MUTE: MULTIMODAL UNCERTAINTY-TEMPERED MUTUAL TEACHING FOR LIFELONG TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) is crucial for deploying multimodal models in real-world scenarios where test distributions differ from training data. Multimodal test-time adaptation (MMTTA) further leverages complementary information across multiple modalities to improve robustness. In this work, we focus on audio-video MMTTA, where models process temporally aligned audio and video inputs to produce audio, video, and fused predictions. However, we observe that existing TTA methods suffer significant performance degradation in lifelong MMTTA settings, where models must adapt sequentially to a stream of diverse corruptions without forgetting previous knowledge. To address this challenge, we propose MUTE (Multimodal Uncertainty-Tempered Mutual Teaching), a plug-and-play framework that uses confidence-based routing with a modality-advantage criterion to select the most reliable teacher among the audio, video, and fused heads and transfer its supervision to the others. Confidence-conditioned temperature scheduling and adaptive weight modulation stabilize mutual teaching, while our analysis provides a margin-dependent per-step risk bound and a temperature-dependent teacher-target sensitivity bound. MUTE is buffer-free and composes seamlessly with existing TTA methods. Extensive experiments with twelve base TTA methods across two datasets and both audio and video corruption streams demonstrate that MUTE generally improves average accuracy, with gains of up to 42.9 percentage points for severely degraded base methods and up to 2.1 points for already stable ones. Code is available at `https://anonymous.4open.science/r/MUTE-ICLR27-C2B4/`.

1 INTRODUCTION

Test-time adaptation (TTA) adapts pre-trained models to unlabeled test streams under distribution shift, typically via self-supervised objectives such as entropy minimization and batch-norm updating (Wang et al., 2021; Niu et al., 2022). While effective in single-step settings, TTA performance degrades sharply in lifelong scenarios with sequential corruptions, suffering from spurious updates and catastrophic forgetting. Multimodal TTA (MMTTA) extends TTA to leverage complementary information across multiple modalities. In this work, we focus on audio-video MMTTA, where models process temporally aligned audio and video inputs to produce audio, video, and fused predictions. However, modality reliability asymmetry and the lack of a unified confidence metric for cross-modal comparison make naïve extensions of unimodal TTA unstable, leading to compounded errors over time (Yang et al., 2024; Guo & Jin, 2025; Cao et al., 2023).

Fig. 1 compares Tent (Wang et al., 2021) on KS50 video corruptions (Kay et al., 2017; Yang et al., 2024) under single-step and lifelong adaptation. Its average accuracy falls from 58.94% to 7.1% in the lifelong stream, motivating multimodal mutual teaching for sequential adaptation.

We introduce **MUTE** (*Multimodal Uncertainty-Tempered Mutual Teaching*) to address this gap. MUTE is a plug-and-play framework that enables multimodal mutual teaching to fully exploit the advantages of multimodal learning. At its core, MUTE selects the most reliable teacher from the audio, video, and fused heads using a confidence score and a modality-advantage criterion. The routing mechanism enables direction-adaptive knowledge transfer, allowing the most reliable head to

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.

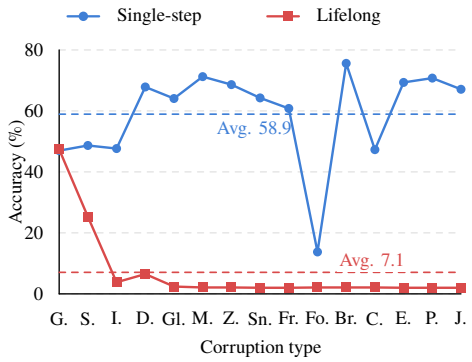

Figure 1: Performance comparison of Tent on KS50 video corruptions under single-step vs lifelong adaptation settings. Single-step adaptation achieves 58.94% average accuracy, while lifelong adaptation catastrophically fails at 7.1%, demonstrating the severe degradation of existing TTA methods in lifelong multimodal scenarios.

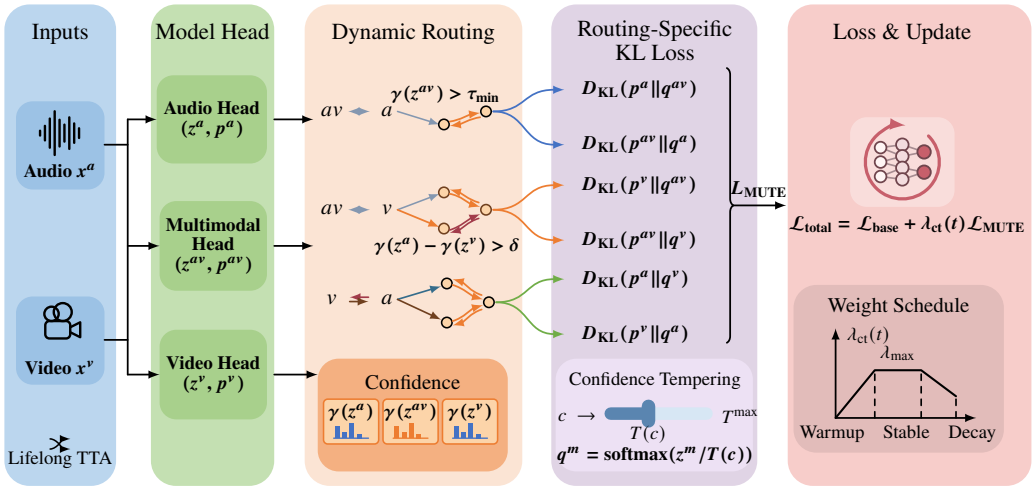

Figure 2: MUTE pipeline. Given temporally aligned audio-video inputs, three heads (audio, video, multimodal) produce logits and compute confidence scores. The routing mechanism selects the highest-confidence head as teacher and activates up to two of six possible directed paths when it meets the confidence gate $\tau_{\min}$ and exceeds the corresponding student by more than δ. Teacher distributions are tempered with confidence-conditioned temperature to produce softer targets under uncertainty. The mutual-teaching loss aggregates KL divergences over all active routing paths, weighted by a three-phase schedule (warmup, stable, decay), and combines with the base TTA loss to update model parameters.

supervise the other heads. We employ confidence-conditioned temperature scheduling and adaptive weight modulation to stabilize the mutual-teaching process. MUTE is buffer-free and composes seamlessly with existing TTA methods.

Our main contributions are summarized as follows:

- We reveal severe degradation of existing TTA methods in lifelong multimodal settings and propose **MUTE**, a plug-and-play multimodal mutual-teaching head that selects the most reliable audio, video, or fused head and routes its supervision to eligible student heads using confidence and modality-advantage criteria. We also employ confidence-conditioned temperature scheduling and adaptive weight modulation to further stabilize the mutual-teaching process.

- We derive a margin-dependent per-step risk bound and a temperature-dependent teacher-target sensitivity bound that characterize MUTE's routing and tempering mechanisms.
- Experiments with twelve base TTA methods across two datasets and both audio and video corruption streams show that MUTE generally improves average accuracy, with gains of up to 42.9 percentage points for severely degraded base methods and up to 2.1 points for already stable ones.

2 Related Works

TTA enables pre-trained models to adapt to unlabeled test streams under distribution shift without access to source data. Early approaches update normalization affine parameters through entropy minimization (Tent) (Wang et al., 2021), filter reliable samples and preserve important source parameters (EATA) (Niu et al., 2022), or minimize marginal entropy across augmentations (MEMO) (Zhang et al., 2022). Continual TTA methods such as CoTTA (Wang et al., 2022), RoTTA (Yuan et al., 2023), and SAR (Niu et al., 2023) use teacher averaging, time-aware memory, or sharpness-aware optimization to limit error accumulation and forgetting. Recent methods further address realistic constraints: TRIBE (Su et al., 2024) combines balanced normalization with anchored tri-net self-training; DeYO (Lee et al., 2024) supplements entropy with PLPD-based sample selection and weighting; SURGEON (Ma et al., 2025) controls adaptation memory through dynamic activation sparsity; and TACT (Liu et al., 2025) trims estimated non-causal representation directions. These methods strengthen general or unimodal TTA, but do not explicitly coordinate reliability among audio, video, and fused predictions.

Multimodal TTA (MMTTA) aims to exploit complementary cues while accounting for asymmetric shifts across modalities. Earlier work studies continual multimodal adaptation in 3D perception and two-level adaptation (Cao et al., 2023; Lei & Pernkopf, 2024). READ (Yang et al., 2024) addresses reliability bias through self-adaptive attention-based fusion and a confidence-aware objective, whereas SuMi (Guo & Jin, 2025) targets complex multimodal noise through IQR-smoothed sample identification, unimodal assistance, and mutual-information sharing. More recent approaches selectively route updates to shifted modalities (Chen et al., 2025), partition reliable and unreliable samples to reduce prediction bias (Wang et al., 2025), progressively realign unimodal and cross-modal representations (Li & Feng, 2026), decouple stability from plasticity (He et al., 2026), or calibrate category-conditional distributions under modality asymmetry (Xu et al., 2026). These works demonstrate rapidly growing interest in MMTTA, but are primarily standalone adaptation algorithms. In contrast, MUTE composes with diverse base TTA methods and introduces confidence-routed mutual teaching among audio, video, and fused heads for lifelong corruption streams.

3 Methodology

We first provide a brief overview of MUTE before detailing each component. Given temporally aligned audio-video inputs, three heads (audio, video, multimodal) produce logits and confidence scores. A dynamic routing module picks the most reliable teacher and activates up to two of six possible directed paths according to a confidence gate $\tau_{\min}$ and an advantage margin δ. The selected teacher distribution is tempered by a confidence-conditioned temperature and distilled via a scheduled mutual-teaching loss that plugs into any base TTA objective. The full workflow is illustrated in Fig. 2.

3.1 Problem Setup

We consider paired inputs (x^a, x^v) for audio and video modalities and a multimodal model with heads producing logits $z^a, z^v, z^{av} \in \mathbb{R}^K$. We denote heads by superscripts $m \in \{a, v, av\}$, where a = audio head, v = video head, and av = fused audio-video (multimodal) head. Let

$$p^m = \mathrm{softmax}(z^m) \in \Delta^{K-1}, \quad m \in \{a, v, av\}. \tag{1}$$

Here, Δ^{K-1} denotes the $(K-1)$-dimensional probability simplex over K classes. At test time, inputs arrive sequentially under a lifelong corruption stream $c_1, c_2, \ldots$. The goal is to update model

Table 1: KS50 lifelong video-corruption accuracy at severity 5, reported as mean$_{\pm\text{std}}$ (%). Full per-corruption results are in Appendix C.1. Bold marks improvements over Base.

Method	Base	+MUTE
EATA	$63.3_{\pm0.3}$	**$63.9_{\pm0.1}$**
Tent	$7.1_{\pm0.2}$	**$10.3_{\pm1.2}$**
READ	$43.0_{\pm3.2}$	**$64.1_{\pm0.1}$**
SuMi	$59.6_{\pm0.2}$	**$61.7_{\pm0.2}$**
TRIBE	$61.1_{\pm0.1}$	**$62.0_{\pm0.2}$**
DeYO	$53.6_{\pm6.9}$	**$59.6_{\pm0.3}$**
SURGEON	$10.7_{\pm0.4}$	**$53.6_{\pm0.9}$**
TACT	$53.9_{\pm0.7}$	**$58.0_{\pm0.1}$**
CoTTA	$60.0_{\pm0.0}$	**$60.5_{\pm0.1}$**
MEMO	$51.0_{\pm0.0}$	**$51.0_{\pm0.0}$**
SAR	$60.3_{\pm0.0}$	**$61.1_{\pm0.0}$**
RoTTA	$59.7_{\pm0.2}$	**$60.4_{\pm0.1}$**

Table 2: KS50 lifelong audio-corruption accuracy at severity 5, reported as mean$_{\pm\text{std}}$ (%). Full per-corruption results are in Appendix C.1. Bold marks improvements over Base.

Method	Base	+MUTE
EATA	$71.4_{\pm0.1}$	**$71.5_{\pm0.1}$**
Tent	$71.3_{\pm0.1}$	**$71.7_{\pm0.1}$**
READ	$72.2_{\pm0.1}$	$71.8_{\pm0.1}$
SuMi	$67.4_{\pm0.2}$	**$68.9_{\pm0.2}$**
TRIBE	$71.6_{\pm0.2}$	**$71.9_{\pm0.1}$**
DeYO	$70.8_{\pm0.1}$	**$71.2_{\pm0.1}$**
SURGEON	$70.9_{\pm0.1}$	**$71.3_{\pm0.0}$**
TACT	$71.6_{\pm0.0}$	**$71.9_{\pm0.1}$**
CoTTA	$69.1_{\pm0.0}$	**$69.2_{\pm0.0}$**
MEMO	$70.9_{\pm0.0}$	**$71.3_{\pm0.0}$**
SAR	$70.6_{\pm0.0}$	**$71.1_{\pm0.0}$**
RoTTA	$71.0_{\pm0.2}$	**$71.2_{\pm0.0}$**

Table 3: VGG-Sound lifelong video-corruption accuracy at severity 5, reported as mean$_{\pm\text{std}}$ (%). Full per-corruption results are in Appendix C.1. Bold marks improvements over Base.

Method	Base	+MUTE
EATA	$55.9_{\pm0.0}$	**$56.5_{\pm0.0}$**
Tent	$55.2_{\pm0.0}$	**$56.3_{\pm0.0}$**
READ	$55.8_{\pm0.0}$	**$56.4_{\pm0.0}$**
SuMi	$53.9_{\pm0.0}$	**$55.5_{\pm0.0}$**
TRIBE	$55.7_{\pm0.0}$	**$55.8_{\pm0.0}$**
DeYO	$55.3_{\pm0.0}$	**$55.6_{\pm0.0}$**
SURGEON	$54.4_{\pm0.0}$	**$54.5_{\pm0.0}$**
TACT	$54.7_{\pm0.0}$	**$55.2_{\pm0.0}$**
CoTTA	$55.4_{\pm0.0}$	**$55.7_{\pm0.0}$**
MEMO	$55.0_{\pm0.0}$	**$55.4_{\pm0.0}$**
SAR	$55.7_{\pm0.0}$	**$56.3_{\pm0.0}$**
RoTTA	$55.7_{\pm0.0}$	**$55.8_{\pm0.0}$**

Table 4: VGG-Sound lifelong audio-corruption accuracy at severity 5, reported as mean$_{\pm\text{std}}$ (%). Full per-corruption results are in Appendix C.1. Bold marks improvements over Base.

Method	Base	+MUTE
EATA	$30.0_{\pm0.1}$	**$30.1_{\pm0.0}$**
Tent	$16.6_{\pm0.0}$	**$27.5_{\pm0.2}$**
READ	$31.4_{\pm0.0}$	**$32.5_{\pm0.0}$**
SuMi	$29.5_{\pm0.0}$	**$30.0_{\pm0.0}$**
TRIBE	$30.9_{\pm0.2}$	**$31.2_{\pm0.2}$**
DeYO	$23.9_{\pm0.1}$	**$24.7_{\pm0.1}$**
SURGEON	$11.6_{\pm0.1}$	**$11.7_{\pm0.1}$**
TACT	$11.1_{\pm0.0}$	**$26.5_{\pm0.0}$**
CoTTA	$27.5_{\pm0.0}$	**$28.2_{\pm0.0}$**
MEMO	$17.2_{\pm0.0}$	**$27.3_{\pm0.0}$**
SAR	$26.4_{\pm0.0}$	**$27.2_{\pm0.0}$**
RoTTA	$30.6_{\pm0.0}$	**$31.7_{\pm0.0}$**

parameters online to improve accuracy on the current corruption while avoiding cumulative drift and forgetting of robustness acquired in earlier stages. MUTE augments any base TTA objective with a multimodal mutual-teaching strategy.

3.2 Multimodal Mutual Teaching Head

For each head $m \in \{a, v, av\}$, we use a standard bounded confidence proxy (Lee et al., 2022; Katsumata et al., 2024):

$$\gamma(z^m) = 1 - \frac{H(p^m)}{\ln(K)}, \tag{2}$$

where $H(p^m) = -\sum_{k=1}^{K} p_k^m \ln(p_k^m)$ denotes the Shannon entropy. Since it is a monotone transformation of predictive entropy, selecting the highest-confidence head is equivalent to selecting the lowest-entropy head.

Let $\mathcal{M} = \{a, v, av\}$, $\tau_{\min} > 0$ be a confidence gate, and $\delta > 0$ the modality-advantage margin. For sample i, define $\gamma_i^m := \gamma(z_i^m)$ and select a single teacher $m_i^\star := \arg\max_{m \in \mathcal{M}} \gamma_i^m$, with ties broken deterministically. Routing abstains if the selected teacher's confidence is below $\tau_{\min}$; otherwise, it activates $m_i^\star \to m_s$ only when the confidence advantage over student m_s is strictly greater than δ.

Algorithm 1 MUTE: Multimodal Uncertainty-Tempered Mutual Teaching at Test Time

Input stream: paired samples (x^a, x^v).
Model heads: produce logits z^a, z^v, z^{av}.
Hyperparameters: $\tau_{\min}$, δ, $T_{\min}$, $T_{\max}$, $\lambda_{\max}$, α, β.
for each minibatch $\{(x_i^a, x_i^v)\}_{i=1}^B$ at iteration t **do**
  Forward: Obtain z^a, z^v, z^{av} and probabilities p^a, p^v, p^{av}.
  Confidence: for each i, compute $\gamma_i^a, \gamma_i^v, \gamma_i^{av}$ using Eq. (2).
  for each sample i **do**
    Teacher: $m_i^\star \leftarrow \arg\max_{m\in\{a,v,av\}} \gamma_i^m$.
    if $\gamma_i^{m_i^\star} \geq \tau_{\min}$ **then**
      for each potential student $m_s \in \{a, v, av\} \setminus \{m_i^\star\}$ **do**
        if $\gamma_i^{m_i^\star} - \gamma_i^{m_s} > \delta$ **then**
          Add i to routing set $\mathcal{I}_{m_i^\star \to m_s}$.
        end if
      end for
    end if
  end for
  Temperature: for each routed sample $i \in \mathcal{I}_{m_t\to m_s}$, set $c = \gamma_i^{m_t}$ and compute $T(c)$ using Eq. (4).
  Teacher target: form tempered $q_i^{m_t}$ using Eq. (5).
  Mutual-teaching loss using Eq. (6):
  $\mathcal{L}_{\text{MUTE}} = \sum_{(m_t,m_s)} \mathbb{E}_{i\in\mathcal{I}_{m_t\to m_s}} D_{\text{KL}}(p_i^{m_s} \,\|\, q_i^{m_t})$.
  Schedule: compute $\lambda_{\text{ct}}(t)$ using Eq. (7).
  Base loss: compute the base TTA loss $\mathcal{L}_{\text{base}}$.
  Total objective: set $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{ct}}(t)\,\mathcal{L}_{\text{MUTE}}$.
  Update: update model parameters by one optimizer step on $\mathcal{L}_{\text{total}}$.
end for

Table 5: Ablation studies on the core components of MUTE on the KS50 video lifelong corruption stream with READ. The full MUTE configuration achieves the highest average accuracy.

Variant	Avg. (%)	Δ (%)
Full MUTE	**64.1**	—
No unimodal cross-routing ($a \leftrightarrow v$)	63.9	-0.2
No multimodal-related routes ($av \leftrightarrow a, v$)	58.2	-5.9
No Adaptive Weight scheduling (λ_{ct})	63.7	-0.4
No adaptive temperature scaling $T(c)$	62.1	-2.0

For $m_t, m_s \in \mathcal{M}$ with $m_t \neq m_s$, the routing index set is

$$\mathcal{I}_{m_t\to m_s} = \{i : m_t = m_i^\star,\ \gamma_i^{m_t} \geq \tau_{\min},\ \gamma_i^{m_t} - \gamma_i^{m_s} > \delta\}. \tag{3}$$

$\mathcal{I}_{m_t\to m_s}$ denotes the set of sample indices where head m_t teaches head m_s. The mechanism operates over six possible directed paths, (av, a), (av, v), (a, v), (v, a), (a, av), and (v, av); because each sample has one selected teacher, at most two paths are active for that sample.

3.3 Confidence-Conditioned Tempering and Adaptive Weighting

For a teacher with confidence $c \in [0, 1]$, we temper the teacher distribution by

$$T(c) = T_{\min} + (T_{\max} - T_{\min})(1 - c) \tag{4}$$

Let q^m denote the tempered teacher distribution:

$$q^m = \text{softmax}\left(z^m / T(c)\right) \tag{5}$$

The mutual-teaching loss aggregates KL divergences over all routing pairs (m_t, m_s):

$$\mathcal{L}_{\text{MUTE}} = \sum_{(m_t,m_s)} \mathbb{E}_{i\in\mathcal{I}_{m_t\to m_s}} D_{\text{KL}}\big(p_i^{m_s} \,\|\, q_i^{m_t}\big) \tag{6}$$

High-confidence teachers (small T) impose sharper targets, while low-confidence teachers (large T) yield softer supervision, reducing negative transfer when the teacher is uncertain. Direction-adaptive routing allows the most reliable head to supervise the other heads, improving adaptation robustness. We further modulate the mutual-teaching weight λ_{ct} with a three-phase schedule over iteration $t \in [0, \text{iters})$: Warmup (linear to $\lambda_{\max}$), stable (constant), and decay (to $0.5\,\lambda_{\max}$):

$$\lambda_{\text{ct}}(t) = \begin{cases} \lambda_{\max} \cdot \dfrac{t}{\alpha}, & t < \alpha\,, \\ \lambda_{\max}, & \alpha \le t < \beta\,, \\ \lambda_{\max} \cdot \Big(1 - 0.5\,\dfrac{t-\beta}{(1-\beta)}\Big), & \text{otherwise} \end{cases} \tag{7}$$

with hyperparameters $\alpha \in (0,1)$ and $\beta \in (\alpha, 1)$, where α marks the end of the warmup fraction of iterations and β marks the end of the stable plateau (both measured as fractions of total iterations iters). For a base TTA loss $\mathcal{L}_{\text{base}}$ (e.g., entropy minimization), the total objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{ct}}(t)\,\mathcal{L}_{\text{MUTE}} \tag{8}$$

Algorithm 1 summarizes the complete procedure.

3.4 Theoretical Analysis

In this subsection, we analyze the two core pillars of MUTE: the modality-advantage routing mechanism and confidence-conditioned tempering. We derive a margin-dependent upper bound on the student's per-step risk and show that temperature scaling reduces the local sensitivity of teacher targets to logit perturbations.

Preliminaries and Assumptions

Let $\mathcal{D}$ be the latent ground-truth distribution over K classes. We define the risk (expected cross-entropy loss) of a prediction p as $\mathcal{R}(p) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[-\log p_y(x)]$. We assume the loss is bounded by a constant $L < \infty$, enforced in practice through logit clipping. Let $p_{\text{curr}}^{m_s}$ denote the student's probability distribution at the current iteration before the update.

We enforce hyperparameters $\tau_{\min} \in (0,1)$ and $0 < T_{\min} \le T(c) \le T_{\max}$. The advantage margin δ is chosen such that $\delta \ln K$ represents a non-trivial entropy reduction (i.e., $\delta \ln K < H(p_{\text{curr}}^{m_s})$) to ensure directional correctness.

Assumption 1 (Lipschitz Risk-Confidence Calibration). *We assume the confidence score $\gamma(\cdot)$ acts as a proxy for correctness. Specifically, there exists a calibration constant $\kappa > 0$ such that for any two heads m_t, m_s, if $\gamma(z^{m_t}) \ge \gamma(z^{m_s}) + \delta$, then:*

$$\mathcal{R}(p^{m_t}) \le \mathcal{R}(p_{\text{curr}}^{m_s}) - \kappa\delta. \tag{9}$$

This assumption formalizes the reliability of the confidence score in Eq. (2). Empirically, κ could be calibrated (e.g., via reliability diagrams), but here we treat it as a theoretical constant linking confidence gap to risk gap.

Risk Bound via Mutual Teaching

We will analyze why our routing strategy (Eq. (3)) leads to effective adaptation. Consider a routed pair (m_t, m_s) selected by MUTE.

Theorem 1 (Margin-Dependent Per-Step Risk Bound)**.** *Let $\mathcal{Z}$ denote the bounded logit domain induced by clipping, and define the maximum risk increase from temperature scaling as*

$$\Delta_{\textit{temp}} = \sup_{\substack{z\in\mathcal{Z} \\ T\in[T_{\min},T_{\max}]}} \big[\mathcal{R}(\sigma(z/T)) - \mathcal{R}(\sigma(z))\big]. \tag{10}$$

Under Assumption 1, minimizing the mutual-teaching loss tightens the following upper bound on the student's risk:

$$\mathcal{R}(p^{m_s}) \leq \underbrace{\mathcal{R}(p^{m_s}_{curr}) - \kappa\delta}_{\textit{Margin Term}} + \Delta_{\textit{temp}} + \underbrace{L\sqrt{\frac{1}{2}D_{\mathrm{KL}}(p^{m_s}\|q^{m_t})}}_{\textit{Optimization Gap}} \tag{11}$$

The proof is provided in Appendix B.1.

In Eq. (11), the confidence advantage contributes the improvement term $-\kappa\delta$, while the tempering shift Δ_{temp} and residual KL term offset this reduction. Under Assumption 1, the bound certifies a strict per-step risk decrease when $\kappa\delta > \Delta_{\text{temp}} + L\sqrt{\frac{1}{2}D_{\mathrm{KL}}(p^{m_s}\|q^{m_t})}$; otherwise, it identifies the terms that must be controlled for the routed update to improve risk.

Teacher-Target Sensitivity via Confidence-Conditioned Tempering

Finally, we analyze how the adaptive temperature $T(c)$ defined in Eq. (4) controls the sensitivity of a teacher target to perturbations in its logits.

Proposition 1 (Temperature-Dependent Target Sensitivity)**.** *Let $J(z) = \frac{\partial q^{m_t}}{\partial z^{m_t}}$ be the Jacobian of the tempered softmax. The spectral norm is bounded by $\|J(z)\|_2 \leq \frac{1}{T(c)}(1 - \frac{1}{K})$.*

The proof is provided in Appendix B.2.

Proposition 1 shows that increasing T for low-confidence teachers reduces the upper bound on the local sensitivity of the teacher target at a rate proportional to $1/T$. Consequently, uncertain teachers yield less perturbation-sensitive targets; in the high-temperature limit, the target approaches the uniform distribution and provides less directional supervision.

4 Experiments

Following READ (Yang et al., 2024), we evaluate Kinetics50-C (KS50) and VGGSound-C, derived from Kinetics (Kay et al., 2017) and VGG-Sound (Chen et al., 2020). We compare twelve base TTA methods with and without MUTE augmentation. All experiments follow the lifelong protocol where models adapt sequentially to a stream of heterogeneous corruptions without labels or replay buffers.

4.1 Experimental Setup

Datasets. KS50 (50 classes) and VGG-Sound (309 classes) (Yang et al., 2024) provide complementary scale and category coverage. We evaluate on audio corruptions (gaussian noise, traffic, crowd, rain, thunder, wind) and video corruptions (15 types from ImageNet-C (Hendrycks & Dietterich, 2019)).

Baselines. We compare EATA (Niu et al., 2022), Tent (Wang et al., 2021), READ (Yang et al., 2024), SuMi (Guo & Jin, 2025), TRIBE (Su et al., 2024), DeYO (Lee et al., 2024), SURGEON (Ma et al., 2025), TACT (Liu et al., 2025), CoTTA (Wang et al., 2022), MEMO (Zhang et al., 2022), SAR (Niu et al., 2023), and RoTTA (Yuan et al., 2023) as base methods. Appendix A summarizes their core mechanisms.

Implementation. Using PyTorch, experiments run a pre-trained CAV-MAE backbone (Gong et al., 2023) on one NVIDIA A100-80GB GPU and average five seeds. Detailed efficiency measurements are reported in Appendix C.2.

Evaluation Protocol. Models adapt sequentially to each corruption in the stream. We report average accuracy across all corruptions in the lifelong setting. For compact presentation, all four main tables report stream-average accuracy as $\text{mean}_{\pm\text{std}}$ for Base and +MUTE; complete per-corruption $\text{mean}_{\pm\text{std}}$ results are provided in Appendix C.1. Following Guo & Jin (2025), we use batch size 16 and learning rate 1×10^{-4} for KS50, and batch size 64 and learning rate 1×10^{-5} for VGG-Sound. To simulate valid-set-free TTA, we use a single fixed configuration across all datasets and base methods without target-domain validation: $\tau_{\min} = 0.3$, $\delta = 0.1$, $T_{\min} = 0.4$, $T_{\max} = 0.9$,

$\lambda_{\max} = 0.1$, $\alpha = 0.2$, and $\beta = 0.6$. Each lifelong run uses one initialization and carries both model and optimizer states across corruptions without reset, following the corruption order displayed in Appendix C.1.

4.2 RESULTS

Overall Trends. Across twelve base methods, two datasets, and both audio and video corruption streams, MUTE generally improves average accuracy (Tables 1 to 4). The improvements span both datasets and corruption modalities, indicating that the same mutual-teaching module can complement adapters with substantially different update rules.

Uneven Gains across Base Adapters. The magnitude of the improvement varies with the behavior of the underlying adapter. The largest gains occur in streams with pronounced sequential degradation, including SURGEON and READ on KS50-V (+42.9 and +21.1 points) and several unstable adapters on VGG-A (+10.1 to +15.4 points). When the base method remains stable, the gains are typically smaller; for example, improvements on VGG-V range from +0.1 to +1.6 points. READ on KS50-A is the only negative average case, changing from 72.2% to 71.8%. Together, these results are consistent with MUTE acting primarily as a stabilizer against accumulated drift, with less marginal benefit in already-stable regimes.

Appendix C.1 provides the complete per-corruption $\text{mean}_{\pm\text{std}}$ results. Appendix C.2 presents the efficiency analysis, covering runtime, throughput, peak memory, and extra computation; joint-corruption, missing-modality, and corruption-severity results are discussed in Appendix C.3.

4.3 ABLATION STUDIES

In this section, we conduct extensive ablation studies to isolate and validate the effectiveness of the core components proposed in our MUTE framework. All ablation experiments are performed on the KS50 Video sequential corruption dataset, using READ as the base TTA learner.

CONTRIBUTION OF ALGORITHMIC COMPONENTS.

To assess the specific impact of the mutual-teaching pathways and the uncertainty-tempering mechanisms, we evaluate four degraded variants of our framework. The results are summarized in Table 5. First, we analyze the routing mechanisms. In this READ/KS50-V setting, removing direct unimodal cross-routing ($a \to v$ and $v \to a$) causes only a 0.2-point drop, whereas removing routes involving the fused head ($av \leftrightarrow a$ and $av \leftrightarrow v$) reduces accuracy by 5.9 points. This indicates that the gains in this setting are driven primarily by interactions involving the fused head. Next, we evaluate our stabilization mechanisms: (3) Disabling the Adaptive Weight scheduling (i.e., applying a constant mutual-teaching weight throughout adaptation) reduces accuracy by 0.4%, confirming that a warmup-stable-decay schedule helps prevent early-stage overfitting to noisy pseudolabels. (4) Disabling the confidence-conditioned temperature adaptation $T(c)$ (using a fixed temperature instead) leads to a 2.0% drop. This result is consistent with Proposition 1: assigning higher temperatures to low-confidence teachers reduces the sensitivity of their targets to logit perturbations.

Additional hyperparameter sensitivity and stress-test results are moved to Appendix C.3. Across the tested ranges, moderate teaching weights and confidence thresholds perform best, while the advantage margin is stable over a broad interval; the full sweep and its original discussion are retained there.

5 CONCLUSION

We introduce MUTE as a buffer-free, plug-and-play mutual-teaching framework for lifelong multimodal TTA. It uses confidence-based routing with a modality-advantage criterion to transfer supervision among audio, video, and fused predictions. We derive a margin-dependent per-step risk bound and a temperature-dependent teacher-target sensitivity bound. Across two datasets and both audio and video corruption streams, MUTE generally improves average accuracy over a broad set of strong TTA baselines, with gains of up to 42.9 percentage points for severely degraded base methods and up to 2.1 points for already stable ones. Future work will study online hyperparameter adaptation and extend MUTE to additional modalities.

AI USE STATEMENT

In this work, we used generative AI tools for grammar and sentence polishing and for literature survey. We have reviewed all AI-assisted content and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendix A summarizes the twelve base TTA methods, and Appendix B gives the proofs of Theorem 1 and Proposition 1. Appendix C.1 reports the complete per-corruption results, Appendix C.2 the computational overhead, and Appendix C.3 the hyperparameter sweeps and stress tests. The anonymous repository linked in the abstract contains the MUTE implementation, the twelve base adapters with their MUTE integration points, the CAV-MAE model interface, data preprocessing code, the four protocol configurations, a lifelong evaluation harness, and behavioral tests; its README notes that the public datasets and pretrained checkpoints are not redistributed and must be obtained from their original providers.

REFERENCES

Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, and Lihua Xie. Multi-modal continual test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18809–18819, 2023.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

Mingcai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multimodal data. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 9711–9727, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 2 edition, 2006. doi: 10.1002/047174882X.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *CoRR*, abs/1704.00805, 2017. URL https://arxiv.org/abs/1704.00805.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QPtMRyk5rb.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *International Conference on Learning Representations*, 2025.

Yongbo He, Zirun Guo, and Tao Jin. Decoupling stability and plasticity for multi-modal test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15020–15029, 2026.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Kai Katsumata, Duc Minh Vo, Tatsuya Harada, and Hideki Nakayama. Soft curriculum for learning conditional GANs with noisy-labeled and uncurated unlabeled data. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5323–5332, 2024.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. URL `https://arxiv.org/abs/1705.06950`.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 12365–12377, 2022.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024.

Jixiang Lei and Franz Pernkopf. Two-level test-time adaptation in multimodal learning. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Jiacheng Li and Songhe Feng. Bridging modalities via progressive re-alignment for multimodal test-time adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22931–22939, 2026.

Yingnan Liu, Rui Qiao, Mong-Li Lee, and Wynne Hsu. Test-time adaptation by causal trimming. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ke Ma, Jiaqi Tang, Bin Guo, Fan Dang, Sicong Liu, Zhui Zhu, Lei Wu, Cheng Fang, Ying-Cong Chen, Zhiwen Yu, and Yunhao Liu. SURGEON: Memory-adaptive fully test-time adaptation via dynamic activation sparsity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30514–30523, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Yongyi Su, Xun Xu, and Kui Jia. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13): 15126–15135, 2024.

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.

Guowei Wang, Fan Lyu, and Changxing Ding. Partition-then-adapt: Combating prediction bias for reliable multi-modal test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

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 *The twelfth international conference on learning representations*, 2024.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

A Baseline Methods

We summarize the twelve TTA methods used as base adapters in our evaluation.

EATA. EATA (Niu et al., 2022) selects reliable, non-redundant test samples for entropy minimization and applies Fisher-based regularization to preserve important source parameters. **Tent.** Tent (Wang et al., 2021) minimizes prediction entropy by updating normalization affine parameters online. **READ.** READ (Yang et al., 2024) handles multimodal reliability bias through self-adaptive attention-based fusion and a confidence-aware objective that emphasizes reliable predictions. **SuMi.** SuMi (Guo & Jin, 2025) combines IQR-smoothed sample identification and unimodal assistance with mutual-information sharing across modalities.

TRIBE. TRIBE (Su et al., 2024) uses tri-net self-training, balanced normalization, and an anchored loss for non-i.i.d., class-imbalanced, continual streams. **DeYO.** DeYO (Lee et al., 2024) combines entropy and pseudo-label probability difference (PLPD) for sample selection and weighting, favoring predictions supported by object-shape cues. **SURGEON.** SURGEON (Ma et al., 2025) reduces fully test-time adaptation memory by pruning activations with layer-wise ratios derived from gradient importance and activation memory. **TACT.** TACT (Liu et al., 2025) uses augmentations and principal component analysis to estimate high-variance non-causal directions and removes their projections from the representation.

CoTTA. CoTTA (Wang et al., 2022) uses weight- and augmentation-averaged teacher predictions together with stochastic restoration of source weights. **MEMO.** MEMO (Zhang et al., 2022) adapts to each test input by minimizing marginal entropy over augmented views. **SAR.** SAR (Niu et al., 2023) filters unreliable, high-gradient samples and applies sharpness-aware entropy minimization to avoid collapse. **RoTTA.** RoTTA (Yuan et al., 2023) combines robust batch normalization, a timeliness- and uncertainty-aware category-balanced memory, and teacher-student updates.

B Theoretical Proofs

B.1 Proof of Theorem 1

We decompose the risk bound in three steps:

1. **TV-Lipschitz Continuity:** By the Lipschitz continuity of the cross-entropy risk with respect to the Total Variation (TV) distance (Gibbs & Su, 2002),
$$|\mathcal{R}(p^{m_s}) - \mathcal{R}(q^{m_t})| \leq L \cdot \delta_{TV}(p^{m_s}, q^{m_t}). \tag{12}$$
2. **Pinsker's Inequality:** We bound the TV distance using the KL divergence (Cover & Thomas, 2006):
$$\delta_{TV}(p^{m_s}, q^{m_t}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(p^{m_s} \| q^{m_t})}. \tag{13}$$
3. **Calibration & Tempering:** The teacher target's risk $\mathcal{R}(q^{m_t})$ is bounded by its base risk plus the tempering shift: $\mathcal{R}(q^{m_t}) \leq \mathcal{R}(p^{m_t}) + \Delta_{\text{temp}}$. For a routed pair satisfying the advantage margin, Assumption 1 gives $\mathcal{R}(p^{m_t}) \leq \mathcal{R}(p^{m_s}_{\text{curr}}) - \kappa\delta$.

Table 6: Complete KS50 sequential audio-corruption results at severity 5. Accuracy is reported as mean$_{\pm\text{std}}$ (%) for all twelve baselines and their +MUTE variants. Bold values indicate an improvement over the corresponding base method.

Method	Avg.	Gauss.	Traffic	Crowd	Rain	Thunder	Wind
EATA	$71.4_{\pm0.1}$	$74.2_{\pm0.1}$	$68.0_{\pm0.3}$	$71.2_{\pm0.2}$	$70.9_{\pm0.2}$	$73.6_{\pm0.2}$	$70.2_{\pm0.3}$
+MUTE	$\mathbf{71.5}_{\pm\mathbf{0.1}}$	$74.2_{\pm0.1}$	$\mathbf{68.1}_{\pm\mathbf{0.2}}$	$\mathbf{71.3}_{\pm\mathbf{0.3}}$	$\mathbf{71.0}_{\pm\mathbf{0.3}}$	$\mathbf{73.7}_{\pm\mathbf{0.3}}$	$\mathbf{70.7}_{\pm\mathbf{0.4}}$
Tent	$71.3_{\pm0.1}$	$73.9_{\pm0.2}$	$69.7_{\pm0.4}$	$71.1_{\pm0.2}$	$69.6_{\pm0.3}$	$73.5_{\pm0.1}$	$69.9_{\pm0.4}$
+MUTE	$\mathbf{71.7}_{\pm\mathbf{0.1}}$	$\mathbf{74.0}_{\pm\mathbf{0.1}}$	$69.2_{\pm0.4}$	$\mathbf{71.3}_{\pm\mathbf{0.2}}$	$\mathbf{71.3}_{\pm\mathbf{0.2}}$	$\mathbf{73.6}_{\pm\mathbf{0.2}}$	$\mathbf{70.9}_{\pm\mathbf{0.3}}$
READ	$72.2_{\pm0.1}$	$74.3_{\pm0.1}$	$69.6_{\pm0.4}$	$72.0_{\pm0.1}$	$71.8_{\pm0.3}$	$73.9_{\pm0.1}$	$71.4_{\pm0.5}$
+MUTE	$71.8_{\pm0.1}$	$74.0_{\pm0.1}$	$68.6_{\pm0.3}$	$71.0_{\pm0.2}$	$\mathbf{72.3}_{\pm\mathbf{0.1}}$	$72.7_{\pm0.2}$	$\mathbf{71.9}_{\pm\mathbf{0.3}}$
SuMi	$67.4_{\pm0.2}$	$73.8_{\pm0.1}$	$66.4_{\pm0.4}$	$67.4_{\pm0.3}$	$68.2_{\pm0.2}$	$65.6_{\pm0.4}$	$63.3_{\pm0.4}$
+MUTE	$\mathbf{68.9}_{\pm\mathbf{0.2}}$	$73.8_{\pm0.1}$	$\mathbf{66.7}_{\pm\mathbf{0.5}}$	$\mathbf{67.7}_{\pm\mathbf{0.3}}$	$\mathbf{69.5}_{\pm\mathbf{0.2}}$	$\mathbf{69.1}_{\pm\mathbf{0.3}}$	$\mathbf{66.6}_{\pm\mathbf{0.4}}$
TRIBE	$71.6_{\pm0.2}$	$74.2_{\pm0.2}$	$70.1_{\pm0.1}$	$71.4_{\pm0.4}$	$70.8_{\pm0.2}$	$72.7_{\pm0.3}$	$70.7_{\pm0.4}$
+MUTE	$\mathbf{71.9}_{\pm\mathbf{0.1}}$	$74.2_{\pm0.3}$	$69.6_{\pm0.5}$	$\mathbf{71.8}_{\pm\mathbf{0.2}}$	$\mathbf{71.5}_{\pm\mathbf{0.3}}$	$\mathbf{73.2}_{\pm\mathbf{0.2}}$	$\mathbf{71.0}_{\pm\mathbf{0.3}}$
DeYO	$70.8_{\pm0.1}$	$73.6_{\pm0.0}$	$68.5_{\pm0.4}$	$70.0_{\pm0.3}$	$69.4_{\pm0.2}$	$73.2_{\pm0.1}$	$69.9_{\pm0.0}$
+MUTE	$\mathbf{71.2}_{\pm\mathbf{0.1}}$	$73.6_{\pm0.1}$	$\mathbf{68.7}_{\pm\mathbf{0.2}}$	$\mathbf{70.0}_{\pm\mathbf{0.3}}$	$\mathbf{70.9}_{\pm\mathbf{0.1}}$	$\mathbf{73.3}_{\pm\mathbf{0.2}}$	$\mathbf{70.8}_{\pm\mathbf{0.4}}$
SURGEON	$70.9_{\pm0.1}$	$73.8_{\pm0.1}$	$69.2_{\pm0.3}$	$70.5_{\pm0.1}$	$69.5_{\pm0.2}$	$73.1_{\pm0.3}$	$69.4_{\pm0.2}$
+MUTE	$\mathbf{71.3}_{\pm\mathbf{0.0}}$	$\mathbf{74.0}_{\pm\mathbf{0.0}}$	$\mathbf{69.3}_{\pm\mathbf{0.2}}$	$\mathbf{70.7}_{\pm\mathbf{0.1}}$	$\mathbf{70.0}_{\pm\mathbf{0.3}}$	$\mathbf{73.4}_{\pm\mathbf{0.3}}$	$\mathbf{70.4}_{\pm\mathbf{0.1}}$
TACT	$71.6_{\pm0.0}$	$73.6_{\pm0.2}$	$70.2_{\pm0.2}$	$71.7_{\pm0.2}$	$70.0_{\pm0.1}$	$73.3_{\pm0.4}$	$70.8_{\pm0.2}$
+MUTE	$\mathbf{71.9}_{\pm\mathbf{0.1}}$	$\mathbf{73.7}_{\pm\mathbf{0.1}}$	$69.9_{\pm0.1}$	$71.4_{\pm0.3}$	$\mathbf{71.2}_{\pm\mathbf{0.2}}$	$\mathbf{74.0}_{\pm\mathbf{0.0}}$	$\mathbf{71.3}_{\pm\mathbf{0.4}}$
CoTTA	$69.1_{\pm0.0}$	$73.3_{\pm0.0}$	$65.3_{\pm0.0}$	$67.9_{\pm0.0}$	$70.1_{\pm0.1}$	$67.9_{\pm0.2}$	$70.4_{\pm0.1}$
+MUTE	$\mathbf{69.2}_{\pm\mathbf{0.0}}$	$\mathbf{73.3}_{\pm\mathbf{0.1}}$	$\mathbf{65.6}_{\pm\mathbf{0.2}}$	$67.1_{\pm0.2}$	$\mathbf{70.2}_{\pm\mathbf{0.1}}$	$\mathbf{68.3}_{\pm\mathbf{0.2}}$	$\mathbf{70.8}_{\pm\mathbf{0.0}}$
MEMO	$70.9_{\pm0.0}$	$73.8_{\pm0.1}$	$69.1_{\pm0.0}$	$70.4_{\pm0.0}$	$69.8_{\pm0.0}$	$72.8_{\pm0.0}$	$69.5_{\pm0.0}$
+MUTE	$\mathbf{71.3}_{\pm\mathbf{0.0}}$	$73.5_{\pm0.1}$	$68.9_{\pm0.1}$	$\mathbf{71.2}_{\pm\mathbf{0.0}}$	$\mathbf{70.1}_{\pm\mathbf{0.0}}$	$\mathbf{73.3}_{\pm\mathbf{0.0}}$	$\mathbf{70.6}_{\pm\mathbf{0.0}}$
SAR	$70.6_{\pm0.0}$	$73.2_{\pm0.0}$	$67.4_{\pm0.2}$	$69.9_{\pm0.1}$	$71.1_{\pm0.1}$	$72.0_{\pm0.0}$	$70.1_{\pm0.0}$
+MUTE	$\mathbf{71.1}_{\pm\mathbf{0.0}}$	$\mathbf{73.6}_{\pm\mathbf{0.1}}$	$\mathbf{68.7}_{\pm\mathbf{0.2}}$	$\mathbf{71.0}_{\pm\mathbf{0.1}}$	$70.7_{\pm0.2}$	$\mathbf{73.0}_{\pm\mathbf{0.2}}$	$69.7_{\pm0.0}$
RoTTA	$71.0_{\pm0.2}$	$73.6_{\pm0.2}$	$67.6_{\pm0.2}$	$70.4_{\pm0.3}$	$70.5_{\pm0.0}$	$72.5_{\pm0.3}$	$71.7_{\pm0.2}$
+MUTE	$\mathbf{71.2}_{\pm\mathbf{0.0}}$	$73.5_{\pm0.0}$	$\mathbf{68.5}_{\pm\mathbf{0.2}}$	$\mathbf{71.0}_{\pm\mathbf{0.2}}$	$\mathbf{70.7}_{\pm\mathbf{0.2}}$	$72.1_{\pm0.3}$	$71.5_{\pm0.2}$

Combining these inequalities yields Eq. (11), completing the proof.

B.2 Proof of Proposition 1

The Jacobian of softmax with temperature T is (Gao & Pavel, 2017)

$$J = \frac{1}{T}\big(\mathrm{diag}(q) - qq^\top\big). \tag{14}$$

Its spectral norm is bounded by its largest eigenvalue,

$$\|J\|_2 \le \frac{1}{T}\Big(1 - \frac{1}{K}\Big). \tag{15}$$

As uncertainty increases ($c \to 0$, $T(c) \to T_{\max}$), the sensitivity of q^{m_t} to logit perturbations is scaled down by approximately $1/T_{\max}$. In the limit $T \to \infty$, $q^{m_t} \to \mathbf{1}/K$, so the teacher target becomes less directional and less sensitive to perturbations in the teacher logits.

C Additional Experiments

C.1 Full Per-Corruption Results

Tables 6 to 11 retain complete per-corruption mean$_{\pm\text{std}}$ results for all twelve baselines and their +MUTE variants, extending the compact results in Tables 1 to 4.

Table 7: Complete KS50 sequential video-corruption results at severity 5: average, noise, and blur corruptions. Accuracy is reported as mean$_{\pm\text{std}}$ (%) for all twelve baselines and their +MUTE variants. Bold values indicate an improvement over the corresponding base method.

Method	Avg.	Noise			Blur			
		G.	S.	I.	D.	Gl.	M.	Z.
EATA	$63.3_{\pm0.3}$	$49.2_{\pm0.2}$	$50.9_{\pm0.1}$	$51.2_{\pm0.2}$	$67.5_{\pm0.3}$	$66.7_{\pm0.3}$	$70.9_{\pm0.3}$	$69.9_{\pm0.6}$
+MUTE	$\mathbf{63.9}_{\pm\mathbf{0.1}}$	$\mathbf{49.3}_{\pm\mathbf{0.2}}$	$\mathbf{50.9}_{\pm\mathbf{0.3}}$	$51.0_{\pm0.2}$	$\mathbf{67.8}_{\pm\mathbf{0.3}}$	$66.4_{\pm0.2}$	$\mathbf{71.0}_{\pm\mathbf{0.4}}$	$\mathbf{70.6}_{\pm\mathbf{0.3}}$
Tent	$7.1_{\pm0.2}$	$47.4_{\pm0.4}$	$25.3_{\pm2.2}$	$3.8_{\pm0.5}$	$6.5_{\pm0.8}$	$2.4_{\pm0.2}$	$2.1_{\pm0.0}$	$2.1_{\pm0.0}$
+MUTE	$\mathbf{10.3}_{\pm\mathbf{1.2}}$	$\mathbf{49.2}_{\pm\mathbf{0.6}}$	$\mathbf{45.2}_{\pm\mathbf{1.7}}$	$\mathbf{25.9}_{\pm\mathbf{8.5}}$	$\mathbf{11.4}_{\pm\mathbf{8.0}}$	$\mathbf{2.5}_{\pm\mathbf{0.6}}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$
READ	$43.0_{\pm3.2}$	$51.2_{\pm0.1}$	$53.8_{\pm0.2}$	$53.5_{\pm0.5}$	$67.0_{\pm0.7}$	$66.2_{\pm1.0}$	$61.2_{\pm3.7}$	$55.1_{\pm4.7}$
+MUTE	$\mathbf{64.1}_{\pm\mathbf{0.1}}$	$\mathbf{51.7}_{\pm\mathbf{0.4}}$	$\mathbf{54.2}_{\pm\mathbf{0.2}}$	$\mathbf{54.2}_{\pm\mathbf{0.2}}$	$\mathbf{67.6}_{\pm\mathbf{0.5}}$	$\mathbf{68.6}_{\pm\mathbf{0.3}}$	$\mathbf{70.0}_{\pm\mathbf{0.2}}$	$\mathbf{70.4}_{\pm\mathbf{0.3}}$
SuMi	$59.6_{\pm0.2}$	$49.4_{\pm0.2}$	$50.5_{\pm0.1}$	$49.0_{\pm0.3}$	$64.7_{\pm0.4}$	$64.2_{\pm0.4}$	$68.6_{\pm0.3}$	$67.9_{\pm0.5}$
+MUTE	$\mathbf{61.7}_{\pm\mathbf{0.2}}$	$49.4_{\pm0.2}$	$50.5_{\pm0.1}$	$\mathbf{49.4}_{\pm\mathbf{0.3}}$	$64.6_{\pm0.5}$	$\mathbf{64.8}_{\pm\mathbf{0.2}}$	$\mathbf{68.9}_{\pm\mathbf{0.2}}$	$\mathbf{68.6}_{\pm\mathbf{0.6}}$
TRIBE	$61.1_{\pm0.1}$	$47.5_{\pm0.2}$	$48.1_{\pm0.4}$	$48.0_{\pm0.5}$	$62.0_{\pm0.5}$	$66.2_{\pm0.1}$	$67.9_{\pm0.7}$	$67.2_{\pm0.5}$
+MUTE	$\mathbf{62.0}_{\pm\mathbf{0.2}}$	$\mathbf{48.0}_{\pm\mathbf{0.1}}$	$48.0_{\pm0.3}$	$\mathbf{48.3}_{\pm\mathbf{0.4}}$	$\mathbf{63.4}_{\pm\mathbf{0.2}}$	$\mathbf{67.7}_{\pm\mathbf{0.7}}$	$\mathbf{69.9}_{\pm\mathbf{0.6}}$	$\mathbf{68.8}_{\pm\mathbf{0.2}}$
DeYO	$53.6_{\pm6.9}$	$45.6_{\pm0.1}$	$42.0_{\pm0.4}$	$36.9_{\pm2.1}$	$55.7_{\pm3.8}$	$58.0_{\pm4.8}$	$65.1_{\pm6.2}$	$63.2_{\pm5.6}$
+MUTE	$\mathbf{59.6}_{\pm\mathbf{0.3}}$	$\mathbf{45.9}_{\pm\mathbf{0.1}}$	$\mathbf{44.5}_{\pm\mathbf{0.2}}$	$\mathbf{41.0}_{\pm\mathbf{0.1}}$	$\mathbf{60.8}_{\pm\mathbf{0.1}}$	$\mathbf{65.2}_{\pm\mathbf{0.5}}$	$\mathbf{69.4}_{\pm\mathbf{0.6}}$	$\mathbf{67.6}_{\pm\mathbf{0.8}}$
SURGEON	$10.7_{\pm0.4}$	$46.2_{\pm0.3}$	$44.5_{\pm0.1}$	$32.3_{\pm2.0}$	$15.0_{\pm4.0}$	$2.6_{\pm0.5}$	$2.1_{\pm0.0}$	$2.1_{\pm0.0}$
+MUTE	$\mathbf{53.6}_{\pm\mathbf{0.9}}$	$\mathbf{46.5}_{\pm\mathbf{0.3}}$	$\mathbf{45.6}_{\pm\mathbf{0.3}}$	$\mathbf{43.6}_{\pm\mathbf{0.4}}$	$\mathbf{61.2}_{\pm\mathbf{1.8}}$	$\mathbf{65.2}_{\pm\mathbf{8.0}}$	$\mathbf{68.7}_{\pm\mathbf{3.3}}$	$\mathbf{66.1}_{\pm\mathbf{0.1}}$
TACT	$53.9_{\pm0.7}$	$47.2_{\pm0.3}$	$46.8_{\pm0.1}$	$45.5_{\pm0.3}$	$63.8_{\pm0.1}$	$49.5_{\pm2.8}$	$62.0_{\pm1.9}$	$55.5_{\pm1.3}$
+MUTE	$\mathbf{58.0}_{\pm\mathbf{0.1}}$	$47.1_{\pm0.2}$	$46.5_{\pm0.3}$	$45.4_{\pm0.2}$	$63.5_{\pm0.6}$	$\mathbf{66.2}_{\pm\mathbf{0.1}}$	$\mathbf{69.5}_{\pm\mathbf{0.2}}$	$\mathbf{67.6}_{\pm\mathbf{0.5}}$
CoTTA	$60.0_{\pm0.0}$	$47.1_{\pm0.1}$	$48.2_{\pm0.0}$	$47.6_{\pm0.0}$	$67.3_{\pm0.1}$	$61.7_{\pm0.1}$	$72.2_{\pm0.1}$	$66.8_{\pm0.0}$
+MUTE	$\mathbf{60.5}_{\pm\mathbf{0.1}}$	$\mathbf{47.2}_{\pm\mathbf{0.1}}$	$\mathbf{48.4}_{\pm\mathbf{0.1}}$	$\mathbf{48.3}_{\pm\mathbf{0.0}}$	$\mathbf{68.1}_{\pm\mathbf{0.1}}$	$\mathbf{63.6}_{\pm\mathbf{0.0}}$	$\mathbf{72.4}_{\pm\mathbf{0.1}}$	$\mathbf{67.7}_{\pm\mathbf{0.2}}$
MEMO	$51.0_{\pm0.0}$	$38.6_{\pm0.0}$	$38.9_{\pm0.0}$	$38.5_{\pm0.0}$	$54.8_{\pm0.0}$	$54.7_{\pm0.0}$	$59.8_{\pm0.0}$	$56.7_{\pm0.0}$
+MUTE	$\mathbf{51.0}_{\pm\mathbf{0.0}}$	$\mathbf{38.7}_{\pm\mathbf{0.0}}$	$38.9_{\pm0.0}$	$\mathbf{38.6}_{\pm\mathbf{0.0}}$	$54.8_{\pm0.0}$	$54.7_{\pm0.0}$	$\mathbf{59.9}_{\pm\mathbf{0.0}}$	$\mathbf{56.8}_{\pm\mathbf{0.1}}$
SAR	$60.3_{\pm0.0}$	$47.1_{\pm0.1}$	$47.9_{\pm0.0}$	$47.2_{\pm0.0}$	$66.6_{\pm0.1}$	$64.3_{\pm0.1}$	$70.9_{\pm0.1}$	$68.5_{\pm0.1}$
+MUTE	$\mathbf{61.1}_{\pm\mathbf{0.0}}$	$47.1_{\pm0.1}$	$47.7_{\pm0.0}$	$46.9_{\pm0.0}$	$66.3_{\pm0.1}$	$\mathbf{65.2}_{\pm\mathbf{0.1}}$	$\mathbf{71.1}_{\pm\mathbf{0.1}}$	$\mathbf{68.7}_{\pm\mathbf{0.1}}$
RoTTA	$59.7_{\pm0.2}$	$47.1_{\pm0.2}$	$48.5_{\pm0.5}$	$47.8_{\pm0.5}$	$64.2_{\pm0.1}$	$64.6_{\pm0.2}$	$69.0_{\pm0.2}$	$66.4_{\pm0.3}$
+MUTE	$\mathbf{60.4}_{\pm\mathbf{0.1}}$	$47.0_{\pm0.2}$	$48.2_{\pm0.4}$	$\mathbf{48.1}_{\pm\mathbf{0.4}}$	$\mathbf{65.3}_{\pm\mathbf{0.0}}$	$\mathbf{64.9}_{\pm\mathbf{0.2}}$	$\mathbf{70.2}_{\pm\mathbf{0.2}}$	$\mathbf{67.5}_{\pm\mathbf{0.1}}$

C.2 Computational Overhead

Table 12 profiles runtime and peak memory on an A100-80GB GPU. Time, throughput, and memory entries report Base/+MUTE. MUTE introduces no trainable parameters, reuses existing logits for confidence routing and the KL loss, and requires no extra forward pass or model copy. The confidence, routing, and KL operations add 0.010/0.059 MFLOPs on KS50/VGG, compared with a 132.46-GFLOP CAV-MAE forward (Gong et al., 2023) ($< 0.0001\%$); peak memory is unchanged, while measured throughput ranges from 0.88–$1.01\times$ the base method. The VGG-A SAR case incurs a measurable 13.9% runtime increase.

C.3 Sensitivity and Stress Tests

Hyperparameter Sensitivity

We sweep the three key hyperparameters of MUTE—mutual-teaching weight λ_{max}, advantage margin δ, and confidence threshold $\tau_{\min}$—as reported in Table 13. For the **teaching weight** λ_{max}, we observe that a moderate value (0.10) yields the best performance. As λ_{max} increases beyond 0.50, performance gradually declines, indicating that excessive mutual-teaching can over-regularize the base adaptation loss and hinder the model's plasticity to new distributions. However, it is imperative to note that even under the most extreme and suboptimal setting ($\lambda_{max} = 2.0$, yielding 59.6%), MUTE still provides a massive +16.6% absolute improvement over the collapsed READ baseline

Table 8: Complete KS50 sequential video-corruption results at severity 5: weather and digital corruptions. Accuracy is reported as mean$_{\pm\text{std}}$ (%) for all twelve baselines and their +MUTE variants. Bold values indicate an improvement over the corresponding base method.

Method	**Weather**			**Digital**				
	Sn.	**Fr.**	**Fo.**	**Br.**	**C.**	**E.**	**P.**	**J.**
EATA	$63.6_{\pm0.4}$	$66.3_{\pm0.5}$	$63.4_{\pm0.7}$	$71.2_{\pm0.4}$	$53.4_{\pm0.5}$	$72.4_{\pm0.4}$	$66.8_{\pm0.7}$	$66.3_{\pm0.7}$
+MUTE	$\mathbf{64.9}_{\pm\mathbf{0.1}}$	$\mathbf{67.6}_{\pm\mathbf{0.4}}$	$\mathbf{64.2}_{\pm\mathbf{0.3}}$	$\mathbf{72.4}_{\pm\mathbf{0.4}}$	$\mathbf{54.6}_{\pm\mathbf{0.3}}$	$72.4_{\pm0.5}$	$\mathbf{67.8}_{\pm\mathbf{0.5}}$	$\mathbf{67.1}_{\pm\mathbf{0.5}}$
Tent	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.1_{\pm0.1}$	$2.1_{\pm0.1}$	$2.1_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$
+MUTE	$2.0_{\pm0.1}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$
READ	$47.9_{\pm5.2}$	$43.7_{\pm5.6}$	$33.3_{\pm5.8}$	$34.3_{\pm7.6}$	$21.5_{\pm4.7}$	$20.8_{\pm4.6}$	$18.4_{\pm2.8}$	$16.6_{\pm1.9}$
+MUTE	$\mathbf{64.1}_{\pm\mathbf{0.2}}$	$\mathbf{65.8}_{\pm\mathbf{0.2}}$	$\mathbf{62.2}_{\pm\mathbf{0.6}}$	$\mathbf{72.4}_{\pm\mathbf{0.3}}$	$\mathbf{52.8}_{\pm\mathbf{0.4}}$	$\mathbf{70.1}_{\pm\mathbf{0.4}}$	$\mathbf{69.9}_{\pm\mathbf{0.2}}$	$\mathbf{68.0}_{\pm\mathbf{0.2}}$
SuMi	$63.1_{\pm0.5}$	$64.3_{\pm0.1}$	$58.6_{\pm0.6}$	$69.5_{\pm0.7}$	$46.8_{\pm0.7}$	$65.6_{\pm0.4}$	$57.0_{\pm0.6}$	$53.9_{\pm0.4}$
+MUTE	$\mathbf{64.2}_{\pm\mathbf{0.4}}$	$\mathbf{65.2}_{\pm\mathbf{0.1}}$	$\mathbf{60.4}_{\pm\mathbf{0.5}}$	$\mathbf{70.8}_{\pm\mathbf{0.7}}$	$\mathbf{51.3}_{\pm\mathbf{0.5}}$	$\mathbf{69.7}_{\pm\mathbf{0.3}}$	$\mathbf{64.7}_{\pm\mathbf{0.6}}$	$\mathbf{63.1}_{\pm\mathbf{0.3}}$
TRIBE	$63.1_{\pm0.4}$	$65.7_{\pm0.4}$	$61.2_{\pm0.7}$	$69.0_{\pm0.5}$	$52.9_{\pm0.6}$	$70.0_{\pm0.6}$	$64.4_{\pm0.2}$	$62.9_{\pm1.0}$
+MUTE	$\mathbf{64.2}_{\pm\mathbf{0.4}}$	$\mathbf{66.4}_{\pm\mathbf{0.3}}$	$\mathbf{61.4}_{\pm\mathbf{0.5}}$	$\mathbf{70.2}_{\pm\mathbf{0.1}}$	$\mathbf{53.5}_{\pm\mathbf{0.5}}$	$\mathbf{70.2}_{\pm\mathbf{0.1}}$	$\mathbf{65.9}_{\pm\mathbf{0.2}}$	$\mathbf{63.6}_{\pm\mathbf{0.1}}$
DeYO	$54.3_{\pm8.0}$	$57.6_{\pm8.4}$	$50.6_{\pm12.0}$	$62.9_{\pm10.1}$	$43.4_{\pm7.1}$	$62.8_{\pm8.9}$	$54.5_{\pm13.1}$	$51.4_{\pm13.3}$
+MUTE	$\mathbf{61.9}_{\pm\mathbf{0.5}}$	$\mathbf{64.0}_{\pm\mathbf{0.1}}$	$\mathbf{53.7}_{\pm\mathbf{0.9}}$	$\mathbf{71.2}_{\pm\mathbf{0.3}}$	$\mathbf{52.4}_{\pm\mathbf{0.4}}$	$\mathbf{69.3}_{\pm\mathbf{0.2}}$	$\mathbf{64.9}_{\pm\mathbf{0.5}}$	$\mathbf{62.9}_{\pm\mathbf{0.3}}$
SURGEON	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.1_{\pm0.0}$	$2.0_{\pm0.0}$	$2.1_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$	$2.0_{\pm0.0}$
+MUTE	$\mathbf{59.5}_{\pm\mathbf{0.1}}$	$\mathbf{63.4}_{\pm\mathbf{0.0}}$	$\mathbf{56.6}_{\pm\mathbf{0.0}}$	$\mathbf{66.1}_{\pm\mathbf{0.0}}$	$\mathbf{46.2}_{\pm\mathbf{0.0}}$	$\mathbf{62.2}_{\pm\mathbf{0.0}}$	$\mathbf{41.0}_{\pm\mathbf{0.0}}$	$\mathbf{12.0}_{\pm\mathbf{0.0}}$
TACT	$53.3_{\pm1.4}$	$50.3_{\pm3.7}$	$36.6_{\pm3.8}$	$68.3_{\pm0.2}$	$50.2_{\pm0.2}$	$61.5_{\pm1.3}$	$55.7_{\pm0.3}$	$62.7_{\pm0.4}$
+MUTE	$\mathbf{63.3}_{\pm\mathbf{0.2}}$	$\mathbf{63.3}_{\pm\mathbf{0.2}}$	$\mathbf{48.9}_{\pm\mathbf{0.7}}$	$66.8_{\pm0.3}$	$48.9_{\pm0.4}$	$\mathbf{65.9}_{\pm\mathbf{0.3}}$	$53.8_{\pm0.3}$	$56.7_{\pm0.4}$
CoTTA	$59.1_{\pm0.1}$	$60.2_{\pm0.0}$	$46.0_{\pm0.0}$	$75.4_{\pm0.1}$	$51.8_{\pm0.1}$	$65.9_{\pm0.1}$	$65.7_{\pm0.2}$	$64.2_{\pm0.1}$
+MUTE	$\mathbf{60.6}_{\pm\mathbf{0.2}}$	$\mathbf{61.7}_{\pm\mathbf{0.3}}$	$\mathbf{47.8}_{\pm\mathbf{0.1}}$	$74.9_{\pm0.1}$	$50.4_{\pm0.1}$	$\mathbf{65.9}_{\pm\mathbf{0.3}}$	$\mathbf{66.3}_{\pm\mathbf{0.2}}$	$\mathbf{64.2}_{\pm\mathbf{0.1}}$
MEMO	$51.9_{\pm0.0}$	$52.9_{\pm0.0}$	$45.1_{\pm0.0}$	$61.9_{\pm0.0}$	$44.9_{\pm0.0}$	$58.4_{\pm0.0}$	$53.7_{\pm0.0}$	$53.6_{\pm0.0}$
+MUTE	$51.9_{\pm0.0}$	$52.8_{\pm0.0}$	$45.0_{\pm0.0}$	$\mathbf{62.0}_{\pm\mathbf{0.0}}$	$\mathbf{45.0}_{\pm\mathbf{0.0}}$	$58.3_{\pm0.0}$	$\mathbf{53.7}_{\pm\mathbf{0.0}}$	$53.6_{\pm0.0}$
SAR	$58.8_{\pm0.2}$	$63.4_{\pm0.2}$	$55.4_{\pm0.0}$	$72.8_{\pm0.2}$	$50.4_{\pm0.1}$	$68.7_{\pm0.1}$	$62.4_{\pm0.1}$	$60.2_{\pm0.0}$
+MUTE	$\mathbf{60.4}_{\pm\mathbf{0.1}}$	$\mathbf{65.1}_{\pm\mathbf{0.2}}$	$\mathbf{57.4}_{\pm\mathbf{0.1}}$	$\mathbf{73.2}_{\pm\mathbf{0.1}}$	$\mathbf{51.9}_{\pm\mathbf{0.2}}$	$\mathbf{70.2}_{\pm\mathbf{0.1}}$	$\mathbf{63.7}_{\pm\mathbf{0.1}}$	$\mathbf{61.9}_{\pm\mathbf{0.1}}$
RoTTA	$56.8_{\pm0.5}$	$61.4_{\pm0.4}$	$54.9_{\pm0.3}$	$71.4_{\pm0.3}$	$49.5_{\pm0.5}$	$69.0_{\pm0.5}$	$63.2_{\pm0.2}$	$62.0_{\pm0.7}$
+MUTE	$\mathbf{57.7}_{\pm\mathbf{0.3}}$	$\mathbf{62.1}_{\pm\mathbf{0.6}}$	$\mathbf{55.5}_{\pm\mathbf{0.3}}$	$\mathbf{72.7}_{\pm\mathbf{0.3}}$	$\mathbf{51.2}_{\pm\mathbf{0.4}}$	$\mathbf{69.4}_{\pm\mathbf{0.2}}$	$\mathbf{64.0}_{\pm\mathbf{0.0}}$	$\mathbf{62.7}_{\pm\mathbf{0.3}}$

(43.0%). This demonstrates that MUTE is highly robust and beneficial even without precise tuning. For the **advantage margin** δ, performance remains highly stable across the tested range of 0.1 to 0.5. This empirical plateau is consistent with the margin-dependent term in Theorem 1 and indicates limited sensitivity to δ within the evaluated range. Finally, for the **confidence threshold** $\tau_{\min}$, moderate values (0.2–0.3) strike an optimal balance between filtering out unreliable predictions and maintaining a sufficient volume of routing opportunities. Extremely high thresholds (0.4–0.5) overly restrict the mutual teaching process. Consequently, we confidently adopt $\lambda_{max} = 0.10$, $\delta = 0.1$, and $\tau_{\min} = 0.3$ as the universally fixed configuration across all datasets and base methods, bypassing the need for target-domain validation.

JOINT CORRUPTIONS

As shown in Table 14(a), MUTE remains beneficial in these joint-corruption stress settings and can be combined with specialized methods for stronger joint-shift handling.

MISSING MODALITY

Table 14(b) evaluates complete modality absence, a substantially harsher shift than corruption. The completed boundary cases show gains, but do not support extrapolation to arbitrary sensor failures.

Table 9: Complete VGG-Sound sequential audio-corruption results at severity 5. Accuracy is reported as mean$_{\pm\text{std}}$ (%) for all twelve baselines and their +MUTE variants. Bold values indicate an improvement over the corresponding base method.

Method	Avg.	Gauss.	Traffic	Crowd	Rain	Thunder	Wind
EATA	$30.0_{\pm0.1}$	$36.2_{\pm0.1}$	$26.1_{\pm0.0}$	$21.7_{\pm0.2}$	$30.2_{\pm0.1}$	$35.9_{\pm0.1}$	$29.9_{\pm0.0}$
+MUTE	$\mathbf{30.1}_{\pm\mathbf{0.0}}$	$\mathbf{36.3}_{\pm\mathbf{0.0}}$	$\mathbf{26.3}_{\pm\mathbf{0.0}}$	$\mathbf{22.1}_{\pm\mathbf{0.1}}$	$\mathbf{30.2}_{\pm\mathbf{0.1}}$	$35.7_{\pm0.1}$	$29.7_{\pm0.1}$
Tent	$16.6_{\pm0.0}$	$32.9_{\pm0.1}$	$16.6_{\pm0.1}$	$9.3_{\pm0.1}$	$11.2_{\pm0.2}$	$21.2_{\pm0.1}$	$8.4_{\pm0.1}$
+MUTE	$\mathbf{27.5}_{\pm\mathbf{0.2}}$	$\mathbf{36.4}_{\pm\mathbf{0.0}}$	$\mathbf{24.9}_{\pm\mathbf{0.1}}$	$\mathbf{17.2}_{\pm\mathbf{0.7}}$	$\mathbf{30.3}_{\pm\mathbf{0.1}}$	$\mathbf{35.1}_{\pm\mathbf{0.2}}$	$\mathbf{20.9}_{\pm\mathbf{0.2}}$
READ	$31.4_{\pm0.0}$	$37.3_{\pm0.1}$	$28.5_{\pm0.1}$	$25.3_{\pm0.1}$	$30.7_{\pm0.1}$	$35.3_{\pm0.0}$	$31.1_{\pm0.1}$
+MUTE	$\mathbf{32.5}_{\pm\mathbf{0.0}}$	$\mathbf{37.6}_{\pm\mathbf{0.1}}$	$\mathbf{28.7}_{\pm\mathbf{0.1}}$	$\mathbf{26.4}_{\pm\mathbf{0.1}}$	$\mathbf{32.0}_{\pm\mathbf{0.1}}$	$\mathbf{37.4}_{\pm\mathbf{0.0}}$	$\mathbf{33.0}_{\pm\mathbf{0.1}}$
SuMi	$29.5_{\pm0.0}$	$36.0_{\pm0.0}$	$26.3_{\pm0.1}$	$23.1_{\pm0.1}$	$29.5_{\pm0.1}$	$34.4_{\pm0.0}$	$27.5_{\pm0.1}$
+MUTE	$\mathbf{30.0}_{\pm\mathbf{0.0}}$	$\mathbf{36.2}_{\pm\mathbf{0.0}}$	$\mathbf{26.7}_{\pm\mathbf{0.1}}$	$\mathbf{23.4}_{\pm\mathbf{0.1}}$	$\mathbf{30.2}_{\pm\mathbf{0.1}}$	$\mathbf{35.2}_{\pm\mathbf{0.1}}$	$\mathbf{28.2}_{\pm\mathbf{0.1}}$
TRIBE	$30.9_{\pm0.2}$	$36.7_{\pm0.1}$	$27.5_{\pm0.1}$	$19.7_{\pm1.0}$	$31.7_{\pm0.0}$	$37.7_{\pm0.0}$	$31.9_{\pm0.1}$
+MUTE	$\mathbf{31.2}_{\pm\mathbf{0.2}}$	$\mathbf{36.8}_{\pm\mathbf{0.1}}$	$\mathbf{27.7}_{\pm\mathbf{0.1}}$	$\mathbf{20.7}_{\pm\mathbf{1.0}}$	$\mathbf{31.9}_{\pm\mathbf{0.0}}$	$\mathbf{37.9}_{\pm\mathbf{0.0}}$	$\mathbf{32.0}_{\pm\mathbf{0.0}}$
DeYO	$23.9_{\pm0.1}$	$36.1_{\pm0.0}$	$23.8_{\pm0.2}$	$8.4_{\pm0.2}$	$27.0_{\pm0.1}$	$27.2_{\pm0.2}$	$20.7_{\pm0.4}$
+MUTE	$\mathbf{24.7}_{\pm\mathbf{0.1}}$	$\mathbf{36.3}_{\pm\mathbf{0.0}}$	$\mathbf{24.8}_{\pm\mathbf{0.1}}$	$\mathbf{9.0}_{\pm\mathbf{0.2}}$	$\mathbf{28.3}_{\pm\mathbf{0.1}}$	$\mathbf{27.8}_{\pm\mathbf{0.3}}$	$\mathbf{22.0}_{\pm\mathbf{0.3}}$
SURGEON	$11.6_{\pm0.1}$	$33.5_{\pm0.0}$	$12.5_{\pm0.2}$	$2.4_{\pm1.2}$	$5.9_{\pm0.3}$	$10.5_{\pm0.1}$	$4.9_{\pm0.3}$
+MUTE	$\mathbf{11.7}_{\pm\mathbf{0.1}}$	$\mathbf{33.7}_{\pm\mathbf{0.0}}$	$\mathbf{12.6}_{\pm\mathbf{0.2}}$	$\mathbf{2.5}_{\pm\mathbf{0.9}}$	$\mathbf{6.2}_{\pm\mathbf{0.2}}$	$10.4_{\pm0.1}$	$4.9_{\pm0.3}$
TACT	$11.1_{\pm0.0}$	$30.4_{\pm0.0}$	$15.9_{\pm0.2}$	$6.8_{\pm0.3}$	$5.4_{\pm0.1}$	$7.6_{\pm0.1}$	$0.5_{\pm0.1}$
+MUTE	$\mathbf{26.5}_{\pm\mathbf{0.0}}$	$\mathbf{34.9}_{\pm\mathbf{0.1}}$	$\mathbf{24.6}_{\pm\mathbf{0.2}}$	$\mathbf{18.9}_{\pm\mathbf{0.1}}$	$\mathbf{27.0}_{\pm\mathbf{0.1}}$	$\mathbf{32.3}_{\pm\mathbf{0.0}}$	$\mathbf{21.2}_{\pm\mathbf{0.1}}$
CoTTA	$27.5_{\pm0.0}$	$35.1_{\pm0.0}$	$25.3_{\pm0.0}$	$21.1_{\pm0.0}$	$27.5_{\pm0.0}$	$32.7_{\pm0.0}$	$23.2_{\pm0.0}$
+MUTE	$\mathbf{28.2}_{\pm\mathbf{0.0}}$	$\mathbf{35.8}_{\pm\mathbf{0.0}}$	$\mathbf{25.7}_{\pm\mathbf{0.0}}$	$\mathbf{21.8}_{\pm\mathbf{0.1}}$	$\mathbf{28.6}_{\pm\mathbf{0.0}}$	$\mathbf{33.8}_{\pm\mathbf{0.1}}$	$\mathbf{23.7}_{\pm\mathbf{0.1}}$
MEMO	$17.2_{\pm0.0}$	$33.5_{\pm0.0}$	$17.5_{\pm0.0}$	$10.0_{\pm0.0}$	$11.7_{\pm0.0}$	$21.7_{\pm0.0}$	$9.1_{\pm0.0}$
+MUTE	$\mathbf{27.3}_{\pm\mathbf{0.0}}$	$\mathbf{35.0}_{\pm\mathbf{0.0}}$	$\mathbf{25.1}_{\pm\mathbf{0.0}}$	$\mathbf{21.0}_{\pm\mathbf{0.0}}$	$\mathbf{27.3}_{\pm\mathbf{0.0}}$	$\mathbf{32.6}_{\pm\mathbf{0.0}}$	$\mathbf{23.0}_{\pm\mathbf{0.0}}$
SAR	$26.4_{\pm0.0}$	$36.2_{\pm0.0}$	$25.7_{\pm0.0}$	$12.3_{\pm0.1}$	$28.4_{\pm0.1}$	$31.3_{\pm0.0}$	$24.5_{\pm0.1}$
+MUTE	$\mathbf{27.2}_{\pm\mathbf{0.0}}$	$\mathbf{36.3}_{\pm\mathbf{0.0}}$	$\mathbf{26.3}_{\pm\mathbf{0.0}}$	$\mathbf{14.1}_{\pm\mathbf{0.0}}$	$\mathbf{29.6}_{\pm\mathbf{0.1}}$	$\mathbf{31.8}_{\pm\mathbf{0.0}}$	$\mathbf{25.3}_{\pm\mathbf{0.1}}$
RoTTA	$30.6_{\pm0.0}$	$36.3_{\pm0.0}$	$26.6_{\pm0.1}$	$22.0_{\pm0.2}$	$31.3_{\pm0.0}$	$36.6_{\pm0.0}$	$31.0_{\pm0.1}$
+MUTE	$\mathbf{31.7}_{\pm\mathbf{0.0}}$	$\mathbf{37.0}_{\pm\mathbf{0.0}}$	$\mathbf{27.4}_{\pm\mathbf{0.1}}$	$\mathbf{22.8}_{\pm\mathbf{0.1}}$	$\mathbf{32.4}_{\pm\mathbf{0.0}}$	$\mathbf{37.8}_{\pm\mathbf{0.0}}$	$\mathbf{32.9}_{\pm\mathbf{0.1}}$

Corruption Severity

Table 14(c) shows positive gains across the reported S1–S4 cells, with especially large gains when the underlying adapter begins to collapse. For READ KS50-V, S1–S2 use the repaired mild-severity configuration, whereas S3–S4 use separately tuned completed configurations. These results provide additional evidence of robustness across the reported severity settings.

Table 10: Complete VGG-Sound sequential video-corruption results at severity 5: average, noise, and blur corruptions. Accuracy is reported as mean$_{\pm\text{std}}$ (%) for all twelve baselines and their +MUTE variants. Bold values indicate an improvement over the corresponding base method.

Method	Avg.	Noise			Blur			
		G.	S.	I.	D.	Gl.	M.	Z.
EATA	$55.9_{\pm0.0}$	$53.2_{\pm0.0}$	$53.2_{\pm0.0}$	$53.5_{\pm0.0}$	$56.9_{\pm0.0}$	$56.7_{\pm0.0}$	$57.6_{\pm0.0}$	$57.8_{\pm0.1}$
+MUTE	$\mathbf{56.5}_{\pm\mathbf{0.0}}$	$\mathbf{53.2}_{\pm\mathbf{0.0}}$	$\mathbf{53.3}_{\pm\mathbf{0.0}}$	$\mathbf{53.6}_{\pm\mathbf{0.0}}$	$\mathbf{57.4}_{\pm\mathbf{0.0}}$	$\mathbf{57.1}_{\pm\mathbf{0.0}}$	$\mathbf{58.3}_{\pm\mathbf{0.0}}$	$\mathbf{58.3}_{\pm\mathbf{0.0}}$
Tent	$55.2_{\pm0.0}$	$53.2_{\pm0.0}$	$53.1_{\pm0.0}$	$53.3_{\pm0.0}$	$56.5_{\pm0.0}$	$56.3_{\pm0.0}$	$57.1_{\pm0.0}$	$57.1_{\pm0.0}$
+MUTE	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$53.0_{\pm0.0}$	$52.9_{\pm0.0}$	$53.2_{\pm0.0}$	$\mathbf{57.1}_{\pm\mathbf{0.0}}$	$\mathbf{56.8}_{\pm\mathbf{0.0}}$	$\mathbf{58.0}_{\pm\mathbf{0.0}}$	$\mathbf{58.0}_{\pm\mathbf{0.0}}$
READ	$55.8_{\pm0.0}$	$53.2_{\pm0.0}$	$53.2_{\pm0.0}$	$53.5_{\pm0.0}$	$56.9_{\pm0.0}$	$56.7_{\pm0.0}$	$57.6_{\pm0.0}$	$57.7_{\pm0.0}$
+MUTE	$\mathbf{56.4}_{\pm\mathbf{0.0}}$	$53.0_{\pm0.0}$	$53.0_{\pm0.0}$	$53.2_{\pm0.0}$	$\mathbf{57.2}_{\pm\mathbf{0.0}}$	$\mathbf{56.9}_{\pm\mathbf{0.0}}$	$\mathbf{58.2}_{\pm\mathbf{0.0}}$	$\mathbf{58.1}_{\pm\mathbf{0.0}}$
SuMi	$53.9_{\pm0.0}$	$52.9_{\pm0.0}$	$52.7_{\pm0.0}$	$52.6_{\pm0.0}$	$56.7_{\pm0.0}$	$55.6_{\pm0.0}$	$56.5_{\pm0.0}$	$55.2_{\pm0.0}$
+MUTE	$\mathbf{55.5}_{\pm\mathbf{0.0}}$	$\mathbf{53.0}_{\pm\mathbf{0.0}}$	$\mathbf{52.9}_{\pm\mathbf{0.0}}$	$\mathbf{53.1}_{\pm\mathbf{0.0}}$	$\mathbf{57.0}_{\pm\mathbf{0.0}}$	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$\mathbf{57.4}_{\pm\mathbf{0.0}}$	$\mathbf{57.0}_{\pm\mathbf{0.0}}$
TRIBE	$55.7_{\pm0.0}$	$52.4_{\pm0.0}$	$52.4_{\pm0.0}$	$52.7_{\pm0.0}$	$56.5_{\pm0.1}$	$56.5_{\pm0.1}$	$57.9_{\pm0.0}$	$57.7_{\pm0.0}$
+MUTE	$\mathbf{55.8}_{\pm\mathbf{0.0}}$	$52.4_{\pm0.0}$	$52.4_{\pm0.0}$	$52.7_{\pm0.0}$	$\mathbf{56.6}_{\pm\mathbf{0.1}}$	$\mathbf{56.5}_{\pm\mathbf{0.0}}$	$\mathbf{58.1}_{\pm\mathbf{0.0}}$	$\mathbf{57.8}_{\pm\mathbf{0.0}}$
DeYO	$55.3_{\pm0.0}$	$52.5_{\pm0.1}$	$52.5_{\pm0.0}$	$52.7_{\pm0.0}$	$56.5_{\pm0.0}$	$56.3_{\pm0.1}$	$57.5_{\pm0.0}$	$57.3_{\pm0.0}$
+MUTE	$\mathbf{55.6}_{\pm\mathbf{0.0}}$	$52.5_{\pm0.1}$	$52.5_{\pm0.0}$	$\mathbf{52.7}_{\pm\mathbf{0.0}}$	$56.5_{\pm0.1}$	$\mathbf{56.5}_{\pm\mathbf{0.0}}$	$\mathbf{57.8}_{\pm\mathbf{0.0}}$	$\mathbf{57.6}_{\pm\mathbf{0.0}}$
SURGEON	$54.4_{\pm0.0}$	$52.6_{\pm0.1}$	$52.4_{\pm0.0}$	$52.5_{\pm0.0}$	$56.0_{\pm0.0}$	$55.6_{\pm0.0}$	$56.6_{\pm0.0}$	$56.4_{\pm0.0}$
+MUTE	$\mathbf{54.5}_{\pm\mathbf{0.0}}$	$52.6_{\pm0.1}$	$52.4_{\pm0.0}$	$52.5_{\pm0.0}$	$56.0_{\pm0.0}$	$\mathbf{55.7}_{\pm\mathbf{0.0}}$	$\mathbf{56.7}_{\pm\mathbf{0.0}}$	$\mathbf{56.5}_{\pm\mathbf{0.0}}$
TACT	$54.7_{\pm0.0}$	$52.6_{\pm0.0}$	$52.4_{\pm0.0}$	$52.7_{\pm0.0}$	$56.2_{\pm0.0}$	$55.9_{\pm0.0}$	$56.9_{\pm0.0}$	$56.8_{\pm0.0}$
+MUTE	$\mathbf{55.2}_{\pm\mathbf{0.0}}$	$52.6_{\pm0.0}$	$52.4_{\pm0.0}$	$\mathbf{52.8}_{\pm\mathbf{0.0}}$	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$\mathbf{56.4}_{\pm\mathbf{0.0}}$	$\mathbf{57.7}_{\pm\mathbf{0.0}}$	$\mathbf{57.4}_{\pm\mathbf{0.0}}$
CoTTA	$55.4_{\pm0.0}$	$52.3_{\pm0.0}$	$52.2_{\pm0.0}$	$52.4_{\pm0.0}$	$56.8_{\pm0.0}$	$56.2_{\pm0.0}$	$57.9_{\pm0.0}$	$57.1_{\pm0.0}$
+MUTE	$\mathbf{55.7}_{\pm\mathbf{0.0}}$	$\mathbf{52.4}_{\pm\mathbf{0.0}}$	$\mathbf{52.2}_{\pm\mathbf{0.0}}$	$\mathbf{52.4}_{\pm\mathbf{0.0}}$	$\mathbf{56.9}_{\pm\mathbf{0.0}}$	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$\mathbf{58.1}_{\pm\mathbf{0.0}}$	$\mathbf{57.5}_{\pm\mathbf{0.0}}$
MEMO	$55.0_{\pm0.0}$	$52.5_{\pm0.0}$	$52.4_{\pm0.0}$	$52.7_{\pm0.0}$	$56.3_{\pm0.0}$	$56.1_{\pm0.0}$	$57.3_{\pm0.0}$	$57.0_{\pm0.0}$
+MUTE	$\mathbf{55.4}_{\pm\mathbf{0.0}}$	$52.3_{\pm0.0}$	$52.2_{\pm0.0}$	$52.3_{\pm0.0}$	$\mathbf{56.8}_{\pm\mathbf{0.0}}$	$\mathbf{56.2}_{\pm\mathbf{0.0}}$	$\mathbf{57.9}_{\pm\mathbf{0.0}}$	$\mathbf{57.1}_{\pm\mathbf{0.0}}$
SAR	$55.7_{\pm0.0}$	$52.9_{\pm0.0}$	$52.9_{\pm0.0}$	$53.0_{\pm0.0}$	$56.9_{\pm0.0}$	$56.7_{\pm0.0}$	$58.3_{\pm0.0}$	$57.2_{\pm0.0}$
+MUTE	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$\mathbf{52.9}_{\pm\mathbf{0.1}}$	$52.9_{\pm0.0}$	$53.0_{\pm0.0}$	$\mathbf{57.2}_{\pm\mathbf{0.0}}$	$\mathbf{57.4}_{\pm\mathbf{0.1}}$	$\mathbf{59.1}_{\pm\mathbf{0.0}}$	$\mathbf{57.9}_{\pm\mathbf{0.0}}$
RoTTA	$55.7_{\pm0.0}$	$52.5_{\pm0.1}$	$52.5_{\pm0.0}$	$52.8_{\pm0.0}$	$56.8_{\pm0.0}$	$56.5_{\pm0.0}$	$57.9_{\pm0.0}$	$57.6_{\pm0.0}$
+MUTE	$\mathbf{55.8}_{\pm\mathbf{0.0}}$	$52.5_{\pm0.0}$	$52.4_{\pm0.0}$	$52.6_{\pm0.0}$	$56.8_{\pm0.0}$	$\mathbf{56.5}_{\pm\mathbf{0.0}}$	$\mathbf{57.9}_{\pm\mathbf{0.0}}$	$\mathbf{57.6}_{\pm\mathbf{0.0}}$

Table 11: Complete VGG-Sound sequential video-corruption results at severity 5: weather and digital corruptions. Accuracy is reported as mean$_{\pm\text{std}}$ (%) for all twelve baselines and their +MUTE variants. Bold values indicate an improvement over the corresponding base method.

Method	Weather			Digital				
	Sn.	Fr.	Fo.	Br.	C.	E.	P.	J.
EATA	$55.4_{\pm0.0}$	$56.1_{\pm0.1}$	$56.1_{\pm0.1}$	$58.2_{\pm0.0}$	$54.2_{\pm0.1}$	$56.9_{\pm0.0}$	$55.9_{\pm0.0}$	$56.0_{\pm0.0}$
+MUTE	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$\mathbf{56.8}_{\pm\mathbf{0.1}}$	$\mathbf{56.8}_{\pm\mathbf{0.0}}$	$\mathbf{59.1}_{\pm\mathbf{0.0}}$	$\mathbf{54.8}_{\pm\mathbf{0.1}}$	$\mathbf{58.4}_{\pm\mathbf{0.0}}$	$\mathbf{56.7}_{\pm\mathbf{0.0}}$	$\mathbf{57.1}_{\pm\mathbf{0.0}}$
Tent	$54.8_{\pm0.0}$	$55.4_{\pm0.1}$	$55.4_{\pm0.0}$	$57.2_{\pm0.0}$	$53.2_{\pm0.0}$	$56.1_{\pm0.0}$	$54.8_{\pm0.0}$	$54.5_{\pm0.0}$
+MUTE	$\mathbf{56.4}_{\pm\mathbf{0.0}}$	$\mathbf{56.5}_{\pm\mathbf{0.1}}$	$\mathbf{56.2}_{\pm\mathbf{0.1}}$	$\mathbf{59.3}_{\pm\mathbf{0.0}}$	$\mathbf{54.6}_{\pm\mathbf{0.0}}$	$\mathbf{58.3}_{\pm\mathbf{0.0}}$	$\mathbf{56.6}_{\pm\mathbf{0.0}}$	$\mathbf{57.0}_{\pm\mathbf{0.0}}$
READ	$55.4_{\pm0.0}$	$56.0_{\pm0.1}$	$56.1_{\pm0.0}$	$58.1_{\pm0.0}$	$54.1_{\pm0.0}$	$57.0_{\pm0.0}$	$55.8_{\pm0.0}$	$56.0_{\pm0.0}$
+MUTE	$\mathbf{56.5}_{\pm\mathbf{0.0}}$	$\mathbf{56.7}_{\pm\mathbf{0.0}}$	$\mathbf{56.6}_{\pm\mathbf{0.1}}$	$\mathbf{59.4}_{\pm\mathbf{0.0}}$	$\mathbf{54.9}_{\pm\mathbf{0.0}}$	$\mathbf{58.5}_{\pm\mathbf{0.0}}$	$\mathbf{56.8}_{\pm\mathbf{0.0}}$	$\mathbf{57.2}_{\pm\mathbf{0.0}}$
SuMi	$54.3_{\pm0.0}$	$52.9_{\pm0.1}$	$50.4_{\pm0.1}$	$57.9_{\pm0.0}$	$50.1_{\pm0.1}$	$54.0_{\pm0.0}$	$51.2_{\pm0.0}$	$55.1_{\pm0.0}$
+MUTE	$\mathbf{55.7}_{\pm\mathbf{0.0}}$	$\mathbf{55.2}_{\pm\mathbf{0.0}}$	$\mathbf{54.2}_{\pm\mathbf{0.0}}$	$\mathbf{58.9}_{\pm\mathbf{0.0}}$	$\mathbf{53.0}_{\pm\mathbf{0.1}}$	$\mathbf{57.2}_{\pm\mathbf{0.0}}$	$\mathbf{55.7}_{\pm\mathbf{0.0}}$	$\mathbf{56.6}_{\pm\mathbf{0.0}}$
TRIBE	$55.6_{\pm0.0}$	$56.6_{\pm0.0}$	$56.1_{\pm0.1}$	$58.4_{\pm0.0}$	$54.1_{\pm0.1}$	$57.6_{\pm0.1}$	$55.3_{\pm0.0}$	$55.8_{\pm0.0}$
+MUTE	$\mathbf{55.8}_{\pm\mathbf{0.0}}$	$\mathbf{56.7}_{\pm\mathbf{0.1}}$	$\mathbf{56.2}_{\pm\mathbf{0.0}}$	$\mathbf{58.6}_{\pm\mathbf{0.0}}$	$\mathbf{54.4}_{\pm\mathbf{0.0}}$	$\mathbf{57.9}_{\pm\mathbf{0.0}}$	$\mathbf{55.5}_{\pm\mathbf{0.0}}$	$\mathbf{56.0}_{\pm\mathbf{0.0}}$
DeYO	$54.9_{\pm0.0}$	$55.7_{\pm0.0}$	$55.9_{\pm0.0}$	$57.5_{\pm0.0}$	$53.8_{\pm0.0}$	$56.5_{\pm0.0}$	$54.8_{\pm0.0}$	$55.2_{\pm0.0}$
+MUTE	$\mathbf{55.2}_{\pm\mathbf{0.0}}$	$\mathbf{56.1}_{\pm\mathbf{0.1}}$	$\mathbf{56.1}_{\pm\mathbf{0.0}}$	$\mathbf{57.9}_{\pm\mathbf{0.0}}$	$\mathbf{54.0}_{\pm\mathbf{0.0}}$	$\mathbf{57.1}_{\pm\mathbf{0.0}}$	$\mathbf{55.2}_{\pm\mathbf{0.0}}$	$\mathbf{55.7}_{\pm\mathbf{0.0}}$
SURGEON	$54.0_{\pm0.0}$	$55.0_{\pm0.1}$	$54.6_{\pm0.0}$	$56.0_{\pm0.0}$	$52.5_{\pm0.1}$	$55.2_{\pm0.1}$	$53.4_{\pm0.0}$	$53.5_{\pm0.0}$
+MUTE	$\mathbf{54.0}_{\pm\mathbf{0.0}}$	$\mathbf{55.1}_{\pm\mathbf{0.1}}$	$\mathbf{54.6}_{\pm\mathbf{0.0}}$	$\mathbf{56.2}_{\pm\mathbf{0.0}}$	$\mathbf{52.6}_{\pm\mathbf{0.0}}$	$\mathbf{55.4}_{\pm\mathbf{0.0}}$	$\mathbf{53.5}_{\pm\mathbf{0.0}}$	$\mathbf{53.7}_{\pm\mathbf{0.0}}$
TACT	$54.4_{\pm0.0}$	$55.5_{\pm0.0}$	$55.0_{\pm0.0}$	$56.5_{\pm0.0}$	$52.5_{\pm0.0}$	$55.5_{\pm0.1}$	$53.8_{\pm0.0}$	$54.0_{\pm0.0}$
+MUTE	$\mathbf{55.0}_{\pm\mathbf{0.0}}$	$\mathbf{56.2}_{\pm\mathbf{0.0}}$	$\mathbf{55.6}_{\pm\mathbf{0.0}}$	$\mathbf{57.2}_{\pm\mathbf{0.0}}$	$\mathbf{53.1}_{\pm\mathbf{0.0}}$	$\mathbf{56.2}_{\pm\mathbf{0.0}}$	$\mathbf{54.2}_{\pm\mathbf{0.0}}$	$\mathbf{54.5}_{\pm\mathbf{0.0}}$
CoTTA	$55.4_{\pm0.0}$	$55.1_{\pm0.0}$	$54.4_{\pm0.0}$	$58.9_{\pm0.0}$	$53.3_{\pm0.0}$	$57.0_{\pm0.0}$	$55.4_{\pm0.0}$	$56.6_{\pm0.0}$
+MUTE	$\mathbf{55.7}_{\pm\mathbf{0.0}}$	$\mathbf{55.9}_{\pm\mathbf{0.0}}$	$\mathbf{55.0}_{\pm\mathbf{0.0}}$	$\mathbf{59.3}_{\pm\mathbf{0.0}}$	$\mathbf{53.8}_{\pm\mathbf{0.0}}$	$\mathbf{57.7}_{\pm\mathbf{0.0}}$	$\mathbf{55.8}_{\pm\mathbf{0.0}}$	$\mathbf{56.8}_{\pm\mathbf{0.0}}$
MEMO	$54.7_{\pm0.0}$	$55.6_{\pm0.0}$	$55.5_{\pm0.0}$	$57.2_{\pm0.0}$	$53.3_{\pm0.0}$	$56.2_{\pm0.0}$	$54.3_{\pm0.0}$	$54.5_{\pm0.0}$
+MUTE	$\mathbf{55.4}_{\pm\mathbf{0.0}}$	$55.1_{\pm0.0}$	$54.4_{\pm0.0}$	$\mathbf{58.9}_{\pm\mathbf{0.0}}$	$53.3_{\pm0.0}$	$\mathbf{57.0}_{\pm\mathbf{0.0}}$	$\mathbf{55.4}_{\pm\mathbf{0.0}}$	$\mathbf{56.6}_{\pm\mathbf{0.0}}$
SAR	$55.8_{\pm0.0}$	$56.1_{\pm0.0}$	$56.4_{\pm0.0}$	$57.5_{\pm0.0}$	$53.8_{\pm0.0}$	$56.6_{\pm0.0}$	$55.1_{\pm0.0}$	$55.5_{\pm0.0}$
+MUTE	$\mathbf{56.5}_{\pm\mathbf{0.0}}$	$\mathbf{56.7}_{\pm\mathbf{0.0}}$	$\mathbf{57.0}_{\pm\mathbf{0.0}}$	$\mathbf{58.9}_{\pm\mathbf{0.0}}$	$\mathbf{54.8}_{\pm\mathbf{0.0}}$	$\mathbf{57.6}_{\pm\mathbf{0.0}}$	$\mathbf{56.1}_{\pm\mathbf{0.0}}$	$\mathbf{57.0}_{\pm\mathbf{0.0}}$
RoTTA	$55.3_{\pm0.0}$	$56.3_{\pm0.1}$	$56.3_{\pm0.1}$	$58.3_{\pm0.0}$	$54.3_{\pm0.1}$	$57.2_{\pm0.0}$	$55.5_{\pm0.0}$	$56.1_{\pm0.0}$
+MUTE	$\mathbf{55.5}_{\pm\mathbf{0.0}}$	$56.1_{\pm0.0}$	$\mathbf{56.3}_{\pm\mathbf{0.0}}$	$\mathbf{58.7}_{\pm\mathbf{0.0}}$	$\mathbf{54.3}_{\pm\mathbf{0.0}}$	$\mathbf{57.7}_{\pm\mathbf{0.0}}$	$\mathbf{55.8}_{\pm\mathbf{0.0}}$	$\mathbf{56.4}_{\pm\mathbf{0.0}}$

Table 12: Computational overhead on an A100-80GB GPU. Runtime, throughput, and memory are Base/+MUTE; +MFLOPs is the extra computation.

Case	Time/corr.	Samples/s	ΔTime	Memory (GiB)	+MFLOPs
KS50-V EATA	127/128s	19.4/19.3	+0.6%	22.1/22.1	0.010
KS50-V SAR	161/160s	15.3/15.4	-0.9%	22.4/22.4	0.010
KS50-V TRIBE	171/170s	14.4/14.5	-0.6%	16.1/16.1	0.010
VGG-A SAR	912/1038s	15.4/13.5	+13.9%	22.4/22.4	0.059
VGG-A TRIBE	1172/1185s	12.0/11.8	+1.1%	16.1/16.1	0.059

Table 13: **Hyperparameter sensitivity analysis.** We evaluate the mutual-teaching weight $\lambda_{\max}$, advantage margin δ, and confidence threshold $\tau_{\min}$ on the KS50 video-corruption stream with READ as the base adapter, reporting average accuracy (%). Bold denotes the best result within each parameter sweep.

Parameter	**Evaluated Values**				
Teaching Weight (λ_{max})	0.05	0.10	0.50	1.00	2.00
Avg. Accuracy (%)	63.2	**64.1**	61.7	60.6	59.6
Advantage Margin (δ)	0.05	0.10	0.20	0.30	0.50
Avg. Accuracy (%)	63.7	**64.1**	64.0	63.9	63.9
Confidence Threshold ($\tau_{\min}$)	0.1	0.2	0.3	0.4	0.5
Avg. Accuracy (%)	64.0	64.1	**64.1**	63.7	62.4

Table 14: Stress tests. (a) Joint corruptions on KS50-C. (b) Missing-modality tests on KS50. Entries in (a)–(b) are Base/+MUTE accuracy with absolute changes in parentheses. (c) Corruption-severity accuracy reported as Base/+MUTE.

(a) Joint corruptions

Method	**Joint shift**	**Base/+MUTE (Δ)**
READ	crowd+motion	51.21/51.98 (+0.77)
SAR	crowd+motion	46.17/46.94 (+0.77)
TRIBE	rain+snow	21.61/24.92 (+3.31)
RoTTA	rain+snow	27.06/27.86 (+0.80)
RoTTA	traffic+gnoise	8.10/8.59 (+0.49)
TRIBE	traffic+gnoise	7.14/7.58 (+0.44)

(b) Missing modality

Method	**Condition**	**Base/+MUTE (Δ)**
READ	KS50-V drop audio	18.24/22.24 (+4.00)
READ	KS50-A drop video	12.07/12.20 (+0.13)
RoTTA	KS50-V drop audio	17.17/24.33 (+7.16)
RoTTA	KS50-A drop video	9.25/9.93 (+0.68)
SAR	KS50-A drop video	7.56/7.71 (+0.15)
TRIBE	KS50-A drop video	3.30/3.54 (+0.24)

(c) Corruption severity

Setting	**S1**	**S2**	**S3**	**S4**
READ KS50-V	77.77/77.86	74.66/74.79	65.72/71.68	42.45/67.70
READ VGG-A	36.74/37.52	35.81/36.71	34.48/35.52	32.99/33.97
READ VGG-V	60.02/60.77	58.57/59.42	57.64/58.38	56.57/57.25
SAR KS50-V	77.20/77.46	73.10/73.42	69.20/69.52	65.13/65.60