An Emotion-Preserving Conditional Information Bottleneck for Domain-Generalizable Speech Emotion Recognition

Zhichen Yuan[1] , **C. L. Philip Chen**[1] , **Shuzhen Li**[2] and **Tong Zhang**[1*]
[1]South China University of Technology
[2]Lund University
yyyzc922@gmail.com, philip.chen@ieee.org, shuzhen.li@math.lth.se, tony@scut.edu.cn

Abstract

Domain-generalizable speech emotion recognition (DG-SER) aims to ensure the robustness of SER models across unknown domains, which is essential for real-world human-machine interaction systems. Most DG-SER approaches employ alignment or adversarial strategies with domain labels to promote generalization. However, these strategies often confine generalization to predefined domains, limiting robustness under diverse real-world speech variations. To address these challenges, this paper proposes an emotion-preserving conditional information bottleneck framework (EP-CIB) for domain-free DG-SER. Specifically, EP-CIB introduces a nuisance proxy representation learning module to learn a nuisance proxy as the broad non-emotional variability without requiring any domain annotations, covering both defined and previously unseen domains. It then extracts coarse-grained emotion features via the consistency-aware emotion representation learning module. EP-CIB further introduces an emotion-preserving conditional information bottleneck that, conditioned on the emotion label, disentangles the nuisance proxy from the coarse-grained emotion representation, improving domain-free generalization under open-ended domain shifts. EP-CIB departs from implicit domain surrogates in prior domain-free methods by explicitly learning a proxy for nuisance domains and disentangling it from emotion features, enabling domain-free emotion representation learning. The state-of-the-art performance in both speaker-independent and cross-corpus settings, including an 18% improvement on EmoDB-to-CASIA transfer, demonstrates the effectiveness of EP-CIB for DG-SER.

1 Introduction

Domain-generalization speech emotion recognition (DG-SER) aims to learn emotion recognition models that preserve robust performance across different domains containing new speakers, languages, or recording conditions [Li *et al.*, 2019; Wen *et al.*, 2022]. In real-world human-computer interaction systems, speech signals exhibit pronounced variability, and such variability cannot be exhaustively covered during training, which in turn renders a mismatch between the training distribution and the deployment distribution unavoidable. This variability is further driven by multiple interacting latent factors, a circumstance that blurs domain boundaries while causing distribution shifts to exhibit, in essence, open-ended and compositional characteristics [Koh *et al.*, 2021]. Consequently, DG-SER studies robust emotion representations beyond any predefined or annotated domains.

A core challenge in DG-SER is how to learn domain-invariant emotion representations from diverse domains with different speakers, languages, and recording conditions [Huang and Ren, 2024]. The approaches to address such challenge can be classified into alignment-based and adversarial-based strategies. Zhao et al. and Tao et al. [Zhao *et al.*, 2024b; Tao *et al.*, 2022] propose alignment-based domain generalization methods that narrow inter-domain gaps by leveraging corpus or speaker labels, thereby improving the cross-domain performance of speech emotion recognition, but their effectiveness remains contingent on domain annotations. Zhao et al. and Fadhil et al. [Zhao *et al.*, 2024a; Fadhil and Zahra, 2025] adopt adversarial-based strategies to encourage representations invariant to domain shifts, but this may inadvertently diminish cues that are essential for distinguishing emotional states. In real-world speech emotion recognition, distribution shifts rarely come with explicit domain labels. They instead exhibit long-tailed patterns and arise from complex combinations of factors. As a result, approaches that depend on predefined domain annotations struggle to generalize under practical conditions. Motivated by this limitation, we propose a domain-free DG-SER framework that removes the need for domain definitions while remaining effective on unseen environments.

From a broader perspective, the fundamental challenge of the domain-free DG-SER lies in the entanglement of latent factors and the temporal variability of emotional expressions [Tjandra *et al.*, 2021; Cai *et al.*, 2019]. These nuisance factors are unannotated, entangled with emotional expression, and often share acoustic carriers such as pitch and energy, making it challenging to suppress the nuisance information without compromising the ability to distinguish emotions. This

*Corresponding author.

challenge is exacerbated at the utterance-level supervision, as emotional evidence is sparsely and unevenly distributed in time. These characteristics indicate that an effective domain-free DG-SER requires the ability to suppress diverse nuisance variations while retaining the acoustic structures related to emotions, thereby achieving domain-free emotional representation in weak supervision. Therefore, we aim to develop a DG-SER framework to disentangle nuisance factors from emotional representations in a domain-free manner.

To address the above challenges, we propose EP-CIB, a domain-free DG-SER framework, improving robustness beyond predefined domain annotations. EP-CIB first introduces a Nuisance Proxy Representation Learning Module provides a concrete handle on open-ended nuisance factors by learning an emotion-suppressed nuisance proxy. This proxy is stabilized through teacher–student multi-resolution consistency and further protected against emotion leakage via contrastive regularization and adversarial emotion removal. EP-CIB then employs a Consistency-Aware Emotion Representation Learning Module to model temporally sparse and uneven emotional evidence, preventing weakly emotional or ambiguous segments from dominating the utterance-level embedding. Finally, EP-CIB designs an Emotion-Preserving Conditional Information Bottleneck. Conditioned on the emotion label, it disentangles the nuisance proxy from the emotional representations within each class, yet preserves inter-class separability supported by shared acoustic carriers. Overall, EP-CIB provides a domain-free instantiation of conditional nuisance invariance and delivers more reliable performance under speaker-independent and cross-corpus shifts.

Our contributions are threefold:

- We propose EP-CIB, a domain-free DG-SER framework for generalizing SER models to unseen domains. It supports emotion representation learning beyond predefined domain annotations by explicitly disentangling nuisance factors from emotional representations under label conditioning, reducing the reliance on implicit nuisance modeling in prior domain-free approaches and thereby improving robustness under open-ended deployment shifts.
- We introduce an emotion-suppressed nuisance proxy as an explicit surrogate for unannotated nuisance factors, enabling domain-free nuisance modeling beyond predefined domain taxonomies.
- We formulate an emotion-preserving conditional information bottleneck that performs within-class disentanglement between nuisance proxy and emotional representations, reducing nuisance-driven dispersion while retaining emotion-discriminative structure supported by shared acoustic correlates.

2 Related Work

Domain generalization in SER. Robust SER under distribution shift is commonly studied in speaker-independent and cross-corpus settings. Representative approaches use alignment or adversarial strategies to match representations across predefined domains [Gideon *et al.*, 2021]. Although they show effectiveness under controlled domains, these methods assume access to domain annotations. As emphasized in broader domain generalization benchmarks [Koh *et al.*, 2021], real-world shifts are often open-ended and combinatorial, which limits the applicability of domain-dependent formulations. This gap has prompted the development of the DG-SER framework, which is not limited by domains and can generalize beyond predefined domain annotations.

Emotion representation learning. Recent work improves SER transferability through self-supervised pretraining and weakly supervised learning [Chen *et al.*, 2024]. General SSL models, such as HuBERT [Hsu *et al.*, 2021], provide strong acoustic representations, while emotion-oriented pretraining [Ma *et al.*, 2024] and adversarial learning with unlabeled data [Latif *et al.*, 2023] further enhance cross-corpus robustness. However, these methods mainly focus on representation transfer. They rarely address how to suppress heterogeneous nuisance variables and retain emotion-discriminative acoustic carriers under coarse utterance-level supervision. In contrast, our work explicitly models the consistency at the fragment level and conditional disentanglement to alleviate this mismatch.

Information-theoretic learning. The Information Bottleneck (IB) theory [Tishby *et al.*, 2000] was proposed by Tishby et al. It provides a theoretically reliable framework for extracting task-relevant feature representations and eliminating irrelevant information [Yu *et al.*, 2021]. Traditional methods have difficulty calculating the mutual information in deep neural networks, but methods based on kernel functions such as the Hilbert-Schmidt Independence Criterion (HSIC) [Kurt Ma *et al.*, 2020; Wang *et al.*, 2021] can serve as effective alternatives. Although the existing IB methods have achieved good results, there are still some issues: Most of them enforce unconditional invariance, but have difficulty handling potential task-irrelevant nuisance. Our research expands this research direction by introducing a nuisance proxy with emotional suppression and implementing nuisance suppression under label conditions, which is crucial for preserving the emotional discrimination structure in domain-free SER.

3 Method

To address reliance on predefined domain annotations in DG-SER, we propose EP-CIB, an emotion-preserving conditional information bottleneck framework. EP-CIB uses only utterance-level emotion labels and requires no domain or nuisance annotations. As illustrated in Fig. 1, EP-CIB adopts a two-stream design: a nuisance proxy encoder learns an emotion-suppressed proxy z_1 that summarizes broad non-emotional variability from multiple views, while a consistency-aware emotion encoder extracts a coarse-grained emotion representation z_2 by modeling segment-level consistency under utterance-level supervision. A label-conditioned information bottleneck then suppresses the within-class dependence between z_1 and z_2, so that nuisance-induced dispersion is reduced without discarding emotion-discriminative structure carried by shared acoustic correlates. EP-CIB addresses the coupled challenges of open-ended nuisance vari-

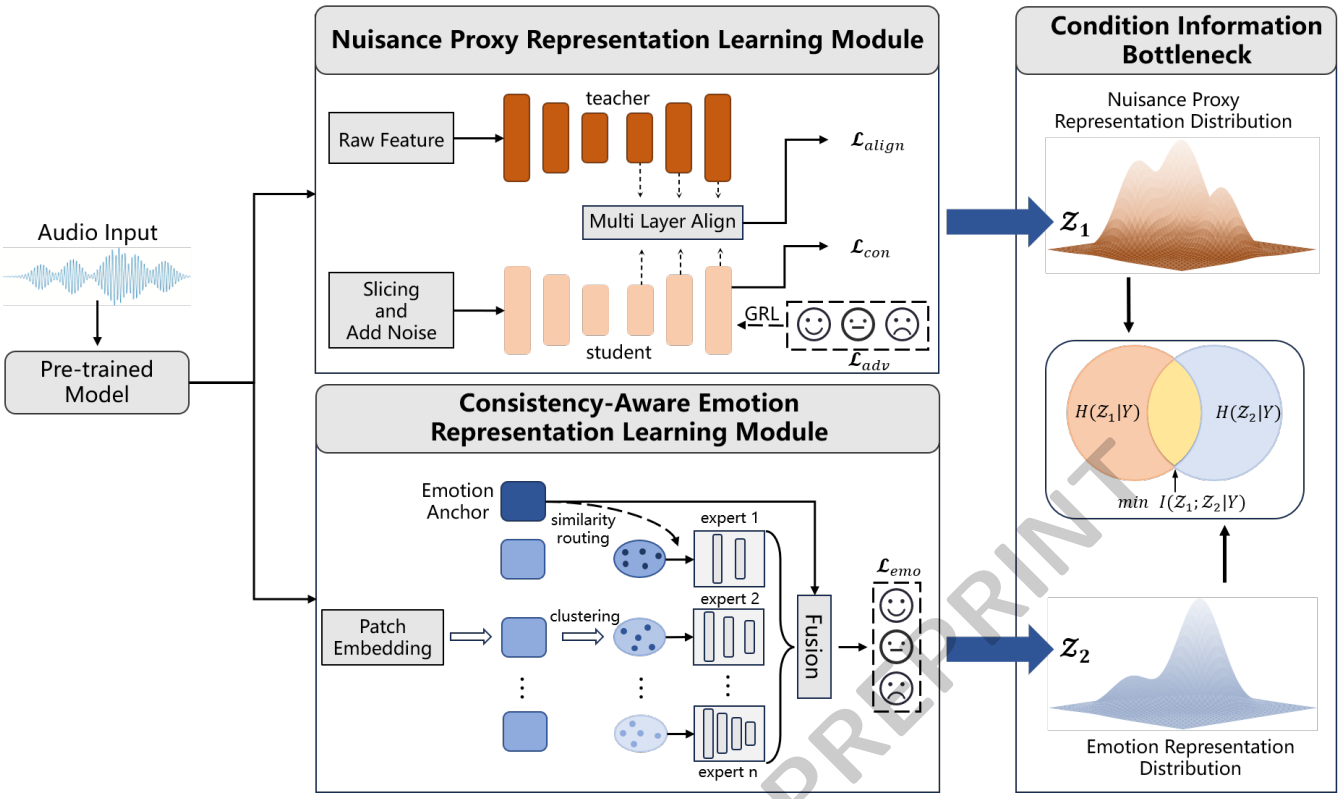


Figure 1: Framework of EP-CIB. From an input utterance, the model learns an emotion-suppressed nuisance proxy z_1 and a coarse emotion representation z_2 built from segment-level consistency. We then apply a label-conditioned bottleneck to reduce within-class dependence between z_1 and z_2, improving generalization without domain labels.

ability and weak supervision in domain-free SER, providing a principled framework that preserves emotion-discriminative structure while improving robustness under unseen domain shifts.

3.1 Nuisance Proxy Representation Learning Module

The role of z_1 is to provide a domain-free handle for nuisance variability by aggregating broad non-emotional factors, while remaining minimally informative about emotion so that it can serve as a reliable reference for conditional disentanglement. Our design is inspired by voice cloning, where compact representations preserve speaker- and style-stable attributes across diverse content, suggesting a practical way to approximate latent nuisance factors in speech without explicit annotations. Guided by this intuition, we instantiate z_1 through a domain-free training scheme that jointly enforces content-agnostic and view-consistent encoding via multi-scale teacher–student alignment, emotion-suppressed contrastive regularization, and multi-level adversarial emotion removal.

Multi-scale teacher–student alignment. We first encourage z_1 to be stable across partial observations of the same utterance. To this end, we build the nuisance proxy encoder on an ECAPA-TDNN [Desplanques *et al.*, 2020] backbone and adopt a teacher–student architecture: the teacher takes the full utterance x, whereas the student processes K cropped views $\{x^{(k)}\}_{k=1}^{K}$. Both networks expose intermediate features $\{h_\ell^T(x)\}_{\ell\in\mathcal{L}}$ and $\{h_\ell^S(x^{(k)})\}_{\ell\in\mathcal{L}}$, and the teacher is updated by exponential moving average (EMA) to provide a slowly moving target. Instead of aligning only the final embedding, we align multi-level features so that the proxy becomes consistent across time crops and abstraction levels; we use the following KL-based alignment loss:

$$\mathcal{L}_{\text{align}} = \sum_{\ell\in\mathcal{L}} \mathrm{KL}\left(\sigma(h_\ell^T(x)) \,\middle\|\, \frac{1}{K}\sum_{k=1}^{K}\sigma(h_\ell^S(x^{(k)}))\right), \quad (1)$$

where $\sigma(\cdot)$ denotes a feature normalization operator applied at each layer.

Emotion-suppressed contrastive regularization. View consistency alone does not prevent the proxy from picking up emotion cues that are shared across views. We therefore push z_1 to represent *non-emotional* differences *within the same emotion class*: negatives are sampled from the same label, so the proxy must discriminate using nuisance variation rather than emotion. Positives are constructed from complementary temporal slices and prosodic perturbations of the same utterance, and we optimize the following multi-positive InfoNCE objective:

$$\mathcal{L}_{\text{con}} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\sum_{p\in\mathcal{P}_i}\exp\left(\mathrm{sim}(\hat{u}_i,p)/\tau\right)}{\sum_{c\in\mathcal{C}_i}\exp\left(\mathrm{sim}(\hat{u}_i,c)/\tau\right)}, \quad (2)$$

where $\hat{u}_i$ is the proxy feature of an anchor view, $\mathcal{P}_i$ is the set of its positives, and $\mathcal{C}_i$ contains candidates sampled within the same emotion class.

Multi-level adversarial emotion removal. Even with class-restricted contrastive learning, emotion can still leak into intermediate features. We make this failure mode explicit and remove it directly: we attach auxiliary emotion classifiers to selected student features $\{h_\ell^S\}_{\ell \in \mathcal{L}_{\text{adv}}}$ and apply a gradient reversal layer (GRL). This trains the classifiers to predict emotion while forcing the proxy encoder to erase emotion-predictive signals across multiple abstraction levels:

$$\mathcal{L}_{\text{adv}} = \sum_{\ell \in \mathcal{L}_{\text{adv}}} \text{CE}\big(C_\ell(\text{GRL}(h_\ell^S)),\, y\big)\,. \tag{3}$$

Proxy objective. These three constraints shape z_1 into a view-consistent, nuisance-dominant, and emotion-suppressed proxy: alignment stabilizes it across views, contrastive learning biases it toward within-class nuisance variation, and adversarial training removes residual emotion leakage. We optimize:

$$\mathcal{L}_{\text{proxy}} = \mathcal{L}_{\text{align}} + \lambda_c \mathcal{L}_{\text{con}} + \lambda_a \mathcal{L}_{\text{adv}}, \tag{4}$$

yielding the nuisance proxy z_1 that we will later use as the conditioning reference for label-conditioned disentanglement.

3.2 Consistency-Aware Emotion Representation Learning Module

We allow z_2 to retain some nuisance information at the early stage. Given temporally sparse and uneven emotion evidence under utterance-level supervision, we first model segment-level consistency to obtain a stable coarse-grained emotion embedding, which serves as the input to subsequent conditional disentanglement.

Structuring emotional segments with a global anchor. We embed frame-level features into overlapping patches and prepend a learnable `[CLS]` token as an utterance-level anchor:

$$[\mathbf{c}, p_1, \ldots, p_N] = \text{T}\big([\texttt{[CLS]};\, \text{patch}(x; L, S)]\big), \tag{5}$$

where $\mathbf{c}$ summarizes global evidence and p_i are contextualized patch embeddings.

Scoring emotional consistency of segments. We cluster patch tokens into M groups and compute centroids

$$\mu_m = \frac{1}{|\mathcal{C}_m|} \sum_{p_i \in \mathcal{C}_m} p_i, \tag{6}$$

then quantify each cluster's label consistency via similarity to the anchor:

$$s_m = \text{sim}(\mu_m, \mathbf{c}). \tag{7}$$

This score distinguishes clusters strongly aligned with the utterance label from weakly aligned or contradictory regions.

Extracting emotion representation by consistency. We use a pool of experts with graded capacity and route clusters by the rank of s_m, assigning higher-capacity experts to low-consistency clusters to avoid under-modeling ambiguous regions. Let $\pi(\cdot)$ map the rank of s_m to an expert index:

$$\tilde{\mu}_m = E_{\pi(m)}(\mu_m). \tag{8}$$

Finally, we fuse the expert outputs with the anchor via cross-attention to obtain z_2:

$$z_2 = \text{Fuse}\big(\mathbf{c}, \{\tilde{\mu}_m\}_{m=1}^M\big)\,. \tag{9}$$

By down-weighting ambiguous or weakly emotional regions, this module produces a coarse-grained emotion representation that is well suited for label-conditioned disentanglement.

3.3 Emotion-Preserving Conditional Information Bottleneck

Since emotion and nuisance factors are entangled and often share acoustic carriers, enforcing unconditional independence between z_1 and z_2 would discard useful emotion cues. We therefore adopt label-conditioned disentanglement to decouple z_2 from nuisance variability captured by z_1 while preserving emotion-discriminative structure.

From IB to conditional nuisance suppression. We start from the standard information bottleneck (IB) objective [Tishby *et al.*, 2000]

$$\mathcal{L}_{\text{IB}} = I(x; z_2) - \beta I(z_2; y), \tag{10}$$

where $z_2 = f_\theta(x)$ denotes the emotion representation. Under the Markov chain $y \to x \to z_2$, we have $I(y; z_2 \mid x) = 0$. Applying the chain rule yields

$$I(x; z_2) = I(x; z_2 \mid y) + I(z_2; y), \tag{11}$$

and therefore

$$\mathcal{L}_{\text{IB}} = I(x; z_2 \mid y) + (1 - \beta) I(z_2; y). \tag{12}$$

For $\beta > 1$, minimizing $\mathcal{L}_{\text{IB}}$ encourages large mutual information between z_2 and y, while penalizing the conditional dependence $I(x; z_2 \mid y)$.

Importantly, for any nuisance factor N that is independent of y and satisfies $N \to x \to z_2 \mid y$, the data processing inequality gives

$$I(N; z_2 \mid y) \le I(x; z_2 \mid y). \tag{13}$$

Thus, minimizing $I(x; z_2 \mid y)$ implicitly suppresses within-class nuisance dependence.

Nuisance proxy formulation. In domain-free SER, nuisance factors are latent and cannot be accessed explicitly. We therefore introduce an emotion-suppressed nuisance proxy $z_1 = g_\phi(x)$ as a tractable surrogate. Assuming that z_1 is approximately sufficient to capture within-class nuisance variability, we reformulate the objective as the following conditional bottleneck loss:

$$\mathcal{L}_{\text{CIB}} = -\, I(z_2; y) + \lambda\, I(z_1; z_2 \mid y), \tag{14}$$

which aims to improve emotion discrimination while alleviating within-class nuisance dependence.

CASIA				**EmoDB**			
Reference	Method	WAR (%)	UAR (%)	Reference	Method	WAR (%)	UAR (%)
[Ye *et al.*, 2023]	TIMNET	94.67	94.67	[Yuan *et al.*, 2024]	DTNET	95.79	95.50
[Li *et al.*, 2024]	SENET	95.16	94.95	[Li *et al.*, 2024]	SENET	96.40	96.83
[Li *et al.*, 2025]	AMHNET	95.97	95.21	[Li *et al.*, 2025]	AMHNET	98.11	98.15
Our approach (w/o CIB)	EP-CIB	93.82	93.35	Our approach (w/o CIB)	EP-CIB	97.36	97.41
Our approach	**EP-CIB**	**96.11**	**96.16**	**Our approach**	**EP-CIB**	**99.20**	**99.06**
IEMOCAP				**RAVDESS**			
Reference	Method	WAR (%)	UAR (%)	Reference	Method	WAR (%)	UAR (%)
[Yuan *et al.*, 2024]	DTNET	78.65	77.92	[Ye *et al.*, 2023]	TIMNET	92.08	91.93
[Li *et al.*, 2024]	SENET	73.38	73.67	[Li *et al.*, 2024]	SENET	93.56	93.28
[Vijayan and Judy, 2025]	EFNET	79.30	79.14	[Li *et al.*, 2025]	AMHNET	95.16	95.13
Our approach (w/o CIB)	EP-CIB	78.97	78.91	Our approach (w/o CIB)	EP-CIB	94.42	95.07
Our approach	**EP-CIB**	**82.34**	**81.48**	**Our approach**	**EP-CIB**	**97.38**	**97.40**

Table 1: Speaker-dependent comparison on CASIA, EmoDB, IEMOCAP, and RAVDESS using WAR/UAR (%). Best results are in bold.

Conditional Residualized HSIC Bottleneck. Direct estimation of $I(z_1; z_2 \mid y)$ is unstable in high-dimensional neural representations. Following recent information-theoretic practice, we adopt a residualized dependence surrogate. Specifically, we remove class-conditional components:

$$\tilde{z}_1 = z_1 - r_1(y), \qquad \tilde{z}_2 = z_2 - r_2(y), \tag{15}$$

where $r_1(y) \approx \mathbb{E}[z_1 \mid y]$ and $r_2(y) \approx \mathbb{E}[z_2 \mid y]$ are learnable residualizers.

The dependence between $(\tilde{z}_1, \tilde{z}_2)$ therefore reflects within-class correlation induced by nuisance factors. We measure this dependence using a linear-kernel HSIC, yielding the Conditional Residualized HSIC Bottleneck (C-RHB):

$$\mathcal{L}_{\text{C-RHB}} = \mathrm{f}_{\text{HSIC}}(\tilde{z}_1, \tilde{z}_2). \tag{16}$$

Final objective and training protocol. The emotion encoder is optimized by

$$\mathcal{L} = \mathcal{L}_{\text{emo}} + \beta\, \mathcal{L}_{\text{C-RHB}}, \tag{17}$$

which preserves emotion-discriminative structure and mitigates nuisance-induced within-class variability under open-ended domain shifts.

4 Experiments

4.1 Experiments Setup

Datasets. We evaluate our method on four widely used emotional speech corpora: IEMOCAP (English) [Busso *et al.*, 2008], CASIA (Mandarin) [Zhang and Jia, 2008], EmoDB (German) [Burkhardt *et al.*, 2005], and RAVDESS (English) [Livingstone and Russo, 2018]. These datasets differ substantially in speakers, languages, recording conditions, and emotion inventories, providing a diverse test bed for domain generalization. The numbers of utterances in IEMOCAP, CASIA, EmoDB, and RAVDESS are 5531, 1200, 535, and 1440, respectively.

Feature Extraction. We use a pre-trained HuBERT model to extract deep acoustic representations, which serve as the input features for our framework. For baseline methods, we directly follow the feature configurations and evaluation protocols reported in their original papers.

Experimental setting. All experiments are conducted on an NVIDIA A100 GPU. Following common practice in SER evaluation, we adopt three standard protocols to assess generalization performance: speaker-dependent (SD), speaker-independent (SID), and cross-corpus (CC) settings. These protocols are consistent with those used in previous studies. In the SD setting, samples within each corpus are randomly split into ten folds for cross-validation. In the SID setting, data are partitioned by speaker, with one speaker held out for testing in each split. In the CC setting, one corpus is used for training and a different corpus is used entirely for testing, in order to evaluate robustness under cross-domain shifts.

Evaluation Metrics. We report Weighted Average Recall (WAR) and Unweighted Average Recall (UAR). WAR reflects recall under the empirical class distribution, while UAR averages recall across classes to characterize class-balanced performance.

4.2 Comparison to State-of-the-art

Tables 1–3 report the performance of EP-CIB under SD, SID, and CC protocols, which correspond to progressively stronger and more open-ended nuisance shifts.

Under the SD setting, EP-CIB achieves performance comparable to existing methods across all datasets in Table 1. The full model consistently outperforms its w/o CIB variant, indicating that the conditional bottleneck does not compromise in-domain emotion discrimination.

Under the SID setting, EP-CIB shows larger gains on all corpora in Table 2, where speaker-related variability dominates. Compared with w/o CIB, the full model consistently improves performance, suggesting effective suppression of within-class speaker nuisance while preserving emotion-discriminative structure.

Under the CC setting, where language, recording conditions, and speakers shift simultaneously, EP-CIB achieves the best performance in most transfer directions and the highest average results in Table 3. TF-DWFormer reports a higher UAR on CASIA→EmoDB, likely because its fine-grained local time–frequency modeling benefits from the clear and

CASIA				**EmoDB**			
Reference	Method	WAR (%)	UAR (%)	Reference	Method	WAR (%)	UAR (%)
[Liu *et al.*, 2025]	AEDD	67.05	67.05	[Yuan *et al.*, 2024]	DTNET	91.43	91.12
[Yan *et al.*, 2025]	MMMAE	55.25	55.25	[Liu *et al.*, 2025]	AEDD	90.78	90.16
[Lu *et al.*, 2026]	GDDA	62.58	62.58	[Lu *et al.*, 2026]	GDDA	90.01	88.75
Our approach (w/o CIB)	EP-CIB	62.54	62.54	Our approach (w/o CIB)	EP-CIB	88.12	87.51
Our approach	**EP-CIB**	**72.33**	**72.33**	**Our approach**	**EP-CIB**	**95.72**	**95.64**
IEMOCAP				**RAVDESS**			
Reference	Method	WAR (%)	UAR (%)	Reference	Method	WAR (%)	UAR (%)
[Yuan *et al.*, 2024]	DTNET	75.11	74.80	[Xue *et al.*, 2025]	MLDSF	70.28	69.73
[Liu *et al.*, 2025]	AEDD	76.22	75.47	[Soltani *et al.*, 2025]	DRESN	75.45	74.79
[Lu *et al.*, 2026]	GDDA	75.26	69.52	[George and others, 2025]	Whisper	-	79.56
Our approach (w/o CIB)	EP-CIB	70.12	70.54	Our approach (w/o CIB)	EP-CIB	69.27	66.41
Our approach	**EP-CIB**	**78.05**	**75.15**	**Our approach**	**EP-CIB**	**85.32**	**84.31**

Table 2: Speaker-independent comparison on CASIA, EmoDB, IEMOCAP, and RAVDESS using WAR/UAR (%). Best results are in bold.

Source Corpus			CASIA	EmoDB	CASIA	IEMOCAP	EmoDB	IEMOCAP	Average
Target Corpus			EmoDB	CASIA	IEMOCAP	CASIA	IEMOCAP	EmoDB	
[Zhu *et al.*, 2020]	DSAN	UAR	50.69	40.30	33.01	43.86	37.07	49.15	42.35
		WAR	51.81	40.30	50.44	43.86	51.37	53.39	48.53
[Fan *et al.*, 2025]	TFDWF	UAR	**76.44**	47.50	39.66	47.50	46.01	53.14	56.99
		WAR	-	-	-	-	-	-	-
[Lu *et al.*, 2026]	GDDA	UAR	52.43	42.70	40.12	47.25	41.23	48.88	45.44
		WAR	57.11	42.70	53.90	47.25	54.82	56.93	52.12
Our approach (w/o CIB)	EP-CIB	UAR	43.15	47.89	28.09	38.26	32.88	55.77	41.01
		WAR	46.98	47.89	38.76	38.26	36.65	58.92	44.58
Our approach	**EP-CIB**	UAR	63.72	**65.66**	**40.17**	**52.50**	**48.33**	**69.91**	**57.05**
		WAR	**64.60**	**65.66**	**55.43**	**52.50**	**59.72**	**72.27**	**61.70**

Table 3: Cross-corpus comparison on paired source and target corpora using WAR/UAR (%). Best results are in bold.

structured emotional patterns in acted corpora. However, this advantage becomes less stable for transfers involving IEMOCAP, where scripted and improvised speech produce more dispersed emotional evidence and less reliable local cues. EP-CIB instead models broad nuisance shifts and segment-level inconsistency through nuisance proxy learning, consistency-aware emotion modeling, and within-class nuisance suppression, leading to more stable cross-corpus robustness.

Overall, EP-CIB maintains strong in-domain accuracy and yields progressively larger robustness gains from SD to SID and CC, validating its effectiveness for domain-free generalization in SER.

4.3 The Ablation Study

We design ablation studies on the sub-components of EP-CIB to quantify their individual contributions under weak utterance-level supervision, with the results reported in Tables 4 and 5.

In Table 4, removing any constraint degrades performance, with the largest drop when disabling adversarial emotion removal (Adv-GRL), suggesting it is important for keeping z_1 minimally emotion-informative. Multi-scale teacher–student alignment enforces view-consistent encoding across crops/layers, and the class-restricted contrastive term further discourages emotion-driven similarity within each class, together supporting z_1 as a reliable nuisance surrogate for label-conditioned disentanglement.

Variant	SD		SID	
	WAR	UAR	WAR	UAR
Full model	**99.20**	**99.06**	**95.72**	**95.64**
w/o Multi-Align	97.84	97.51	92.43	91.98
w/o Contrastive	97.36	97.02	91.61	90.87
w/o Adv-GRL	92.05	92.78	86.72	86.94

Table 4: Ablation study of the nuisance proxy representation learning module on EmoDB.

In Table 5, both clustering-based consistency estimation and capacity-aware expert routing contribute to robustness, and removing both causes the most severe drop. This supports our motivation that emotion evidence is temporally sparse and uneven: clustering separates regions by label consistency, while routing allocates capacity to ambiguous segments to stabilize the coarse-grained z_2 for subsequent disentanglement.

Variant	SD		SID	
	WAR	UAR	WAR	UAR
Full model	**99.20**	**99.06**	**95.72**	**95.64**
w/o Clustering	96.92	96.48	90.81	90.02
w/o Multi-Expert	97.18	96.83	91.57	90.94
w/o Clust. & Expert	93.67	93.21	88.64	87.92

Table 5: Ablation study of the consistency-aware emotion representation learning module on EmoDB.

Representation	Probe Accuracy (%)			
	Speaker ID	Gender	Emotion	Language
z_1	**91.96**	**98.03**	27.38	**88.14**
z_2 (w/o CIB)	45.72	74.29	73.26	51.71
z_2 (with CIB)	17.96	35.15	**84.57**	29.83

Table 6: Linear probing results on frozen nuisance proxy z_1 and emotion representation z_2.

Overall, the ablations are consistent with our narrative: a proxy that captures broad non-emotional variability and a consistency-aware emotion encoder jointly enable the conditional bottleneck to mitigate within-class nuisance effects while retaining emotion-discriminative structure.

4.4 Representation Probing Analysis

We run linear probes on frozen embeddings to diagnose what each representation encodes. A linear SVM is used to restrict classifier capacity, so that the results mainly reflect the information already stored in the embedding. Under the intended emotion–nuisance factorization, the nuisance proxy should predict nuisance attributes well but remain weak on emotion, while the emotion representation should exhibit the opposite pattern.

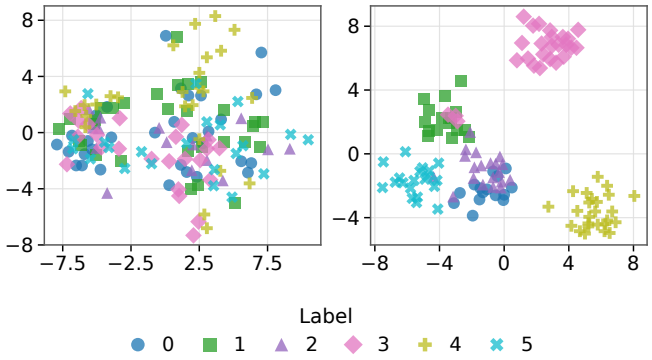
Table 6 shows that the nuisance proxy z_1 is highly predictive of speaker identity, gender, and language, yet weak on emotion, suggesting it captures nuisance-dominant variation. In contrast, z_2 without CIB retains substantial nuisance predictability. With CIB, nuisance predictability in z_2 drops markedly while emotion predictability improves, indicating that the conditional bottleneck can mitigate within-class nuisance dependence without sacrificing emotion-discriminative structure.

4.5 Representation Visualization

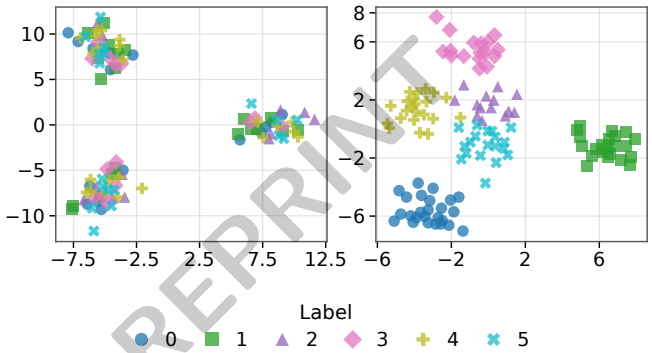
We visualize the nuisance proxy z_1 and the emotion representation z_2 using t-SNE to assess the effect of the conditional bottleneck on domain-free disentanglement.

As shown in Fig. 2(a), without CIB, z_1 shows no clear clustering structure. In contrast, z_2 forms loosely organized emotion clusters with evident overlap, indicating that nuisance variation still affects emotion discrimination.

Fig. 2(b) shows that, with CIB, z_1 separates into three distinct clusters that correspond to the three training speakers. Meanwhile, emotion clusters in z_2 become more compact and better separated.



(a) Model w/o CIB: t-SNE visualization of z_1 and z_2.



(b) Full Model: t-SNE visualization of z_1 and z_2.

Figure 2: Comparison of t-SNE visualizations of the nuisance proxy z_1 and the emotion representation z_2 with and without CIB.

These results suggest that the conditional bottleneck reduces within-class nuisance influence while preserving emotion-discriminative structure, leading to a more stable and emotion-oriented embedding space. This finding is consistent with the probing analysis and explains the improved robustness under speaker-independent and cross-corpus settings.

5 Conclusion

This paper studies domain-free DG-SER under domain shift. To address this challenge, we proposes EP-CIB for domain-free generalization in SER. Firstly, it uses a nuisance proxy representation learning module to represent nuisance information, thus enabling the modeling of nuisance factors without the need for domain annotation. Secondly, it constructs a consistency-aware emotion representation module to extract a coarse-grained emotion representation. Finally, it introduces a conditional information bottleneck mechanism to disentangle the nuisance proxy from the coarse-grained emotion representation while retaining the effective acoustic carriers. EP-CIB achieves the state-of-the-art performance in the SD, SID, and CC experimental setups, and still maintains good robustness under stronger domain shifts. Overall, EP-CIB offers a practical formulation of label-conditioned nuisance suppression for domain-free SER, improving robustness when deployment shifts are open-ended.

Acknowledgments

This work was funded in part by the National Natural Science Foundation of China grant under number 62536004, and in part by the Key-Area Research and Development Program of Guangdong Province under number 2023B0303030001, and in part by the Science and Technology Program of Guangzhou under number 2024A04J6310, and in part by the Fundamental Research Funds for the Central Universities 2025ZYGXZR021. Tong Zhang is the corresponding author.

References

[Burkhardt *et al.*, 2005] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[Busso *et al.*, 2008] 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.

[Cai *et al.*, 2019] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, page 2060, 2019.

[Chen *et al.*, 2024] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):1711–1724, 2024.

[Desplanques *et al.*, 2020] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[Fadhil and Zahra, 2025] Muhammad Farhan Fadhil and Amalia Zahra. The use of generative adversarial network as a domain adaptation method for cross-corpus speech emotion recognition. *Bulletin of Electrical Engineering and Informatics*, 14(1):297–306, 2025.

[Fan *et al.*, 2025] Yonghong Fan, Heming Huang, Huiyun Zhang, and Ziqi Zhou. Temporal-frequency joint hierarchical transformer with dynamic windows for speech emotion recognition. *Engineering Applications of Artificial Intelligence*, 161:112152, 2025.

[George and others, 2025] Swapna Mol George et al. Advancing speech emotion recognition with whisper model embeddings and hand-crafted audio descriptors. *Franklin Open*, page 100403, 2025.

[Gideon *et al.*, 2021] John Gideon, Melvin G McInnis, and Emily Mower Provost. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog). *IEEE Transactions on Affective Computing*, 12(4):1055–1068, 2021.

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[Huang and Ren, 2024] Zixian Huang and Chuan-Xian Ren. Rethinking correlation learning via label prior for open set domain adaptation. In *IJCAI: proceedings of the conference*, pages 884–892, 2024.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *2021 International Conference on Machine Learning*, July 2021.

[Kurt Ma *et al.*, 2020] Wan-Duo Kurt Ma, J.P. Lewis, and W. Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 5085–5092, 2020. Cited by: 123.

[Latif *et al.*, 2023] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Bjorn Schuller. Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):1912–1926, 2023.

[Li *et al.*, 2019] Runnan Li, Zhiyong Wu, Jia Jia, Yaohua Bu, Sheng Zhao, and Helen Meng. Towards discriminative representation learning for speech emotion recognition. In *IJCAI*, volume 2019, pages 5060–5066, 2019.

[Li *et al.*, 2024] Mengbo Li, Yuanzhong Zheng, Dichucheng Li, Yulun Wu, Yaoxuan Wang, and Haojun Fei. Ms-senet: Enhancing speech emotion recognition through multi-scale feature fusion with squeeze-and-excitation blocks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12271–12275. IEEE, 2024.

[Li *et al.*, 2025] Hengrui Li, Yongbing Zhang, and Shaohui Liu. Amh-net: Adaptive multi-band hybrid-aware network for emotion recognition in speech. *IEEE Signal Processing Letters*, 2025.

[Liu *et al.*, 2025] Yang Liu, Xin Chen, Yarong Li, Jie Ma, Xiaoqi Yang, Yuan Song, Xiaolei Meng, Yongwei Li, Xingfeng Li, and Zhen Zhao. Enhanced speech emotion recognition in noisy environments: Adaptive emotion denoising diffusion approach with iterative confidence learning strategy. *IEEE Internet of Things Journal*, 2025.

[Livingstone and Russo, 2018] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[Lu *et al.*, 2026] Cheng Lu, Yuan Zong, Hailun Lian, Sunan Li, Yan Zhao, Björn W. Schuller, and Wenming Zheng. Generalizable dynamic domain adaptation for speaker-independent speech emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 34:230–245, 2026.

[Ma *et al.*, 2024] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, 2024.

[Soltani *et al.*, 2025] Rebh Soltani, Emna Benmohamed, and Hela Ltifi. Topology-adaptive bayesian optimization for deep ring echo state networks in speech emotion recognition. *Neural Computing and Applications*, 37(1):399–416, 2025.

[Tao *et al.*, 2022] Huawei Tao, Yang Wang, Zhihao Zhuang, Hongliang Fu, Xinying Guo, and Shuguang Zou. Cross-corpus speech emotion recognition based on transfer learning and multi-loss dynamic adjustment. *Computational Intelligence and Neuroscience*, 2022, 2022. Cited by: 5; All Open Access, Green Open Access, Hybrid Gold Open Access.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[Tjandra *et al.*, 2021] Andros Tjandra, Ruoming Pang, Yu Zhang, and Shigeki Karita. Unsupervised learning of disentangled speech content and style representation. In *Interspeech 2021*, pages 4089–4093, 2021.

[Vijayan and Judy, 2025] Bineetha Vijayan and MV Judy. Emofusionnet: A unified approach for robust speech emotion recognition. *Digital Signal Processing*, 162:105173, 2025.

[Wang *et al.*, 2021] Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, and Jennifer Dy. Revisiting hilbert-schmidt information bottleneck for adversarial robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 586–597. Curran Associates, Inc., 2021.

[Wen *et al.*, 2022] Xin-Cheng Wen, JiaXin Ye, Yan Luo, Yong Xu, Xuan-Ze Wang, Chang-Li Wu, and Kun-Hong Liu. Ctl-mtnet: A novel capsnet and transfer learning-based mixed task net for single-corpus and cross-corpus speech emotion recognition. In *IJCAI*, pages 2305–2311. International Joint Conferences on Artificial Intelligence Organization, 2022.

[Xue *et al.*, 2025] Peiyun Xue, Xiang Gao, Jing Bai, Zhenan Dong, Zhiyu Wang, and Jiangshuai Xu. A dynamic-static feature fusion learning network for speech emotion recognition. *Neurocomputing*, 633:129836, 2025.

[Yan *et al.*, 2025] Jingjie Yan, Wenjing Sun, Boyan Sun, Xiaoyang Zhou, Guanming Lu, Xiaonan Li, and Xianlan Zheng. Neonatal pain speech emotion recognition based on multiscale multilevel mae network. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS*, 2025.

[Ye *et al.*, 2023] Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Yu *et al.*, 2021] Shujian Yu, Luis G Sánchez Giraldo, and José C Príncipe. Information-theoretic methods in deep neural networks: Recent advances and emerging opportunities. In *IJCAI*, pages 4669–4678, 2021.

[Yuan *et al.*, 2024] Zhichen Yuan, CL Philip Chen, Shuzhen Li, and Tong Zhang. Disentanglement network: Disentangle the emotional features from acoustic features for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11686–11690. IEEE, 2024.

[Zhang and Jia, 2008] JTFLM Zhang and Huibin Jia. Design of speech corpus for mandarin text to speech. In *The blizzard challenge 2008 workshop*, 2008.

[Zhao *et al.*, 2024a] Yan Zhao, Jincen Wang, Cheng Lu, Sunan Li, Björn W Schuller, Yuan Zong, and Wenming Zheng. Emotion-aware contrastive adaptation network for source-free cross-corpus speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11846–11850. IEEE, 2024.

[Zhao *et al.*, 2024b] Yan Zhao, Yuan Zong, Hailun Lian, Cheng Lu, Jingang Shi, and Wenming Zheng. Towards domain-specific cross-corpus speech emotion recognition approach. *IEEE Transactions on Computational Social Systems*, 2024.

[Zhu *et al.*, 2020] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems*, 32(4):1713–1722, 2020.