FORCED FUSION: MULTIMODAL MODELS DO NOT KNOW WHEN NOT TO BIND

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models benefit from complementary observations, but unrelated evidence can redirect predictions away from the intended target. We call this failure *Forced Fusion*. We demonstrate it in seven task-specific fusion architectures on a synthetic task and five audio-visual datasets, and in pretrained general-purpose multimodal models on audio-visual and audio-text tasks. Our analysis of matched training motivates learning source correspondence to determine when auxiliary evidence should influence prediction. We propose a module based on Bayesian causal inference (BCI) that explicitly learns correspondence and uses it to balance fused and target-only predictions. Evaluations show that our approach mitigates Forced Fusion and improves average target accuracy over the implicit Mixed baseline in most audio-visual settings. Furthermore, analyses of task-specific and adapted general-purpose models show that explicit supervision helps models distinguish matched from mismatched inputs, even when gains in target accuracy are modest. Controlled interventions keep the module's branch predictions fixed and vary only its mixing weight, showing that correspondence-dependent weighting improves mismatched prediction on average. These findings highlight the importance of evaluating target accuracy alongside the ability to recognize source correspondence and use it to guide multimodal fusion.

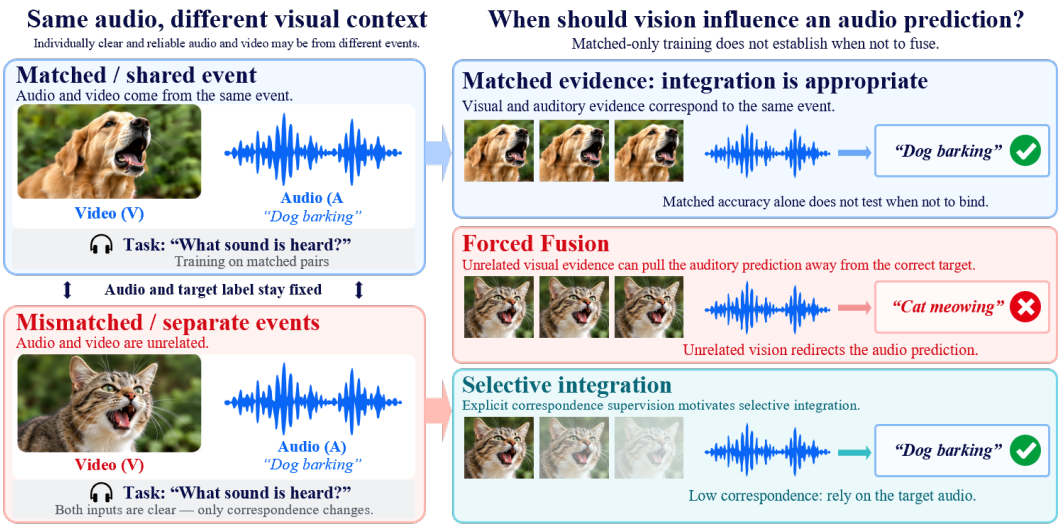


Figure 1: Forced Fusion and Selective Integration. **Left:** the task is to identify the sound. The same dog-barking audio (A) is paired with matching dog video or unrelated cat video (V); the correct auditory label remains unchanged. **Right:** matched evidence supports "dog barking." With the unrelated cat video, Forced Fusion produces "cat meowing," whereas selective integration suppresses visual influence and retains "dog barking."

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.

1 INTRODUCTION

Human perception combines sensory evidence while remaining sensitive to whether the signals originate from a common source. The ventriloquist illusion illustrates this interaction: a puppet's visible mouth movements bias the perceived location of a speaker's voice. At sufficiently large spatial discrepancies between auditory and visual cues, however, integration weakens. Bayesian causal inference (BCI) accounts for this transition by weighting common-source and separate-source interpretations according to the observed cues and their uncertainty (Körding et al., 2007). This dependence on source correspondence highlights a central challenge for multimodal learning: determining when evidence from another modality should influence a prediction.

In typical multimodal training settings, models learn to extract complementary evidence from matched observations of a shared target. Reliability-aware fusion weights modalities according to input quality or predictive uncertainty (Zhang et al., 2023; Cao et al., 2024). Yet individually informative observations can originate from different events, so reliability alone does not determine whether they should be integrated. Prior work has examined audio-visual correspondence (Arandjelovic & Zisserman, 2017) and addressed audio-text conflict and audio-visual mismatch (Wang et al., 2025; Xu et al., 2025c). For selective integration, however, correspondence judgments must guide how the accompanying evidence affects the prediction. Matched accuracy and robustness to degraded inputs alone do not establish this ability.

We define *Forced Fusion* as a failure of selective integration in which unrelated crossmodal evidence distorts a prediction about the source specified by the task (illustrated in Fig. 1). Our analysis distinguishes the *target modality*, whose source defines the prediction target, from the *auxiliary modality*, whose relevance to that source can vary. Our evaluation extends from *task-specific fusion models* trained on matched pairs to pretrained *general-purpose multimodal models*, revealing Forced Fusion in both model families.

To address this behavior, we propose a BCI-inspired approach that explicitly supervises source correspondence during training. For task-specific models, a BCI-based module uses the learned correspondence estimate to balance fused and target-only predictions. We also derive an upper bound on the module's prediction log-loss that depends on correspondence estimation. In both model families, incorporating correspondence into learning mitigates Forced Fusion and improves mismatched prediction relative to matched training. Controlled ablations further show that explicit supervision improves correspondence discrimination. Interventions on the fusion module, with branch predictions fixed, show that correspondence-sensitive weights improve mismatched prediction on average.

Our contributions are threefold:

- We define and demonstrate Forced Fusion across task-specific and general-purpose multimodal models, and characterize a limitation of matched training.
- We propose a BCI-based module for selective fusion with explicit correspondence supervision and derive an upper bound on its prediction log-loss.
- We demonstrate improved prediction under mismatch and quantify the contributions of correspondence supervision and learned fusion weights through controlled analyses.

2 MATCHED TRAINING AND SELECTIVE INTEGRATION

We first formalize shared and separate sources while keeping observation reliability distinct from source correspondence. We then examine what training on shared-source pairs constrains under mismatch and derive a BCI reference that adjusts prediction to the inferred source structure.

2.1 CAUSAL STRUCTURE AND OBSERVATIONS

Let x_t and x_a denote target and auxiliary observations of underlying source states s_t and s_a. A source state specifies the task-relevant quantity, such as a location or semantic category. The target label is $y = h(s_t)$, where h specifies the prediction task, and a classifier predicts $\hat{y} = f_\theta(x_t, x_a)$ from both observations.

Let $C \in \{1, 2\}$ denote the latent causal structure underlying the two observations. Under $C = 1$, a single source state is sampled and shared by both observations; under $C = 2$, two source states are drawn independently from the same prior p_s (Körding et al., 2007):

$$\begin{aligned} C = 1: &\quad s_t = s_a = s, \qquad s \sim p_s, \\ C = 2: &\quad s_t, s_a \overset{\text{iid}}{\sim} p_s. \end{aligned} \tag{1}$$

The observations follow modality-specific likelihoods p_t and p_a, independently given their source states:

$$p(x_t, x_a \mid s_t, s_a, C) = p_t(x_t \mid s_t)\, p_a(x_a \mid s_a). \tag{2}$$

Thus, a shared source can produce observations with different forms and values. Changing C changes source dependence while preserving each modality's marginal distribution; the observation likelihoods separately specify reliability. The synthetic evaluation instantiates this model with Gaussian source and observation distributions (specified in Appendix B.1.1). For real-data evaluations, Sec. 3 specifies the task-dependent criteria used to construct matched and mismatched pairs.

2.2 Limits of Matched Training

Within this model, matched training uses shared-source pairs ($C = 1$), although prediction may also be required for separate sources ($C = 2$). The following bound quantifies how much predictions in regions that are rare under matched sampling can affect matched risk.

Proposition 1 (Matched Risk on Low-Probability Inputs)**.** *Let P_{mat} be any matched distribution of (X, Y), where $X = (X_t, X_a)$, and define $R_{\text{mat}}(f) = \mathbb{E}_{P_{\text{mat}}}[\mathbf{1}\{f(X) \neq Y\}]$. For any measurable input region $\mathcal{B}$, if classifiers f and g agree outside $\mathcal{B}$, then*

$$|R_{\text{mat}}(f) - R_{\text{mat}}(g)| \leq P_{\text{mat}}(X \in \mathcal{B}). \tag{3}$$

Predictions can therefore differ substantially within a region of low matched probability while producing similar matched risks. In the Gaussian instance of the source model, the probability of large observation discrepancies under $C = 1$ has an exponential upper bound determined by the discrepancy relative to observation noise (derived in Appendix A.1). Consequently, when separate-source inputs fall in such regions, matched accuracy alone may not distinguish a predictor that continues to follow the auxiliary from one that relies on the target.

2.3 Selective Integration through BCI

To specify how predictions should respond to these two source structures, BCI infers their probabilities from the observations and averages the corresponding predictions. When the source prior and observation likelihoods are known, a superscript $\star$ denotes the resulting reference posteriors. With common-source prior $\pi = \Pr(C = 1)$, Bayes' rule gives

$$c^\star := \Pr(C = 1 \mid x_t, x_a) = \frac{\pi\, p(x_t, x_a \mid C = 1)}{\pi\, p(x_t, x_a \mid C = 1) + (1 - \pi)\, p(x_t, x_a \mid C = 2)}. \tag{4}$$

Under $C = 1$, define the fused posterior $p^\star_F(y \mid x_t, x_a) = p(y \mid x_t, x_a, C = 1)$. Under $C = 2$, source independence makes the auxiliary uninformative about y given x_t, yielding the target-only posterior $p^\star_{S,t}(y \mid x_t) := p(y \mid x_t, C = 2)$. Averaging over the unknown source structure gives

$$p^\star(y \mid x_t, x_a) = c^\star p^\star_F(y \mid x_t, x_a) + (1 - c^\star) p^\star_{S,t}(y \mid x_t). \tag{5}$$

When the observations favor separate sources, $c^\star$ decreases and the prediction shifts toward the target-only posterior.

3 Empirical Evidence of Forced Fusion

To test whether Forced Fusion extends beyond task-specific training, we evaluate task-specific fusion models on synthetic and real audio-visual tasks, then pretrained general-purpose models on audio-visual and audio-text tasks. Throughout, we retain the target task while varying the accompanying evidence. In the real-data evaluations, audio is the target modality, with images, video, or text as auxiliary inputs.

3.1 Evaluation Design

Task-Specific Fusion Models. The synthetic task uses population-code vectors of noisy source observations for 41-class target-position classification, with models trained on matched, shared-source pairs. We compare basic fusion structures, Concatenation and Cross-attention; reliability- or uncertainty-based methods, TMC (Han et al., 2022), QMF (Zhang et al., 2023), and PDF (Cao et al., 2024); and conflict-aware methods, CoRiM (Zou & Wei, 2026) and DCR (Yu et al., 2026) (detailed in Appendix B.1.1).

We next apply the same methods to real audio-visual classification: digit recognition on AV-MNIST (Becker et al., 2024; LeCun et al., 1998); event recognition on AVE and Kinetics-Sounds (Tian et al., 2018; Arandjelovic & Zisserman, 2017); and emotion recognition on CREMA-D and RAVDESS (Cao et al., 2014; Livingstone & Russo, 2018). We train each method with five random seeds per condition across synthetic and real audio-visual tasks; for real data, matched pairs share a digit class in AV-MNIST or a source clip in the other datasets, while mismatched evaluation pairs audio with different-class visual inputs (detailed in Appendix B.1.2).

General-Purpose Multimodal Models. We evaluate pretrained Qwen2.5 (Xu et al., 2025a) and Qwen3 (Xu et al., 2025b) on the same five datasets, comparing matched and mismatched inputs with model weights frozen (detailed in Appendix B.2.1). We also evaluate these models on an independent audio-text benchmark (Wang et al., 2025) covering audio question answering (AQA), speech emotion recognition (SER), and vocal sound classification (VSC), with matched, conflicting, unrelated, or no auxiliary text (detailed in Appendix B.2.2).

3.2 Observed Behavior

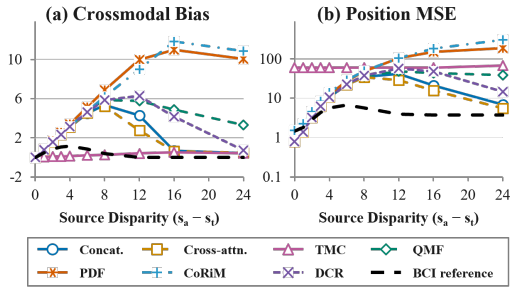


Figure 2: Forced Fusion in the synthetic task. **(a)** Crossmodal bias and **(b)** position MSE (log scale) after Matched-only training. At large disparities, models retain bias or high error, while BCI (black dashed) limits both through selective integration.

Fig. 2 shows the crossmodal bias $\mathbb{E}[\hat{s}_t - s_t]$ and position MSE $\mathbb{E}[(\hat{s}_t - s_t)^2]$ of the target-location estimate $\hat{s}_t$ as source disparity $s_a - s_t$ increases in the synthetic task. For nearby sources, a high correspondence probability $c^\star$ favors fused inference under Eq. (5), drawing the BCI estimate toward the auxiliary. As the sources separate, the observations increasingly favor independent sources: $c^\star$ decreases and $1 - c^\star$ increases, placing more weight on the target-only posterior and returning bias toward zero. At large disparities, BCI combines near-zero bias with lower MSE than models trained on matched pairs, supporting its use as a reference for selective integration in this controlled setting.

Table 1(a) compares matched and mismatched accuracy on real audio-visual data, showing lower mean target accuracy under mismatch on all five datasets. Pretrained models show the same decline relative to matched inputs across all tested datasets and tasks, with mismatched visual inputs (see Table 1(a)) and conflicting or unrelated text (see Table 1(b)). Together, these evaluations demonstrate Forced Fusion under mismatch in both task-specific models trained on matched pairs and pretrained general-purpose models.

4 BCI-Inspired Selective Fusion

To mitigate Forced Fusion, we draw on the BCI reference in the synthetic evaluation, which reduces auxiliary influence as the evidence favors separate sources. We use Mixed training as an implicit baseline and augment it with explicit correspondence supervision.

Table 1: Forced Fusion on real data. **(a)** Audio-visual; **(b)** audio-text. Mat./Mis.: matched/mismatched accuracy (%); Drop: Mat. minus Mis. (pp). Unrelated or conflicting auxiliary inputs reduce target accuracy across model families and tasks.

(a) Audio-visual	**Task-specific models**			**General-purpose models**					
	Seven-method mean			Qwen2.5			Qwen3		
Dataset	Mat.	Mis.	Drop	Mat.	Mis.	Drop	Mat.	Mis.	Drop
AV-MNIST	99.22	72.18	27.05	100.00	48.67	51.33	100.00	29.33	70.67
AVE	60.74	6.02	54.72	82.99	47.65	35.34	92.52	39.60	52.92
Kinetics-Sounds	65.79	3.54	62.25	74.67	28.00	46.67	70.67	28.00	42.67
CREMA-D	60.93	14.43	46.50	36.67	22.67	14.00	22.00	15.33	6.67
RAVDESS	62.14	16.30	45.84	83.52	65.34	18.18	26.70	16.48	10.23
Mean	69.77	22.49	47.27	75.57	42.47	33.10	62.38	25.75	36.63

(b) Audio-text		**General-purpose models**					
		Qwen2.5			Qwen3		
Task	Mismatched text	Mat.	Mis.	Drop	Mat.	Mis.	Drop
AQA	Conflicting	99.00	26.67	72.33	96.33	50.33	46.00
	Unrelated		69.00	30.00		84.00	12.33
SER	Conflicting	100.00	0.00	100.00	85.00	34.67	50.33
	Unrelated		40.00	60.00		47.67	37.33
VSC	Conflicting	100.00	6.00	94.00	99.00	58.00	41.00
	Unrelated		88.00	12.00		90.33	8.67
Mean		99.67	38.28	61.39	93.44	60.83	32.61

4.1 Implicit Learning through Target Prediction

One implicit approach is *Mixed* training: learning target prediction from both matched and mismatched pairs, as explored in prior work (Xu et al., 2025c; Chen et al., 2026). By default, we retain or reassign auxiliary inputs with equal probability. For task-specific fusion models, the target prediction loss discourages reliance on unrelated auxiliary evidence without directly supervising correspondence. Improved accuracy alone, however, does not establish whether the model distinguishes matched from mismatched pairs or uses that distinction to regulate fusion.

4.2 Explicit Correspondence with a BCI-Based Module

For task-specific models, our BCI-based module builds on Mixed training by directly supervising a correspondence estimate and using it to weight fused and target-only predictions.

Design. Let $p_X(y \mid x_t, x_a)$ denote the fused class distribution of the original model X. The module adds a target-only predictor $p_{t,\theta}(y \mid x_t)$ and an MLP-based sigmoid gate $c_\theta(x_t, x_a) \in (0, 1)$. Writing c_θ without its inputs for brevity, the augmented prediction is

$$p_{X+\mathrm{BCI}}(y \mid x_t, x_a) = c_\theta p_X(y \mid x_t, x_a) + (1 - c_\theta) p_{t,\theta}(y \mid x_t). \tag{6}$$

Higher c_θ places more weight on the fused prediction, while lower values favor the target-only branch. The learned gate implements the mixing role of $c^\star$ in Eq. (5) without requiring a known source model.

Theoretical Motivation. The following bound relates correspondence estimation to the mixture's prediction log-loss. For an input pair (x_t, x_a) with target label y, define $\ell_X = -\log p_X(y \mid x_t, x_a)$, $\ell_t = -\log p_{t,\theta}(y \mid x_t)$, and $\ell_{\mathrm{mix}} = -\log p_{X+\mathrm{BCI}}(y \mid x_t, x_a)$.

Proposition 2 (Prediction and Correspondence Loss)**.** *For any two positive class distributions and* $0 < c_\theta < 1$*,*

$$\ell_{\text{mix}} \leq \min\{\ell_X - \log c_\theta,\ \ell_t - \log(1 - c_\theta)\}. \tag{7}$$

Let a binary label $r \in \{0, 1\}$ *select the fused branch for* $r = 1$ *and the target-only branch for* $r = 0$*, and define* $\ell_{\text{route}} = r\ell_X + (1 - r)\ell_t$*. Let* $R_{\text{mix}} = \mathbb{E}[\ell_{\text{mix}}]$ *and* $R_{\text{route}} = \mathbb{E}[\ell_{\text{route}}]$*, with all expectations taken over the same distribution and assumed finite. Then*

$$R_{\text{mix}} \leq R_{\text{route}} + R_{\text{corr}}, \qquad R_{\text{corr}} = \mathbb{E}[-r \log c_\theta - (1 - r) \log(1 - c_\theta)]. \tag{8}$$

When r denotes pair correspondence, R_{route} measures the expected log-loss of using the fused branch for matched pairs and the target-only branch for mismatched pairs. The correspondence loss R_{corr} bounds the mixture's additional expected log-loss relative to this selection (proof in Appendix A.2). For fixed branches, reducing R_{corr} lowers this upper bound, providing a theoretical motivation for correspondence supervision alongside target prediction.

Training Objective. For task-specific fusion models, classification and correspondence supervision use the same task pairs as Mixed. The classification loss $\mathcal{L}_{\text{cls}}$ combines the original objective of X with cross-entropy losses for the mixture, the target-only branch, and an auxiliary-only training branch. Each unimodal branch uses its own input label, while the mixture retains the target label. Correspondence supervision assigns $r = 1$ to original pairs and $r = 0$ to pairs formed by reassigning auxiliary inputs. These labels describe pair construction, since reassignment can preserve the original partner or class. We jointly optimize all components with

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{corr}}, \tag{9}$$

where $\mathcal{L}_{\text{corr}}$ is the batch-average binary cross-entropy between gate estimates c_θ and construction labels r, and λ controls its contribution ($\lambda = 1$ by default).

For general-purpose models, Mixed+BCI provides explicit correspondence supervision through yes/no questions about whether the inputs correspond. Architectures, training objectives, and experimental protocols are detailed in Appendix B.

5 Results and Analysis

We compare three training conditions: the *Matched* and *Mixed* baselines and *Mixed+BCI*, labeled Matched-only, Mixed (Implicit), and Mixed+BCI (Ours; +Explicit) in tables. We evaluate task-specific models on synthetic and real audio-visual tasks, and general-purpose models on audio-visual tasks before and after LoRA fine-tuning (Hu et al., 2021) (training details in Appendix B.2.3). As in Sec. 3, audio is the target modality in the real-data evaluations, and task-specific results are averaged over five training seeds. We first report matched and mismatched performance, then analyze correspondence-sensitive fusion and the effects of training exposure and supervision.

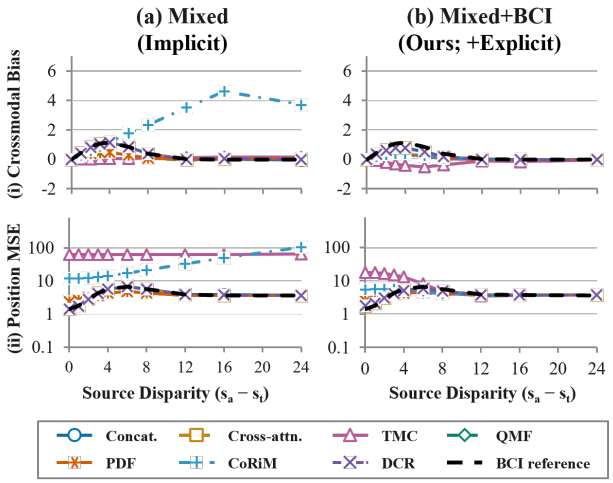


Figure 3: Synthetic prediction. **(a)** Mixed and **(b)** Mixed+BCI: **(i)** bias and **(ii)** MSE (log scale). Black dashed: BCI reference. Mixed+BCI yields more stable predictions at large disparities.

5.1 Predictive Performance

Task-Specific Models. On the synthetic task, Fig. 3(a) shows that Mixed training alone substantially alleviates the Forced Fusion observed in Fig. 2. At source separation $s_a - s_t = 24$, Mixed

Table 2: Audio-visual target accuracy. **(a)** Task-specific models; **(b)** general-purpose models. Mat./Mis.: matched/mismatched accuracy (%); Avg.: their mean; bold: best Avg. per row. Mixed recovers mismatched accuracy; Mixed+BCI further improves Avg. on most datasets.

(a) Task-specific	**Training strategy**								
	(i) Matched-only			(ii) Mixed (Implicit)			(iii) Mixed+BCI (Ours; +Explicit)		
Dataset	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.
AV-MNIST	99.22	72.18	85.70	97.24	88.36	92.80	98.03	95.18	**96.60**
AVE	60.74	6.02	33.38	61.89	41.35	51.62	56.81	48.10	**52.45**
Kinetics-Sounds	65.79	3.54	34.67	61.37	24.01	42.69	57.89	27.89	**42.89**
CREMA-D	60.93	14.43	37.68	57.99	47.77	52.88	57.41	51.62	**54.51**
RAVDESS	62.14	16.30	39.22	47.03	35.69	41.36	43.34	40.84	**42.09**
Mean	69.77	22.49	46.13	65.10	47.44	56.27	62.70	52.73	**57.71**

(b) General-purpose

Dataset	(i) Pretrained			(ii) Matched-only			(iii) Mixed (Implicit)			(iv) Mixed+BCI (Ours; +Explicit)		
	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.
AV-MNIST	100.00	39.00	69.50	100.00	99.00	99.50	100.00	100.00	**100.00**	100.00	100.00	**100.00**
AVE	87.76	43.62	65.69	93.54	15.77	54.65	87.41	73.15	80.28	89.46	73.83	**81.64**
Kinetics-Sounds	72.67	28.00	50.33	86.67	10.67	48.67	73.00	52.00	62.50	74.67	53.33	**64.00**
CREMA-D	29.33	19.00	24.17	71.33	36.67	54.00	71.67	60.33	**66.00**	65.00	53.00	59.00
RAVDESS	55.11	40.91	48.01	88.64	69.32	78.98	88.35	86.65	87.50	90.62	87.78	**89.20**
Mean	68.97	34.11	51.54	88.03	46.28	67.16	84.09	74.43	**79.26**	83.95	73.59	78.77

training brings five of the seven methods close to the BCI reference. Mixed+BCI further reduces the residual error in TMC and CoRiM at this separation, bringing all seven methods within 2.0% of the reference MSE (shown in Fig. 3(b)(ii)). These improvements are clearest at large source separations, where auxiliary evidence should have little influence on the target prediction.

Table 2(a) reports matched and mismatched accuracy and their average on five real audio-visual datasets, aggregated across the seven methods. As in the synthetic task, Mixed substantially improves mismatched accuracy over Matched training. Mixed+BCI shows a further modest increase in accuracy averaged over the matched and mismatched tests.

General-Purpose Models. Table 2(b) reports accuracy averaged over Qwen2.5 and Qwen3 for the Pretrained baseline and the three adaptation conditions. Mixed training again substantially improves mismatched accuracy relative to Matched adaptation. Mixed+BCI has similar overall mean accuracy to Mixed, with modest gains in both test conditions on some datasets. Qualitative examples for both model families are provided in Appendix C.1.

5.2 Analysis

Mixed recovers much of the lost accuracy. We now examine how explicit supervision shapes correspondence sensitivity and its contribution to prediction.

Correspondence-Sensitive Fusion. To test whether BCI adds correspondence sensitivity to reliability-based fusion, we examine QMF and PDF. Fig. 4(a) and Fig. 4(b) compare how fusion weights respond to input noise and changes in correspondence. For Matched and Mixed, we use normalized measures of the models' own fusion weights; for Mixed+BCI, we use the BCI mixing

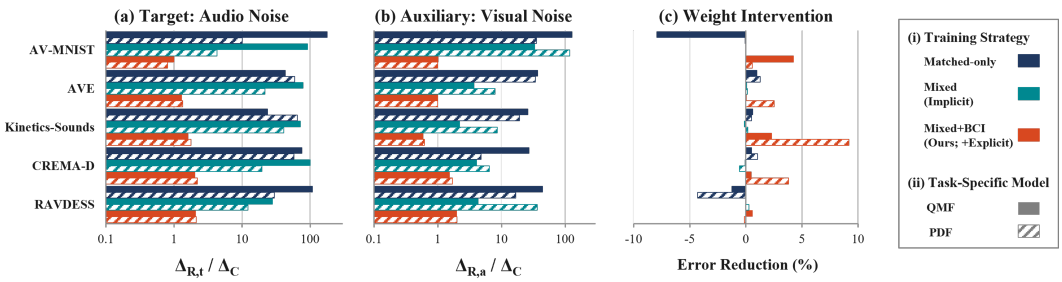


Figure 4: Correspondence sensitivity and predictive effects. **(a)** Audio-noise and **(b)** visual-noise response relative to correspondence changes (log scale). **(c)** Relative target-error reduction versus matched-counterpart weights, with branch predictions fixed. BCI increases relative correspondence sensitivity, and its learned weights reduce mismatched error on average.

Table 3: General-purpose correspondence judgments. **(i)–(iv)** Pretrained, Matched-only, Mixed, and Mixed+BCI. Mat./Mis.: matched/mismatched accuracy (%); Avg.: their mean; bold: best Avg. Mean averages both models. Mixed+BCI strengthens correspondence judgments, distinguishing matched from mismatched inputs more accurately.

Model	**Training strategy**											
	(i) Pretrained			(ii) Matched-only			(iii) Mixed (Implicit)			(iv) Mixed+BCI (Ours; +Explicit)		
	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.	Mat.	Mis.	Avg.
Qwen2.5	93.04	23.72	58.38	93.54	17.37	55.45	77.93	86.16	82.05	90.79	82.14	**86.47**
Qwen3	67.32	73.26	70.29	94.42	24.15	59.28	73.13	86.15	79.64	93.51	80.29	**86.90**
Mean	80.18	48.49	64.33	93.98	20.76	57.37	75.53	86.16	80.84	92.15	81.22	**86.68**

weight c_θ. We degrade one modality at a time using Gaussian noise: 0 dB SNR on audio log-mel features, or standard deviation 1 on visual pixels with clipping to $[0, 1]$. Relative to clean matched inputs, let Δ_C denote the absolute change in mean weight under mismatch, and $\Delta_{R,t}$ and $\Delta_{R,a}$ the corresponding changes under audio and visual degradation. We average these response magnitudes over seeds (as detailed in Appendix B.1.4). The ratios $\Delta_{R,t}/\Delta_C$ and $\Delta_{R,a}/\Delta_C$ show that native weights remain more responsive to the tested quality perturbations than to correspondence changes, even after Mixed training. With BCI, correspondence changes have a much larger effect relative to the same quality perturbations, indicating that the mixing weight reflects the relationship between the inputs.

Contribution to Target Prediction. To test whether this sensitivity affects prediction, we keep each mismatched input and its branch predictions fixed, replacing only the weight with one computed from the same target and a matched auxiliary input. Figure 4(c) reports the relative reduction in target-classification error when the original mismatched-input weight is restored. Restoring the BCI weight reduces error in nine of ten method–dataset combinations on average over seeds. For QMF and PDF, Mixed recovers accuracy despite little change in target error under the native-weight intervention, whereas correspondence-dependent weighting in Mixed+BCI contributes directly to target prediction.

Correspondence Judgments in General-Purpose Models. Table 3 reports correspondence accuracy on matched and mismatched pairs and their average across all five audio-visual datasets, using the same evaluation pairs across conditions (evaluation detailed in Appendix B.2.4). Correct answers are "yes" for matched pairs and "no" for mismatched pairs; Avg. gives the two test conditions equal weight. Relative to Mixed, Mixed+BCI improves this average by 4.42 pp for Qwen2.5 and 7.26 pp for Qwen3. These gains show that explicit correspondence supervision strengthens the models' ability to distinguish matched from mismatched inputs.

Training Exposure. We return to task-specific models to examine training exposure: Fig. 5(a) varies the probability p_{swap} of reassigning auxiliary inputs, with $\lambda = 1$ for Mixed+BCI. Even at $p_{\text{swap}} = 1\%$, mean mismatched accuracy improves by 5.17 pp for Mixed and 13.07 pp for Mixed+BCI relative to Matched training, without reducing mean matched accuracy (see Fig. 5(a)(i)). Across the tested levels, mismatched gains exceed matched accuracy costs on average. For Mixed+BCI, gate AUROC measures how well c_θ separates matched from mismatched pairs across thresholds (0.5: chance; 1: perfect discrimination; computation in Appendix B.1.4). It improves with exposure, with diminishing gains at higher levels (see Fig. 5(a)(ii)). At every tested exposure level, Mixed+BCI achieves larger mean mismatched-accuracy gains than Mixed, indicating that it uses mixed training more effectively for mismatched prediction.

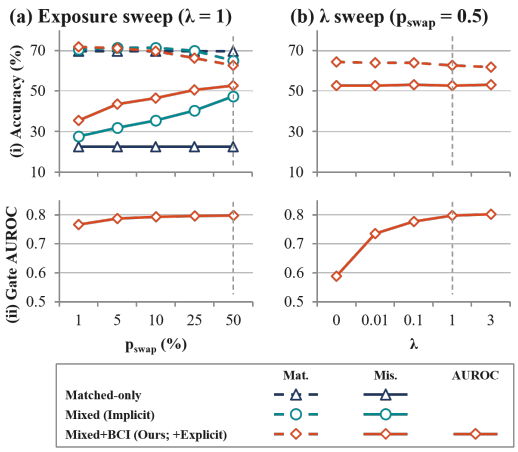


Figure 5: Exposure and supervision effects. **(a)** Exposure and **(b)** λ sweeps: **(i)** accuracy and **(ii)** gate AUROC. Across exposure levels, mismatched gains outweigh matched costs. Explicit supervision chiefly improves gate discrimination, with diminishing gains near $\lambda = 1$.

Role of Correspondence Supervision. To isolate explicit supervision, Fig. 5(b) varies λ with exposure fixed at 50% and the Mixed+BCI architecture unchanged. At $\lambda = 0$, the gate still learns through the mixture classification loss, providing a control without explicit correspondence supervision. Increasing λ strengthens correspondence discrimination: mean gate AUROC rises from 0.588 at $\lambda = 0$ to 0.798 at $\lambda = 1$, with little further change at $\lambda = 3$ (see Fig. 5(b)(ii)). Mismatched accuracy changes little, while matched accuracy decreases gradually (see Fig. 5(b)(i)). The default $\lambda = 1$ therefore captures most of the gain in correspondence discrimination across the tested weights, with little additional benefit from stronger supervision. Explicit supervision thus aligns the mixing weight with pair correspondence. Together with the fixed-branch intervention in Fig. 4(c), this links the ability to recognize mismatch to its use in selective prediction.

6 Conclusion

We demonstrated Forced Fusion across task-specific and general-purpose models and showed why matched performance alone does not establish selective integration under mismatch. To mitigate Forced Fusion, our BCI-based module balances fused and target-only predictions, with a log-loss bound motivating correspondence supervision. Mixed training recovers much of the lost accuracy, while explicit supervision strengthens correspondence discrimination even when target accuracies are similar. Fusion-weight analyses distinguish responses to input degradation from responses to changes in correspondence. Interventions that hold branch predictions fixed further show that correspondence-dependent weighting reduces mismatched target error on average. Reliable auxiliary evidence can still be irrelevant to the target when it comes from a different source. Our findings therefore support evaluating multimodal fusion through both target accuracy and whether models recognize source correspondence and use it to regulate the contribution of auxiliary evidence to the final prediction.

To further examine the limits of correspondence discrimination, we analyze task-specific models on mismatched pairs from different classes within the same broad semantic category (see Appendix C.2). Mixed+BCI improves mean target accuracy over Mixed even for these semantically similar pairs. On the event datasets, however, correspondence judgments are less accurate for these semantically related pairs than for pairs from different categories. These findings therefore call for further research on correspondence learning to distinguish separate events more precisely, even when their semantic content is similar.

Reproducibility Statement

We provide experimental details in Appendix B. For task-specific models, Appendix B.1 specifies data generation and splits, input processing, shared architectures and BCI components, training objectives and hyperparameters, random seeds, and analysis procedures. For general-purpose models, Appendix B.2 describes the evaluation data and pairing rules, question templates, decoding and scoring procedures, LoRA training settings, and correspondence-evaluation protocol.

AI Use Statement

Generative AI tools were used to assist with manuscript writing and language refinement, as well as with organizing and visualizing author-generated experimental results in figures and tables and preparing and refining visual materials. These tools were not used to conduct experiments or generate experimental results. All AI-assisted content was reviewed and verified by the authors.

References

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617. IEEE, 2017.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. *arXiv preprint arXiv:2406.04802*, 2024.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Tianle Chen, Chaitanya Chakka, Arjun Reddy Akula, Xavier Thomas, and Deepti Ghadiyaram. Some modalities are more equal than others: Decoding and architecting multimodal integration in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2151, 2026.

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155. IEEE, 2022.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Konrad P Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS one*, 2(9):e943, 2007.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1140–1144. IEEE, 2022.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. When audio and text disagree: Revealing text bias in large audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4878–4888, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Le Xu, Chenxing Li, Yong Ren, Yujie Chen, Yu Gu, Ruibo Fu, Shan Yang, and Dong Yu. Mitigating audiovisual mismatch in visual-guide audio captioning. *arXiv preprint arXiv:2505.22045*, 2025c.

Yangchen Yu, Qian Chen, Jia Li, Zhenzhen Hu, Jinpeng Hu, Lizi Liao, Erik Cambria, and Richang Hong. To fuse or to drop? dual-path learning for resolving modality conflicts in multimodal emotion recognition. *arXiv preprint arXiv:2605.04877*, 2026.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Shihao Zou and Wei Wei. Corim: Conflict-driven risk minimization for dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37821–37830, 2026.

A PROOFS OF PROPOSITIONS

This appendix proves the matched-risk and mixture-loss bounds used in Secs. 2 and 4. Appendix A.1 gives the matched-risk proof for general input–target distributions and its Gaussian illustration. Appendix A.2 proves the mixture-loss bound for arbitrary probabilistic branches and discusses correspondence weighting and target-only prediction.

A.1 MATCHED RISK ON LOW-PROBABILITY INPUTS

Proof of Proposition 1. Let $L_f(X,Y) = \mathbf{1}\{f(X) \neq Y\}$, with L_g defined analogously. Because f and g agree outside $\mathcal{B}$, their losses satisfy $|L_f - L_g| \leq \mathbf{1}\{X \in \mathcal{B}\}$. Taking expectations under P_{mat} gives

$$\begin{aligned} |R_{\mathrm{mat}}(f) - R_{\mathrm{mat}}(g)| &= \left|\mathbb{E}_{P_{\mathrm{mat}}}[L_f - L_g]\right| \\ &\leq \mathbb{E}_{P_{\mathrm{mat}}}[|L_f - L_g|] \leq P_{\mathrm{mat}}(X \in \mathcal{B}), \end{aligned}$$

which proves Eq. (3). □

The bound applies to any modalities and label space: changing predictions only on a region of small matched probability has a correspondingly small effect on matched risk.

Gaussian Illustration. For the source model in Appendix B.1.1, let $\mathcal{B}_d = \{(x_t, x_a) : |x_t - x_a| \geq d\}$ for $d > 0$ and take $P_{\mathrm{mat}} = P(\cdot \mid C = 1)$. Under Eqs. (11) and (12), the shared source cancels from $D = x_t - x_a = \epsilon_t - \epsilon_a$ when $C = 1$. Thus, $D \mid C = 1 \sim \mathcal{N}(0, v)$ with $v = \sigma_t^2 + \sigma_a^2$. For any $\eta > 0$, the exponential Markov inequality gives

$$\Pr(D \geq d \mid C = 1) \leq e^{-\eta d}\mathbb{E}[e^{\eta D} \mid C = 1] = \exp(-\eta d + \eta^2 v/2).$$

Choosing $\eta = d/v$, applying symmetry, and using the union bound yields the following specialization for classifiers that agree outside $\mathcal{B}_d$:

$$\begin{aligned} |R_{\mathrm{mat}}(f) - R_{\mathrm{mat}}(g)| &\leq P_{\mathrm{mat}}(X \in \mathcal{B}_d) \\ &\leq 2\exp\left[-\frac{d^2}{2(\sigma_t^2 + \sigma_a^2)}\right]. \end{aligned} \tag{10}$$

Large observation discrepancies can therefore carry little matched probability. This exponential expression applies to scalar observations, rather than directly to the noisy population-code inputs used by the learned classifiers.

A.2 PREDICTION AND CORRESPONDENCE LOSS

Proof of Proposition 2. Suppressing the input arguments in Eq. (6), write the mixture prediction as $p_{\mathrm{mix}}(y) = c_\theta p_X(y) + (1 - c_\theta)p_{t,\theta}(y)$. Both terms are nonnegative, giving

$$p_{\mathrm{mix}}(y) \geq c_\theta p_X(y), \qquad p_{\mathrm{mix}}(y) \geq (1 - c_\theta)p_{t,\theta}(y).$$

Applying the decreasing function $-\log$ to both inequalities gives Eq. (7). For $r = 1$, select the first bound; for $r = 0$, select the second. For either value,

$$\ell_{\mathrm{mix}} \leq r\ell_X + (1 - r)\ell_t - r\log c_\theta - (1 - r)\log(1 - c_\theta).$$

Taking expectations proves Eq. (8), assuming the displayed expectations are finite. □

The comparison label r may encode either task correspondence or the empirical pair-construction label. The bound applies to both choices without identifying the learned weight with the Bayesian correspondence posterior $c^\star$. All three risks in Eq. (8) refer to the same distribution; a correspondence loss measured on balanced constructed pairs is not automatically the deployment correspondence risk.

The bound also gives a sufficient condition for lower expected log-loss than the fused branch. Let $R_X = \mathbb{E}[\ell_X]$ and $\rho = \Pr(r = 0) > 0$. Since

$$R_X - R_{\mathrm{route}} = \rho\,\mathbb{E}[\ell_X - \ell_t \mid r = 0],$$

we have $R_{\text{mix}} < R_X$ whenever

$$R_{\text{corr}} < \rho\, \mathbb{E}[\ell_X - \ell_t \mid r = 0].$$

This condition compares the target-only branch's expected advantage on pairs labeled $r = 0$, weighted by their probability, with the correspondence loss R_{corr}. It is a sufficient log-loss condition, not a guarantee of improved classification accuracy or optimization convergence.

The source model in Sec. 2.1 provides a further rationale for the target-only branch. Under Eqs. (1) and (2), the marginal source prior and target observation model are identical for $C = 1$ and $C = 2$. Because the class label depends only on the target source,

$$p(y, x_t \mid C = 1) = p(y, x_t \mid C = 2), \qquad p(y \mid x_t, C = 1) = p(y \mid x_t, C = 2).$$

The target-only population prediction problem is therefore unchanged by correspondence in this model; under $C = 2$, the independent auxiliary input provides no additional target information. For natural data, class-based pairing is an empirical construction and need not satisfy the exact conditional-independence assumptions of this reference model.

B Experimental Details

This appendix details the data, architectures, training objectives, and experimental protocols used in Secs. 3 and 5. Appendix B.1 covers task-specific fusion models, and Appendix B.2 covers general-purpose multimodal models. Within each model family, we describe the initial evaluation settings, training interventions, and analyses in that order.

B.1 Task-Specific Fusion Models

Across synthetic and real audio-visual tasks, all seven methods are trained with five random seeds (0–4) per condition.

B.1.1 Synthetic Source-Position Classification

The synthetic task instantiates the source model in Sec. 2.1 with scalar locations; discretizing the target location gives the class label $y = h(s_t)$. Observations follow

$$x_m = s_m + \epsilon_m, \qquad \epsilon_m \sim \mathcal{N}(0, \sigma_m^2), \qquad m \in \{t, a\}, \tag{11}$$

where $\sigma_m^2 > 0$ and the observation noises are independent of one another and of the sources. With $p_s = \mathcal{N}(\mu_p, \sigma_p^2)$, the source structure is

$$\begin{aligned} C = 1: &\quad s_t = s_a = s, \qquad s \sim \mathcal{N}(\mu_p, \sigma_p^2), \\ C = 2: &\quad s_t, s_a \overset{\text{iid}}{\sim} \mathcal{N}(\mu_p, \sigma_p^2). \end{aligned} \tag{12}$$

Sources and observations are not truncated. Each observation is converted into a population-code vector with independent additive channel noise. The auxiliary observation has lower noise than the target observation, so shared-source integration can improve target estimation. Table 4 summarizes the inputs, training, and evaluation settings.

Matched training uses shared-source pairs and target-label cross-entropy, retaining each method's additional objectives. Evaluation reuses the same observation and channel noise samples across disparities. Classifiers estimate position by weighting class centers with their predicted probabilities. Bias and position MSE are averaged over trials and anchors, then over training seeds.

BCI Reference. For the Gaussian model above, we evaluate $c^\star$ in Eq. (4) by integrating over the source locations. The shared-source and target-only posterior means are

$$\mu_F = \frac{\mu_p/\sigma_p^2 + x_t/\sigma_t^2 + x_a/\sigma_a^2}{\sigma_p^{-2} + \sigma_t^{-2} + \sigma_a^{-2}}, \qquad \mu_{S,t} = \frac{\mu_p/\sigma_p^2 + x_t/\sigma_t^2}{\sigma_p^{-2} + \sigma_t^{-2}}. \tag{13}$$

The reference estimate is $\hat{s}_t^\star = c^\star \mu_F + (1 - c^\star)\mu_{S,t}$. For population-code evaluation, we set $\pi = 0.5$ and account for channel noise through effective observation standard deviations. Template decoding gives $(2.0036, 1.0020)$ for the target and auxiliary modalities, respectively.

Table 4: Synthetic inputs, encoders, training, and evaluation settings.

Component	Configuration
Sources and observations	$\mu_p = 0$, $\sigma_p = 8$; fixed $(\sigma_t, \sigma_a) = (2, 1)$.
Population codes	Target / auxiliary: 48 / 64 channels; tuning widths 4 / 1.5. Preferred locations evenly spaced over $[-30, 30]$. Gaussian channel noise: standard deviation 0.05.
Target classes	41 centers spaced by 1.5 over $[-30, 30]$; midpoint boundaries and endpoint classes for outlying locations.
Encoders	Target $48 \to 128 \to 128$ and auxiliary $64 \to 128 \to 128$ MLPs with GELU. Cross-attention projects input channels into tokens.
Training / validation	200,000 / 20,000 examples per seed, shared across methods.
Optimization	25 epochs; batch size 1,024; Adam (learning rate 10^{-3}). Cosine decay; no weight decay.
Evaluation	Target anchors: $\{-10, 0, 10\}$. Source disparity $\delta = s_a - s_t$: $\{0, 1, 2, 3, 4, 6, 8, 12, 16, 24\}$. 4,000 trials per anchor–disparity pair.

B.1.2 AUDIO-VISUAL CLASSIFICATION

Audio is the target modality, with digit images or video as auxiliary inputs. Table 5 summarizes the five datasets.

Table 5: Dataset information and input durations for task-specific audio-visual models.

Dataset	Contents	Classes	Train	Val.	Test	Audio (s)
AV-MNIST	Digit images and spoken digits	10	24,000	3,000	3,000	1
AVE	Event sounds and video	28	3,314	383	390	4
Kinetics-Sounds	Action sounds and video	32	21,011	2,389	1,594	4
CREMA-D	Emotional speech and video	6	6,130	654	656	2
RAVDESS	Emotional speech and video	6	792	88	176	2

Splits and Pairing. AV-MNIST combines AudioMNIST utterances with MNIST images (Becker et al., 2024; LeCun et al., 1998); speakers 1–48, 49–54, and 55–60 define training, validation, and testing, respectively. AVE follows the released splits (Tian et al., 2018); Kinetics-Sounds uses available released test clips and a video-ID hash to reserve approximately 10% of training clips for validation (Arandjelovic & Zisserman, 2017). CREMA-D assigns actors 1001–1075, 1076–1083, and 1084–1091 to training, validation, and testing, respectively (Cao et al., 2014). RAVDESS uses actors 1–18, 19–20, and 21–24 for the corresponding splits (Livingstone & Russo, 2018). RAVDESS includes audio-visual speech from six emotion classes, excluding calm and surprised.

Training batches sample audio uniformly with replacement, without class balancing. Matched AV-MNIST pairs use same-digit images; the other datasets use frames from the source clip. Mismatched evaluation pairs each target audio with a different-class visual input. AV-MNIST and RAVDESS first sample the visual class uniformly from the other classes, then an example within that class; AVE, Kinetics-Sounds, and CREMA-D sample uniformly from different-class test clips. Evaluation pairs are sampled with replacement from the test splits and fixed across models.

Inputs and Training. Table 6 summarizes the shared input processing and encoder architectures; parameters are learned separately for each method. Table 7 gives the training budgets. Matched training uses the pairs defined above and target-label cross-entropy, retaining each method's additional objectives.

Table 6: Shared inputs, encoders, and optimization for audio-visual fusion models.

Component	Configuration
Audio input	Mono, 16 kHz; cropped or zero-padded to the durations in Table 5, then converted to 64-band log-mel spectrograms.
Visual input	AV-MNIST: 28×28 grayscale images. Other datasets: three 112×112 RGB frames. Pixels scaled to $[0, 1]$.
Audio / digit encoders	Four-layer audio CNN and three-layer digit CNN, both with GELU, pooling, and 128-dimensional outputs.
Video-frame encoder	ImageNet-pretrained ResNet-18 (He et al., 2016); features from three frames are averaged and projected to 128 dimensions.
Optimization	Adam, cosine learning-rate decay, and no weight decay.

Table 7: Training budgets for audio-visual fusion models. Matched CoRiM on CREMA-D uses a learning rate of 10^{-3}.

Dataset	Epochs	Steps/epoch	Batch	Learning rate
AV-MNIST	12	400	256	10^{-3}
AVE	20	120	96	5×10^{-4}
Kinetics-Sounds	20	120	96	5×10^{-4}
CREMA-D	20	150	128	5×10^{-4}
RAVDESS	20	60	64	10^{-3}

B.1.3 MIXED AND MIXED+BCI TRAINING

Training Objective. Given a matched batch of size B, we draw a permutation τ of its indices and independent indicators $b_i \sim \text{Bernoulli}(p_{\text{swap}})$, with $p_{\text{swap}} = 0.5$ by default. We set $\tilde{x}_{a,i} = x_{a,i}$ when $b_i = 0$ and $\tilde{x}_{a,i} = x_{a,\tau(i)}$ otherwise; $\tilde{y}_{a,i}$ is the label of the selected auxiliary input. The target input and label y_i remain unchanged. Mixed and Mixed+BCI use this pair construction and the training budgets in Tables 4 and 7.

Mixed optimizes the base method's objective $\mathcal{L}_X$ on these pairs. For Mixed+BCI, let $\ell_{t,i}(y_i)$, $\ell_{a,i}(\tilde{y}_{a,i})$, and $\ell_{\text{mix},i}(y_i)$ denote the target-branch, auxiliary-branch, and mixture cross-entropies for example i. The classification objective is

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_X + \frac{1}{B}\sum_{i=1}^{B}\big[\ell_{t,i}(y_i) + \ell_{a,i}(\tilde{y}_{a,i}) + \ell_{\text{mix},i}(y_i)\big]. \tag{14}$$

The auxiliary classifier contributes only to training; prediction uses the fused and target-only distributions in Eq. (6). With construction label $r_i = 1 - b_i$ and $c_{\theta,i} = c_\theta(x_{t,i}, \tilde{x}_{a,i})$, correspondence supervision is

$$\mathcal{L}_{\text{corr}} = -\frac{1}{B}\sum_{i=1}^{B}\left[r_i \log c_{\theta,i} + (1 - r_i)\log(1 - c_{\theta,i})\right]. \tag{15}$$

All components are jointly optimized with $\mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{corr}}$, using $\lambda = 1$ by default. Both losses use the same task pairs; no separate negative batch is added. Since reassignment can retain the original partner or class, r_i indicates how the pair was constructed.

Implementation. The base method supplies the fused distribution in Eq. (6); Table 8 lists the additional BCI components.

Table 8: BCI components for task-specific fusion models. K is the number of classes; hidden layers use GELU.

Added component	Synthetic	Audio-visual
Encoders	Target: $48 \to 128 \to 128$ Auxiliary: $64 \to 128 \to 128$	$128 \to 128 \to 128$ per modality
Unimodal classifiers	$128 \to 128 \to K$ per modality	
Correspondence gate	$256 \to 128 \to 1$, followed by a sigmoid	

B.1.4 Analysis Protocols

The task-specific analyses in Sec. 5.2 use the audio-visual datasets and inputs in Tables 5 and 6 under the training conditions described above.

Weight Responses. For the QMF and PDF analyses in Fig. 4(a) and Fig. 4(b), we extract a scalar weight w from each input pair. Under Matched and Mixed training, this summarizes the model's native weighting of auxiliary evidence. Let $\boldsymbol{u}_m \in \mathbb{R}^K$ denote modality m's class logits before softmax, with $m \in \{t, a\}$ and K classes. QMF computes energy-based fusion weights from these logits; its auxiliary-weight summary is

$$e_m = \frac{1}{10} \log\left(\sum_{k=1}^{K} \exp(u_{m,k})\right), \qquad w_{\mathrm{QMF}} = \frac{|e_a|}{|e_t| + |e_a|}. \tag{16}$$

These energy weights combine the unimodal logits as $e_t \boldsymbol{u}_t + e_a \boldsymbol{u}_a$. The absolute-value normalization is used only for analysis and keeps w_{QMF} in $[0, 1]$ even when an energy weight is negative. PDF produces two normalized modality weights whose sum is one; w is its auxiliary weight at inference. For Mixed+BCI, $w = c_\theta$ controls the contribution of the fused distribution relative to the target-only distribution in Eq. (6).

For each method, dataset, and training condition, let j index the $J = 5$ checkpoints obtained from the training seeds. Within checkpoint j, $\bar{w}_{j,\mathrm{mat}}$ and $\bar{w}_{j,\mathrm{mis}}$ are mean weights over clean matched and mismatched pairs. Similarly, $\bar{w}_{j,\mathrm{noise},m}$ is the mean over matched pairs with noise applied only to modality m, as specified in Sec. 5.2. We take the absolute change from the clean matched mean within each checkpoint, then average across checkpoints:

$$\Delta_C = \frac{1}{J} \sum_{j=1}^{J} |\bar{w}_{j,\mathrm{mis}} - \bar{w}_{j,\mathrm{mat}}|, \qquad \Delta_{R,m} = \frac{1}{J} \sum_{j=1}^{J} |\bar{w}_{j,\mathrm{noise},m} - \bar{w}_{j,\mathrm{mat}}|. \tag{17}$$

The reported ratio $\Delta_{R,m}/\Delta_C$ divides the mean noise response by the mean correspondence response.

Matched-Weight Intervention. For Fig. 4(c), we compare (x_t, x_a^-) and (x_t, x_a^+) with the same target input. Superscripts $-$ and $+$ denote mismatched and matched auxiliary inputs, respectively. We hold the branch predictions from the mismatched pair fixed and replace only the fusion weights with those computed from its matched counterpart. For Matched and Mixed, both native modality weights are replaced while the unimodal logits remain fixed. For Mixed+BCI, we replace $c_\theta(x_t, x_a^-)$ by $c_\theta(x_t, x_a^+)$ in Eq. (6), keeping the mismatched fused distribution and target-only distribution fixed.

For checkpoint j, let $E_{\mathrm{actual},j}$ and $E_{\mathrm{replaced},j}$ be the fractions of incorrect target-class predictions using the original and replaced weights, respectively. We compute the relative error reduction within each checkpoint and average:

$$\mathrm{ErrorReduction}\ (\%) = \frac{100}{J} \sum_{j=1}^{J} \left(1 - \frac{E_{\mathrm{actual},j}}{E_{\mathrm{replaced},j}}\right). \tag{18}$$

A positive value indicates a relative error reduction from restoring the original mismatched-input weights, averaged across checkpoints.

Gate AUROC. For Fig. 5, we use c_θ as the correspondence score, with matched pairs labeled positive and mismatched pairs negative. We compute rank-based AUROC separately for each checkpoint, then average equally over checkpoints, methods, and datasets at each sweep setting.

B.2 General-Purpose Multimodal Models

Section 3 evaluates frozen Qwen2.5 and Qwen3 (Xu et al., 2025a;b) on audio-visual and audio-text tasks. Section 5 uses the same audio-visual evaluation protocol to compare LoRA training conditions and analyze correspondence judgments.

B.2.1 Audio-Visual Data and Evaluation

Table 9 summarizes the audio-visual data used for evaluation and subsequent adaptation. AV-MNIST evaluation uses AudioMNIST speakers 55–60 and MNIST test images (Becker et al., 2024; LeCun et al., 1998). The other datasets use the test splits in Appendix B.1.2.

Table 9: Audio-visual dataset information for general-purpose models.

Dataset	Classes	Training pool	Evaluation pairs
AV-MNIST	10	27,000	150
AVE	28	3,277	147
Kinetics-Sounds	32	5,832	150
CREMA-D	6	6,124	150
RAVDESS	6	792	176

LoRA training uses classification questions for both the target (auditory) and auxiliary (visual) modalities, with correspondence questions added in Mixed+BCI. We report target-classification and correspondence performance. Table 10 presents the question templates used in training and evaluation, each issued in a separate conversation.

Both models receive the system message "You are a careful assistant. Answer concisely." and use greedy decoding. Matched pairs share a digit class in AV-MNIST or a recording in the other datasets; mismatched pairs use different-class visual inputs. Evaluation pairs are fixed across models and adaptation conditions. For AV-MNIST, we score the first digit in the response; for the other datasets, we compare the normalized answer with the listed class names or fixed aliases. Unparsed answers are incorrect.

Table 10: Audio-visual question templates for training and evaluation. $\mathcal{Y}$ lists the class names in Table 11, separated by semicolons.

Question	Prompt
Digits: AV-MNIST	
Auditory	What digit do you hear in the audio? Answer with only a single digit.
Visual	What digit is written in the image? Answer with only a single digit.
Correspondence	Does the digit written in the image match the digit spoken in the audio? Answer only yes or no.
Events: AVE and Kinetics-Sounds	
Auditory	What is the source of the sound in the audio? Choose exactly one from this list and answer with only that name: $\mathcal{Y}$.
Visual	What do the video frames show? Choose exactly one from this list and answer with only that name: $\mathcal{Y}$.
Correspondence	Do the video frames show the same event that you hear in the audio? Answer only yes or no.
Emotions: CREMA-D and RAVDESS	
Auditory	What emotion does the speaker's voice express in the audio? Choose exactly one from this list and answer with only that word: $\mathcal{Y}$.
Visual	What emotion does the person's face show in the video frames? Choose exactly one from this list and answer with only that word: $\mathcal{Y}$.
Correspondence	Do the video frames show the same emotion that you hear in the audio? Answer only yes or no.

Table 11: Class labels used in the five audio-visual datasets.

Dataset	Class labels
AV-MNIST	0; 1; 2; 3; 4; 5; 6; 7; 8; 9.
AVE	Accordion; Acoustic guitar; Baby cry, infant cry; Banjo; Bark; Bus; Cat; Chainsaw; Church bell; Clock; Female speech, woman speaking; Fixed-wing aircraft, airplane; Flute; Frying (food); Goat; Helicopter; Horse; Male speech, man speaking; Mandolin; Motorcycle; Race car, auto racing; Rodents, rats, mice; Shofar; Toilet flush; Train horn; Truck; Ukulele; Violin, fiddle.
Kinetics-Sounds	blowing nose; blowing out candles; bowling; chopping wood; dribbling basketball; laughing; mowing lawn; playing accordion; playing bagpipes; playing bass guitar; playing clarinet; playing cymbals; playing drums; playing guitar; playing harmonica; playing harp; playing keyboard; playing organ; playing piano; playing saxophone; playing trombone; playing trumpet; playing violin; playing xylophone; ripping paper; shoveling snow; shuffling cards; singing; stomping grapes; tap dancing; tapping pen; tickling.
CREMA-D	angry; disgusted; fearful; happy; neutral; sad.
RAVDESS	angry; disgusted; fearful; happy; neutral; sad.

B.2.2 AUDIO-TEXT EVALUATION

Both frozen models use the first 300 released MCR-Bench items per task (Wang et al., 2025) (tasks listed in Table 12). Keeping the audio and reference answer fixed, we vary the accompanying description: absent, aligned with the audio, conflicting with it, or unrelated to it.

Table 12: Audio-text tasks from MCR-Bench.

Task	Audio source	Prediction target
Audio question answering (AQA)	ClothoAQA (Lipping et al., 2022)	Yes/no answer
Speech emotion recognition (SER)	MELD (Poria et al., 2019)	Seven emotion classes
Vocal sound classification (VSC)	VocalSound (Gong et al., 2022)	Six sound classes

The description preamble in Table 13 precedes the task instruction. The no-description condition omits the preamble and, for AQA, the literal "Question:" prefix. We use greedy decoding and count a response as correct if it contains the reference answer as a substring, ignoring case.

Table 13: Audio-text prompt components. Placeholders contain the released description and item-specific question.

Component	Prompt
Description preamble	The following text provides descriptions derived from audio content. This information may be accurate, inaccurate, or irrelevant. Answer the question with discretion regarding how you use these details. Description: ⟨description⟩
AQA	Question: ⟨question⟩ Please answer with yes or no.
SER	Classify this emotion into one of the following categories: sad, happy, fearful, angry, surprised, disgusted, or neutral.
VSC	Classify this vocal sound into one of the following categories: Laughter, Sigh, Cough, Throat clearing, Sneeze, or Sniff.

B.2.3 LoRA Training Conditions

For the training comparisons in Sec. 5, we apply LoRA (Hu et al., 2021) using the shared settings in Table 14; the pretrained baseline remains frozen. Training instructions are constructed from examples sampled with replacement from the pools in Table 9. The AV-MNIST training pool combines AudioMNIST speakers 1–54 with MNIST training images; the other datasets follow the training splits in Appendix B.1.2.

Training Conditions and Objective. All LoRA conditions optimize answer-token cross-entropy using the questions in Table 10. Matched uses only matched pairs, while Mixed samples matched and different-class pairs with equal probability. Both select auditory or visual classification questions with probability $1/2$ each. Mixed+BCI selects auditory classification, visual classification, or yes/no correspondence questions with probability $1/3$ each, retaining equal probabilities for matched and different-class pairs within every question type.

B.2.4 Correspondence Judgments

Table 15 details the evaluation of correspondence answers from the pretrained models and all three LoRA conditions, reported in Table 3.

Table 14: Shared LoRA training settings for Matched-only, Mixed (Implicit), and Mixed+BCI (Ours; +Explicit).

Setting	Configuration
Instruction budget	4,000 sampled instructions; one pass; one adaptation run per model–dataset–condition.
LoRA	Rank 16, scaling 32, dropout 0.05; base model weights frozen.
Optimizer	AdamW, learning rate 10^{-4}, no weight decay.
Updates	Gradient accumulation eight; 500 updates; gradients clipped at norm one.

Table 15: Protocol for correspondence judgments in general-purpose models.

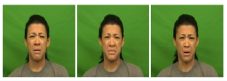
Component	Protocol
Data and prompts	The five audio-visual datasets and fixed evaluation pairs in Appendix B.2.1, with correspondence prompts from Table 10.
Answer parsing	Case-insensitive yes/no prefixes; unparsed answers count as incorrect.
Balanced accuracy	Mean of the correct-answer rate on matched pairs (“yes”) and the correct-answer rate on mismatched pairs (“no”).

C SUPPLEMENTARY RESULTS

Appendix C.1 presents prediction examples, and Appendix C.2 examines correspondence judgments and target-accuracy gains across semantic groups.

C.1 QUALITATIVE EXAMPLES

Figure 6 illustrates corrections on mismatched input pairs for task-specific models.

Input pair	Model	Matched-only	Mixed (Implicit)	Mixed+BCI (Ours; +Explicit)	c_θ
(a) AV-MNIST					
Target (audio) **1** Auxiliary (visual) 9 	Concat.	9	3	**1**	0.016
	Cross-attn.	9	**1**	**1**	0.271
	TMC	**1**	9	**1**	0.007
	QMF	9	9	**1**	0.012
	PDF	9	**1**	**1**	0.112
	CoRiM	9	9	**1**	0.163
	DCR	**1**	**1**	**1**	0.084
(b) AVE					
Target (audio) **Male speech** Auxiliary (visual) Church bell 	Concat.	Church bell	**Male speech**	**Male speech**	0.003
	Cross-attn.	Church bell	**Male speech**	**Male speech**	0.000
	TMC	Church bell	Church bell	**Male speech**	0.008
	QMF	Church bell	**Male speech**	**Male speech**	0.000
	PDF	Church bell	Female speech	**Male speech**	0.005
	CoRiM	Church bell	Church bell	**Male speech**	0.002
	DCR	Church bell	**Male speech**	**Male speech**	0.003
(c) Kinetics-Sounds					
Target (audio) **Playing accordion** Auxiliary (visual) Playing cymbals 	Concat.	Playing cymbals	**Playing accordion**	**Playing accordion**	0.028
	Cross-attn.	Playing cymbals	**Playing accordion**	**Playing accordion**	0.002
	TMC	Playing cymbals	Playing violin	**Playing accordion**	0.024
	QMF	Playing cymbals	**Playing accordion**	**Playing accordion**	0.006
	PDF	Playing cymbals	Playing violin	**Playing accordion**	0.015
	CoRiM	Playing cymbals	Playing cymbals	**Playing accordion**	0.013
	DCR	Playing cymbals	Playing violin	**Playing accordion**	0.007
(d) CREMA-D					
Target (audio) **Sad** Auxiliary (visual) Disgust 	Concat.	Disgust	Fear	**Sad**	0.027
	Cross-attn.	Disgust	Disgust	**Sad**	0.029
	TMC	Disgust	Disgust	**Sad**	0.026
	QMF	Disgust	**Sad**	**Sad**	0.018
	PDF	Disgust	**Sad**	**Sad**	0.025
	CoRiM	Disgust	**Sad**	**Sad**	0.005
	DCR	Disgust	Disgust	**Sad**	0.005
(e) RAVDESS					
Target (audio) **Neutral** Auxiliary (visual) Happy 	Concat.	Happy	**Neutral**	**Neutral**	0.000
	Cross-attn.	Happy	Sad	**Neutral**	0.000
	TMC	Happy	Sad	**Neutral**	0.016
	QMF	Happy	**Neutral**	**Neutral**	0.002
	PDF	Happy	**Neutral**	**Neutral**	0.000
	CoRiM	Happy	**Neutral**	**Neutral**	0.000
	DCR	Happy	Sad	**Neutral**	0.001

Target label (correct) Auxiliary label Other label c_θ: fusion weight

Figure 6: Task-specific predictions on mismatched input pairs. One pair per dataset is fixed across all seven methods and three training conditions. Prediction colors identify target, auxiliary, and other classes; c_θ is the Mixed+BCI fusion weight.

Figure 7 illustrates corrections on mismatched input pairs for general-purpose models.

(a) AV-MNIST

Question — **Target:** What digit do you hear in the audio?
Correspondence: Does the digit written in the image match the digit spoken in the audio?

Input pair	Model	Question	Pretrained	Matched-only	Mixed (Implicit)	Mixed+BCI (Ours; +Explicit)
Target: **6** Auxiliary: 3 	Qwen2.5	Target	3	**6**	**6**	**6**
		Correspondence	**No**	Yes	**No**	**No**
	Qwen3	Target	3	3	**6**	**6**
		Correspondence	Yes	Yes	**No**	**No**

(b) AVE

Question — **Target:** What is the source of the sound in the audio?
Correspondence: Do the video frames show the same event that you hear in the audio?

Input pair	Model	Question	Pretrained	Matched-only	Mixed (Implicit)	Mixed+BCI (Ours; +Explicit)
Target: **Ukulele** Auxiliary: Train horn 	Qwen2.5	Target	**Ukulele**	Train horn	Mandolin	**Ukulele**
		Correspondence	Yes	**No**	**No**	**No**
	Qwen3	Target	Banjo	Mandolin	Mandolin	**Ukulele**
		Correspondence	**No**	**No**	**No**	**No**

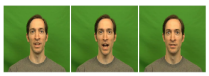
(c) Kinetics-Sounds

Question — **Target:** What is the source of the sound in the audio?
Correspondence: Do the video frames show the same event that you hear in the audio?

Input pair	Model	Question	Pretrained	Matched-only	Mixed (Implicit)	Mixed+BCI (Ours; +Explicit)
Target: **Laughing** Auxiliary: Playing cymbals 	Qwen2.5	Target	Playing drums	Playing drums	Tickling	**Laughing**
		Correspondence	Yes	Yes	**No**	**No**
	Qwen3	Target	Playing drums	Playing drums	Tickling	**Laughing**
		Correspondence	**No**	Yes	**No**	**No**

(d) CREMA-D

Question — **Target:** What emotion does the speaker's voice express in the audio?
Correspondence: Do the video frames show the same emotion that you hear in the audio?

Input pair	Model	Question	Pretrained	Matched-only	Mixed (Implicit)	Mixed+BCI (Ours; +Explicit)
Target: **Neutral** Auxiliary: Fearful 	Qwen2.5	Target	Sad	**Neutral**	Angry	**Neutral**
		Correspondence	Yes	Yes	Yes	**No**
	Qwen3	Target	Sad	**Neutral**	**Neutral**	**Neutral**
		Correspondence	**No**	**No**	**No**	**No**

(e) RAVDESS

Question — **Target:** What emotion does the speaker's voice express in the audio?
Correspondence: Do the video frames show the same emotion that you hear in the audio?

Input pair	Model	Question	Pretrained	Matched-only	Mixed (Implicit)	Mixed+BCI (Ours; +Explicit)
Target: **Fearful** Auxiliary: Happy 	Qwen2.5	Target	Happy	Happy	Disgusted	**Fearful**
		Correspondence	Yes	Yes	Yes	**No**
	Qwen3	Target	Neutral	Happy	Happy	**Fearful**
		Correspondence	**No**	Yes	**No**	**No**

Target label (correct) Auxiliary label Other label

Figure 7: General-purpose predictions on mismatched input pairs. Each pair is fixed across both Qwen models and four conditions. Target rows show predicted classes; Correspondence rows show separately queried answers (correct: No). Class lists and answer-format instructions are omitted from this figure; full question templates are provided in Table 10.

C.2 Semantic Categories and Correspondence Discrimination

This analysis examines how broad semantic category sharing relates to correspondence judgments and target-accuracy gains from explicit supervision. We compare Mixed and Mixed+BCI on AVE, Kinetics-Sounds, CREMA-D, and RAVDESS in task-specific models. The prediction target remains auditory.

Defining Semantic Groups. Table 16 lists the manually specified broad categories and their class assignments. Event labels are grouped by sound source or activity, while emotion labels are grouped by valence. We compare pairs whose target and auxiliary labels belong to the same category with pairs from different categories. For example, acoustic guitar and flute share the music category, while acoustic guitar and bark belong to different categories. For the emotion datasets, different-class pairs sharing a category occur only within the negative-valence group, because the positive and neutral groups each contain one class.

Table 16: Broad semantic categories used to group mismatched class pairs. Emotion labels use the canonical names adopted in the analysis.

Broad category	Class labels
(a) AVE	
Human voices	Baby cry, infant cry; Female speech, woman speaking; Male speech, man speaking
Animals	Bark; Cat; Goat; Horse; Rodents, rats, mice
Music	Accordion; Acoustic guitar; Banjo; Church bell; Flute; Mandolin; Shofar; Ukulele; Violin, fiddle
Vehicles	Bus; Fixed-wing aircraft, airplane; Helicopter; Motorcycle; Race car, auto racing; Train horn; Truck
Household/tools	Chainsaw; Clock; Frying (food); Toilet flush
(b) Kinetics-Sounds	
Instruments	playing accordion; playing bagpipes; playing bass guitar; playing clarinet; playing cymbals; playing drums; playing guitar; playing harmonica; playing harp; playing keyboard; playing organ; playing piano; playing saxophone; playing trombone; playing trumpet; playing violin; playing xylophone
Human sounds	blowing nose; laughing; singing; tickling
Object-related activities	blowing out candles; bowling; chopping wood; dribbling basketball; mowing lawn; ripping paper; shoveling snow; shuffling cards; stomping grapes; tap dancing; tapping pen
(c) CREMA-D and RAVDESS	
Negative valence	anger; disgust; fear; sad
Positive valence	happy
Neutral valence	neutral

Selecting Comparison Pairs. We retain mismatched pairs with different target and auxiliary class labels. Sharing a broad category therefore does not change a pair's mismatched status. Assignments use reference class labels and remain fixed across models and training conditions. For example, AVE contains 159 shared-category pairs and 608 different-category pairs.

Measurements. Table 17 separates correspondence judgments from target prediction. Correspondence is the percentage of mismatched pairs correctly identified by Mixed+BCI, using $c_\theta < 0.5$. Target accuracy gain is the change in auditory classification accuracy from Mixed to Mixed+BCI, in

percentage points. We report both measures separately for shared-category and different-category pairs.

Table 17: Task-specific correspondence judgments and target-prediction gains on mismatched pairs. **(i)–(ii)** report correct correspondence judgments by Mixed+BCI (Ours; +Explicit), in %. **(iii)–(iv)** report its auditory target-classification accuracy minus that of Mixed (Implicit), in pp; positive values indicate an accuracy gain. Same and Different denote whether the pair shares a broad category.

	Correspondence (%)		Target accuracy gain (pp)	
Dataset	**(i)** Same	**(ii)** Different	**(iii)** Same	**(iv)** Different
AVE	71.82	94.95	+5.98	+6.97
Kinetics-Sounds	59.81	75.41	+3.87	+3.90
CREMA-D	80.33	81.28	+2.62	+4.74
RAVDESS	87.43	86.78	+1.34	+8.37
All mean	74.85	84.61	+3.45	+6.00

Correspondence Judgments. Table 17(i)–(ii) shows that shared-category event pairs are harder to identify as mismatched. On AVE, models correctly identify 71.82% of shared-category pairs versus 94.95% of different-category pairs. Kinetics-Sounds shows the same direction, while the differences are smaller and less consistent in the emotion datasets.

Target Prediction. Table 17(iii)–(iv) shows that the task-specific averages improve over Mixed in both semantic groups on all four datasets, with mean gains of 3.45 and 6.00 pp for shared-category and different-category pairs. Target prediction can therefore improve even when semantically related mismatches remain difficult to distinguish.