DISAGREEMENT IS NOT ALWAYS INFORMATION: TASK-RELEVANT COUNTERFACTUAL CONSENSUS REPAIR UNDER MULTI-VIEW CORRUPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Disagreement between multiple observations is often treated as evidence of corruption or complementary information, but disagreement need not be useful for task-relevant repair. We ask when a sample's own disagreement pattern adds value beyond the pair consensus and scalar disagreement severity. Given clean representation z_0 and corrupted representations z_1, z_2, we define $M = (z_1 + z_2)/2$ and repair target $T = M - z_0$. Under squared loss, disagreement has incremental value only when conditioning on it changes the task-visible conditional mean of T. We introduce Counterfactual Consensus Repair (CCR), which compares the target's own disagreement detail with a severity-matched shuffled control while holding fixed the target consensus, scalar severity, global parent, pair architecture, and optimization. Our primary contrast is $\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$, so negative values indicate value from the target's own disagreement detail. Across Speech Commands V2, ESC-50, and UrbanSound8K, this contrast is consistently favorable in asymmetric, heteroscedastic, and structured cross-family regimes; in the cross-family regime, $\Delta_{\text{detail}} = -0.755, -0.903$, and -0.562, respectively. Exact representation-space probes recover the predicted IID Gaussian null and heteroscedastic boundary, and the matched effect replicates on ESC-50 with a frozen pretrained BEATs representation. Zero-shot DHAuDS-Pair experiments, with no repair retraining, show that the learned disagreement-to-repair mapping transfers selectively: external corruption families can be beneficial, near null, or adverse. These results support a conditional rather than universal view of disagreement.

1 INTRODUCTION

Multiple observations of the same semantic instance can improve robustness because they provide both shared evidence and view-specific variation. In multi-view learning, uncertainty-aware fusion, and test-time adaptation, disagreement is often used as a cue for corruption, reliability, or complementary information (Christoudias et al., 2008; Liu et al., 2022; Yang et al., 2024; Xu et al., 2026). But the presence of disagreement does not imply that its detailed structure is useful for correcting a corrupted representation. Two views may disagree in ways that are already explained by their consensus, reveal only corruption magnitude, or affect directions irrelevant to the downstream task. We therefore ask a narrower question: *when does a sample's own disagreement pattern provide task-relevant information beyond the pair consensus and scalar disagreement severity?* Given a clean training representation z_0 and two corrupted representations z_1, z_2, we define

$$M = \frac{z_1 + z_2}{2}, \qquad T = M - z_0$$

where M is the pair consensus and T is the counterfactual displacement to be repaired. A consensus-only estimator predicts T from M; a pair-aware estimator additionally observes disagreement D. Under squared loss, the classical conditional-expectation decomposition implies that disagreement can reduce Bayes risk only when conditioning on D changes the conditional mean of T. For classification, we further restrict this change to directions visible through the centered logits of the frozen classifier. This yields a task-visible estimand that CCR isolates experimentally beyond consensus, scalar severity, and matched model capacity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We operationalize this question with *Counterfactual Consensus Repair* (CCR). CCR first trains a consensus-only global repair and then freezes it before learning a pair residual from disagreement. The central design is not the additional pair network itself, but a matched control that preserves the target consensus, the target's scalar disagreement severity, the global parent, architecture, objective, and optimization while replacing only the target-specific disagreement detail with that of a severity-matched donor. Our primary contrast is

$$\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$$

so $\Delta_{\text{detail}} < 0$ indicates that the target sample's own disagreement detail contributes information beyond consensus, severity, and matched model capacity. Figure 1 summarizes the paired counterfactual setup, nested repair architecture, and matched identification control.

This formulation produces falsifiable positive and negative cases. An exact additive IID Gaussian model in representation space gives a no-value reference, whereas heteroscedastic corruption breaks the independence underlying that null. We test these predictions together with asymmetric, structured, and common-mode corruptions on Speech Commands V2, ESC-50, and UrbanSound8K. Correct sample-specific disagreement consistently improves centered-logit reconstruction in the asymmetric, heteroscedastic, and structured cross-family regimes across all three datasets, while null-like and adverse cases remain. In the cross-family regime, Δ_{detail} is -0.755, -0.903, and -0.562 on Speech Commands, ESC-50, and UrbanSound8K, respectively. Exact representation-space probes recover the predicted IID null and heteroscedastic boundary, and the matched effect also replicates on ESC-50 with a frozen pretrained BEATs representation.

Finally, we test whether a disagreement-to-repair mapping learned under controlled corruptions transfers without retraining to externally designed DHAuDS corruptions. It does not do so universally: some external corruption families benefit, others are near null, and others are adverse. This boundary is central to the paper rather than an exception to it. Together, the experiments support a conditional view of disagreement: its value depends on the joint structure of the representation, nuisance, and task, rather than on the mere existence of disagreement.

Our contributions are:

- We formulate the incremental value of multi-view disagreement as a task-visible conditional-estimation problem, separating disagreement *existence* from disagreement *value*.
- We establish an exact representation-space IID Gaussian no-value result and characterize how heteroscedastic structure breaks the independence underlying that null, including the effect of randomizing the high/low-noise view.
- We introduce CCR and a severity-matched disagreement control that isolate sample-specific disagreement detail from consensus, scalar severity, and additional pair-model capacity.
- Across three controlled audio datasets, representation-space probes, a frozen-BEATs replication, and zero-shot DHAuDS-Pair transfer, we find beneficial, null-like, and adverse disagreement effects, supporting a conditional rather than universal account of disagreement value.

2 Related Work

Multi-view disagreement, conflict, and reliability. Multi-view learning typically exploits complementary information across views, but disagreement can also indicate corruption, conflict, or unequal reliability. Christoudias et al. (2008) detect corruption-induced disagreement and filter inconsistent examples. Trusted multi-view methods instead model view-specific uncertainty and combine uncertain predictions during fusion (Liu et al., 2022). More directly, Xu et al. (2024) model conflicting instances through view-specific evidence and reliability-aware opinion aggregation, while Lu et al. (2025) use instance-wise computational trust to discount unreliable views when resolving cross-view conflicts. These approaches use disagreement, uncertainty, or trust to determine how views should be retained, weighted, or fused. CCR asks a different question: after conditioning on pair consensus and scalar disagreement severity, does the *target sample's own detailed disagreement* contain additional information about the task-visible repair target? We therefore study disagreement as an incremental conditional-estimation signal rather than assuming that conflict itself should be suppressed or reliability-weighted.

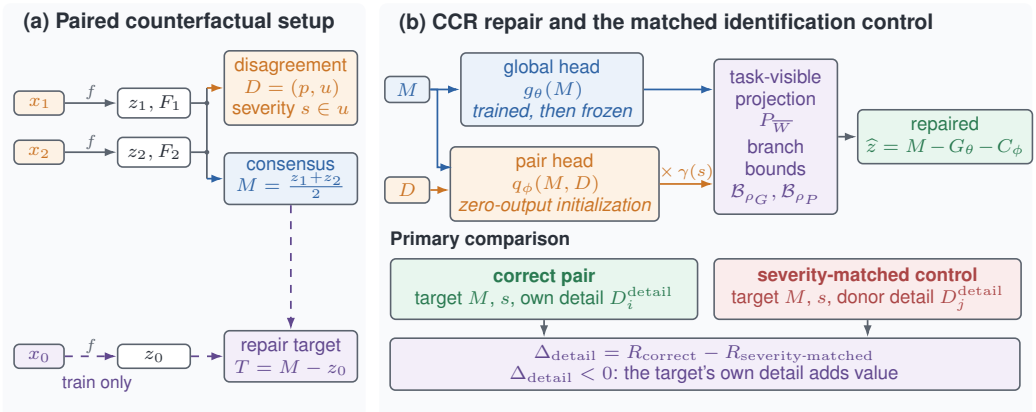


Figure 1: CCR architecture and matched identification control. **(a)** Two corrupted views define the consensus M and disagreement descriptor $D = (p, u)$; the clean view is used only during training to construct the repair target $T = M - z_0$. **(b)** CCR first learns a consensus-only global correction and then, with that parent frozen, learns a gated pair-specific residual from disagreement. The primary comparison keeps the target consensus and exact severity fixed while replacing only the target's disagreement detail with severity-matched donor detail. Thus $\Delta_{\text{detail}} < 0$ indicates that the target's own detailed disagreement contributes task-visible repair information beyond consensus, severity, and matched pair-model capacity.

Multimodal imbalance, robust fusion, and adaptation. A related literature studies why modalities contribute unequally and how training or fusion should compensate. Wang et al. (2020) show that modalities can overfit and generalize at different rates during joint training and introduce gradient blending, while Peng et al. (2022) identify modality optimization imbalance and adaptively modulate gradients to prevent a dominant modality from suppressing others. Under corrupted or shifted inputs, READ adapts cross-modal attention and predictions to modality reliability imbalance (Yang et al., 2024), and related methods use probabilistic calibration or instance-level adapt-then-fuse mechanisms (Xu et al., 2026; Zhang et al., 2026). CAV2vec learns audio-visual representations robust to corrupted inputs by predicting clean teacher targets (Kim et al., 2025). These methods intervene on training, adaptation, or fusion to improve multimodal robustness. CCR instead freezes the semantic encoder and classifier and isolates whether sample-specific disagreement detail changes a task-visible repair estimate beyond consensus and severity.

Paired denoising and counterfactual repair. Noise2Noise learns from paired noisy observations under suitable expectation assumptions, while Noise2Self exploits conditional independence for self-supervised denoising (Lehtinen et al., 2018; Batson & Royer, 2019). CCR is not a clean-target-free denoising method: a clean training observation defines the counterfactual representation target $T = M - z_0$. The repaired object is a frozen semantic representation, and our primary estimand is the incremental task-visible value of disagreement after conditioning on consensus and scalar severity.

Audio robustness and external shift. BEATs provides a strong pretrained audio representation (Chen et al., 2023); we use it as a representation intervention to test whether the matched disagreement effect persists with a stronger frozen encoder. DHAuDS studies dynamic heterogeneous audio corruptions for test-time adaptation (Shao et al., 2025). Our DHAuDS-Pair experiment instead keeps the learned repair heads frozen and tests zero-shot transfer of the disagreement-to-repair mapping rather than adaptation to the DHAuDS target distribution.

3 Problem and Theory

3.1 Paired counterfactual repair and task geometry

Let x_0 be a clean training observation and x_1, x_2 two semantic-preserving corruptions of the same instance. A frozen encoder gives $z_i = f(x_i) \in \mathbb{R}^d$, and a frozen K-class linear classifier is $h(z) = Wz + b$, with $W \in \mathbb{R}^{K \times d}$ and $b \in \mathbb{R}^K$. Define the consensus and repair target as

$$M = \frac{z_1 + z_2}{2}, \qquad T = M - z_0 \tag{1}$$

A repair r returns $\widehat{z} = M - r$, so exact representation recovery corresponds to $r = T$.

Uniform logit shifts do not affect softmax or argmax. Let

$$H_K = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \qquad \overline{W} = H_K W \tag{2}$$

and define the task-visible repair risk

$$\mathcal{R}_{\overline{W}}(r) = \frac{1}{K}\mathbb{E}\left\|\overline{W}(T - r)\right\|_2^2 \tag{3}$$

Only corrections in $\mathrm{row}(\overline{W})$ can affect this risk; the projection argument is given in Appendix B.1. A sufficient condition under which the repaired representation preserves the clean predicted class is given in Appendix B.8.

3.2 Conditional value of disagreement

Let

$$\mathcal{G} = \sigma(M), \qquad \mathcal{D} = \sigma(M, D), \qquad \mu_G = \mathbb{E}[T \mid \mathcal{G}], \qquad \mu_D = \mathbb{E}[T \mid \mathcal{D}]$$

The following is the standard L_2 decomposition for nested conditional expectations; we use it to define the quantity measured by our experiments.

Theorem 1 (Nested conditional-risk identity)**.** *Assume $\mathbb{E}\|T\|_2^2 < \infty$. For any fixed matrix A,*

$$\mathbb{E}\,\|A(T - \mu_G)\|_2^2 - \mathbb{E}\,\|A(T - \mu_D)\|_2^2 = \mathbb{E}\,\|A(\mu_D - \mu_G)\|_2^2 \tag{4}$$

In particular,

$$\mathcal{R}_{\overline{W}}(\mu_G) - \mathcal{R}_{\overline{W}}(\mu_D) = \frac{1}{K}\mathbb{E}\left\|\overline{W}(\mu_D - \mu_G)\right\|_2^2 \tag{5}$$

Hence disagreement has zero incremental task-visible Bayes value iff $\overline{W}(\mu_D - \mu_G) = 0$ almost surely. Positive dependence or conditional mutual information alone is therefore insufficient: disagreement must change the conditional mean in a classifier-visible direction. The full decomposition and proof are in Appendix B.3.

3.3 Exact Gaussian null and heteroscedastic boundary

Theorem 2 (IID additive Gaussian null)**.** *Assume $z_i = z_0 + \varepsilon_i$ with $\varepsilon_1, \varepsilon_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ independent of z_0. Let*

$$T = \frac{\varepsilon_1 + \varepsilon_2}{2}, \qquad Q = \varepsilon_1 - \varepsilon_2, \qquad M = z_0 + T$$

For any measurable representation-space descriptor $D_z = d(M, Q)$,

$$\mathbb{E}[T \mid M, D_z] = \mathbb{E}[T \mid M] \tag{6}$$

Thus D_z has zero incremental Bayes value under latent or centered-logit squared loss.

Here $\mathrm{Cov}(T, Q) = 0$ and joint Gaussianity gives independence. This is an exact *representation-space* null; it does not require input-space Gaussian corruption to remain a null after a nonlinear encoder. The complete proof is in Appendix B.4.

For independent heteroscedastic errors,

$$\mathrm{Cov}(T, Q) = \frac{1}{2}(\Sigma_1 - \Sigma_2), \tag{7}$$

so unequal covariance breaks the IID argument. The derivation of Eq. (7) is given in Appendix B.5. Randomizing the high/low view removes the signed slot cue but can retain higher-order dependence; the Gaussian derivation yielding $\mathrm{Cov}(T^2, Q_S^2) = \frac{1}{2}(v_1 - v_2)^2$ is given in Appendix B.6. Such dependence still does not guarantee that a particular symmetric descriptor improves task risk.

4 COUNTERFACTUAL CONSENSUS REPAIR

4.1 SYMMETRIC DISAGREEMENT DESCRIPTOR

CCR separates disagreement magnitude from its detailed structure. From normalized pooled-representation and feature-energy differences we construct a symmetric raw channel pattern r. Its scalar severity and severity-normalized pattern are

$$s = \frac{1}{C}\sum_c r_c, \qquad p_c = \frac{C r_c}{\sum_j r_j} \quad \text{for nonzero disagreement} \tag{8}$$

with $p = 0$ otherwise. The practical descriptor is $D = (p, u)$, where u contains s and seven non-severity shape/balance statistics. Exact definitions of r, the pooled and feature-energy components, u, and the correction bound $\mathcal{B}_\rho$ are in Appendix A.

The exact Gaussian null of Theorem 2 concerns descriptors measurable from (M, Q); the practical descriptor additionally uses nonlinear intermediate feature-map energies.

4.2 NESTED GLOBAL-TO-PAIR REPAIR

CCR first trains the consensus-only parent $G_\theta(M)$, freezes it, and then learns a zero-initialized pair residual $C_\phi(M, D)$ to predict the remaining repair target $T - G_\theta(M)$. Let $P_{\overline{W}}$ denote the orthogonal projector onto $\mathrm{row}(\overline{W})$.

$$G_\theta(M) = \mathcal{B}_{\rho_G}(P_{\overline{W}} g_\theta(M); M) \tag{9}$$

$$C_\phi(M, D) = \mathcal{B}_{\rho_P}(\gamma(s) P_{\overline{W}} q_\phi(M, D); M)\,, \qquad \gamma(s) = \frac{s}{s+\tau} \tag{10}$$

$$\widehat{z} = M - G_\theta(M) - C_\phi(M, D) \tag{11}$$

The global head internally uses $\mathrm{LN}(M)$, while the pair head receives $[\mathrm{LN}(M), \mathrm{LN}(p), u]$, where LN denotes feature-wise LayerNorm (Ba et al., 2016) applied independently to each example over the final feature dimension, without learned affine parameters and with numerical stabilizer $\epsilon_{\mathrm{LN}} = 10^{-5}$. Zero initialization makes every pair child start exactly from its trained global parent, while $\gamma(0) = 0$ forces pair abstention when the views agree. Architecture, optimization, and nesting details are in Appendix C; the corresponding invariance and initialization arguments are in Appendix B.7.

4.3 OBJECTIVE AND MATCHED CONTROL

The global head predicts T and the pair head predicts the residual $T - G_\theta(M)$ using centered-logit squared loss. Exact training objectives, clean-example weighting, and the latent ablation are specified in Appendix C.4.

We define the sample-specific disagreement detail as

$$D_i^{\mathrm{detail}} = \left(p_i, u_{i,\setminus s}\right)$$

where $u_{i,\setminus s}$ contains the seven summary statistics other than the scalar severity s_i.

The primary intervention keeps the target consensus and its exact scalar severity fixed while replacing only sample-specific disagreement detail:

$$(M_i, s_i, D_i^{\mathrm{detail}}) \qquad \text{vs.} \qquad (M_i, s_i, D_j^{\mathrm{detail}})$$

where j is a nearby-severity donor. Correct and matched models share the same global parent, pair architecture, objective, and optimization. The exact one-to-one donor construction is given in Appendix C.5.

Our primary contrast is

$$\Delta_{\mathrm{detail}} = R_{\mathrm{correct}} - R_{\text{severity-matched}} \tag{12}$$

Thus $\Delta_{\mathrm{detail}} < 0$ means that the target's own detailed disagreement improves task-visible reconstruction beyond consensus, scalar severity, and matched pair-model capacity.

5 EXPERIMENTAL DESIGN

Controlled evaluation. We evaluate CCR on Speech Commands V2 (GSCv2-35) (Warden, 2018), ESC-50 (Piczak, 2015), and UrbanSound8K (Salamon et al., 2014) using frozen compact CNN/linear-classifier models. We consider identity (I), asymmetric Gaussian (A), IID Gaussian (B), variance-matched heteroscedastic (H), structured cross-family (X), and common-mode (C) pair corruptions. Speech Commands uses 10 seeds; ESC-50 and UrbanSound8K use five and ten official folds, respectively, with five seeds per fold and fold-first inference. The primary endpoint is centered-logit MSE Δ_{detail}; accuracy and macro-F1 are secondary. Implementation and aggregation details are in Appendix C, with corruption definitions in Appendix D.

Mechanism checks. On ESC-50 we test the exact representation-space Gaussian null and heteroscedastic boundary with ridge probes based on M, (M, Q), and (M, D_z). We also repeat the matched-control experiment with frozen pretrained BEATs (Chen et al., 2023). Probe details are in Appendix F, and the BEATs replication is in Appendix G.

External transfer. DHAuDS-Pair uses externally designed DHAuDS corruptions (Shao et al., 2025) to test zero-shot transfer of the learned disagreement-to-repair mapping. The semantic model and repair heads remain frozen, with no retraining on DHAuDS. UrbanSound8K-C uses 10 official folds with five seeds per fold, while SpeechCommandsV2-C uses 10 seeds; Monte-Carlo pair/shuffle repetitions are collapsed before inference. Full protocols and results are in Appendix H. Appendix I separately reports the PANNs-based source, TENT, TTN, and DHAuDS-native results, which are not treated as a CCR head-to-head comparison because the backbone and adaptation problem differ.

6 RESULTS

6.1 DETAILED DISAGREEMENT HAS CONDITIONAL, NOT UNIVERSAL, VALUE

Figure 2 summarizes the primary contrast in Eq. (12). A, H, and X favor correct target-specific disagreement across all three datasets and across the corresponding inferential units. B exposes a different boundary: it is adverse on Speech Commands but favorable on ESC-50 and UrbanSound8K. This does not contradict Theorem 2, whose exact null concerns additive IID Gaussian noise directly in representation space rather than input corruption passed through a nonlinear encoder. Exact numerical aggregates and confidence intervals are reported in Table 2 of Appendix E.

The severity-only control tests whether the pair head merely exploits scalar disagreement magnitude. It retains the same trained global parent and corresponding pair architecture except for the disagreement intervention (Appendix C.4). Negative pair-task-minus-severity-only values favor the full descriptor. On Speech Commands, pair-task minus severity-only centered-logit MSE is -0.1219 (A), $+0.0249$ (B), -0.1330 (H), and -0.7306 (X), with A/H/X favorable in 10/10 seeds. On ESC-50, pair-task MSE is lower than severity-only by approximately 38.0%, 16.8%, 21.5%, and 44.2% for A/B/H/X, respectively. Speech B is therefore adverse under both the severity-matched control (Table 2) and the severity-only control, which agree on the location of that boundary. Full per-regime severity-only results are reported in Appendix E.

Within X, the effect also depends on corruption family. On Speech Commands, frequency-localized noise is nearly null ($\Delta_{\mathrm{detail}} = +0.0003$), whereas time and frequency masking yield -1.0671 and -1.1984. The corresponding ESC-50 values are -0.7865, -0.8002, and -1.1234, and the UrbanSound8K values are -0.1020, -0.6174, and -0.9675. Thus mask-based conditions are markedly more favorable than frequency-localized noise on Speech Commands and UrbanSound8K, whereas ESC-50 shows little separation between frequency-localized noise and time masking. The full decomposition is reported in Table 4 of Appendix E.

6.2 TASK-VISIBLE RECONSTRUCTION AND CLASSIFICATION ARE RELATED BUT NOT IDENTICAL

Pair-task minus matched accuracy on Speech Commands is approximately $+0.085$, $+0.007$, $+0.134$, and $+0.499$ percentage points for A/B/H/X. ESC-50 gives $+0.143$, $+0.273$, $+0.660$, and $+1.096$ points. For UrbanSound8K, A/B/H/X change accuracy by approximately -0.190, -0.017, $+0.123$,

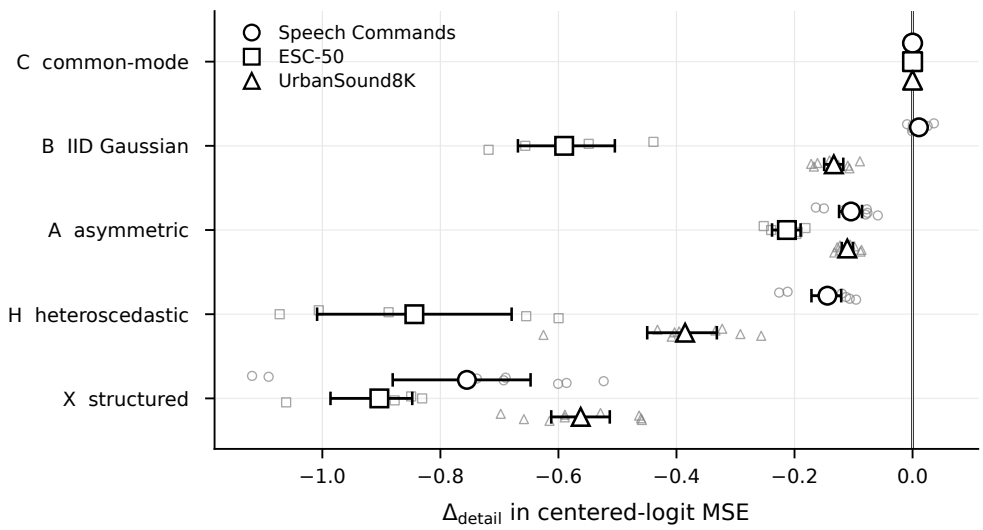


Figure 2: Primary controlled disagreement effect across inferential units. The horizontal axis is $\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$; negative values favor the target sample's own detailed disagreement. Small markers denote inferential units: 10 seeds for Speech Commands, five official folds for ESC-50, and 10 official folds for UrbanSound8K; fold-level values average the five seeds within each fold. Large markers show aggregate estimates and horizontal bars show 95% confidence intervals. Under common-mode corruption (C), the gated pair contribution is exactly zero.

and $+1.063$ percentage points, respectively. Thus lower centered-logit MSE need not translate into an equally large change in top-1 accuracy. Available accuracy and macro-F1 results are reported in Table 3 of Appendix E. The common-mode experiment behaves as intended. When the two views are identical, the gated pair correction is exactly zero, so the pair child reproduces the global parent exactly. Removing the gate increases common-mode MSE from 2.1559 to 2.6845 on Speech Commands, 6.0355 to 12.2637 on ESC-50, and 1.8457 to 4.3611 on UrbanSound8K. The global parent also incurs a measurable clean-accuracy cost: $89.906\% \rightarrow 89.773\%$, $56.87\% \rightarrow 55.90\%$, and $68.21\% \rightarrow 67.32\%$, respectively. The complete common-mode gate and clean-data audit is reported in Table 5 of Appendix E.

6.3 Representation-space probes recover the predicted null and boundary

Under IID representation-space Gaussian noise, $\mathrm{Cov}(T, Q) \approx -1.09 \times 10^{-4}$ and $\mathrm{Cov}(T^2, Q^2) \approx -1.10 \times 10^{-5}$ are near zero. Task-visible ridge risk is 0.01050 using M, 0.01182 using (M, Q), and 0.01188 using (M, D_z); adding disagreement improves none of the 25 fold–seed cases. The finite-sample increase therefore does not constitute negative population information value. Under fixed heteroscedastic noise, the theoretical covariance -0.032 closely matches the empirical -0.0319814. Adding signed Q reduces task-visible ridge risk from 0.01039 to 0.00436, a 58.03% reduction with 25/25 cases improving. With randomized slots, signed covariance cancels while the predicted higher-order dependence (0.002048) closely matches the empirical value (0.002053); linear ridge obtains no gain. These synthetic representation-space probes directly test the null and heteroscedastic boundary of Theorem 2 and Eq. (7). This randomized-slot linear-probe result does not imply that controlled H should be null; the two settings differ in representation, descriptor, and function class (Appendix F).

Figure 3 summarizes this progression from the exact IID null, through the signed fixed-heteroscedastic boundary, to the randomized-slot case, and shows the corresponding ESC-50 fold–seed probe effects. On real representations, adding pair information *worsens* ridge error by 11.97% for B/frequency-noise, but reduces it by 29.08% for H/frequency-noise, 43.56% for X/frequency-noise, 72.63% for X/time-mask, and 73.30% for X/frequency-mask. Table 6 in Appendix F reports the synthetic ridge comparison, followed by the real-representation ridge changes.

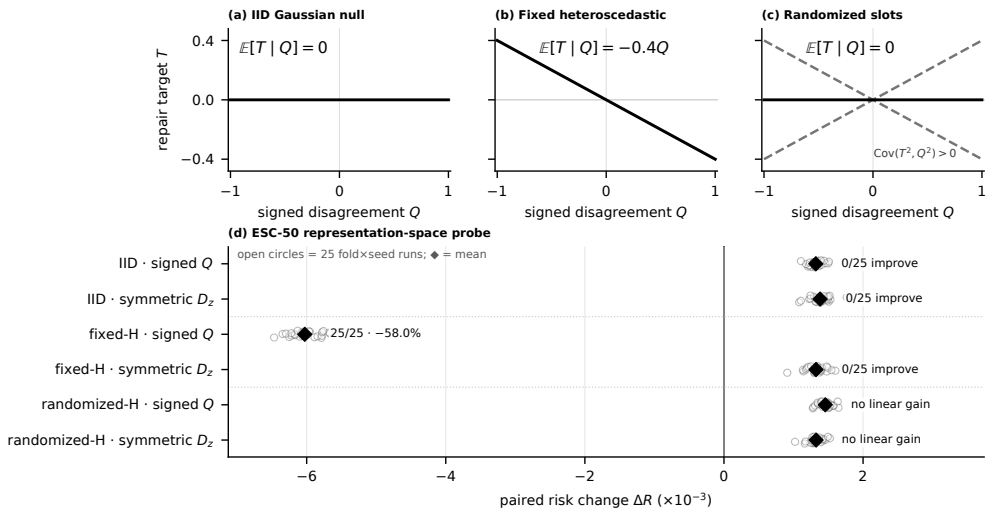


Figure 3: Representation-space null and heteroscedastic boundary. **(a)** Under IID Gaussian errors, $\mathbb{E}[T \mid Q] = 0$, so signed disagreement provides no repair direction. **(b)** A fixed low/high-variance assignment creates a signed conditional-mean cue, $\mathbb{E}[T \mid Q] = -0.4Q$. **(c)** Randomizing the high/low slot cancels this signed cue while retaining higher-order dependence. **(d)** Real ESC-50 representation-space probes across 5 folds $\times$ 5 seeds. Open circles denote fold–seed effects and diamonds denote means. Negative $\Delta R = R(\text{augmented}) - R(M)$ indicates improvement over the M-only probe. Signed Q improves $0/25$ IID cases and $25/25$ fixed-heteroscedastic cases, with a 58.03% mean reduction; after slot randomization, linear ridge obtains no gain.

6.4 Matched disagreement effect replicates with frozen BEATs on ESC-50

With frozen official BEATs on ESC-50, the correct-minus-matched contrast is negative for B (-0.01033), H (-0.01090), and X (-0.00680 averaged over structured X cells). Every evaluated B/H/X severity/family cell is negative and its reported confidence interval excludes zero. We do not compare these MSE magnitudes with those of the compact CNN because the representation and classifier geometries differ. Full BEATs results are in Appendix G.

6.5 Zero-shot transfer reveals a distribution-dependent boundary

DHAuDS-Pair tests whether the disagreement-to-repair mapping learned under the controlled corruption distribution transfers without repair retraining to externally designed corruption families. Table 1 summarizes the family-level zero-shot contrasts. UrbanSound8K-C is predominantly adverse: TST is near null at the family level, whereas WHN, ENSC, and PSH favor the matched control. Its overall contrast is $\Delta_{\text{detail}} = +0.4456$, with pair-task minus matched accuracy of -0.658 percentage points. SpeechCommandsV2-C is mixed: END1 is beneficial in 10/10 seeds, WHN is modestly beneficial, END2 and TST are near null, and ENQ/ENSC/PSH are adverse. Overall, $\Delta_{\text{detail}} = +0.1796$ with 95% CI $[+0.1466, +0.2083]$, and pair-task minus matched accuracy is -0.079 percentage points.

Despite the adverse overall correct-versus-matched contrast on SpeechCommandsV2-C, the correct pair model still improves over the global-only parent:

$$R_{\text{pair}} - R_{\text{global}} = -0.1526 \qquad (95\%\ \text{CI}\ [-0.2042, -0.1050]),$$

with all 10 seeds favorable. Thus pair-head information can remain useful relative to the global parent even when the target-specific detailed correspondence transfers less well than the matched control. WHN also changes sign across datasets, being adverse on UrbanSound8K-C but beneficial on SpeechCommandsV2-C. Because the repair heads are frozen rather than refit on DHAuDS-Pair, these results demonstrate that *transfer of the learned disagreement-to-repair mapping* is distribution dependent; they do not identify the intrinsic Bayes value of disagreement in either external distribution. Condition-level external results are reported in Tables 8 and 9 of Appendix H; Figure 4 in Appendix H.3 visualizes the cross-dataset family geometry and the distinction between the matched-control and global-parent contrasts.

Table 1: Zero-shot DHAuDS-Pair family-level Δ_{detail}. Negative values favor the target's own detailed disagreement; positive values favor the severity-matched control. **Bold** denotes family-level cases whose 95% confidence interval lies entirely below zero. No repair parameter is retrained on DHAuDS-Pair. UrbanSound8K-C uses fold-first inference; SpeechCommandsV2-C uses 10 seeds.

External family	UrbanSound8K-C	SpeechCommandsV2-C
WHN	$+0.2438\ [+0.184, +0.306]$	$\mathbf{-0.0671}\ [-0.118, -0.018]$
ENSC	$+0.4117\ [+0.312, +0.511]$	$+0.4292\ [+0.323, +0.573]$
PSH	$+1.2077\ [+1.009, +1.447]$	$+0.3883\ [+0.287, +0.497]$
TST	$-0.0809\ [-0.187, +0.018]$	$-0.0093\ [-0.045, +0.026]$
END1	–	$\mathbf{-0.2392}\ [-0.296, -0.182]$
END2	–	$-0.0025\ [-0.047, +0.039]$
ENQ	–	$+0.7580\ [+0.655, +0.856]$

7 Discussion and Limitations

Our results support a conditional rather than universal view of disagreement. A/H/X provide task-visible value beyond consensus and scalar severity across all three controlled datasets, but important boundaries remain. In particular, input-space IID Gaussian corruption (B) is adverse on Speech Commands but favorable on ESC-50 and UrbanSound8K. This does not contradict Theorem 2, whose exact null applies to additive IID Gaussian noise in representation space rather than to arbitrary input corruption passed through a nonlinear encoder.

The representation probes clarify this distinction. The IID construction recovers the predicted no-value case, while fixed heteroscedasticity creates a useful signed cue. Randomizing the high/low-noise slot removes that first-order cue while retaining higher-order dependence; for the tested linear ridge probe, this remaining dependence provides no gain. The negative correct-minus-matched contrast also persists with frozen pretrained BEATs on ESC-50, showing that the effect is observed with a substantially different frozen pretrained representation.

DHAuDS-Pair reveals a complementary transfer boundary. The learned disagreement-to-repair mapping transfers beneficially for some external families, is near null for others, and is adverse for others; WHN even changes sign across datasets. On SpeechCommandsV2-C, the correct pair model still outperforms the global-only parent in all ten seeds despite an adverse aggregate correct-versus-matched contrast. Thus usefulness relative to the global-only parent and transfer of target-specific disagreement correspondence are distinct questions.

CCR requires paired observations at deployment and clean anchors during repair training. Pair disagreement also cannot identify common-mode corruption when the views agree. The exact Gaussian null applies to representation-space descriptors measurable from (M, Q), whereas the practical CCR descriptor also uses nonlinear intermediate feature-map energies. The severity-matched contrast is an empirical intervention rather than a direct estimate of population Bayes improvement, and its outcome depends on model capacity, optimization, and descriptor choice. ESC-50 uncertainty is necessarily based on only five official outer folds, which limits the resolution of fold-level uncertainty estimates. Finally, the empirical validation is restricted to audio; whether analogous disagreement descriptors provide the same conditional value in other modalities remains open. Appendix J maps the principal alternative explanations to the corresponding experimental controls, while Appendix K summarizes the scope of the supported claims.

8 Conclusion

We study when sample-specific disagreement adds task-relevant information beyond consensus and scalar severity. CCR combines nested consensus-to-pair repair with a severity-matched control that isolates this incremental value. Controlled experiments, exact representation-space probes, frozen-BEATs replication on ESC-50, and zero-shot DHAuDS-Pair transfer collectively show beneficial, near-null, and adverse cases. The central conclusion is therefore conditional: disagreement has incremental task-visible value only when its detailed structure changes the task-visible estimate of the repair target beyond what consensus and severity already explain.

Reproducibility Statement

Section 5 summarizes the statistical units and the controlled, frozen-BEATs, and zero-shot external evaluation protocols. Appendix A specifies the disagreement descriptor and repair details; Appendix B contains the complete proofs; Appendix C specifies preprocessing, optimization, checkpoint nesting, dataset aggregation, and parent-provenance requirements; Appendix D gives the exact controlled corruption definitions; Appendix E reports the detailed controlled results; Appendix F reports the representation-space probe results; Appendix G reports the frozen-BEATs replication; Appendix H reports the zero-shot DHAuDS-Pair results; and Appendix I provides the separate PANNs/TTA deployment context.

AI Use Statement

Generative AI tools were used for language editing, including grammar and clarity improvements, and for assistance with LaTeX formatting and manuscript organization. Experimental measurements and numerical results reported in the paper were produced by executed project code rather than generated by an AI system. The authors reviewed the final manuscript, experimental results, mathematical content, and citations and retain responsibility for their correctness.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533. PMLR, 2019. URL https://proceedings.mlr.press/v97/batson19a.html.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

C. Mario Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 88–96. AUAI Press, 2008.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Sungnyun Kim, Sungwoo Cho, Sangmin Bae, Kangwook Jang, and Se-Young Yun. Multi-task corrupted prediction for learning robust audio-visual speech representation. In *The Thirteenth International Conference on Learning Representations*, pp. 91645–91666, 2025.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974. PMLR, 2018. URL https://proceedings.mlr.press/v80/lehtinen18a.html.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=EQfeudmWLQ.

Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7585–7593, 2022. doi: 10.1609/aaai.v36i7.20724.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jueqing Lu, Wray Buntine, Yuanyuan Qi, Joanna Dipnall, Belinda Gabbe, and Lan Du. Navigating conflicting views: Harnessing trust for learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40411–40435, 2025.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pp. 2613–2617, 2019. doi: 10.21437/Interspeech.2019-2680.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237, 2022. doi: 10.1109/CVPR52688.2022.00806.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018. ACM, 2015. doi: 10.1145/2733373.2806390.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044. ACM, 2014. doi: 10.1145/2647868.2655045.

Weichuang Shao, Iman Yi Liao, Tomas Henrique Bode Maul, and Tissa Chandesa. DHAuDS: A dynamic and heterogeneous audio benchmark for test-time adaptation. *arXiv preprint arXiv:2511.18421*, 2025. URL https://arxiv.org/abs/2511.18421. Version 3 revised June 2026.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12692–12702, 2020. doi: 10.1109/CVPR42600.2020.01271.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. URL https://arxiv.org/abs/1804.03209.

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16129–16137, 2024. doi: 10.1609/aaai.v38i14.29546.

Jinglin Xu, Yi Li, Chuxiong Sun, Xiao Xu, Jiangmeng Li, and Fanjiang Xu. Multi-modal test-time adaptation via adaptive probabilistic gaussian calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30268–30277, 2026.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations*, volume 2024, pp. 55036–55055, 2024.

Boyuan Zhang, Jie Pan, Shuai Yang, and Lichuan Gu. Adapt-then-fuse: Instance feature adaptation for multi-modal test-time adaptation. *Knowledge-Based Systems*, 348:116427, 2026. doi: 10.1016/j.knosys.2026.116427.

A Descriptor and Repair Details

Let $F_i \in \mathbb{R}^{C\times H\times L}$ denote the last convolutional feature map associated with z_i. Define

$$a_c = \frac{|z_{1c} - z_{2c}|}{|z_{1c}| + |z_{2c}| + \epsilon}$$

For channel RMS energy,

$$m_{ic} = \left(\frac{1}{HL}\sum_{u=1}^{H}\sum_{v=1}^{L} F_i(c,u,v)^2\right)^{1/2}, \qquad e_c = \frac{|m_{1c} - m_{2c}|}{m_{1c} + m_{2c} + \epsilon}$$

The raw disagreement and scalar severity are

$$r_c = \mathrm{clip}_{[0,1]}\left(\frac{a_c + e_c}{2}\right), \qquad s = \frac{1}{C}\sum_c r_c$$

The shape-normalized detailed pattern is

$$p_c = \begin{cases} Cr_c/\sum_j r_j, & \sum_j r_j > \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Likewise

$$a_c^{\mathrm{sh}} = \begin{cases} Ca_c/\sum_j a_j, & \sum_j a_j > \epsilon, \\ 0, & \text{otherwise,} \end{cases} \qquad e_c^{\mathrm{sh}} = \begin{cases} Ce_c/\sum_j e_j, & \sum_j e_j > \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

The pooled-versus-energy balance statistic is

$$\beta = \frac{\mathrm{mean}_c(a_c)}{\mathrm{mean}_c(a_c) + \mathrm{mean}_c(e_c) + \epsilon}$$

The eight summary statistics are

$$u = [s, \mathrm{sd}(p), \max p, \mathrm{sd}(a^{\mathrm{sh}}), \max a^{\mathrm{sh}}, \mathrm{sd}(e^{\mathrm{sh}}), \max e^{\mathrm{sh}}, \beta]$$

Here $\mathrm{sd}(\cdot)$ and $\max(\cdot)$ are taken over the channel dimension, with sd denoting the population standard deviation (denominator C).

The practical descriptor is $D = (p, u)$. Every operation in the descriptor construction is invariant to swapping the two views. Throughout this construction, we use the numerical stabilizer $\epsilon = 10^{-6}$.

The relative-norm operator used for global and pair corrections is

$$\mathcal{B}_\rho(d; M) = d \min\left\{1, \frac{\rho \max(\|M\|_2, \epsilon)}{\max(\|d\|_2, \epsilon)}\right\}$$

The same $\epsilon = 10^{-6}$ is used in the relative-norm bound. Main values are $\rho_G = 0.80$, $\rho_P = 0.50$, and pair-gate $\tau = 0.05$.

B Complete Proofs

B.1 Centered-logit invariance and row-space sufficiency

For $\bar{\ell} = K^{-1}\mathbf{1}^\top \ell$, $H_K \ell = \ell - \bar{\ell}\mathbf{1}$. Subtracting the same scalar from every logit preserves softmax and all pairwise logit differences, hence preserves argmax. If $P = P_{\overline{W}}$, then $(I - P)r \in \mathrm{row}(\overline{W})^\perp = \ker(\overline{W})$, so $\overline{W}r = \overline{W}Pr$.

B.2 Conditional mean under a linear seminorm

For sigma-field $\mathcal{F}$, set $\mu = \mathbb{E}[T \mid \mathcal{F}]$. For any $\mathcal{F}$-measurable r,

$$T - r = (T - \mu) + (\mu - r)$$

Applying a fixed matrix A and expanding gives

$$\mathbb{E}\,\|A(T - r)\|^2 = \mathbb{E}\,\|A(T - \mu)\|^2 + \mathbb{E}\,\|A(\mu - r)\|^2$$

because the cross term has conditional expectation zero. Hence μ is a Bayes minimizer; if A is rank deficient, minimizers are nonunique modulo $\ker(A)$.

B.3 Proof of Theorem 1

Write $T - \mu_G = (T - \mu_D) + (\mu_D - \mu_G)$. Since $\mathcal{G} \subseteq \mathcal{D}$, both conditional means are $\mathcal{D}$-measurable and

$$\mathbb{E}[(T - \mu_D)^\top A^\top A(\mu_D - \mu_G)] = 0$$

Expanding the squared norm yields Eq. (4); setting $A = \overline{W}$ and dividing by K yields Eq. (5).

B.4 Proof of the IID Gaussian null

With $T = (\varepsilon_1 + \varepsilon_2)/2$ and $Q = \varepsilon_1 - \varepsilon_2$, (T, Q) is jointly Gaussian and

$$\mathrm{Cov}(T, Q) = \tfrac{1}{2}(\Sigma - \Sigma) = 0$$

Thus $T \perp Q$. Since z_0 is independent of (T, Q), $Q \perp (z_0, T)$ and therefore $Q \perp (M, T)$ for $M = z_0 + T$. Conditioning on M, any measurable $D_z = d(M, Q)$ cannot change the conditional law of T, proving $\mathbb{E}[T \mid M, D_z] = \mathbb{E}[T \mid M]$.

B.5 Fixed heteroscedastic boundary

For independent zero-mean errors with covariances Σ_1, Σ_2,

$$\mathrm{Cov}(T, Q) = \frac{1}{2}\mathbb{E}[(\varepsilon_1 + \varepsilon_2)(\varepsilon_1 - \varepsilon_2)^\top] = \frac{1}{2}(\Sigma_1 - \Sigma_2)$$

B.6 Randomized-slot higher-order dependence

Let $e_1 \sim \mathcal{N}(0, v_1)$ and $e_2 \sim \mathcal{N}(0, v_2)$ be independent, and define

$$T = \frac{e_1 + e_2}{2}, \qquad Q_S = S(e_1 - e_2)$$

where S is an independent Rademacher sign. Then $\mathbb{E}[TQ_S] = 0$. Since $Q_S^2 = (e_1 - e_2)^2$ and Gaussian fourth moments satisfy $\mathbb{E}[e_i^4] = 3v_i^2$,

$$\mathbb{E}[T^2 Q_S^2] = \frac{1}{4}(3v_1^2 + 3v_2^2 - 2v_1 v_2),$$

while

$$\mathbb{E}[T^2]\mathbb{E}[Q_S^2] = \frac{1}{4}(v_1 + v_2)^2$$

Hence

$$\mathrm{Cov}(T^2, Q_S^2) = \frac{1}{2}(v_1 - v_2)^2$$

For $v_1 = \eta^2/5$ and $v_2 = 9\eta^2/5$, this becomes $32\eta^4/25$.

B.7 Permutation invariance, abstention, and zero initialization

All descriptor primitives use absolute differences and symmetric sums, so swapping the two views leaves D unchanged. If both views are identical at pooled and feature-map levels, $a = e = r = p = s = 0$ and $\gamma(0) = 0$, making the pair residual exactly zero. If the final pair affine layer has zero weights and bias, its raw output is zero for every input; projection, gating, and norm bounding preserve zero, so the pair child initially reproduces the trained global parent exactly.

B.8 A sufficient classification-margin condition

Let y be the clean predicted class and

$$m = \min_{k \neq y}[(Wz_0 + b)_y - (Wz_0 + b)_k] > 0$$

If $\left\|\overline{W}(\hat{z} - z_0)\right\|_\infty < m/2$, then for every $k \neq y$ the change in the y–k logit difference has magnitude less than m, preserving the positive clean margin and hence the argmax.

C Exact Controlled Implementation

C.1 Preprocessing and semantic CNN

Controlled experiments use 16 kHz waveforms and a 64-bin mel spectrogram with

$$n_{\text{fft}} = 1024, \quad n_{\text{hop}} = 320, \quad n_{\text{win}} = 1024, \quad f_{\min} = 50, \quad f_{\max} = 8000$$

power 2, followed by $\log(1 + \cdot)$ and per-example standardization. ESC-50 and UrbanSound8K use 5 s and 4 s waveforms, respectively; after resampling, longer clips are truncated from the end and shorter clips are right-zero-padded.

The compact semantic CNN has three 3×3 Conv2d–ReLU blocks with width 128, stride 1, and padding 1; the first two are followed by 2×2 max pooling with stride 2. Global spatial mean pooling is followed by a linear classifier. All controlled baseline and repair stages use minibatch size 64.

The semantic baseline is trained for up to 80 epochs with AdamW (Loshchilov & Hutter, 2019), learning rate 3×10^{-4}, weight decay 10^{-4}, and Gaussian augmentation standard deviation 0.10. For Speech Commands V2, checkpoint selection uses the official validation split and the predefined robust-mean score. Let $\mathcal{S} = \{0.10, 0.20, 0.30\}$. Then

$$S_{\text{robust}} = \frac{1}{4}\Bigg[A_I + \frac{1}{3}\sum_{\sigma\in\mathcal{S}} A_B(\sigma) + \frac{1}{3}\sum_{\sigma\in\mathcal{S}} A_H(\sigma) + \frac{1}{9}\sum_{f\in\mathcal{F}}\sum_{\sigma\in\mathcal{S}} A_X(f,\sigma)\Bigg]$$

where $\mathcal{F} = \{\text{frequency noise}, \text{time mask}, \text{frequency mask}\}$ and each A is validation accuracy computed from the mean logits of the two views. Thus I/B/H/X receive equal weight despite X containing three structured families. Checkpoint selection uses patience 15 and minimum improvement 10^{-5}, with validation loss breaking ties within that tolerance.

For ESC-50 and UrbanSound8K, no independent validation split is used under the fold-first protocol; training follows the fixed 80-epoch schedule and retains the final state. The selected encoder and classifier are then frozen.

C.2 Repair optimization and provenance

Global and pair MLPs use one hidden layer of width 64 with GELU (Hendrycks & Gimpel, 2016). Their final linear layers are zero initialized. Both stages use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, gradient clipping at 5, and minimum checkpoint improvement 10^{-5}. For Speech Commands V2, each stage is trained for up to 40 epochs and the checkpoint minimizing mean squared centered-logit validation risk is retained, with early-stopping patience 8. For ESC-50 and UrbanSound8K, no held-out outer fold is used for checkpoint selection; each stage follows the fixed 40-epoch schedule and the final state is retained. The main objective has no auxiliary cross-entropy term. Global clean examples have loss weight 2; pair clean examples have weight 1. The bound and gate parameters $\rho_G = 0.80$, $\rho_P = 0.50$, and $\tau = 0.05$, together with the width and loss-weight settings above, are held fixed across the reported controlled datasets and seeds. We do not ascribe these values to a formal hyperparameter-search procedure.

The global parent is trained once per semantic checkpoint. Every pair ablation must load that exact parent. Final provenance checks require stored global fingerprints to match the corresponding global checkpoint exactly across all evaluated conditions. The final Speech Commands archive satisfies this for all 10 seeds, including the corrected seed-6 run.

C.3 Dataset and aggregation protocols

Speech Commands V2 uses the official 35-class split (84,843 train / 9,981 validation / 11,005 test) and 10 scientific seeds. ESC-50 uses all five official folds with five seeds per fold. UrbanSound8K uses all ten official folds with five seeds per fold and no reshuffling across source folds.

For Speech Commands, corruption conditions are first averaged within each seed/regime and the seed is the inferential unit. For ESC-50 and UrbanSound8K, seeds and corruption conditions are first averaged within each official outer fold; the fold is then the inferential unit. Reported 95% confidence intervals for the primary controlled contrasts are percentile-bootstrap intervals for the mean using 20,000 resamples of the corresponding inferential units.

C.4 Repair objectives and ablations

For each corrupted pair,

$$M = \frac{z_1 + z_2}{2}, \qquad T = M - z_0$$

The global stage uses the per-example centered-logit loss

$$\ell_G = \frac{1}{K} \left\| \overline{W}(G_\theta(M) - T) \right\|_2^2$$

After the global parent is trained, it is frozen. The pair stage predicts the remaining residual using

$$\ell_P = \frac{1}{K} \left\| \overline{W}(C_\phi(M, D) - [T - G_\theta(M)]) \right\|_2^2$$

The main configuration uses no auxiliary cross-entropy term. Global clean examples have loss weight 2 and pair clean examples have loss weight 1, as specified above. The `pair_latent` ablation instead trains the pair residual with representation-space MSE and omits row-space projection of that pair residual; the trained global parent remains unchanged.

The severity-only, unrestricted-shuffle, no-gate, and latent variants all retain the same trained global parent and corresponding pair architecture except for the intervention being tested.

C.5 Severity-matched donor construction

Write

$$D_i^{\text{detail}} = \left(p_i, u_{i,\setminus s}\right),$$

where p_i is the severity-normalized disagreement pattern and $u_{i,\setminus s}$ denotes the seven summary statistics other than scalar severity. The correct condition receives

$$(M_i, s_i, D_i^{\text{detail}}),$$

whereas the severity-matched condition receives

$$(M_i, s_i, D_j^{\text{detail}})$$

Thus the target always retains its own consensus M_i and exact scalar severity s_i; only the normalized pattern and non-severity statistics are provided by the donor.

Within each minibatch, examples are sorted by scalar severity. Adjacent examples exchange donors. If the batch size is odd, the final three examples form a 3-cycle. This produces a one-to-one derangement: no target receives its own detailed descriptor and every donor is used exactly once. The legacy `severity_bins` argument is not used by the final V5.2 matching procedure.

Correct and severity-matched models use the same frozen global parent, pair architecture, task objective, optimization settings, and training schedule. The primary empirical contrast is

$$\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$$

D Controlled Corruption Definitions

Let the per-example standardized log-mel spectrogram be $X \in \mathbb{R}^{F \times L}$, and let Ξ contain IID standard-normal entries. All paired views are generated from the same clean semantic instance.

Primitive corruptions. Gaussian corruption is

$$\mathcal{C}_{\mathrm{G}}(X;\eta) = X + \eta\Xi$$

For frequency-localized noise,

$$\alpha_F(\eta) = \min(0.15 + 0.75\eta, 0.65), \qquad w_F = \max(1, \min(F, \mathrm{round}(\alpha_F(\eta)F)))$$

With an independently sampled contiguous frequency-block mask B_F,

$$\mathcal{C}_{\mathrm{FN}}(X;\eta) = X + 1.5\eta(\Xi \odot B_F)$$

For time masking,

$$\alpha_M(\eta) = \min(0.05 + 0.90\eta, 0.55), \qquad w_T = \max(1, \min(L, \mathrm{round}(\alpha_M(\eta)L)))$$

and a uniformly located contiguous time block of width w_T is zeroed. For frequency masking,

$$w_M = \max(1, \min(F, \mathrm{round}(\alpha_M(\eta)F)))$$

and a uniformly located contiguous frequency block of width w_M is zeroed. Time and frequency masking follow the spectrogram-augmentation paradigm of Park et al. (2019); the severity schedules used here are specific to this study.

Controlled pair regimes. We use six paired-corruption regimes.

- **I (identity):** both views are clean.
- **A (asymmetric Gaussian):** one view is clean and the other receives Gaussian corruption $\mathcal{C}_{\mathrm{G}}(X;\eta)$; the corrupted slot is randomized.
- **B (IID Gaussian):** both views receive independent Gaussian corruptions with the same standard deviation η.
- **H (variance-matched heteroscedastic):** the two views receive independent Gaussian corruptions with

  $$\sigma_L = \eta/\sqrt{5}, \qquad \sigma_H = 3\eta/\sqrt{5},$$

  so

  $$\frac{\sigma_L^2 + \sigma_H^2}{2} = \eta^2$$

  matching the mean per-view variance of regime B. The low/high-noise slot assignment is randomized.
- **X (structured cross-family):** one view receives Gaussian corruption and the other receives one structured corruption: frequency-localized noise, time masking, or frequency masking. The Gaussian/structured slot assignment is randomized.
- **C (common mode):** the same Gaussian corruption realization is used for both views.

Training samples regimes I/A/B/H/X; for the corrupted regimes, $\eta \sim U(0.05, 0.30)$. Evaluation uses $\eta \in \{0.10, 0.20, 0.30, 0.40\}$. Regime C is retained as a common-mode diagnostic: because both views are identical, their pair disagreement is zero even though their shared consensus may be corrupted.

E Full Controlled Results

Table 2 presents the primary controlled Δ_{detail} estimates and confidence intervals, while Table 3 reports the corresponding secondary classification results.

E.1 X-family decomposition

Table 4 decomposes the structured X regime into frequency-localized noise, time masking, and frequency masking. For ESC-50, corresponding accuracy changes are $+0.488$, $+0.733$, and $+2.068$ pp for frequency-noise, time-mask, and frequency-mask.

Table 2: Primary controlled contrast Δ_{detail} (centered-logit MSE). Negative values favor the target's own detailed disagreement over the severity-matched control; **bold** denotes cases whose 95% confidence interval lies entirely below zero. Speech Commands intervals are over seeds; ESC-50 and UrbanSound8K use fold-first inference. Under common-mode corruption (C), the gated pair residual is exactly zero.

Dataset	A	B	H	X	C
Speech Commands	$\mathbf{-0.1043}$	$+0.0108$	$\mathbf{-0.1442}$	$\mathbf{-0.7551}$	0
95% CI	$[-0.1245, -0.0856]$	$[+0.0029, +0.0193]$	$[-0.1713, -0.1209]$	$[-0.8808, -0.6472]$	exact
ESC-50	$\mathbf{-0.2129}$	$\mathbf{-0.5903}$	$\mathbf{-0.8441}$	$\mathbf{-0.9034}$	0
95% CI	$[-0.2381, -0.1897]$	$[-0.6686, -0.5044]$	$[-1.0089, -0.6793]$	$[-0.9862, -0.8477]$	exact
UrbanSound8K	$\mathbf{-0.1106}$	$\mathbf{-0.1334}$	$\mathbf{-0.3854}$	$\mathbf{-0.5623}$	0
95% CI	$[-0.1199, -0.1009]$	$[-0.1498, -0.1174]$	$[-0.4496, -0.3316]$	$[-0.6122, -0.5130]$	exact

Table 3: Controlled effects of target-specific disagreement detail. For each dataset, $\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$ is the primary centered-logit MSE contrast; negative values favor the target sample's own disagreement detail. Accuracy and macro-F1 rows report correct-pair minus severity-matched differences in percentage points (pp), for which positive values favor the target-specific detail condition.

Dataset / metric	A	B	H	X
Speech Δ_{detail}	-0.1043	+0.0108	-0.1442	-0.7551
Speech acc. diff. (pp)	+0.0847	+0.0070	+0.1339	+0.4994
Speech macro-F1 diff. (pp)	+0.0872	+0.0174	+0.1413	+0.5324
ESC Δ_{detail}	-0.2129	-0.5903	-0.8441	-0.9034
ESC acc. diff. (pp)	+0.143	+0.273	+0.660	+1.096
ESC macro-F1 diff. (pp)	+0.129	+0.276	+0.729	+1.150
UrbanSound Δ_{detail}	-0.1106	-0.1334	-0.3854	-0.5623
UrbanSound acc. diff. (pp)	-0.190	-0.017	+0.123	+1.063

E.2 Severity-only comparison

On Speech Commands, pair-task minus severity-only MSE is -0.1219 (A), $+0.0249$ (B), -0.1330 (H), and -0.7306 (X); A/H/X are favorable in 10/10 seeds. On ESC-50, pair-task MSE is lower than severity-only by about 38.0% (A), 16.8% (B), 21.5% (H), and 44.2% (X).

E.3 Common-mode and clean trade-off

Table 5 reports the common-mode gate audit and clean-data cost for all three controlled datasets.

The consensus-only global parent also improves common-mode corruption relative to no repair: Speech C MSE drops from about 4.3230 to 2.1559; ESC from 30.104 to 6.036. On UrbanSound8K, common-mode accuracy rises from 51.43% without repair to about 67.49% after the global parent.

F Representation-Space Probe Results

The Gaussian probe uses 5 ESC folds times 5 seeds. Each probe fit uses 1200 training and 400 validation examples. Table 6 summarizes the task-visible ridge risks for the IID and heteroscedastic probe settings.

The randomized-slot ridge result does not imply that controlled H should be null: the probe tests linear recoverability from representation-space Q or D_z, whereas controlled H is applied in input space and CCR uses a nonlinear encoder, intermediate feature-map disagreement, and a nonlinear pair head.

Table 4: Structured cross-family (X) decomposition. Values report the primary centered-logit MSE contrast $\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$ separately for the three structured corruption families. Negative values favor the target sample's own disagreement detail; positive values favor the severity-matched control.

Dataset	freq-noise	time-mask	freq-mask
Speech	+0.0003	-1.0671	-1.1984
ESC-50	-0.7865	-0.8002	-1.1234
UrbanSound8K	-0.1020	-0.6174	-0.9675

Table 5: Common-mode abstention and clean-data trade-off. Under common-mode corruption (C), the two views have zero disagreement, so the gate forces the pair residual to zero and the gated pair model exactly reproduces the global parent. Removing the gate increases centered-logit MSE. The final column reports the change in clean classification accuracy after global repair, in percentage points (pp); negative values denote a clean-data cost.

Dataset	gated/global C MSE	no-gate C MSE	clean acc. cost
Speech	2.1559	2.6845	-0.133 pp
ESC-50	6.0355	12.2637	-0.97 pp
UrbanSound8K	1.8457	4.3611	-0.89 pp

For the variance-matched setting with $\sigma = 0.2$, $a = 1/\sqrt{5}$, $b = 3/\sqrt{5}$,

$$\text{Cov}(T, Q)_{\text{theory}} = -0.032, \qquad \text{Cov}(T, Q)_{\text{empirical}} = -0.0319814.$$

The randomized-slot higher-order prediction is 0.002048, with empirical value 0.00205343. IID empirical values are approximately $\text{Cov}(T, Q) = -1.09 \times 10^{-4}$ and $\text{Cov}(T^2, Q^2) = -1.10 \times 10^{-5}$. In the real-representation ridge probe, labels of the form "regime/family" correspond to the stored $(\texttt{regime}, \texttt{structured_family})$ keys. For B and H, $\texttt{structured_family} = \texttt{freq_noise}$ is an inert bookkeeping value and does not alter the Gaussian corruption specified by the regime. For X, $\texttt{structured_family}$ identifies the structured corruption family actually used.

Real-representation ridge changes from adding pair information are: B/frequency-noise -11.97%, H/frequency-noise $+29.08\%$, X/frequency-noise $+43.56\%$, X/time-mask $+72.63\%$, and X/frequency-mask $+73.30\%$. Positive percentages denote relative reductions in ridge error after adding pair information; negative percentages denote worsening.

G Frozen BEATs Replication

The BEATs experiment uses the official pretrained Iteration-3+ AS2M checkpoint, five official ESC-50 folds times five seeds per fold, and the pooled descriptor as the primary configuration. Confidence intervals use the executed 20,000-draw hierarchical fold–seed bootstrap, resampling outer folds first and seeds within sampled folds (bootstrap seed 2027). The primary correct-minus-matched contrast is negative in every evaluated B/H/X severity/family cell, and all 20 corresponding 95% confidence intervals exclude zero. The full severity/family-level contrasts and confidence intervals are reported in Table 7.

Broad-regime means are approximately

$$\Delta_B = -0.01033, \qquad \Delta_H = -0.01090, \qquad \Delta_X = -0.00680.$$

Absolute MSE values are not compared with those of the compact CNN because the representation and classifier geometries differ.

H Zero-Shot DHAuDS-Pair Results

For both external datasets, each (fold/seed, condition, pair-draw) cell contains one evaluation for each of five non-shuffled run types and five repetitions of the severity-matched shuffle control, yielding 10 archived rows per cell; shuffle repetitions and pair draws are collapsed before inference.

Table 6: Task-visible ridge probe. IID: 0/25 cases improve after adding disagreement. Fixed heteroscedasticity: M, Q improves 25/25 and reduces error by 58.03%.

Case	M ridge	M, Q ridge	M, D_z ridge
IID	0.01050	0.01182	0.01188
fixed heteroscedastic	0.01039	0.00436	0.01171
randomized heteroscedastic	baseline	no linear gain	no linear gain

Table 7: Frozen-BEATs ESC-50 correct-minus-matched contrasts. Intervals are 95% hierarchical fold–seed bootstrap confidence intervals using five outer folds and five seeds per fold. Negative values favor target-specific disagreement.

Regime	Family	η	Δ_{detail}	95% CI
B	Gaussian	0.1	-0.00342	$[-0.00523, -0.00177]$
B	Gaussian	0.2	-0.00891	$[-0.01290, -0.00470]$
B	Gaussian	0.3	-0.01458	$[-0.02448, -0.00447]$
B	Gaussian	0.4	-0.01443	$[-0.02663, -0.00396]$
H	Gaussian	0.1	-0.00284	$[-0.00396, -0.00180]$
H	Gaussian	0.2	-0.00733	$[-0.01023, -0.00468]$
H	Gaussian	0.3	-0.01497	$[-0.01989, -0.01045]$
H	Gaussian	0.4	-0.01847	$[-0.02495, -0.01240]$
X	Freq. mask	0.1	-0.00221	$[-0.00371, -0.00104]$
X	Freq. noise	0.1	-0.00230	$[-0.00402, -0.00085]$
X	Time mask	0.1	-0.00188	$[-0.00304, -0.00088]$
X	Freq. mask	0.2	-0.00598	$[-0.00983, -0.00250]$
X	Freq. noise	0.2	-0.00554	$[-0.00817, -0.00295]$
X	Time mask	0.2	-0.00494	$[-0.00675, -0.00331]$
X	Freq. mask	0.3	-0.00937	$[-0.01390, -0.00511]$
X	Freq. noise	0.3	-0.00921	$[-0.01306, -0.00525]$
X	Time mask	0.3	-0.00877	$[-0.01213, -0.00572]$
X	Freq. mask	0.4	-0.00954	$[-0.01524, -0.00429]$
X	Freq. noise	0.4	-0.01044	$[-0.01474, -0.00635]$
X	Time mask	0.4	-0.01141	$[-0.01524, -0.00773]$

H.1 URBANSOUND8K-C

The final archive contains 10 official folds $\times$ five seeds = 50 runs, 12,000 evaluation rows, eight conditions, three pair draws, and five matched-shuffle repeats, with no missing conditions, duplicate design rows, NaNs, or infinities. No repair parameter is retrained on DHAuDS-Pair. Aggregation is shuffle $\rightarrow$ draw $\rightarrow$ seed $\rightarrow$ fold. Condition-level UrbanSound8K-C contrasts are reported in Table 8.

Family means are TST -0.0809, WHN $+0.2438$, ENSC $+0.4117$, and PSH $+1.2077$. Overall $\Delta_{\text{detail}} = +0.4456$ and pair-task minus matched accuracy is -0.658 pp.

H.2 SPEECHCOMMANDSV2-C

The final archive contains 10/10 seeds, 4,200 rows, 14 conditions, 11,005 test examples per evaluation, three pair draws, five matched-shuffle repeats, and pinned DHAuDS revision `e5a64084a01dd22c1f0caf4616a64867cebb57ef`. The TST policy is `paper_v3_no_slowdown`; all runs have `repair_retraining_on_dhauds=false`. There are no missing conditions, duplicate design rows, NaNs, or infinities.

Family means/CIs are END1 -0.2392 $[-0.2964, -0.1821]$, WHN -0.0671 $[-0.1184, -0.0180]$, TST -0.0093 $[-0.0451, +0.0256]$, END2 -0.0025 $[-0.0465, +0.0390]$, PSH $+0.3883$ $[+0.2869, +0.4966]$, ENSC $+0.4292$ $[+0.3232, +0.5725]$, and ENQ $+0.7580$ $[+0.6547, +0.8564]$. Overall $\Delta_{\text{detail}} = +0.1796$ with 95% CI $[+0.1466, +0.2083]$. Pair-task minus matched accuracy

Table 8: Zero-shot DHAuDS-Pair transfer on UrbanSound8K-C. Condition-level values report $\Delta_{\text{detail}} = R_{\text{correct}} - R_{\text{severity-matched}}$ for frozen CCR repair heads evaluated without retraining on DHAuDS-Pair. Negative values favor the target sample's own disagreement detail, whereas positive values favor the severity-matched control. Reported 95% confidence intervals use fold-first inference after collapsing matched-shuffle repetitions and pair draws.

Condition	Δ_{detail}	95% CI
TST-L1	-0.0957	[-0.196,-0.002]
TST-L2	-0.0661	[-0.182,+0.043]
WHN-L1	+0.2334	[+0.177,+0.292]
WHN-L2	+0.2541	[+0.192,+0.318]
ENSC-L1	+0.3590	[+0.243,+0.477]
ENSC-L2	+0.4644	[+0.374,+0.550]
PSH-L1	+1.1077	[+0.910,+1.345]
PSH-L2	+1.3077	[+1.110,+1.544]

Table 9: SpeechCommandsV2-C DHAuDS-Pair zero-shot condition-level correct-minus-matched contrasts. Confidence intervals are 95% percentile bootstrap intervals over the 10 seed-level inferential units, using 20,000 resamples with bootstrap seed 2027. Negative values favor target-specific disagreement.

Condition	Δ_{detail}	95% CI
END1-L1	-0.2913	$[-0.3638, -0.2207]$
END1-L2	-0.1871	$[-0.2301, -0.1429]$
WHN-L1	-0.0595	$[-0.1068, -0.0136]$
WHN-L2	-0.0746	$[-0.1304, -0.0220]$
TST-L1	-0.0332	$[-0.0698, +0.0006]$
END2-L1	-0.0166	$[-0.0590, +0.0229]$
END2-L2	$+0.0116$	$[-0.0355, +0.0570]$
TST-L2	$+0.0147$	$[-0.0210, +0.0516]$
ENSC-L1	$+0.2302$	$[+0.1354, +0.3389]$
PSH-L1	$+0.3165$	$[+0.2293, +0.4105]$
PSH-L2	$+0.4601$	$[+0.3438, +0.5826]$
ENSC-L2	$+0.6281$	$[+0.4941, +0.8110]$
ENQ-L2	$+0.6473$	$[+0.5540, +0.7395]$
ENQ-L1	$+0.8687$	$[+0.7508, +0.9830]$

is -0.079 pp. Pair-task nevertheless beats the global parent overall by -0.1526 MSE (95% CI $[-0.2042, -0.1050]$, 10/10 seeds favorable). Condition-level SpeechCommandsV2-C contrasts are reported in Table 9.

H.3 Cross-dataset transfer geometry

Figure 4 summarizes the family-level transfer geometry across the two external datasets and separates the correct-versus-matched contrast from the pair-versus-global comparison.

I Separate PANNs / Official DHAuDS TTA Context

This block uses a PANNs-based backbone (Kong et al., 2020) and reports TENT (Wang et al., 2021), TTN (Lim et al., 2023), and DHAuDS-native adaptation (Shao et al., 2025). It uses a different backbone and adaptation problem and must not be treated as a controlled CCR comparison. Table 10 reports the separate deployment context.

All 70/70 DHAuDS-native run–condition cases improve relative to source. Selected gains are WHN-L1/L2 $+11.62/+11.65$ pp, ENQ-L1/L2 about $+2.93/+2.95$, END1 $+1.55/+1.75$, END2

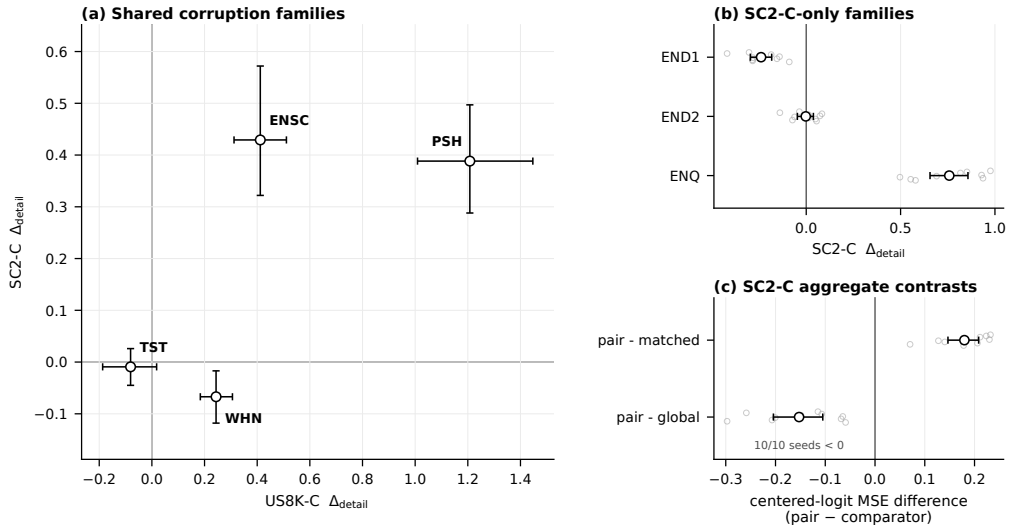


Figure 4: **Zero-shot DHAuDS-Pair transfer geometry. (a)** Shared corruption families plotted by their Δ_{detail} on UrbanSound8K-C and SpeechCommandsV2-C. **(b)** SpeechCommandsV2-C-only families. **(c)** Aggregate SpeechCommandsV2-C pair-versus-matched and pair-versus-global contrasts. Small markers denote inferential units; large markers and error bars denote aggregate estimates and 95% confidence intervals.

Table 10: PANNs-based DHAuDS test-time adaptation context. These results use a separate PANNs backbone and test-time adaptation protocol and are reported only as deployment context, not as a controlled head-to-head comparison with CCR.

Method	source acc. (%)	adapted acc. (%)	change (pp)
TENT	73.228	3.963	-69.265
TTN	73.228	22.965	-50.263
DHAuDS native	73.228	80.184	+6.956

$+1.30/+1.08$, ENSC $+10.86/+7.78$, PSH $+2.68/+7.02$, and TST $+16.32/+17.92$. These provide robustness/TTA context, not a fair head-to-head ranking against CCR.

J Primary Estimand and Control Logic

Alternative explanation	Design element
Pair model wins only because it has extra parameters	Correct and matched-shuffled disagreement use the same pair architecture.
The pair model only estimates scalar corruption severity	Target severity s is retained in the matched control; severity-only is reported separately.
Any disagreement vector is useful	Exact IID representation-space Gaussian probe gives a true no-value null.
H gain is an ordered-slot shortcut	High/low slots are randomized; signed first-order covariance cancels.
The result is a weak-CNN artifact	Frozen BEATs changes the representation while preserving the controlled design.
The result is tailored to author-designed corruptions	DHAuDS-Pair changes the corruption source with no repair retraining.
Common-mode corruption is silently ignored	C is retained explicitly; gated pair repair abstains at zero disagreement.

K SCOPE OF CLAIMS

Our claim is conditional rather than universal. The exact Gaussian null applies to additive representation-space noise and descriptors measurable from (M, Q); it does not extend automatically to input-space corruption passed through a nonlinear encoder. Likewise, heteroscedastic dependence does not guarantee that a particular disagreement descriptor will improve task risk. CCR requires paired observations at deployment and clean anchors during repair training, and cannot diagnose common-mode corruption from disagreement when the views agree.

The frozen-BEATs experiment establishes replication of the matched effect on ESC-50, not comparability of MSE magnitudes across representations. DHAuDS-Pair tests zero-shot transfer of frozen repair heads, not the intrinsic Bayes value of disagreement or the official DHAuDS test-time-adaptation task; the separate PANNs/TTA results provide deployment context only. Overall, the supported claim is that sample-specific disagreement has incremental task-visible value when its detailed structure changes the repair-target estimate beyond consensus and scalar severity.