar{\mathbf{p}}_i] = -\sum_{c=1}^{C} \bar{p}_{ic} \log\left(\bar{p}_{ic} + \epsilon_h\right),$$

where ϵ_h is a small constant used only for numerical stability. This quantity captures the overall uncertainty in the final prediction after marginalizing over the stochastic adapter posterior. Following the standard Bayesian classification decomposition, we further separate total predictive uncertainty into aleatoric and epistemic components. The aleatoric component is estimated as the expected entropy of the class distribution under posterior samples:

$$U_{\text{alea}}(x_i) = \mathbb{E}_{q_\phi(E_i|x_i)}\left[H[p_\theta(y_i \mid x_i, E_i)]\right] \approx \frac{1}{M}\sum_{m=1}^{M} H\left[\mathbf{p}_i^{(m)}\right],$$

with: $H\left[\mathbf{p}_i^{(m)}\right] = -\sum_{c=1}^{C} p_{ic}^{(m)} \log\left(p_{ic}^{(m)} + \epsilon_h\right)$. This term is large when individual stochastic adapter samples are themselves uncertain, which typically corresponds to intrinsically ambiguous examples, overlapping classes, or label noise. The epistemic component is measured by the mutual information between the prediction and the stochastic adapter latent:

$$U_{\text{epi}}(x_i) = I(y_i, E_i \mid x_i, \mathcal{D}) = H[\bar{\mathbf{p}}_i] - \mathbb{E}_{q_\phi(E_i|x_i)}\left[H[p_\theta(y_i \mid x_i, E_i)]\right].$$

Using the Monte Carlo approximation, this becomes

$$U_{\text{epi}}(x_i) \approx U_{\text{total}}(x_i) - U_{\text{alea}}(x_i).$$

Intuitively, $U_{\text{epi}}(x_i)$ is high when different stochastic adapter samples produce confident but disagreeing predictions, indicating uncertainty about the appropriate adapter behavior for the input. In SPECTRA, this quantity should be interpreted as adapter-space epistemic uncertainty, since stochasticity is confined to the compact latent matrix E_i in the low-rank adapter space rather than placed over all backbone parameters.

Because SPECTRA conditions $q_\phi(E_i \mid x_i)$ on both the local text-derived low-rank features and the cross-attended audio context, this decomposition provides a fine-grained view of multimodal uncertainty. When the audio-text evidence is ambiguous, weakly aligned, or unreliable, the posterior over E_i may induce greater variability across stochastic adapter samples, increasing U_{epi}. In contrast, when the input is intrinsically ambiguous but posterior samples agree in their uncertainty, U_{alea} is expected to dominate. For tasks with different numbers of classes, we use normalized uncertainty scores:

$$\widetilde{U}_{\text{total}}(x_i) = \frac{U_{\text{total}}(x_i)}{\log C}, \qquad \widetilde{U}_{\text{alea}}(x_i) = \frac{U_{\text{alea}}(x_i)}{\log C}, \qquad \widetilde{U}_{\text{epi}}(x_i) = \frac{U_{\text{epi}}(x_i)}{\log C}.$$

This normalization maps all entropy-based scores to $[0, 1]$, making uncertainty values comparable across binary and multi-class prediction tasks.

Epistemic and aleatoric uncertainty analysis. In addition to reporting predictive entropy, we decompose the uncertainty of SPECTRA into aleatoric and epistemic components using the same $M = 10$ Monte Carlo stochastic forward passes used for posterior predictive inference. For each held-out example, we compute the total uncertainty U_{total}, the expected predictive entropy U_{alea}, and the mutual-information-based adapter-space epistemic uncertainty U_{epi}. Since our benchmark contains both binary and multi-class tasks, all entropy-based quantities are normalized by $\log C$ before aggregation.

This decomposition allows us to distinguish between two qualitatively different sources of uncertainty. High aleatoric uncertainty indicates that individual sampled adapters remain uncertain, suggesting intrinsically ambiguous, noisy, or difficult-to-label examples. High epistemic uncertainty indicates disagreement across stochastic adapter samples, suggesting that the model is uncertain about the appropriate low-rank adaptation for the given audio-text input. In the context of SPECTRA, this is particularly useful because the posterior over the adapter latent is conditioned on token-level cross-modal audio context. Thus, increased epistemic uncertainty can reflect cases where the text and audio provide weak, conflicting, or unreliable evidence for adaptation. Let

$$\mathcal{I}_{\text{corr}} = \{i : \hat{y}_i = y_i\}, \qquad \mathcal{I}_{\text{err}} = \{i : \hat{y}_i \neq y_i\}.$$

For each uncertainty type $b \in \{\text{total}, \text{alea}, \text{epi}\}$, we summarize the mean uncertainty over correctly and incorrectly classified examples:

$$\bar{U}_b^{\text{corr}} = \frac{1}{|\mathcal{I}_{\text{corr}}|} \sum_{i \in \mathcal{I}_{\text{corr}}} \widetilde{U}_b(x_i),$$

$$\bar{U}_b^{\text{err}} = \frac{1}{|\mathcal{I}_{\text{err}}|} \sum_{i \in \mathcal{I}_{\text{err}}} \widetilde{U}_b(x_i).$$

We also compute the uncertainty gap

$$\Delta U_b = \bar{U}_b^{\text{err}} - \bar{U}_b^{\text{corr}},$$

where larger positive values indicate that the uncertainty score assigns higher uncertainty to misclassified examples. To evaluate whether uncertainty predicts model failure, we compute error-detection AUROC:

$$\text{AUROC}_{\text{err}}(U_b) = \text{AUROC}\left(\{\mathbb{I}[\hat{y}_i \neq y_i]\}_{i=1}^N, \left\{\widetilde{U}_b(x_i)\right\}_{i=1}^N\right),$$

where U_b is one of U_{total}, U_{alea}, or U_{epi}. A higher value indicates that the corresponding uncertainty component better identifies examples on which the model is likely to be incorrect.

Summary of novelties and differences to prior work. SPECTRA preserves C-LoRA's lightweight Bayesianization by restricting stochasticity to $E_{x,t}^{\ell} \in \mathbb{R}^{r \times r}$, keeping A^{ℓ} and B^{ℓ} deterministic and avoiding the $O(d)$-scaled Bayesian overhead typical of Bayesianizing full LoRA factors. Unlike BLoB, the posterior is amortized and input-dependent, yielding sample-specific heteroscedastic uncertainty. Unlike C-LoRA, which conditions the contextual posterior only on internal low-rank features z^{ℓ}, SPECTRA additionally incorporates per-layer, token-level text-audio cross-attention over audio frames, enabling uncertainty to reflect temporally localized cross-modal evidence while remaining parameter-efficient.

B Further Experimental Results

This section describes the experimental protocol used to evaluate the proposed SPECTRA framework and its relationship to prior contextualized low-rank adaptation methods. Our goal is to assess whether token-level cross-modal contextualization improves parameter-efficient multimodal adaptation compared to both text-only PEFT methods and global audio conditioning approaches.

Models and Modalities. We evaluate SPECTRA using multiple combinations of pre-trained text and audio encoders to ensure robustness across backbone architectures. For the text modality, we consider six sentence-level transformer encoders widely used for semantic representation learning: nliRoBERTa, paraTinyBERT, allMiniLM12, nliDistilRoBERTa, allDistilRoBERTa, and allRoBERTa-Large. These models provide diverse trade-offs between representational capacity and computational cost.

For the audio modality, we employ three large pre-trained speech representation models: whisperMedium, hubertLargeFT, and wav2vec2LargeFT. These encoders produce high-level acoustic embeddings that capture prosody, speaker characteristics, and paralinguistic cues.

Compared Methods. We compare three families of approaches representing different levels of multimodal integration and parameter-efficient adaptation:

1. **Text-only PEFT.** such as LoRA and C-LoRA that adapt the frozen LLM using low-rank updates on text only (no audio).
2. **Multimodal transfer learning with feature fusion.** we obtain a pooled audio embedding from an audio encoder and fuse it with the text representation (e.g., concatenation), followed by a task-specific classification head (3-layer MLP, 32 units followed by 16 units followed by a softmax). This is a conventional lightweight multimodal baseline that injects audio at the representation/head level.
3. **Multimodal context-conditioned PEFT (ours).** instead of heavy feature fusion, we use audio as context that conditions the variational low-rank adapter uncertainty and adaptation throughout the transformer.

SPECTRA Variants and Ablations. To better understand the contribution of the proposed contextualized adapter mechanism, we evaluate several variants of SPECTRA that differ in how audio information is incorporated into the LoRA adaptation process. These ablations allow us to isolate the effect of cross-modal contextualization and identify which architectural components contribute most to performance improvements.

1. **SPECTRA-X (Cross-Attention).** The primary variant uses token-level cross-attention between text tokens and frame-level audio representations within each LoRA layer (as in Figure 1). Text token representations act as queries while audio frames provide keys and values. The resulting contextual signal is projected into the LoRA latent space and used to condition the variational adapter parameters. This design enables fine-grained alignment between linguistic tokens and acoustic events.
2. **SPECTRA-X (Shared-KV).** To reduce parameter overhead, we evaluate a variant where the key and value projections of the cross-attention module are shared across transformer layers. In this configuration, the audio representation is projected once at the model level and reused by all LoRA layers, while each layer maintains its own query projection and output mapping. This design significantly reduces the number of additional parameters while preserving layer-specific contextualization.
3. **SPECTRA-G (Global Context).** We use a single global audio embedding to condition the adapters. Comparing SPECTRA-G and -X isolates the benefit of moving from global contextualization to token-level cross-modal alignment.

Together, these variants allow us to analyze the impact of architectural choices on performance, parameter efficiency, and multimodal representation quality.

Training and Evaluation Protocol. All PEFT methods, including LoRA, C-LoRA, and the SPECTRA variants, use rank $r = 8$ and scaling factor $\alpha = 32$. Only the adapter parameters are updated, while all text and audio backbone encoders remain frozen. Models are trained for 50 epochs using AdamW with learning rate 10^{-3} and weight decay 10^{-3}. For SPECTRA-G, pooled audio embeddings are projected into a 16-dimensional context space using a shared MLP, followed by layer-specific linear mappings into the 8-dimensional latent adapter space. For SPECTRA-X, frame-level audio representations are used instead of pooled embeddings, and a lightweight multi-head cross-attention module performs token-level alignment between text tokens and audio frames in a low-dimensional latent space. The Bayesian adapter formulation follows the same variational parameterization as C-LoRA, with prior and regularization hyperparameters set to $\beta = 0.2$, $\varepsilon = 0.05$, and $\gamma = 0.008$. During inference, predictive uncertainty is estimated using Monte Carlo sampling with $M = 10$ stochastic forward passes. All experiments are conducted using 5-fold cross-validation with speaker-level separation to prevent leakage between training and test folds. Performance is reported using area under the ROC curve (AUC%), which is appropriate for the imbalanced label distributions in the clinical interview dataset. Final results are averaged across the five folds for each task.

B.1 Experimenal Results

Modality disagreement and calibration stress tests. A central motivation of SPECTRA is that audio should not be treated merely as an additional feature stream, but also as a contextual reliability signal that modulates adaptation and predictive confidence. In ordinary multimodal evaluation, however, the text and audio modalities are naturally paired and usually contain mutually consistent evidence. Such evaluation measures predictive performance under matched conditions, but it does not directly test whether the model responds appropriately when one modality becomes unreliable or contradicts the other. To evaluate this behavior, we introduce a modality-disagreement stress test that perturbs the relationship between the textual and acoustic inputs at test time. Let the held-out evaluation set be $\mathcal{D}_{\text{test}} = \{(x_i^{\text{text}}, x_i^{\text{audio}}, y_i)\}_{i=1}^{N}$, where x_i^{text} denotes the transcript, x_i^{audio} denotes the corresponding audio segment, and y_i is the original label. We evaluate the same trained SPECTRA model under four input conditions: $\mathcal{C} = \{c_{\text{clean}}, c_{\text{audio-noise}}, c_{\text{audio-mismatch}}, c_{\text{text-corrupt}}\}$.

The clean condition preserves the original paired input as $c_{\text{clean}} : (x_i^{\text{text}}, x_i^{\text{audio}})$. The noisy-audio condition preserves the correct text-audio pairing but corrupts the acoustic signal as $c_{\text{audio-noise}} : (x_i^{\text{text}}, \tau_{\text{noise}}(x_i^{\text{audio}}))$, where $\tau_{\text{noise}}(\cdot)$ denotes the audio corruption operator. This condition tests robustness to degraded but still semantically matched acoustic evidence. The mismatched-audio

Table 6: Modality-disagreement stress-test results. Higher AUC is better, while lower NLL, ECE, and Brier score are better. The same held-out examples are evaluated under each condition.

Condition	**AUC** $\uparrow$	**NLL** $\downarrow$	**ECE** $\downarrow$	**Brier** $\downarrow$
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

condition preserves the clean transcript but replaces the corresponding audio with an audio segment from another held-out example as $c_{\text{audio-mismatch}} : (x_i^{\text{text}}, x_{\pi(i)}^{\text{audio}})$ with $\pi(i) \neq i$. This condition intentionally destroys the correspondence between linguistic and acoustic evidence, thereby testing whether the model over-relies on misleading audio context. Finally, the corrupted-text condition perturbs the transcript while keeping the original audio clean as $c_{\text{text-corrupt}} : (\tau_{\text{text}}(x_i^{\text{text}}), x_i^{\text{audio}})$, where $\tau_{\text{text}}(\cdot)$ denotes the text corruption operator. This condition evaluates whether clean acoustic information can partially compensate when the lexical stream is degraded. Importantly, the target label y_i is kept fixed in all conditions. Therefore, performance differences reflect the model's sensitivity to modality reliability rather than changes in the prediction task.

For each condition c, we compute the posterior predictive distribution using Monte Carlo model averaging over the stochastic adapter latent: $p_{i,c}^{(m)} = p_\theta\left(y \mid x_{i,c}^{\text{text}}, x_{i,c}^{\text{audio}}, E_{i,c}^{(m)}\right)$, with $E_{i,c}^{(m)} \sim q_\phi\left(E \mid x_{i,c}^{\text{text}}, x_{i,c}^{\text{audio}}\right)$, and $\bar{p}_{i,c} = \frac{1}{M}\sum_{m=1}^{M} p_{i,c}^{(m)}$. We then evaluate discrimination, probabilistic accuracy, and calibration using AUC, negative log-likelihood (NLL), expected calibration error (ECE), and Brier score: $\text{NLL}_c = -\frac{1}{N}\sum_{i=1}^{N} \log \bar{p}_{i,c,y_i}$ and $\text{Brier}_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\bar{p}_{i,c,k} - \mathbb{1}\{y_i = k\}\right)^2$.

For calibration, we define the predicted confidence as $\text{conf}_{i,c} = \max_k \bar{p}_{i,c,k}$ and partition the predictions into B confidence bins $\{\mathcal{B}_b\}_{b=1}^{B}$. The expected calibration error is $\text{ECE}_c = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{N} |\text{acc}(\mathcal{B}_b) - \text{conf}(\mathcal{B}_b)|$, where $\text{acc}(\mathcal{B}_b) = \frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b} \mathbb{1}\left\{y_i = \arg\max_k \bar{p}_{i,c,k}\right\}$ and $\text{conf}(\mathcal{B}_b) = \frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b} \text{conf}_{i,c}$.

Uncertainty decomposition. To determine whether modality perturbations primarily affect irreducible data uncertainty or model uncertainty, we decompose predictive uncertainty into total, aleatoric, and epistemic components. Let $H(p) = -\sum_{k=1}^{K} p_k \log p_k$ denote predictive entropy. The normalized total uncertainty is $U_{i,c}^{\text{total}} = \frac{H(\bar{p}_{i,c})}{\log K}$. The normalized aleatoric component is estimated by the expected entropy of individual stochastic predictions as $U_{i,c}^{\text{alea}} = \frac{1}{M}\sum_{m=1}^{M} \frac{H(p_{i,c}^{(m)})}{\log K}$. The epistemic component is then estimated by the mutual information between predictions and the stochastic adapter latent as $U_{i,c}^{\text{epi}} = U_{i,c}^{\text{total}} - U_{i,c}^{\text{alea}}$. This decomposition is particularly relevant for SPECTRA, because the model's stochasticity is concentrated in the low-rank adapter-space latent E. Hence, increases in U^{epi} indicate that the model's predictions are more sensitive to posterior samples of the adapter, while increases in U^{alea} indicate that the individual sampled predictors are themselves less confident.

We also measure the aggregate posterior dispersion of the adapter latent as $U_{i,c}^{E} = \frac{1}{LT_i r^2}\sum_{\ell=1}^{L}\sum_{t=1}^{T_i}\sum_{j=1}^{r^2} \sigma_{E,t,j}^{\ell}\left(x_{i,c}^{\text{text}}, x_{i,c}^{\text{audio}}\right)$, where $\sigma_{E,t,j}^{\ell}$ denotes the posterior standard deviation of the j-th entry of $\text{vec}(E_{i,t}^{\ell})$. This quantity measures whether modality perturbations directly increase the inferred posterior scale in the Bayesian adapter space.

Discrimination and probabilistic performance. Table 6 and Figure 3 show that the clean matched condition achieves the strongest overall performance, with an AUC of 0.835, NLL of 0.419, ECE of 0.042, and Brier score of 0.260. This represents the standard evaluation setting where textual and acoustic evidence are correctly paired.

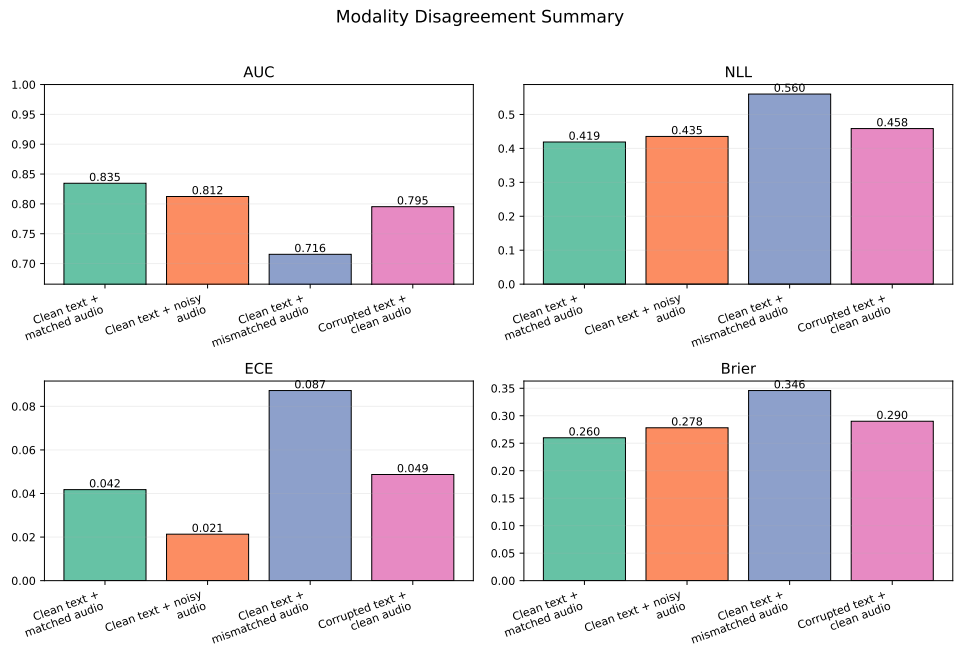


Figure 3: Summary of predictive performance under modality-disagreement stress tests on FORBOW clinical dataset. Clean matched text-audio pairs achieve the best overall performance. Noisy audio causes a moderate decrease in AUC and mild increases in NLL and Brier score. Mismatched audio produces the largest degradation across all metrics, indicating that contradictory cross-modal evidence is substantially more harmful than simple acoustic degradation. Corrupted text with clean audio leads to an intermediate degradation, suggesting that clean acoustic information partially compensates for degraded lexical input.

When the audio is corrupted with noise but remains correctly paired with the text, AUC decreases moderately from 0.835 to 0.812, and both NLL and Brier score increase slightly. This indicates that acoustic degradation reduces discriminative and probabilistic quality, but the model remains relatively robust because the linguistic stream is still clean and the audio retains its original semantic correspondence. Interestingly, ECE decreases from 0.042 to 0.021. This should not be interpreted as an overall improvement in prediction quality, since AUC, NLL, and Brier score all deteriorate. Rather, the lower ECE suggests that the model becomes more conservative under noisy audio, producing confidence estimates that are better aligned with empirical accuracy even though the predictions themselves are less accurate.

The most severe degradation occurs under clean text with mismatched audio. AUC drops to 0.716, NLL increases to 0.560, ECE increases to 0.087, and Brier score increases to 0.346. This condition is substantially harder than noisy audio because the acoustic modality is not merely degraded; it is actively misleading. The model receives clean lexical evidence but audio context that belongs to another example. The large drop in AUC and the sharp increases in NLL and Brier score show that incorrect cross-modal pairing can harm both ranking quality and probability estimation. The increase in ECE further indicates that mismatched audio produces residual overconfidence: the model sometimes assigns high confidence to predictions whose empirical accuracy is lower than expected.

Under corrupted text with clean audio, performance degrades less severely than under mismatched audio. AUC decreases from 0.835 to 0.795, NLL increases to 0.458, ECE increases slightly to 0.049, and Brier score rises to 0.290. This suggests that clean acoustic information can partially compensate for degraded lexical evidence, but not fully recover the performance of the clean matched condition. The asymmetry between the mismatched-audio and corrupted-text conditions is informative: misleading audio is more damaging than degraded text in this setting, suggesting that the model uses audio context strongly enough that incorrect acoustic evidence can distort the posterior adapter behavior.

Calibration under modality disagreement. Figure 4 shows reliability diagrams for all four conditions. In the clean matched setting, predicted confidence generally follows empirical accuracy, yielding an ECE of 0.042. This indicates that SPECTRA's posterior predictive probabilities are reasonably calibrated when the modalities are correctly paired.

The noisy-audio condition achieves the lowest ECE, 0.021, despite lower AUC and higher NLL than the clean condition. The reliability curve remains close to the diagonal, suggesting that the model appropriately reduces confidence when the acoustic signal becomes less reliable. This behavior is desirable in uncertainty-aware multimodal learning: degradation in one modality should not only affect the predicted label, but should also be reflected in predictive confidence.

The mismatched-audio condition exhibits the worst calibration, with ECE increasing to 0.087. The reliability curve lies noticeably below the diagonal in high-confidence regions, indicating overconfidence under cross-modal contradiction. This is an important finding because it reveals a failure mode that is not visible from standard clean evaluation alone. Although SPECTRA incorporates audio as a posterior-conditioning signal, the model can still be miscalibrated when the audio is plausible but semantically inconsistent with the text. Therefore, mismatched audio provides a stronger test of cross-modal reliability than noisy audio.

The corrupted-text condition yields ECE 0.049, only slightly worse than the clean matched condition. This suggests that the model remains moderately calibrated when the lexical stream is degraded but the audio remains valid. The reliability curve is closer to the diagonal than in the mismatched-audio setting, supporting the interpretation that valid acoustic evidence helps stabilize predictions when text quality decreases.

Uncertainty decomposition. Figure 5 decomposes predictive uncertainty into total, aleatoric, and epistemic components. In the clean matched condition, normalized total uncertainty is 0.465, with aleatoric uncertainty 0.325 and epistemic uncertainty 0.139. These values provide the reference point for interpreting the perturbation-induced uncertainty shifts.

Noisy audio produces the largest increase in total uncertainty, from 0.465 to 0.568. This corresponds to a shift of $+0.104$, decomposed into $+0.063$ aleatoric and $+0.040$ epistemic uncertainty. This pattern is consistent with the nature of the perturbation: noise reduces the clarity of the acoustic evidence, making individual sampled predictors less confident, while also increasing disagreement across posterior adapter samples. The simultaneous increase in aleatoric and epistemic uncertainty suggests that the model recognizes noisy audio as both less informative and less stable for adaptation.

Mismatched audio increases total uncertainty more modestly, from 0.465 to 0.488, with an epistemic shift of $+0.016$. This indicates that the model detects some additional posterior uncertainty under cross-modal conflict, but the increase is smaller than the degradation observed in AUC, NLL, ECE, and Brier score. This mismatch between performance degradation and uncertainty increase is important: it suggests that semantically incorrect audio can mislead the model without always producing a sufficiently large uncertainty response. In other words, mismatched audio exposes a residual overconfidence failure mode.

Corrupted text with clean audio produces an intermediate increase in total uncertainty, from 0.465 to 0.513, with aleatoric and epistemic shifts of $+0.020$ and $+0.028$, respectively. The larger epistemic component relative to the aleatoric shift suggests that corrupting the lexical input increases uncertainty about how the adapter should adapt, while clean audio still provides useful evidence that prevents the larger degradation observed under mismatched audio.

Overall, the decomposition shows that all modality perturbations increase predictive uncertainty relative to the clean matched condition. However, the type of uncertainty depends on the perturbation. Noisy audio primarily increases both aleatoric and epistemic uncertainty, corrupted text produces a moderate epistemic increase, and mismatched audio causes the largest performance and calibration degradation despite only a modest uncertainty increase.

Predictive entropy and adapter posterior response. Perturbed conditions shift the uncertainty distribution relative to the clean matched baseline, consistent with the aggregate decomposition in Figure 5.

The right panel reports the mean posterior standard deviation of the stochastic adapter latent E. The aggregate value remains nearly constant across conditions, approximately 0.047-0.048. This result is

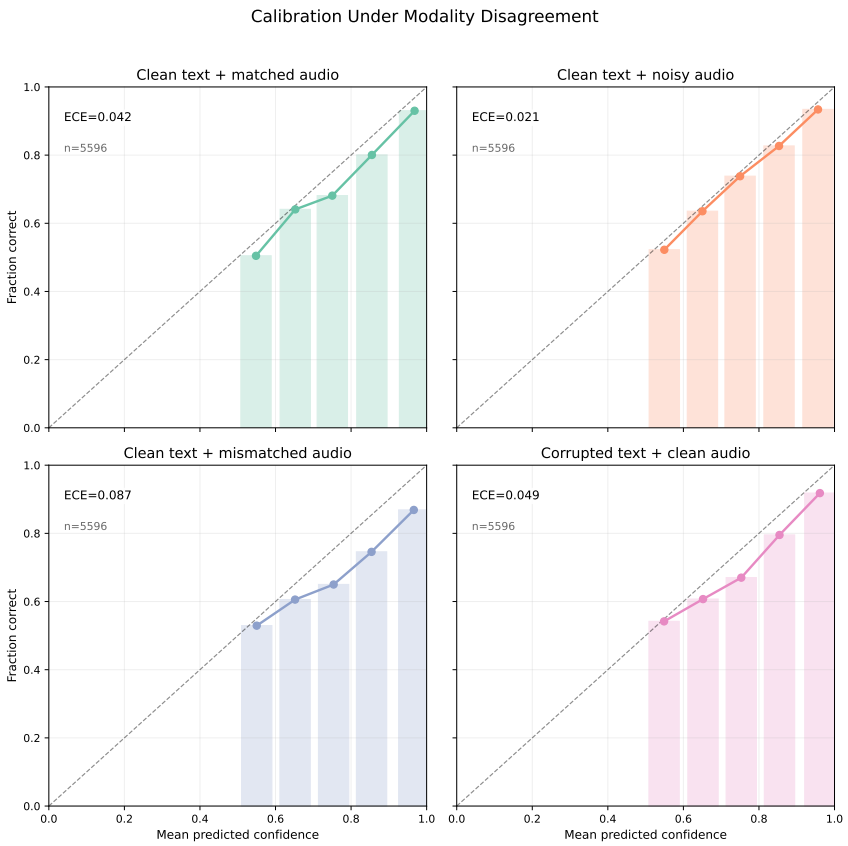


Figure 4: Reliability diagrams under modality disagreement on FORBOW clinical dataset. Each panel reports calibration over the same $N = 5596$ held-out examples. The clean matched condition is well calibrated overall. Noisy audio produces the lowest ECE, suggesting that the model becomes more conservative under acoustic degradation. Mismatched audio leads to the largest ECE and clear overconfidence in high-confidence bins, showing that semantically inconsistent audio is more harmful to calibration than noisy but correctly paired audio. Corrupted text with clean audio produces only a mild calibration degradation.

informative. It suggests that the observed changes in predictive uncertainty are not primarily driven by a large global increase in the average posterior scale σ_E. Instead, the uncertainty response may arise from changes in posterior means, downstream class probabilities, token-specific effects that are averaged out by the global statistic, or localized layer-wise changes in σ_E that are not visible after averaging over all layers, tokens, and rank-space dimensions.

This finding also highlights an important direction for improving SPECTRA. Although the current model produces meaningful predictive uncertainty shifts under modality perturbations, the aggregate adapter posterior variance does not strongly separate clean, noisy, mismatched, and text-corrupted inputs. A more explicit reliability-aware posterior could encourage the adapter variance itself to respond monotonically to cross-modal reliability. For example, one could introduce a learned reliability score $\rho_t^\ell(x) \in [0, 1]$ and define $\sigma_{E,t}^\ell(x)^2 = \sigma_{\phi,t}^\ell(x)^2 + \lambda \left(1 - \rho_t^\ell(x)\right)$, so that unreliable or contradictory cross-modal evidence directly increases the posterior variance in the adapter space. The present stress test therefore serves two purposes: it validates that SPECTRA's predictive uncertainty responds to modality degradation, and it reveals that future versions should make the adapter posterior variance more explicitly reliability-sensitive.

Discussion. The modality-disagreement experiments provide a stronger test of uncertainty-aware multimodal adaptation than standard clean evaluation. The results support three main conclusions. First, SPECTRA is robust to moderate acoustic degradation: noisy audio reduces AUC and increases NLL and Brier score, but the model remains comparatively well calibrated and increases predictive

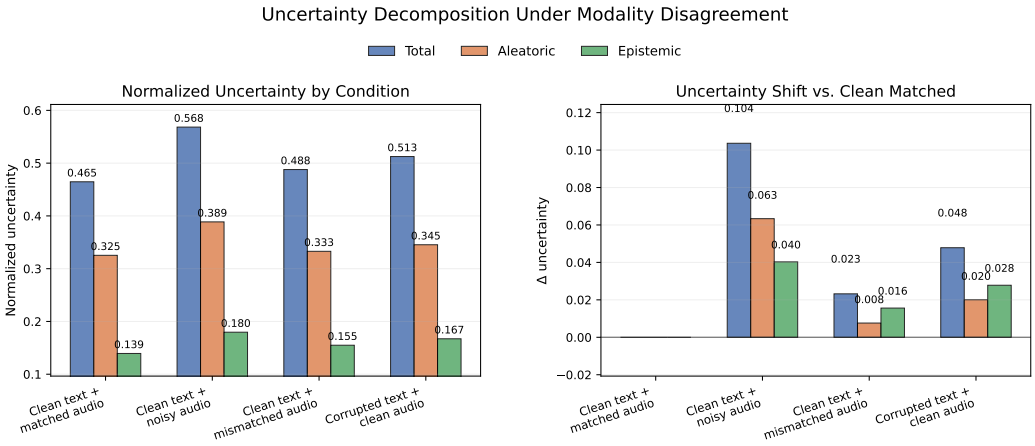


Figure 5: Uncertainty decomposition under modality disagreement on FORBOW clinical dataset. Left: normalized total, aleatoric, and epistemic uncertainty for each condition. Right: uncertainty shift relative to the clean matched condition. Noisy audio produces the largest total uncertainty increase, with increases in both aleatoric and epistemic components. Corrupted text with clean audio yields an intermediate increase, especially in epistemic uncertainty. Mismatched audio produces the largest performance degradation but only a modest uncertainty increase, revealing a residual overconfidence failure mode under semantically inconsistent audio.

uncertainty. Second, semantically mismatched audio is substantially more harmful than noisy but correctly paired audio. This confirms that the model uses cross-modal context in a meaningful way, but it also reveals a calibration failure mode under contradictory modalities. Third, clean audio partially compensates for corrupted text, as the corrupted-text condition performs better than the mismatched-audio condition and remains reasonably calibrated.

These results strengthen the central claim that audio can function as a reliability signal for Bayesian low-rank adaptation. At the same time, they show that the current version of SPECTRA is more responsive at the predictive-distribution level than at the aggregate adapter-posterior-variance level. This distinction is important: a model can become less confident in its class probabilities without necessarily increasing the global posterior standard deviation of the adapter latent. Therefore, future extensions should incorporate explicit cross-modal reliability estimation, layer-wise variance diagnostics, and disagreement-aware posterior regularization to ensure that modality conflict is reflected not only in predictive entropy but also in the Bayesian adapter posterior itself.

Table 7 shows the evaluation results on the offspring data. Results on offspring data reveal a more heterogeneous pattern than the parent setting, but overall it still supports the value of multimodal contextualization, especially for the stronger text backbones. For nliRoBERTa and allDistillRoBERTa, SPECTRA remains consistently competitive and often achieves the best results on several clinically relevant tasks, including fear, self-criticism, criticism, sadness, rumination, and coherence. In particular, SPECTRA-X yields the top performance on 6 tasks, indicating that token-level cross-modal conditioning can again improve over both text-only PEFT and simple transfer-based fusion when the textual backbone is sufficiently expressive. At the same time, the offspring results also highlight a sharper dependence on backbone quality and task difficulty. For weaker encoders such as paraTinyBERT and allMiniLM12, SPECTRA substantially boosts sentiment into the mid-90s, but often degrades performance on sparse affective and psychological targets relative to the transfer baselines, suggesting that the additional flexibility of contextual Bayesian conditioning is harder to exploit when the underlying text representation is weak and the labels are extremely imbalanced. In these settings, strong transfer baselines with whisperMedium remain highly competitive. Overall, the offspring results suggest that SPECTRA is most effective when paired with sufficiently strong language encoders and when the task benefits from localized cross-modal reliability cues, whereas for weaker backbones or highly sparse targets, simpler multimodal fusion may sometimes provide a more stable inductive bias.

Figure 6 visualizes how the model allocates cross-modal attention over an input utterance. The upper panel shows the mel-spectrogram of the speech signal, while the lower panel shows the corresponding

Table 7: AUC (%) on a 5-fold cross-validation on segment-level predictions on offspring data.

Method	LLM	Audio Encoder	**sentiment**	**anger**	**fear**	**joy**	**sadness**	**richness**	**coherence**	**rumination**	**worry**	**anxiousness**	**criticism**	**self-criticism**
LoRA	nRoBERTa	-	93.88	87.36	73.03	83.36	81.42	83.99	78.82	51.05	64.98	66.81	93.12	93.82
C-LoRA	nRoBERTa	-	93.38	85.77	80.65	80.00	82.73	82.57	77.17	66.81	87.35	64.91	91.66	90.02
Transfer	nRoBERTa	whisperM	93.19	**88.37**	87.90	**86.63**	84.88	84.22	82.34	83.46	87.93	80.90	92.52	92.58
		wav2vec2LFT	93.03	87.48	73.82	83.59	82.31	83.54	80.51	73.44	86.72	75.24	91.31	91.70
		hubertLFT	92.59	87.13	86.24	83.29	82.11	83.39	79.64	79.07	87.03	72.57	90.84	92.21
SPECTRA-G	nRoBERTa	whisperM	94.65	86.16	88.83	84.16	73.73	83.83	81.36	75.56	76.99	63.75	92.20	90.19
		wav2vec2LFT	94.06	86.05	80.95	81.00	83.53	83.91	78.79	76.33	87.51	70.86	91.26	92.98
		hubertLFT	94.10	86.95	83.55	81.32	79.61	83.25	78.36	84.15	87.50	65.72	93.27	91.05
SPECTRA-X	nRoBERTa	whisperM	94.27	86.21	**90.12**	84.38	73.81	84.91	81.92	75.86	77.25	64.4	92.46	91.57
		wav2vec2LFT	93.54	86.52	81.17	81.34	84.45	84.8	79.29	77.91	88.76	71.89	91.95	**94.31**
		hubertLFT	93.91	87.05	83.8	81.86	80.02	84.17	79.03	85.43	87.71	66.72	**94.71**	91.61
LoRA	pTinyBERT	-	90.92	53.15	50.25	51.92	50.30	72.80	72.50	60.17	50.03	52.86	50.00	50.00
C-LoRA	pTinyBERT	-	91.79	65.82	51.92	53.92	52.36	76.32	75.83	52.63	51.53	55.78	62.60	55.07
Transfer	pTinyBERT	whisperM	90.86	86.08	84.29	85.66	83.69	85.09	82.22	83.73	83.40	79.47	90.32	89.98
		wav2vec2LFT	90.78	84.83	74.03	81.31	79.98	82.67	79.56	75.44	82.01	71.68	87.18	88.43
		hubertLFT	90.54	84.60	70.41	81.87	78.13	80.51	80.38	78.71	84.84	72.00	88.27	86.56
SPECTRA-G	pTinyBERT	whisperM	**95.06**	63.41	55.11	54.37	58.71	78.99	81.53	53.00	58.51	56.63	67.66	54.78
		wav2vec2LFT	94.86	62.45	55.37	57.72	63.98	78.69	82.25	53.00	55.00	66.73	75.55	61.74
		hubertLFT	94.64	79.64	53.00	56.72	64.00	79.37	81.32	54.50	53.54	60.57	75.86	56.03
SPECTRA-X	pTinyBERT	whisperM	94.66	64.41	56.13	54.97	59.21	79.38	82.04	53.64	59.56	57.44	68.61	54.91
		wav2vec2LFT	94.11	63.59	56.06	58.45	64.32	79.59	82.92	54.03	55.54	67.06	76.25	62.47
		hubertLFT	94.32	80.37	52.66	57.16	64.89	80.17	81.91	55.6	54.01	61.56	76.58	56.62
LoRA	aMiniLM12	-	80.51	50.00	50.00	50.00	52.97	71.40	69.07	59.62	50.20	51.42	50.00	50.00
C-LoRA	aMiniLM12	-	91.64	52.23	53.25	51.19	51.87	74.24	72.82	51.43	50.60	51.24	51.29	51.25
Transfer	aMiniLM12	whisperM	75.96	79.29	80.12	80.81	75.73	**85.11**	81.98	68.45	75.09	79.84	79.47	83.45
		wav2vec2LFT	81.76	75.10	72.52	71.75	64.75	78.89	79.10	62.52	66.45	72.02	77.42	77.76
		hubertLFT	80.84	74.22	69.86	74.06	66.94	79.33	79.18	68.26	66.13	70.23	74.85	76.94
SPECTRA-G	aMiniLM12	whisperM	94.70	57.13	53.79	54.47	54.13	77.39	81.28	53.98	53.89	58.39	53.19	53.80
		wav2vec2LFT	94.10	53.99	53.00	55.45	54.23	77.43	83.44	53.27	58.60	64.31	53.72	53.08
		hubertLFT	94.63	55.76	53.37	55.27	56.06	77.52	83.06	53.69	53.00	59.03	55.67	53.43
SPECTRA-X	aMiniLM12	whisperM	94.94	58.73	54.85	54.77	53.76	78.03	82.29	55.24	54.2	59.28	54.07	54.58
		wav2vec2LFT	93.85	54.81	54.42	55.84	54.41	77.37	83.72	54.22	59.42	65.28	54.35	53.07
		hubertLFT	94.21	56.89	53.77	56.04	57.31	78.21	83.85	53.8	53.54	59.49	55.93	54.16
LoRA	aDRoBERTa	-	92.91	86.13	74.92	82.40	84.16	84.32	79.67	75.47	87.85	73.35	92.17	93.04
C-LoRA	aDRoBERTa	-	92.58	82.76	81.89	78.76	81.86	84.13	78.73	72.91	86.68	69.79	92.01	91.36
Transfer	aDRoBERTa	whisperM	85.62	82.10	83.84	82.06	73.47	83.85	81.77	70.05	80.54	**81.57**	83.90	83.83
		wav2vec2LFT	84.86	77.93	75.00	73.41	69.89	80.11	79.08	61.88	71.29	71.74	80.05	81.22
		hubertLFT	84.47	78.35	70.11	76.10	71.19	79.01	78.18	68.18	65.96	71.80	78.71	76.86
SPECTRA-G	aDRoBERTa	whisperM	93.65	81.82	78.77	83.41	84.27	84.75	81.39	85.99	87.54	66.68	93.94	93.04
		wav2vec2LFT	93.59	85.56	76.64	80.41	83.06	84.68	82.59	75.14	**89.28**	75.99	93.93	93.32
		hubertLFT	93.64	84.14	81.71	79.34	83.32	84.92	81.49	74.64	88.53	70.37	93.58	91.78
SPECTRA-X	aDRoBERTa	whisperM	94.46	82.32	79.72	84.39	**85.12**	84.33	82.39	**86.91**	87.24	66.72	94.4	93.56
		wav2vec2LFT	93.34	86.13	77.46	80.43	83.89	83.15	**83.93**	75.86	88.66	76.48	93.62	94.0
		hubertLFT	94.39	84.87	82.58	80.3	83.52	84.43	82.53	75.58	88.9	71.55	94.44	92.28

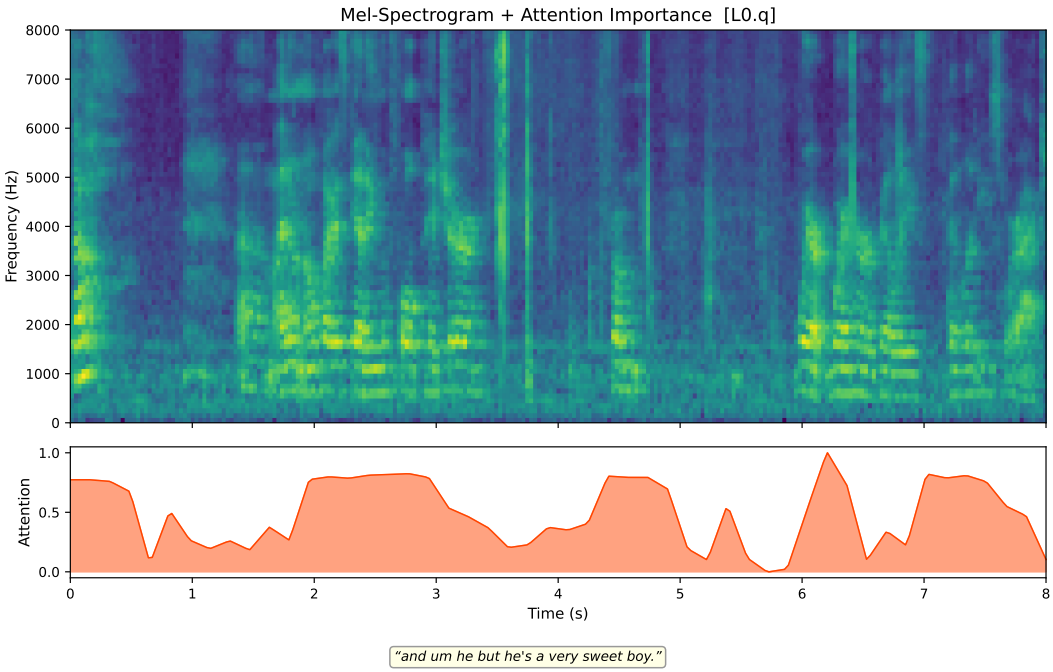


Figure 6: Mel-spectrogram of a random datapoint with token-level cross-modal attention over time.

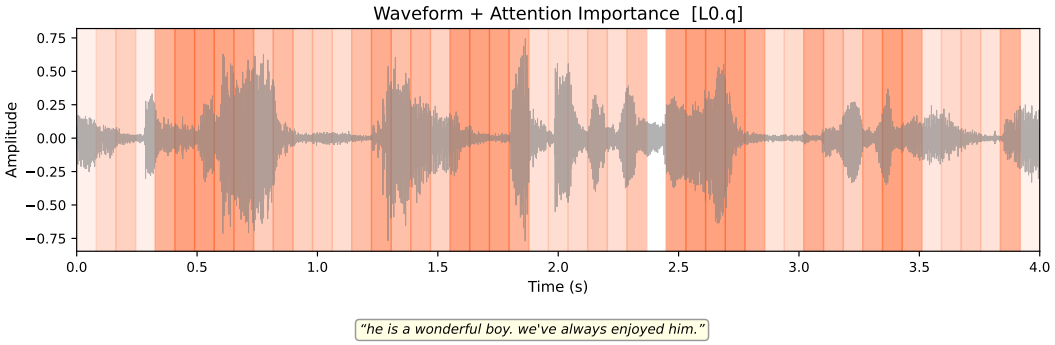


Figure 7: Waveform of a random datapoint with token-level cross-modal attention over time.

token-level attention importance across time. Peaks in the attention curve indicate temporal regions of the audio that the model considers more informative when conditioning the text representation. This illustrates that the proposed cross-attention mechanism is able to focus selectively on localized acoustic segments rather than relying on a single global audio summary.

Figure 7 shows the temporal alignment between the speech waveform and the model's token-level cross-modal attention. The gray trace represents the input audio waveform, while the shaded orange regions indicate the relative attention assigned to different time intervals. Stronger shading highlights portions of the utterance that the model considers more informative when conditioning the text representation. This visualization suggests that the proposed cross-attention mechanism selectively emphasizes localized acoustic regions, supporting the claim that SPECTRA uses fine-grained audio context rather than a single pooled summary.

C Complementary Information on FORBOW Dataset

In this section, we provide more details on the collected data. As mentioned in the main manuscript, we transcribed and broke down each sample into multiple segments based on changes in emotion, sentiment, objectivity/subjectivity, etc. Average word count in a segment is 17 and average audio

Table 8: Dataset information for the segment-level tasks in parent and offspring data.

	Task	# Samples	# Classes	Imbalanced (%)	Label 0	Label 1	Label 2	Label 3
Parent	objective	23693	2	19.08	13096	10597	-	-
Parent	sentiment	23479	3	57.39	4725	11088	7666	-
Parent	anger	23694	2	92.22	21984	1710	-	-
Parent	fear	23694	2	95.29	22628	1066	-	-
Parent	joy	23690	2	66.58	17756	5934	-	-
Parent	sadness	23694	2	92.70	22081	1613	-	-
Parent	neutral	23693	2	16.28	10797	12896	-	-
Parent	cohesion	23694	2	81.32	19964	3730	-	-
Parent	rumination	23694	2	98.95	23447	247	-	-
Parent	overinclusive	23694	2	97.20	23049	645	-	-
Parent	worry	23694	2	91.76	21890	1804	-	-
Parent	criticism	23694	2	88.42	21235	2459	-	-
Offspring	objective	10319	2	1.90	5209	5110	-	-
Offspring	sentiment	10222	3	52.51	2404	5062	2756	-
Offspring	richness	10273	3	97.08	1366	8654	253	-
Offspring	reference	10328	4	77.89	2072	4970	2187	1099
Offspring	irrelevance	10301	2	87.98	9196	1105	-	-
Offspring	anger	10329	2	95.55	9889	440	-	-
Offspring	fear	10329	2	99.04	10231	98	-	-
Offspring	joy	10329	2	94.49	9790	539	-	-
Offspring	sadness	10329	2	94.62	9802	527	-	-
Offspring	neutral	10328	2	80.92	1655	8673	-	-
Offspring	coherence	10329	2	68.35	7846	2483	-	-
Offspring	rumination	10329	2	99.01	10228	101	-	-
Offspring	worry	10329	2	99.13	10240	89	-	-
Offspring	anxiousness	10317	2	96.51	9969	348	-	-
Offspring	aggression	2182	2	99.49	2171	11	-	-
Offspring	criticism	8090	2	95.04	7708	382	-	-
Offspring	self-criticism	10329	2	96.77	10006	323	-	-

length for a segment is 6.47 seconds. Six multidisciplinary researchers rated these segments for subjectivity, emotions, sentiment, criticism, etc. 5,818 segments were rated by two or more researchers and the intraclass correlation for ratings of different researchers was high showing strong agreement in the labeling. Table 8 show detailed information about the dataset.

Researchers also rated affect, warmth, overprotection, cohesion, criticism, and worry at the document-level (i.e. for each audio sample). Document-level assessments are provided as nominal ratings. Table 9 shows the statistics of the dataset, all the tasks, and the distribution of labels. For certain tasks, some documents were missing the corresponding label, or some labels were extremely imbalanced such there were only 1 or 2 examples from one category, and we removed those cases. This is the reason for the differences in the number of samples in Table 9. The imbalanced % is calculated as $100 * (\max_{c \in C} n_c - \min_{c \in C} n_c) / \max_{c \in C} n_c$. As we can see most of the tasks are quite imbalanced.

In addition to the segment-level and document-level tasks that were labeled by the researchers, there are also psychological and cognitive tasks that are labeled by psychiatrists. The 'spectrum' task is a 4-class categorization of normal/control, depression, Bipolar, and Schizophrenia. In addition to that, we have introvert, extrovert, ADHD, anxiety, and separate binary depression labels. The statistics of these tasks, the distribution of labels and imbalancedness are also shown in Table 9.

The summary of information and statistics about the segment-level, document-level, and psychological and cognitive tasks on parent and offspring data are shown in Tables 10 and 11, respectively

Figure 8 illustrates the changes in mood, emotions, and sentiment in four random samples in the parent data. We can also see sample segments in those files shown at the top of the figures which gives an idea of how the segmentation and the labeling is done. To protect participant privacy, all text excerpts shown throughout this paper are paraphrased to remove potentially identifying details. These modifications do not affect the underlying segmentation, emotion labels, or sentiment trajectories used in the analysis.

A comparison between the ground-truth and predicted sentiment-emotion trajectories for sentenceT5XL (the best-performing text-based model) on a random sample is illustrated in Figure 9. The model captures the overall sentiment patterns across segments effectively, including major polarity shifts. The predicted emotion signals are generally consistent with the underlying seman-

Table 9: Dataset information for the document-level and psychological and cognitive prediction tasks in parent and offspring data.

	Level	Task	Num Samples	Num Classes	Imbalanced (%)	Label 0	Label 1	Label 2	Label 3	Label 4
Parent	Document	affect	484	3	93.18	23	124	337	-	-
Parent	Document	warmth	485	3	39.69	117	194	174	-	-
Parent	Document	overprotection	487	3	85.77	267	182	38	-	-
Parent	Document	cohesion	492	5	92.63	14	14	132	190	142
Parent	Document	criticism	492	4	89.22	232	166	69	25	-
Parent	Document	worry	381	4	93.29	164	147	59	11	-
Parent	Cognitive	spectrum	363	4	87.25	129	149	66	19	-
Parent	Cognitive	introvert	369	2	49.39	245	124	-	-	-
Parent	Cognitive	extrovert	369	2	63.84	271	98	-	-	-
Parent	Cognitive	ADHD	369	2	67.74	279	90	-	-	-
Parent	Cognitive	anxiety	369	2	51.81	249	120	-	-	-
Parent	Cognitive	depression	369	2	89.19	333	36	-	-	-
Offspring	Document	affect	621	4	97.73	16	11	109	485	-
Offspring	Document	coherence	616	5	95.83	11	33	109	264	199
Offspring	Document	richness	614	3	96.70	82	515	17	-	-
Offspring	Cognitive	spectrum	85	4	86.11	36	27	17	5	-
Offspring	Cognitive	introvert	85	2	39.62	53	32	-	-	-
Offspring	Cognitive	extrovert	85	2	53.45	58	27	-	-	-
Offspring	Cognitive	ADHD	85	2	53.45	58	27	-	-	-
Offspring	Cognitive	anxiety	85	2	39.62	53	32	-	-	-
Offspring	Cognitive	depression	85	2	83.56	73	12	-	-	-

tic content, although some variations in fine-grained emotional labeling can be observed. This is expected given that predictions are made independently at the segment level using only textual information. Overall, the results demonstrate that text-based transfer learning provides a strong baseline for segment-level affect modeling, while leaving room for potential improvements through the incorporation of additional modalities such as audio.

Figure 10 shows the pairwise Normalized Mutual Information (NMI) between all the segment-level tasks on parent data. As we can see there are some labels that are correlated and have a fairly high mutual information in the parent data. For instance, 'joy' and 'sentiment' have an NMI of 0.282, 'anger' and 'criticism' have an NMI of 0.384, and 'fear' and 'worry' have an NMI of 0.349 which all makes sense as those tasks are intuitively correlated.

Figure 11 illustrates the pairwise Normalized Mutual Information (NMI) between all the psychological and cognitive tasks. As we can see there are some tasks with very high mutual information. Particularly, 'anxiety' and 'introvert' have an NMI of 0.917 which is close to perfect mutual information. Similarly, 'ADHD' and 'extrovert' also have an NMI of 0.842 which is very high. 'depression' and 'introvert' have an NMI of 0.223 which is fairly high but not too strong. We can also observe that the 'spectrum' task has a very low NMI with any other task which show there are many other factors into play and it is a complex task.

Figure 12 illustrates the changes in mood, emotions, and sentiment in four random samples in offspring data. We can also see sample segments in those files shown at the top of the figures which gives an idea of how the segmentation and the labeling is done. The first two figures on top are random samples from the three minutes general interviews, and the bottom figures are random samples from a positive experience and a negative experience. As we can see in the positive experience interview, the dominant emotion is 'joy' and in the negative experience, 'sadness' is the dominant emotion. Often times 'anger' is the dominant emotion in the negative experience interviews.

Figure 13 illustrates the pairwise Normalized Mutual Information (NMI) between all the segment-level tasks on offspring data. As we can see there are some labels that are correlated and have a fairly high mutual information in the offspring data. For instance, 'irrelevance' and 'richness' have an NMI

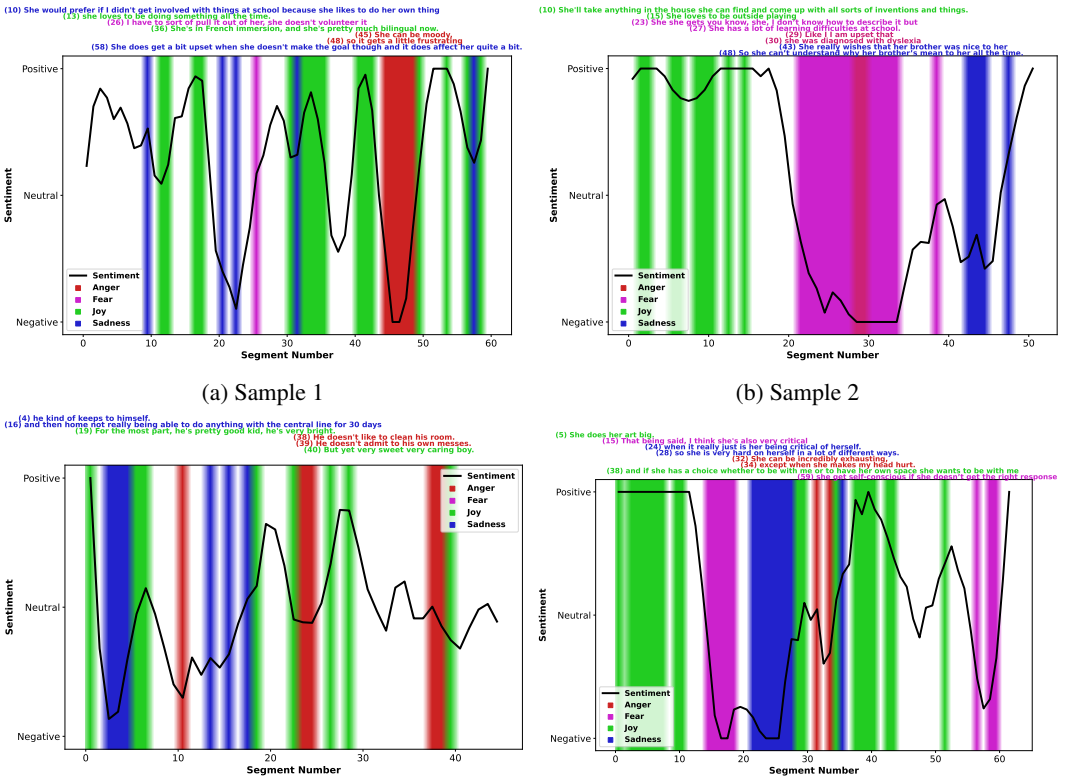
Table 10: Summary statistics of the parent dataset used in this research. The table reports the number of samples, number of classes, class distribution, and degree of imbalance for each task across three levels: segment-level (fine-grained annotations per speech segment), document-level (ratings per full interview), and cognitive/psychological labels (diagnostic categories assigned by clinicians). The imbalance percentage is computed as $100 \times (\max_c n_c - \min_c n_c)/\max_c n_c$, where n_c denotes the number of samples in class c. Missing or extremely sparse labels have been removed, resulting in varying sample sizes across tasks.

Level	Task	Num Samples	Num Classes	Imbalanced (%)	Label 0	Label 1	Label 2	Label 3	Label 4
Segment	objective	23693	2	19.08	13096	10597	-	-	-
Segment	sentiment	23479	3	57.39	4725	11088	7666	-	-
Segment	anger	23694	2	92.22	21984	1710	-	-	-
Segment	fear	23694	2	95.29	22628	1066	-	-	-
Segment	joy	23690	2	66.58	17756	5934	-	-	-
Segment	sadness	23694	2	92.70	22081	1613	-	-	-
Segment	neutral	23693	2	16.28	10797	12896	-	-	-
Segment	cohesion	23694	2	81.32	19964	3730	-	-	-
Segment	rumination	23694	2	98.95	23447	247	-	-	-
Segment	overinclusive	23694	2	97.20	23049	645	-	-	-
Segment	worry	23694	2	91.76	21890	1804	-	-	-
Segment	criticism	23694	2	88.42	21235	2459	-	-	-
Document	affect	484	3	93.18	23	124	337	-	-
Document	warmth	485	3	39.69	117	194	174	-	-
Document	overprotection	487	3	85.77	267	182	38	-	-
Document	cohesion	492	5	92.63	14	14	132	190	142
Document	criticism	492	4	89.22	232	166	69	25	-
Document	worry	381	4	93.29	164	147	59	11	-
Cognitive	spectrum	363	4	87.25	129	149	66	19	-
Cognitive	introvert	369	2	49.39	245	124	-	-	-
Cognitive	extrovert	369	2	63.84	271	98	-	-	-
Cognitive	ADHD	369	2	67.74	279	90	-	-	-
Cognitive	anxiety	369	2	51.81	249	120	-	-	-
Cognitive	depression	369	2	89.19	333	36	-	-	-

of 0.21 which makes sense as they are fairly opposite of each other. The two tasks of 'fear' and 'worry' also have an NMI of 0.36 which again makes sense as those tasks are intuitively correlated.

In terms of document-level tasks there are only three tasks of 'affect', 'coherence', and 'richness' for the offspring data. These are labeled for each of the three interview files of children separately. In terms of psychological and cognitive labels we have the same labels as we saw in the parent data, as those are indeed the children's conditions labeled by psychiatrists. The information and statistics about the tasks on offspring data is shown in Table **??**. Here again for certain tasks, some segments or documents were missing the corresponding label, or some labels were extremely imbalanced such there were only 1 or 2 examples from one category, and we removed those cases. This is the reason for the differences in the number of samples in Table **??**. The imbalanced % is calculated similarly to the parent data, as $100 * (\max_{c \in C} n_c - \min_{c \in C} n_c)/\max_{c \in C} n_c$. As we can see most of the tasks are quite imbalanced.

For predicting the psychological and cognitive labels, we concatenate the three audio files of each child and use the three interviews together to predict the final cognitive labels. In contrary, for segment-level and document-level we use them separately as they are labeled separately in the first place. We can see from Table **??** that we only have 85 samples for the cognitive tasks where we have all 3 audio files, and we have the psychiatrist label. This is a low data setting and poses a lot of challenges.



(a) Sample 1

(b) Sample 2

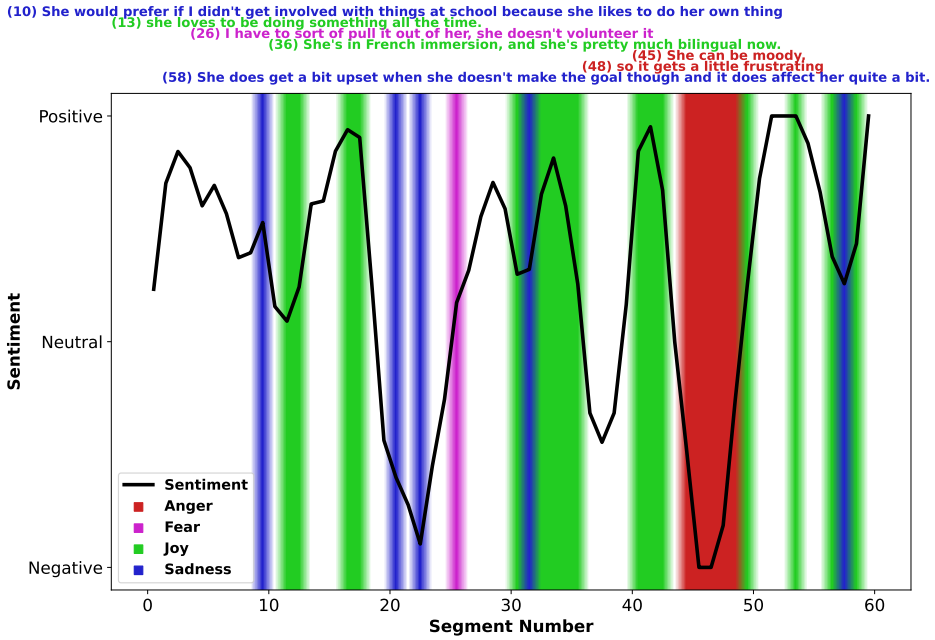
(c) Sample 3

(d) Sample 4

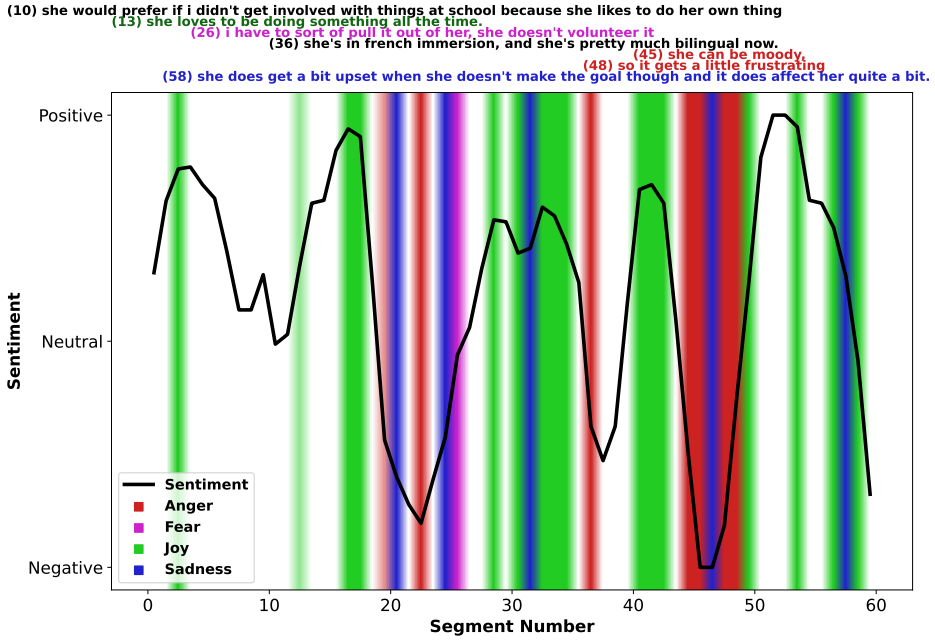
Figure 8: Emotion and sentiment dynamics for four representative parent interview samples from the FORBOW dataset. Each subplot corresponds to a different interview. The horizontal axis represents time (segmented speech units), and the vertical axis represents normalized sentiment scores. The black curve shows segment-level sentiment trajectories, while colored vertical bands indicate discrete emotion labels (anger, fear, joy, sadness, or neutral) assigned to each segment. Segments are obtained via content-aware partitioning based on changes in affective and linguistic cues. Representative text excerpts are displayed above each plot to illustrate the correspondence between raw speech content, segmentation boundaries, and assigned annotations.

Inter-rater agreement at the segment level is evaluated on a per-subject basis by first constructing, for each subject and each task, a ratings matrix in which rows represent segments belonging to that subject and columns represent the raters who annotated those segments. Reliability is quantified using multiple complementary statistics. Three intraclass correlation coefficients, ICC(1), ICC(2), and ICC(3) are derived from an ANOVA decomposition of between-segment, between-rater, and residual variance components, corresponding respectively to a one-way random-effects model, a two-way random-effects model assessing absolute agreement, and a two-way mixed-effects model assessing consistency. Chance-corrected reliability is captured through Krippendorff's α for interval data, computed by contrasting the observed disagreement among raters with the disagreement expected under statistical independence. In addition, average pairwise Pearson correlations are calculated across all rater pairs. This full set of agreement metrics is computed independently for each subject, producing one reliability estimate per metric per subject. Finally, to characterize the overall reliability of each task across the entire dataset, these subject-level values are aggregated by computing their mean and standard deviation, yielding the summary statistics reported in Table 12 and Table 13 for parent and offspring data, respectively. Figures 14 and 15 illustrate the distribution of ICC1 scores across subjects for the different tasks on parent and offspring data.

To contextualize these coefficients, ICC values below 0.40 are typically interpreted as indicating poor agreement, values between 0.40-0.59 as fair, 0.60-0.74 as good, and values of 0.75 or higher as excellent reliability Müller and Büttner [1994]. Under these qualitative bands, the Objective task exhibits excellent reliability for both parent and offspring speech, reflecting its relatively well-defined



(a) Ground truth emotion and sentiment for a random sample



(b) Predicted emotion and sentiment for the same random sample

Figure 9: Comparison of ground-truth and predicted segment-level sentiment and emotion trajectories for a representative parent interview using sentenceT5XL.

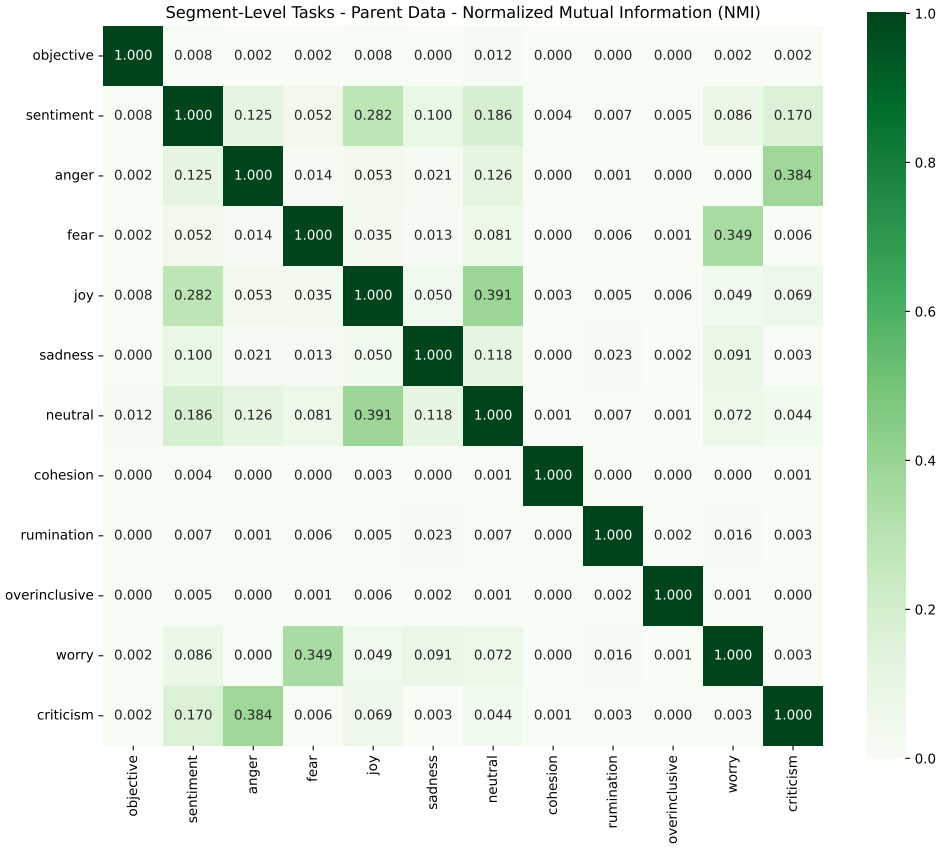


Figure 10: Pairwise Normalized Mutual Information (NMI) between segment-level annotation tasks for parent interview data. Each cell represents the NMI value between two tasks, quantifying the degree of statistical dependency between their label distributions. Higher values indicate stronger relationships between tasks (e.g., co-occurrence or conceptual overlap), while lower values indicate more independent labeling behavior. All values are computed using segment-level annotations across the full parent dataset.

operationalization. In contrast, several affective labels fall into the fair and occasionally poor ranges, including Overinclusive, Cohesion, Rumination, Fear, Neutral, Joy, and Sadness in the parent data, and Rumination, Sadness, Neutral, and Anger in offspring annotations. These labels tend to capture nuanced, introspective, or context-dependent psychosocial constructs, which naturally yield greater rater variability. An additional contributing factor is label imbalance across subjects and tasks. When only a small fraction of segments are assigned a given label, variance across raters becomes dominated by a few sparse decisions, which inflates uncertainty and depresses reliability estimates such as ICC and Krippendorff's α. This effect is particularly evident for rare categories like Overinclusive and Rumination. Consequently, findings derived from these low-frequency, lower-reliability labels should be interpreted with caution in subsequent modeling analyses, as downstream performance will inherently reflect the noisier ground truth associated with them.

D Societal Impacts

In terms of environmental impact, training large models is costly. In this work we used relatively small datasets and small network architecture, so that all the experiments were run on a high-end PC with Titan RTX GPU. However, if someone wants to train a larger version of SPECTRA on larger datasets it could become costly.

As a positive societal impact, this project has great positive impacts in helping psychiatrists in early diagnosis of mental disorders. The goal is to provide decision-support signals and tool that may

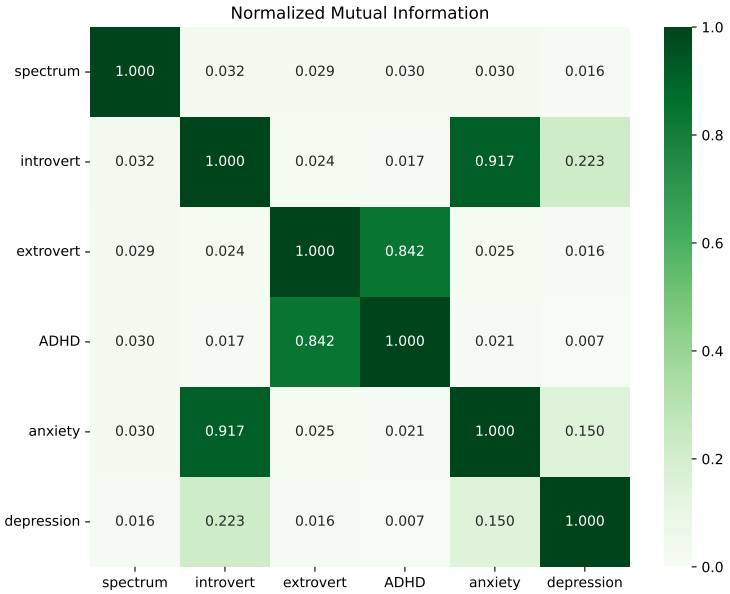
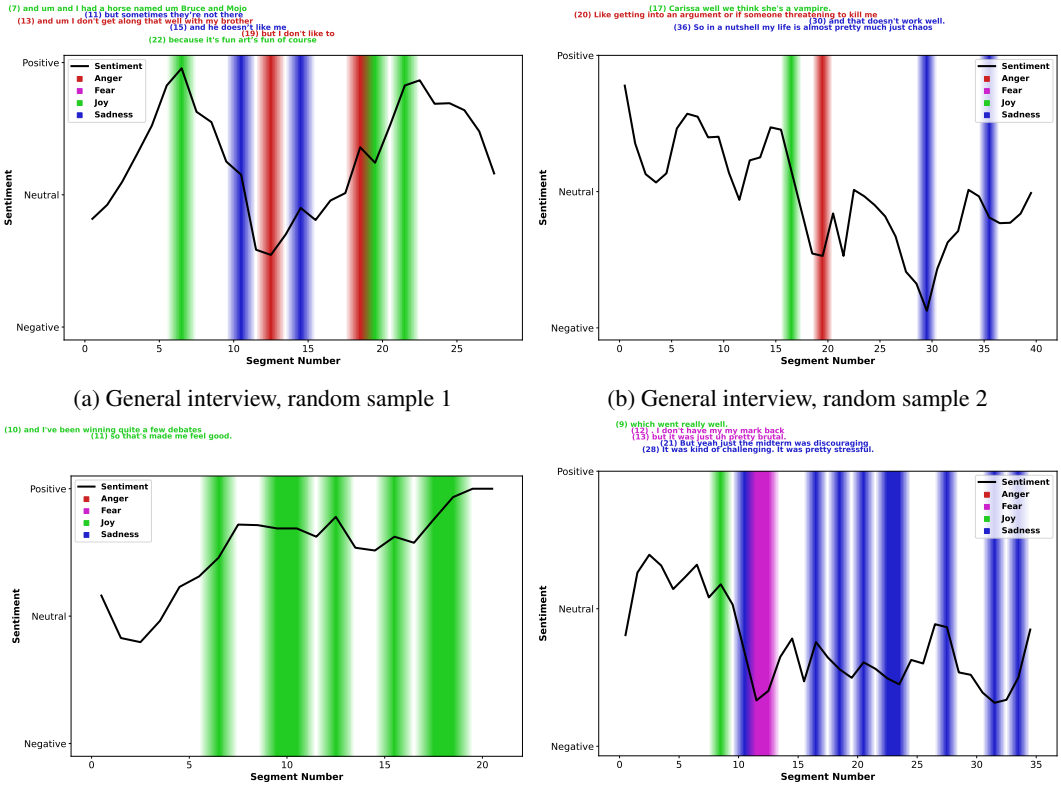


Figure 11: Pairwise Normalized Mutual Information (NMI) between psychological and cognitive labels derived from parent-level annotations. Each cell represents the dependency between two tasks, computed from subject-level labels. High NMI values indicate strong relationships or shared underlying constructs (e.g., introversion and anxiety), while low values indicate weak or independent associations. This matrix highlights redundancy and interdependence among higher-level diagnostic labels.

complement clinical judgment, and not to replace clinicians and human judgment. We advocate for the responsible use of SPECTRA, particularly for potential clinical applications, ensuring consent, de-identification, and human-in-the-loop deployment. Our clinical speech analysis work was approved by the government's Research Ethics Board (file number: 100266).

E Source Code

To facilitate reproducibility, the Python implementation of the proposed SPECTRA framework is included in the supplementary material (spectra.py).



(a) General interview, random sample 1

(b) General interview, random sample 2

(c) Positive experience interview, random sample 3

(d) Negative experience interview, random sample 4

Figure 12: Emotion and sentiment dynamics for representative offspring (child) interview samples. The top row shows segments from general interviews, while the bottom row shows segments from positive and negative experience interviews. The horizontal axis represents segmented time, and the vertical axis represents sentiment scores. The black curve shows sentiment trajectories, while colored vertical bands indicate discrete emotion labels. Segments are derived using content-aware partitioning, and representative text excerpts are displayed above each plot to illustrate alignment between raw speech, segmentation, and annotations.

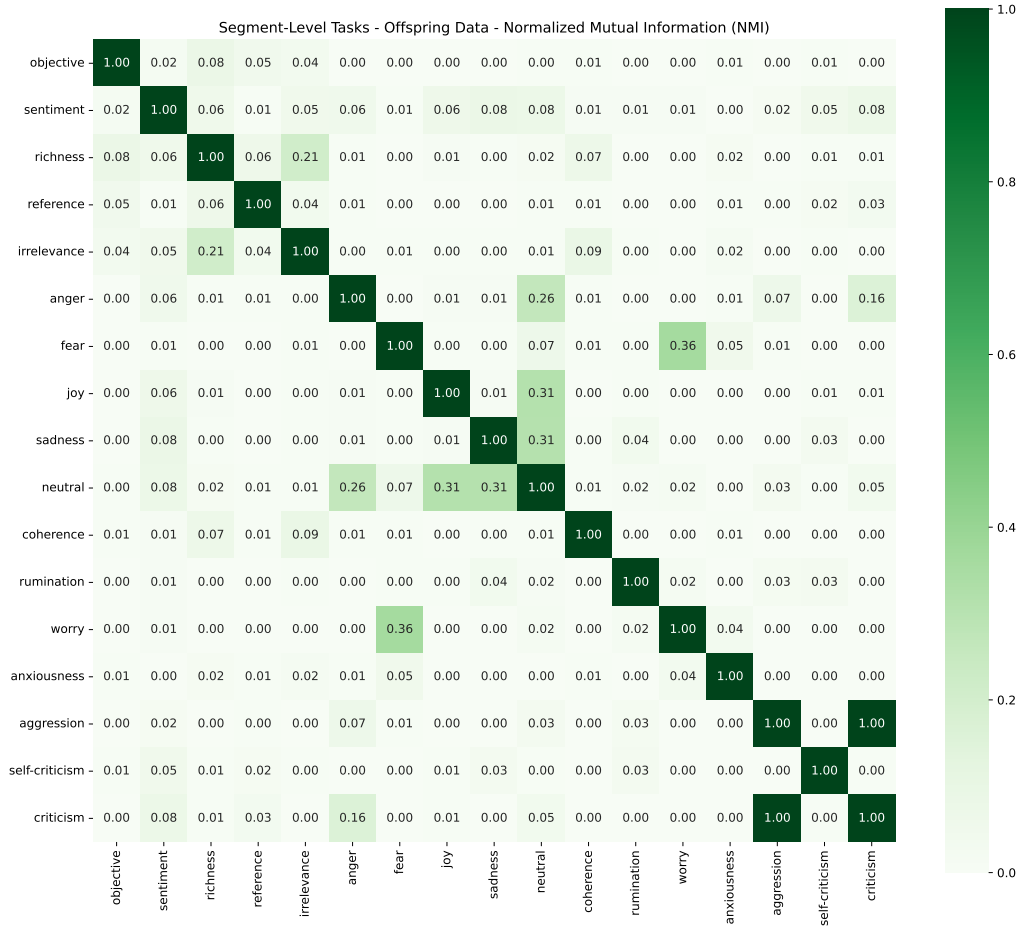


Figure 13: Pairwise Normalized Mutual Information (NMI) between segment-level annotation tasks for offspring (child) interview data. Each cell represents the dependency between two tasks based on their label distributions. Higher values indicate stronger relationships between tasks (e.g., fear and worry), while lower values indicate weaker or independent labeling patterns. All values are computed using segment-level annotations across the offspring dataset.

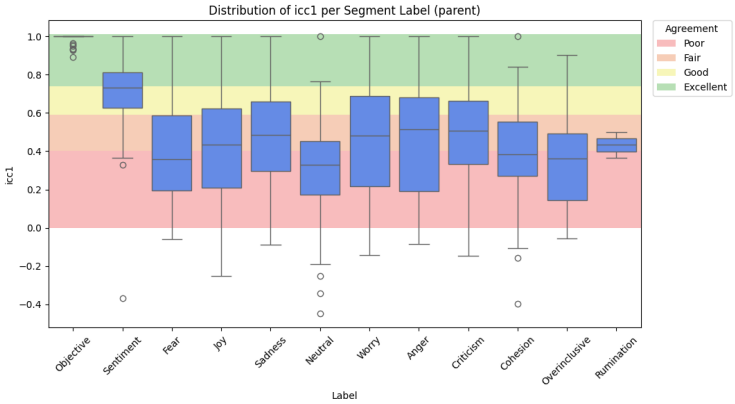


Figure 14: Distribution of Intraclass Correlation Coefficient (ICC1) values across subjects for each segment-level task in parent data. Each boxplot summarizes variability in inter-rater agreement across subjects, showing median, quartiles, and outliers. Higher ICC values indicate stronger agreement among raters, while wider distributions reflect variability in annotation consistency across subjects and tasks.

Table 11: Summary statistics of the offspring (child) dataset. The table reports the number of samples, number of classes, class distribution, and degree of imbalance for each task across segment-level, document-level, and cognitive/psychological categories. Segment-level tasks include additional annotations (e.g., self-criticism, aggression, anxiousness) specific to child interviews. Cognitive labels are based on clinical diagnoses and are available for a subset of subjects. Imbalance percentage is computed as $100 \times (\max_c n_c - \min_c n_c)/\max_c n_c$. Differences in sample counts reflect missing or filtered labels.

Level	Task	Num Samples	Num Classes	Imbalanced (%)	Label 0	Label 1	Label 2	Label 3	Label 4
Segment	objective	10319	2	1.90	5209	5110	-	-	-
Segment	sentiment	10222	3	52.51	2404	5062	2756	-	-
Segment	richness	10273	3	97.08	1366	8654	253	-	-
Segment	reference	10328	4	77.89	2072	4970	2187	1099	-
Segment	irrelevance	10301	2	87.98	9196	1105	-	-	-
Segment	anger	10329	2	95.55	9889	440	-	-	-
Segment	fear	10329	2	99.04	10231	98	-	-	-
Segment	joy	10329	2	94.49	9790	539	-	-	-
Segment	sadness	10329	2	94.62	9802	527	-	-	-
Segment	neutral	10328	2	80.92	1655	8673	-	-	-
Segment	coherence	10329	2	68.35	7846	2483	-	-	-
Segment	rumination	10329	2	99.01	10228	101	-	-	-
Segment	worry	10329	2	99.13	10240	89	-	-	-
Segment	anxiousness	10317	2	96.51	9969	348	-	-	-
Segment	aggression	2182	2	99.49	2171	11	-	-	-
Segment	criticism	8090	2	95.04	7708	382	-	-	-
Segment	self-criticism	10329	2	96.77	10006	323	-	-	-
Document	affect	621	4	97.73	16	11	109	485	-
Document	coherence	616	5	95.83	11	33	109	264	199
Document	richness	614	3	96.70	82	515	17	-	-
Cognitive	spectrum	85	4	86.11	36	27	17	5	-
Cognitive	introvert	85	2	39.62	53	32	-	-	-
Cognitive	extrovert	85	2	53.45	58	27	-	-	-
Cognitive	ADHD	85	2	53.45	58	27	-	-	-
Cognitive	anxiety	85	2	39.62	53	32	-	-	-
Cognitive	depression	85	2	83.56	73	12	-	-	-

Table 12: Inter-rater agreement statistics for segment-level annotation tasks in parent interview data. Each value is reported as mean ± standard deviation across subjects. Metrics include three forms of Intraclass Correlation Coefficient (ICC1: one-way random effects; ICC2: two-way random effects with absolute agreement; ICC3: two-way mixed effects with consistency), Krippendorff's α for interval data, and average pairwise Pearson correlation. These metrics collectively quantify agreement reliability, accounting for both consistency and chance-corrected agreement across multiple annotators.

Task	ICC1	ICC2	ICC3	Krippendorff Alpha	Pearson Correlation
Objective	0.9954 ± 0.0174	0.9954 ± 0.0174	0.9955 ± 0.0173	0.9954 ± 0.0176	0.9957 ± 0.0166
Sentiment	0.7019 ± 0.1725	0.7098 ± 0.1451	0.7229 ± 0.1281	0.6996 ± 0.1726	0.7332 ± 0.1256
Criticism	0.5115 ± 0.2537	0.523 ± 0.2409	0.5345 ± 0.2336	0.5093 ± 0.2536	0.5549 ± 0.2273
Worry	0.4726 ± 0.2814	0.4831 ± 0.2723	0.496 ± 0.2694	0.4706 ± 0.2811	0.5123 ± 0.2745
Anger	0.4684 ± 0.3185	0.4745 ± 0.3135	0.4798 ± 0.3116	0.4666 ± 0.3181	0.4933 ± 0.3132
Sadness	0.4609 ± 0.2678	0.4656 ± 0.2631	0.4693 ± 0.2623	0.459 ± 0.2674	0.4878 ± 0.2667
Joy	0.4152 ± 0.2778	0.4472 ± 0.2414	0.4713 ± 0.2311	0.4136 ± 0.2766	0.4878 ± 0.2251
Rumination	0.4331 ± 0.0946	0.4323 ± 0.0957	0.4306 ± 0.0981	0.4302 ± 0.0953	0.3579 ± 0.0347
Cohesion	0.3994 ± 0.2379	0.4101 ± 0.2272	0.4184 ± 0.2276	0.3973 ± 0.2367	0.4305 ± 0.2305
Fear	0.3742 ± 0.277	0.3839 ± 0.2727	0.3945 ± 0.2717	0.3723 ± 0.2764	0.4088 ± 0.28
Neutral	0.3022 ± 0.2549	0.3594 ± 0.1888	0.3978 ± 0.1783	0.3007 ± 0.253	0.399 ± 0.1787
Overinclusive	0.3463 ± 0.2901	0.3506 ± 0.2886	0.3543 ± 0.2884	0.3448 ± 0.2895	0.3554 ± 0.2998

Table 13: Inter-rater agreement statistics for segment-level annotation tasks in offspring (children) interview data. Values are reported as mean ± standard deviation across subjects. Metrics include ICC1, ICC2, ICC3, Krippendorff's α, and average Pearson correlation. Compared to parent data, these values reflect annotation consistency under increased task difficulty and data variability in child speech.

Task	ICC1	ICC2	ICC3	Krippendorff Alpha	Pearson Correlation
Objective	0.9902 ± 0.0886	0.9913 ± 0.077	0.9928 ± 0.0626	0.9902 ± 0.0888	0.9916 ± 0.074
Sentiment	0.7104 ± 0.1999	0.7156 ± 0.1928	0.7297 ± 0.1836	0.7031 ± 0.2023	0.7391 ± 0.1827
Criticism	0.6758 ± 0.3417	0.69 ± 0.3213	0.7034 ± 0.3058	0.6723 ± 0.3432	0.7184 ± 0.2972
Fear	0.5188 ± 0.2678	0.5302 ± 0.2503	0.5404 ± 0.2377	0.5125 ± 0.2696	0.5132 ± 0.2775
Worry	0.5012 ± 0.3936	0.4972 ± 0.4029	0.5028 ± 0.3992	0.4975 ± 0.3928	0.4667 ± 0.4128
Joy	0.4898 ± 0.3245	0.4932 ± 0.3229	0.4992 ± 0.3216	0.4841 ± 0.3239	0.4952 ± 0.3331
Anger	0.4526 ± 0.3134	0.4592 ± 0.3104	0.4693 ± 0.3049	0.4469 ± 0.312	0.4379 ± 0.3263
Neutral	0.4354 ± 0.2873	0.4396 ± 0.2883	0.45 ± 0.2874	0.4287 ± 0.2851	0.4531 ± 0.2925
Sadness	0.4208 ± 0.2981	0.4172 ± 0.3095	0.425 ± 0.3064	0.4146 ± 0.2956	0.4179 ± 0.3195
Rumination	0.3169 ± 0.3033	0.3254 ± 0.3008	0.3353 ± 0.3005	0.3113 ± 0.301	0.2845 ± 0.2988

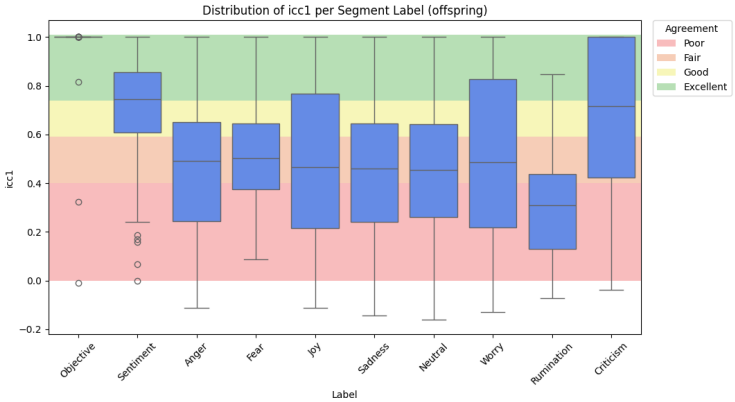


Figure 15: Distribution of Intraclass Correlation Coefficient (ICC1) values across subjects for each segment-level task in offspring data. Boxplots illustrate variability in agreement across annotators, highlighting differences in reliability across tasks. Lower ICC values and wider spreads indicate increased annotation uncertainty, particularly for subjective or low-frequency labels.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims are supported by the empirical results and analysis in both section 4 and appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Monte Carlo sampling could be slow and the method generally addresses uncertainty rather than directly targeting accuracy.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: Our paper does not include theoretical claims.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We have used a combination of public and private datasets. The details of datasets and experiments including hyperparameters are provided in 4. And the code is provided in supplementary material. But we are not allowed to publicly share the mental health data.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: Other than the mental health dataset, the other datasets and the source code are provided.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All the training details and experimental setup is provided in section4.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We have provided the standard deviations of errors for the results on public dataset. However, for the mental health data experiments, we did not include error bars as the data is relatively small.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We have provided the GPU type and memory information in section 4, however we have not measured the runtime for each of the experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This research conforms with code of ethics and does not have harmful impacts. The societal impacts are discussed in appendix D.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We have discussed positive and negative societal impacts in appendix section D.

Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We have provided citations for the datasets and models we compared with. And we are the original owner of our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We have provided complete information on the mental health dataset in section C.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our research involves human subjects. However, we have not collected the dataset ourselves, we have acquired that through collaboration with Psychiatry department and they have followed all the rules and requirements for compensation.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: The paper does not pose such risks. And the dataset was collected with ethics board approvals as explained in section D.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used for final edits and grammar checks on the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.