# ZORA: ZEROTH-ORDER RELIABILITY-CALIBRATED ADAPTATION WITH SHIFT-SENSITIVE UPDATES FOR MULTIMODAL TEST-TIME SHIFTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Audio-visual continual test-time adaptation (AV-CTTA) enables pretrained multimodal models to adapt online to non-stationary target streams. However, most existing AV-CTTA methods rely on backpropagation-based updates, limiting their deployment on edge platforms with quantized backbones, inference-only engines, or strict memory and latency constraints. Zeroth-order optimization (ZOO) provides a forward-only alternative, but its direct application to AV-CTTA faces two challenges: zeroth-order queries must explore a high-dimensional multimodal parameter space without shift awareness, and entropy minimization may reinforce incorrect predictions that exhibit high confidence and low entropy under online multimodal shifts. To address these limitations, we introduce **Z**eroth-**O**rder **R**eliability-Calibrated **A**daptation (**ZORA**), a forward-only AV-CTTA framework that determines both where ZOO should update and which learning signals it should trust. ZORA incorporates Shift-Sensitive Layer Selection (SSLS) to restrict zeroth-order updates to shift-sensitive normalization layers in the audio and visual branches. It further employs Bias-Aware Reliability Calibration (BARC) to assign signed entropy weights based on prediction bias and confidence, while aligning target-window feature statistics with cached source-reference statistics. Experiments on Kinetics50-C and VGGSound-C demonstrate that ZORA consistently outperforms representative CTTA, multimodal adaptation, and forward-only baselines across multiple continual audio-visual shift protocols. Experiments with quantized CAV-MAE models and efficiency analyses further support resource-constrained forward-only deployment.

## 1 INTRODUCTION

Pretrained models often degrade when deployed on target data whose distribution differs from the source training domain. Test-time adaptation (TTA) alleviates this problem by updating a model during inference using unlabeled test samples, enabling adaptation without labeled target data or source-domain retraining Liang et al. (2025); Wang et al. (2021). While early TTA studies mainly consider static target domains, real deployment environments are rarely stationary: illumination, weather, sensor conditions, background noise, and scene dynamics may change continuously over time. Continual test-time adaptation (CTTA) therefore extends TTA to non-stationary target streams, where the model must keep adapting throughout deployment Wang et al. (2022). This problem becomes more challenging in audio-visual perception, since the two modalities may shift with different timing, severity, and reliability Guo & Jin (2025); Zhao et al. (2025b); Zhang et al. (2025). For example, the visual stream can be affected by blur, weather, or lighting changes, while the audio stream can be corrupted by environmental noise or acoustic interference He et al. (2026); Zhao et al. (2025b). Meanwhile, audio-visual (AV) systems are often deployed on edge-side platforms, quantized backbones, inference engines, or hardware accelerators, where online updates must satisfy strict memory, computation, and latency constraints. However, existing AV-CTTA methods largely rely on backpropagation-based updates, which require activation storage and gradient computation Guo & Jin (2025); Zhang et al. (2025); He et al. (2026). In many practical deployment environments, gradients are unavailable or difficult to support Deng et al. (2025); Jia et al. (2024). This creates a fundamental mismatch: AV-CTTA is needed in edge-side online systems, but existing AV-CTTA methods often assume training-time capabilities that such systems cannot provide.



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.

Forward-only adaptation provides a promising direction for closing this gap, but existing methods are usually tied to specific architectures or predefined tunable components. For example, BN-based adaptation relies on updating normalization statistics Nado et al. (2020), while FOA optimizes ViT prompts through derivative-free search Niu et al. (2024). Such interfaces are difficult to uniformly apply to heterogeneous audio-visual models composed of structurally different audio encoders, visual encoders, and fusion modules. In contrast, zeroth-order optimization (ZOO) estimates update directions through forward perturbations and finite differences, without requiring explicit gradients or architecture-specific adaptation modules Nesterov & Spokoiny (2017); Liu et al. (2020); Spall (1992). This makes ZOO a general candidate for forward-only AV-CTTA.

However, simply replacing gradient-based adaptation with ZOO is insufficient for forward-only AV-CTTA. The difficulty is not merely the absence of gradients: ZOO must infer useful update directions from limited forward evaluations, while the target stream is non-stationary and the audio and visual modalities may shift differently. We identify two key bottlenecks. **First, naïve ZOO is not shift-selective.** Audio-visual models contain modality-specific encoders and fusion modules, whereas target shifts often manifest unevenly across branches and layers. Perturbing all adaptable parameters spends limited forward queries on shift-insensitive components and yields noisy update directions under the online budget. **Second, entropy minimization cannot distinguish confidence from correctness.** As audio-visual reliability changes over time, fused predictions can become biased toward unreliable modality cues and produce wrong high-confidence low-entropy (WHLE) outputs. Entropy minimization treats them as reliable adaptation signals, so ZOO may reinforce the bias and accumulate errors. As shown in Figure 1, naïve ZOO exhibits higher WHLE ratios across target domains, while lower WHLE is accompanied by consistently higher accuracy. Therefore, effective forward-only AV-CTTA requires ZOO to be *selective* in its update space and *reliability-aware* in its learning signal.

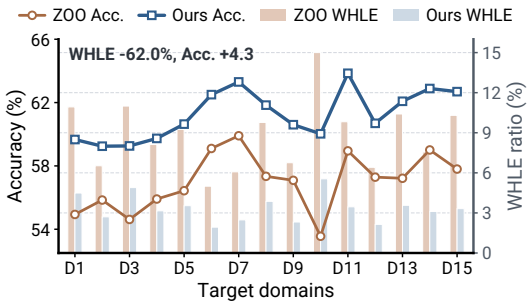



Figure 1: Lines report accuracy, and bars report the ratios of WHLE predictions. Compared with naïve ZOO, our proposed model consistently reduces WHLE ratios and improves accuracy.

To address these challenges, we propose **Z**eroth-**O**rder **R**eliability-calibrated **A**daptation (**ZORA**), a forward-only AV-CTTA framework. ZORA is built on a simple principle: forward queries should be allocated to shift-relevant parameters and guided by reliable learning signals. To this end, ZORA introduces two complementary modules: Shift-Sensitive Layer Selection (SSLS) and Bias-Aware Reliability Calibration (BARC). SSLS makes ZOO selective in its update space. Instead of perturbing all adaptable parameters, SSLS estimates layer-wise source-target discrepancy using cached source-reference statistics and the current target window, and selects the normalization layers that are most sensitive to the deployment shift. This branch-wise selection yields a compact and shift-relevant perturbation space, while keeping the fusion module frozen to avoid disturbing the learned cross-modal interaction structure. As a result, the limited forward-query budget is spent on parameters that are more likely to correct the current multimodal shift.

BARC makes ZOO reliability-aware in its learning signal. It estimates local prediction bias within a streaming window and combines it with confidence to assign signed entropy weights. Low-bias high-confidence samples provide positive adaptation signals, biased high-confidence samples receive suppressive negative weights, and uncertain samples are kept neutral. In addition, BARC aligns target-window feature statistics with cached source-reference statistics on the selected layers, providing a smoother distributional regularizer for forward-only optimization. Together, SSLS and BARC make ZOO both shift-selective in where it updates and reliability-aware in what signal it trusts. The contributions of this work can be summarized as follows:

- We revisit AV-CTTA from a forward-only perspective, motivated by edge-side multimodal deployment where backpropagation may be unavailable or impractical. We identify two key bottlenecks that make naïve ZOO ineffective in this setting: shift-agnostic parameter perturbation and unreliable entropy-driven learning signals under online multimodal shifts.

- We propose ZORA, a **Z**eroth-**O**rder **R**eliability-calibrated **A**daptation framework that makes ZOO selective in where it updates and reliability-aware in what signal it trusts, through Shift-Sensitive Layer Selection and Bias-Aware Reliability Calibration.
- We validate ZORA on Kinetics50-C and VGGSound-C under multiple continual audio-visual shift protocols. ZORA consistently improves over representative CTTA, multimodal adaptation, and forward-only baselines, while maintaining effectiveness under quantized CAV-MAE deployment and achieving favorable query-efficiency trade-offs.

## 2 Related Work

**Continual Test-Time Adaptation (CTTA)** Test-time adaptation (TTA) adapts a pretrained model to unlabeled target data during inference to mitigate source-target distribution shifts Liang et al. (2025); Wang et al. (2021). Continual TTA (CTTA) further considers non-stationary target streams, where the test distribution evolves over time due to mixed domains, gradual transitions, or temporally varying corruptions Wang et al. (2022). Recent work has extended test-time adaptation to multimodal settings, including 2D–3D perception and audio-visual recognition, where modalities can undergo distinct or asynchronous shifts Cao et al. (2023); Guo & Jin (2025); Zhao et al. (2025b); Zhang et al. (2025). In audio-visual settings, prior methods improve robustness through reliability-aware fusion and adaptation Yang et al. (2024); Zhao et al. (2025b), stable multimodal updates Guo & Jin (2025), or analytic continual learning with dynamic modality selection Zhang et al. (2025); complementary audio-only work studies adaptation under background noise Shao et al. (2025). Despite their effectiveness, audio-visual methods mostly rely on backpropagation-based online updates, which require gradient computation and activation storage. This assumption is restrictive for edge-side audio-visual deployment, where models may be quantized, executed by inference-only engines, or constrained by memory, latency, and energy budgets Jia et al. (2024); Lu et al. (2026); Deng et al. (2025). In contrast, our work studies stable continual audio-visual adaptation when only forward passes are supported.

**Zeroth-Order and Forward-Only Optimization** Forward-only adaptation has emerged as a practical alternative to backpropagation-based test-time updates. Existing methods reduce deployment overhead by updating normalization statistics Nado et al. (2020); Lim et al. (2023), adjusting classifier or prototype state Iwasawa & Matsuo (2021); Zhao et al. (2025a), or optimizing lightweight components through derivative-free search Niu et al. (2024); Wang & Wang (2026). However, these methods typically depend on architecture-specific adaptation interfaces, such as BatchNorm statistics, classifier prototypes, or ViT prompts, which are difficult to directly generalize to heterogeneous audio-visual models with modality-specific encoders and fusion modules. Zeroth-order optimization (ZOO) provides a more general forward-only mechanism by estimating update directions from objective-value differences under random perturbations, without requiring explicit gradients or predefined tunable modules Spall (1992); Nesterov & Spokoiny (2017); Liu et al. (2020). Nevertheless, naïvely applying ZOO to AV-CTTA remains ineffective. The multi-branch audio-visual architecture induces a large and coupled perturbation space, making random search inefficient Ma & Huang (2025); Yu et al. (2025). Entropy-based unsupervised objectives can also become unreliable under online shifts Wang et al. (2022); Lee et al. (2024b); Wang et al. (2025); because finite differences are computed from these objective values, that unreliability directly perturbs the estimated direction. Different from architecture-bound forward-only methods and vanilla ZOO, our method develops a selective and reliability-calibrated zeroth-order framework that restricts updates to shift-sensitive modality-specific normalization layers and constructs bias-aware forward objectives.

## 3 Methodology

Forward-only adaptation is not merely a gradient-free substitute for backpropagation; its success depends on whether limited forward queries are allocated to shift-relevant parameters and guided by reliable forward-evaluable objectives. This is particularly important for AV-CTTA, where multimodal models introduce large and heterogeneous parameter spaces, and online cross-modal shifts can make entropy minimization reinforce biased confidence. ZORA addresses these two issues from two perspectives: *where to update* and *what signal to trust*. Specifically, Shift-Sensitive Layer Selection (SSLS) makes ZOO shift-selective by restricting perturbations to shift-sensitive normalization

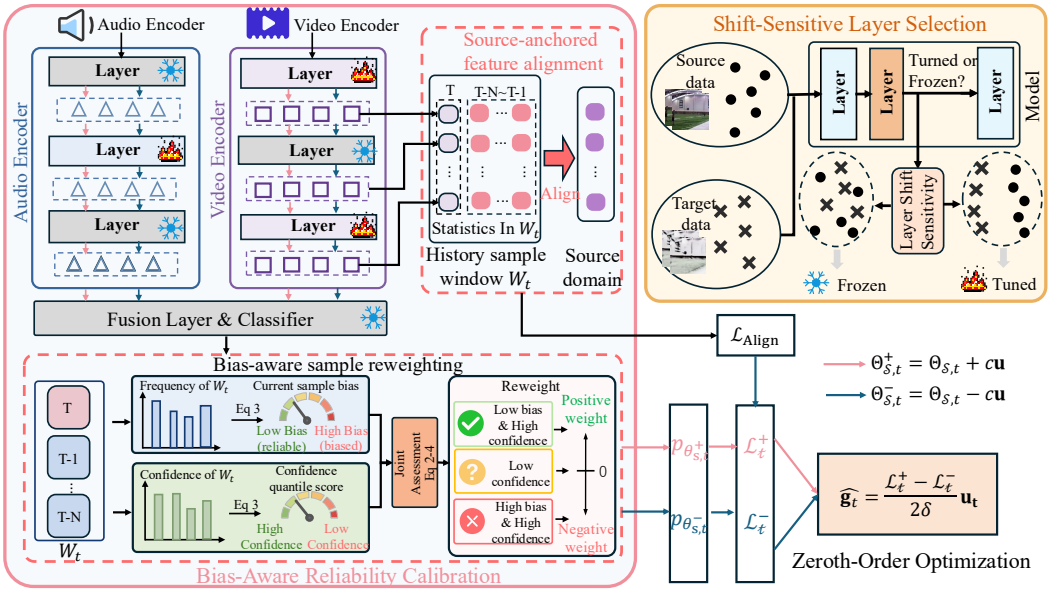



Figure 2: Overview of the proposed ZORA framework for forward-only audio-visual continual test-time adaptation. ZORA consists of two coordinated modules: (a) SSLS, which selects shift-sensitive normalization layers in audio and visual branches as a compact zeroth-order update space; and (b) BARC, which constructs reliability-calibrated objectives by suppressing biased confident errors and aligning source-reference and target-window statistics. Together, SSLS and BARC respectively address the two bottlenecks of zeroth-order online adaptation: query efficiency and signal reliability.

layers in modality branches, while Bias-Aware Reliability Calibration (BARC) constructs reliability-calibrated objectives by suppressing biased confident predictions and anchoring target-window feature statistics to cached source-reference statistics. The overall framework of ZORA is in Figure 2.

### 3.1 Problem Definition

We formulate audio-visual continual test-time adaptation (AV-CTTA) as an online classification problem over a non-stationary unlabeled target stream. Let $f_\Theta$ denote a pretrained audio-visual model consisting of an audio encoder $f^a_{\Theta_a}$, a visual encoder $f^v_{\Theta_v}$, a multimodal fusion module $f^u_{\Theta_u}$, and a classifier $f^c_{\Theta_c}$, with parameters $\Theta = \{\Theta_a, \Theta_v, \Theta_u, \Theta_c\}$. The model is trained on a labeled source domain $\mathcal{D}_s = \{(\mathbf{x}^a_i, \mathbf{x}^v_i, y_i)\}_{i=1}^{N_s} \sim P_s(\mathbf{x}, y)$, where each sample contains paired audio and visual inputs. During deployment, the model receives an unlabeled target stream $\mathcal{D}_t = \{(\mathbf{x}^a_t, \mathbf{x}^v_t)\}_{t=1}^{T}$, where samples arrive sequentially from a time-varying distribution $Q_t(\mathbf{x})$. The target distribution may evolve over time, and the audio and visual modalities may shift with different severities and temporal patterns. Under the forward-only constraint, the model can evaluate unsupervised objectives through forward passes, but cannot compute gradients or store activations for backpropagation. Accordingly, online adaptation must rely on zeroth-order objective evaluations.

### 3.2 Shift-Sensitive Layer Selection (SSLS)

ZOO estimates update directions from objective-value differences under random perturbations. Its efficiency therefore depends on both the dimensionality and the relevance of the perturbed parameter space. For audio-visual models, perturbing all adaptable parameters is inefficient because the audio encoder, visual encoder, and fusion module form a large and heterogeneous search space. However, merely reducing the number of updated layers is also insufficient: perturbing shift-insensitive layers may introduce noisy objective variations and ineffective update directions. This motivates the key principle of SSLS: the ZOO update space should be not only compact, but also shift-relevant.

To this end, SSLS selects normalization layers that exhibit strong source-target feature separability and uses them as the ZOO update space. We focus on normalization layers in the audio and visual branches for two reasons. First, deployment shifts are largely reflected in modality-specific encoders,

making their normalization parameters effective handles for test-time correction. Second, the fusion module encodes source-trained cross-modal interaction patterns; directly perturbing it may distort the learned fusion relationship and destabilize the fused predictions used for reliability estimation. Therefore, SSLS freezes the fusion module and classifier, and performs branch-wise layer selection only within the audio and visual encoders.

**Layer Shift Sensitivity** SSLS measures how strongly each candidate layer responds to the deployment shift using a lightweight source-target separability proxy. Before online adaptation, we use a small source-reference subset and a small early-target subset: $\mathcal{D}_s^{\mathrm{SR}} = \{\mathbf{x}_i^s\}_{i=1}^N$ and $\mathcal{D}_t^{\mathrm{ET}} = \{\mathbf{x}_i^t\}_{i=1}^N$. The source-reference subset can be cached before deployment, while the early-target subset is collected from the beginning of the unlabeled target stream and requires no labels. For each candidate normalization layer $l$ in modality branch $m \in \{a, v\}$, we extract token-level features $\mathbf{Z}_{i,l}^{s,m} \in \mathbb{R}^{M\times d}$ and $\mathbf{Z}_{i,l}^{t,m} \in \mathbb{R}^{M\times d}$, where $M$ and $d$ denote the number of tokens and the feature dimension.

For each token position $r$, we collect the corresponding source and target features and partition them into two clusters. If a layer is sensitive to the target shift, its source and target features should be more separable, leading to higher clustering purity. We define token-wise Layer Shift Sensitivity (LSS):

$$\mathrm{LSS}_{l,r}^m = \frac{1}{2N}\sum_{q=1}^{2}\max_{c\in\{s,t\}}\left|\mathcal{C}_{q,l,r}^m \cap \mathcal{Y}_{c,l,r}^m\right|, \tag{1}$$

where $\mathcal{C}_{q,l,r}^m$ denotes the $q$-th cluster at layer $l$ and token position $r$ in branch $m$, and $\mathcal{Y}_{c,l,r}^m$ denotes the feature set from domain $c \in \{s, t\}$ at the same layer and token position. The layer-level sensitivity is obtained by averaging over token positions: $\mathrm{LSS}_l^m = \frac{1}{M}\sum_{r=1}^{M}\mathrm{LSS}_{l,r}^m$. A larger $\mathrm{LSS}_l^m$ indicates stronger source-target separability at layer $l$ in branch $m$, suggesting that this layer is more responsive to the deployment shift. SSLS ranks candidate normalization layers by their LSS scores within each modality branch and retains the top-$K$ layers as the selected set $\mathcal{S}_m$. Together, $\mathcal{S}_a$ and $\mathcal{S}_v$ define the test-time perturbation subspace: ZOO perturbs and updates only the affine normalization parameters in these selected layers, while all remaining parameters are frozen. In this way, SSLS converts the original high-dimensional adaptation problem into a compact and shift-focused search space for query-efficient forward-only adaptation. The selected sets remain fixed online; Appendices D.1–D.2 provide the full procedure and robustness/cost analyses.

### 3.3 Bias-Aware Reliability Calibration (BARC)

After SSLS determines a compact update space, the remaining challenge is to construct a forward-evaluable objective that yields reliable zeroth-order directions. This is non-trivial because ZOO estimates directions from differences in unsupervised objective values; noisy or misleading objectives can therefore be amplified into erroneous parameter updates. In AV-CTTA, online cross-modal shifts may bias fused predictions toward locally over-amplified classes, producing wrong yet confident low-entropy outputs. Direct entropy minimization treats such predictions as reliable signals and may reinforce the error. BARC addresses this issue with two components: bias-aware signed entropy reweighting, which calibrates sample-level entropy signals, and source-anchored feature alignment, which provides a smoother window-level structural constraint for ZOO.

**Bias-Aware Signed Entropy Reweighting** Given the current sample $\mathbf{x}_t$, we maintain a local history window $\mathcal{W}_t = \{\mathbf{x}_{t-W+1}, \ldots, \mathbf{x}_t\}$. For each sample $\mathbf{x}_i \in \mathcal{W}_t$, let $p_i = p(\mathbf{x}_i)$ denote its predictive distribution and $\hat{y}_i = \arg\max_c p_{i,c}$ denote its predicted class. To identify locally over-amplified predictions, we define the prediction frequency of class $c$ in the window as

$$F_t(c) = \frac{1}{|\mathcal{W}_t|}\sum_{\mathbf{x}_i\in\mathcal{W}_t}\mathbb{1}(\hat{y}_i = c). \tag{2}$$

The local prediction-bias score of $\mathbf{x}_i$ is then defined as

$$B_i = F_t(\hat{y}_i) - \bar{F}_t, \qquad \bar{F}_t = \frac{1}{|\mathcal{W}_t|}\sum_{\mathbf{x}_j\in\mathcal{W}_t}F_t(\hat{y}_j), \tag{3}$$

where $\bar{F}_t$ is the average predicted-class frequency within the window. A larger $B_i$ indicates that the predicted class of $\mathbf{x}_i$ appears more frequently than the local average, suggesting a higher risk of prediction bias. We also define the confidence score as $C_i = \max_c p_{i,c}$.

BARC assigns a signed entropy weight by jointly considering local prediction bias and confidence. Low-bias high-confidence samples are used as reliable positive signals, high-bias high-confidence samples are treated as biased confident predictions and explicitly suppressed, and the remaining samples are kept neutral to avoid uncertain entropy signals. We define the high-confidence set and its positive/negative subsets as $\mathcal{H}_t = \{\mathbf{x}_i \in \mathcal{W}_t \mid C_i \geq \tau_c\}$, $\mathcal{R}_t^+ = \{\mathbf{x}_i \in \mathcal{H}_t \mid B_i \leq \tau_b\}$, and $\mathcal{R}_t^- = \{\mathbf{x}_i \in \mathcal{H}_t \mid B_i > \tau_b\}$, where $\tau_b$ and $\tau_c$ are the bias and confidence thresholds. For a scalar score $s_i$ and a sample set $\mathcal{A}$, let $q_{\mathcal{A}}(s_i) \in [0,1]$ denote the percentile-rank score of $s_i$ within $\mathcal{A}$. We compute $Q_C(\mathbf{x}_i) = q_{\mathcal{H}_t}(C_i)$, $Q_B^+(\mathbf{x}_i) = q_{\mathcal{R}_t^+}(-B_i)$, and $Q_B^-(\mathbf{x}_i) = q_{\mathcal{R}_t^-}(B_i)$, so that higher confidence receives larger $Q_C$, lower bias receives larger $Q_B^+$, and stronger bias receives larger $Q_B^-$.

For the current sample $\mathbf{x}_t$, the signed weighting indicator is

$$I_t = \begin{cases} Q_C(\mathbf{x}_t)Q_B^+(\mathbf{x}_t), & \mathbf{x}_t \in \mathcal{R}_t^+, \\ -\gamma Q_C(\mathbf{x}_t)Q_B^-(\mathbf{x}_t), & \mathbf{x}_t \in \mathcal{R}_t^-, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\gamma \geq 0$ controls the maximum suppression strength. The signed entropy objective is

$$\mathcal{L}_t^{\text{ent}} = -I_t \sum_{c \in \mathcal{C}} p_c(\mathbf{x}_t) \log p_c(\mathbf{x}_t). \tag{5}$$

When $I_t > 0$, minimizing $\mathcal{L}_t^{\text{ent}}$ performs confidence- and bias-weighted entropy minimization on reliable samples. When $I_t < 0$, the objective reverses the entropy signal and discourages biased confident predictions, with stronger suppression for higher confidence and stronger local prediction bias. When $I_t = 0$, the sample does not contribute to the entropy-driven ZOO direction. Prediction frequency is a risk signal, not a correctness criterion; Appendices D.3–D.4 report weighting variants and non-IID stress tests.

**Source-Anchored Feature Alignment** Although signed entropy reweighting reduces misleading sample-level signals, entropy objectives may still fluctuate under continual shifts. Since ZOO relies on objective-value differences, such fluctuations can produce high-variance direction estimates. BARC therefore introduces a feature-statistics alignment term that anchors the target window to cached source-reference statistics on the SSLS-selected layers.

Specifically, we cache source-reference mean and standard deviation statistics, denoted by $\mu_l^{S,m}$ and $\sigma_l^{S,m}$, for candidate layers in each modality branch $m \in \{a, v\}$. After SSLS selects $\mathcal{S}_m$, BARC uses the corresponding source statistics for layers $l \in \mathcal{S}_m$. At test time, we compute the target-window statistics $\mu_l^m(\mathcal{W}_t)$ and $\sigma_l^m(\mathcal{W}_t)$ using the local window. The source-anchored alignment objective is

$$\mathcal{L}_t^{\text{align}} = \sum_{m \in \{a,v\}} \sum_{l \in \mathcal{S}_m} \left( \left\| \mu_l^m(\mathcal{W}_t) - \mu_l^{S,m} \right\|_2^2 + \left\| \sigma_l^m(\mathcal{W}_t) - \sigma_l^{S,m} \right\|_2^2 \right). \tag{6}$$

This term requires no target labels and is computed only on the selected layers. It complements the signed entropy objective by providing a smoother forward-evaluable signal for zeroth-order adaptation.

### 3.4 Forward-Only Zeroth-Order Update

With the SSLS-selected update space and the BARC-calibrated objective, ZORA performs online adaptation through zeroth-order optimization without backpropagation. At each time step $t$, the forward-evaluable objective is

$$\mathcal{L}_t = \mathcal{L}_t^{\text{ent}} + \lambda \mathcal{L}_t^{\text{align}}, \tag{7}$$

where $\lambda$ balances the bias-aware entropy signal and the source-anchored feature alignment term.

Let $\Theta_{\mathcal{S}}$ denote the affine normalization parameters associated with the SSLS-selected layer sets $\mathcal{S}_a$ and $\mathcal{S}_v$. During online adaptation, ZORA perturbs and updates only $\Theta_{\mathcal{S}}$, while keeping all remaining branch parameters, the fusion module, and the classifier frozen. Given a random perturbation direction $\mathbf{u}_t$ with the same dimensionality as $\Theta_{\mathcal{S}}$, the two-point zeroth-order update is

$$\hat{\mathbf{g}}_t = \frac{\mathcal{L}_t(\Theta_{\mathcal{S}} + \delta \mathbf{u}_t) - \mathcal{L}_t(\Theta_{\mathcal{S}} - \delta \mathbf{u}_t)}{2\delta} \mathbf{u}_t, \qquad \Theta_{\mathcal{S}} \leftarrow \Theta_{\mathcal{S}} - \eta \hat{\mathbf{g}}_t, \tag{8}$$

Table 1: Five-seed accuracy (mean ± standard deviation) on Kinetics50-C and VGGSound-C. ICS: Interleaved corruption streams. PACS: PairTrack-Async corruption streams.

| **Method** | Source | T3A | SAR | TSA | PTA | BECoTTA | Tent | READ | FOZO | CMAE | DeYO | DASP | FOA | ZORA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ICS: K50-C | 61.47$_{\pm.47}$ | 49.70$_{\pm1.06}$ | 41.59$_{\pm1.48}$ | 55.51$_{\pm.91}$ | 63.58$_{\pm.90}$ | 63.54$_{\pm.63}$ | 63.62$_{\pm.60}$ | 64.10$_{\pm.24}$ | 63.73$_{\pm.16}$ | 63.62$_{\pm.44}$ | 63.65$_{\pm.73}$ | 63.54$_{\pm.47}$ | 63.36$_{\pm.51}$ | **66.59**$_{\pm.74}$ |
| ICS: VGG-C | 41.54$_{\pm.09}$ | 17.67$_{\pm.59}$ | 24.85$_{\pm.94}$ | 24.93$_{\pm.65}$ | 43.75$_{\pm.32}$ | 44.08$_{\pm.28}$ | 45.72$_{\pm1.07}$ | 44.98$_{\pm1.01}$ | 44.91$_{\pm.84}$ | 45.54$_{\pm1.10}$ | 44.80$_{\pm.42}$ | 44.11$_{\pm.17}$ | 44.21$_{\pm.99}$ | **47.58**$_{\pm.54}$ |
| PACS: K50-C | 44.47$_{\pm.02}$ | 30.02$_{\pm.47}$ | 27.45$_{\pm.98}$ | 43.28$_{\pm1.36}$ | 35.23$_{\pm2.73}$ | 44.85$_{\pm.21}$ | 42.13$_{\pm.17}$ | 46.71$_{\pm.13}$ | 44.12$_{\pm.25}$ | 44.39$_{\pm.04}$ | 45.51$_{\pm.19}$ | 44.40$_{\pm.02}$ | 50.68$_{\pm.36}$ | **55.64**$_{\pm.67}$ |
| PACS: VGG-C | 12.62$_{\pm.45}$ | 3.58$_{\pm.16}$ | 7.19$_{\pm.10}$ | 12.68$_{\pm.50}$ | 4.83$_{\pm.10}$ | 7.57$_{\pm1.32}$ | 8.76$_{\pm.24}$ | 5.27$_{\pm.22}$ | 10.70$_{\pm.35}$ | 12.50$_{\pm.40}$ | 12.56$_{\pm.34}$ | 12.63$_{\pm.42}$ | 17.93$_{\pm.44}$ | **22.06**$_{\pm.45}$ |

where $\delta$ is the perturbation radius and $\eta$ is the update step size. All objective values in Eq. (8) are obtained by forward passes only, without activation storage or gradient computation. Thus, ZORA couples a compact shift-sensitive perturbation space with a reliability-calibrated forward objective to perform online AV-CTTA.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets & Protocols** We evaluate ZORA on two audio-visual corruption benchmarks, Kinetics50-C and VGGSound-C, following prior multimodal TTA settings Kay et al. (2017); Chen et al. (2020); Guo & Jin (2025). Each benchmark contains 15 video corruptions and 6 audio corruptions with five severity levels. To evaluate continual multimodal adaptation, we construct two streaming protocols: interleaved corruption streams (ICS), where audio and video corruptions appear alternately while the other modality remains clean, and PairTrack-Async corruption streams (PACS), where both modalities are corrupted and their severities evolve asynchronously within fixed audio-video corruption pairs. ICS evaluates adaptation under alternating modality-specific shifts, while PACS further tests stability under dynamic dual-modality shifts.

**Baselines & Implementation** We compare ZORA with representative CTTA, multimodal adaptation, and backpropagation-free methods, including T3A Iwasawa & Matsuo (2021), SAR Niu et al. (2023), TSA Chen et al. (2025), PTA Wang et al. (2025), BECoTTA Lee et al. (2024a), Tent Wang et al. (2021), READ Yang et al. (2024), ContinualMAE Liu et al. (2024), DASP He et al. (2026), FOZO Wang & Wang (2026), DeYO Lee et al. (2024b), and FOA Niu et al. (2024). All methods are evaluated with the same CAV-MAE-based backbone and the same single-sample streaming protocol. For ZORA, we use a local window length of $W = 24$, select $K = 3$ layers in each modality branch, and perform one ZOO step per test sample. We set the perturbation radius, update step size, and alignment coefficient to $\delta = 10^{-3}$, $\eta = 10^{-2}$, and $\lambda = 0.5$, respectively. All experiments are repeated with five random seeds. Appendices A–C detail stream construction, checkpoints, and baseline configurations.

## 4.2 Main Results

Table 1 summarizes four settings over five seeds; confidence intervals and domain-wise results appear in Appendices E.1 and F. Under ICS, ZORA reaches 66.59% on Kinetics50-C and 47.58%

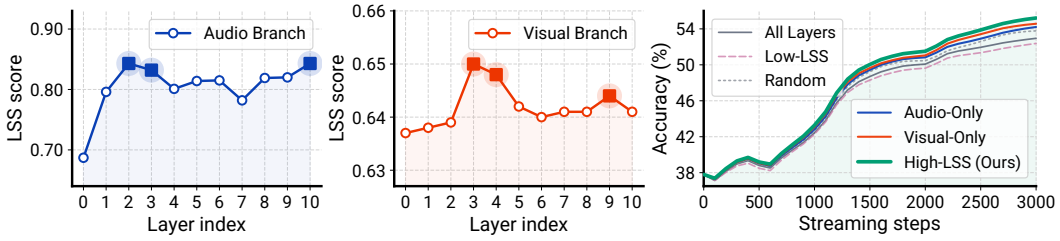



Figure 3: SSLS analysis on Kinetics50-C PairTrack-Async: high-LSS branch-wise selection identifies shift-sensitive layers and achieves the best convergence.

on VGGSound-C, exceeding the strongest baselines by 2.49 and 1.86 points and confirming its effectiveness under alternating single-sample modality shifts.

Under the harder PACS protocol, ZORA reaches 55.64% and 22.06%, surpassing FOA by 4.96 and 4.13 points and the frozen source model by 11.17 and 9.44 points, respectively, which validates shift-sensitive, reliability-calibrated adaptation under dynamic multimodal shifts.

### 4.3 Ablation and Analysis

**Ablation Study** Table 2 evaluates the contribution of each component under the PACS protocol. RF, SR, and DA denote reliable filtering, signed reweighting, and distribution alignment, respectively. Naïve entropy-driven ZOO obtains only 44.31% on Kinetics50-C and 11.96% on VGGSound-C, showing that entropy minimization alone provides unreliable guidance under asynchronous multimodal shifts.

Table 2: Ablation study under PACS.

| Method | RF | SR | DA | SSLS | KS50 | VGG |
|---|---|---|---|---|---|---|
| Naïve ZOO | | | | | 44.31 | 11.96 |
| + RF | ✓ | | | | 46.50 | 13.23 |
| + RF + SR | ✓ | ✓ | | | 50.50 | 17.49 |
| + BARC | ✓ | ✓ | ✓ | | 54.37 | 21.53 |
| ZORA | ✓ | ✓ | ✓ | ✓ | **55.64** | **22.06** |

Adding RF improves the results to 46.50% and 13.23%, indicating that filtering unreliable samples can reduce misleading updates. Introducing SR further increases the accuracy to 50.50% and 17.49%, suggesting that explicitly suppressing biased confident samples is more effective than only removing unreliable ones. With DA, the full BARC objective reaches 54.37% and 21.53%, showing that source-anchored feature alignment provides a smoother window-level signal for ZOO. Finally, adding SSLS achieves the best performance, demonstrating that BARC and SSLS are complementary: BARC calibrates the forward learning signal, while SSLS restricts perturbations to a compact and shift-relevant update space.

**Analysis of SSLS** Figure 3 examines whether LSS can identify effective update layers for ZOO. The left and middle panels show the layer-wise LSS scores in the audio and visual branches, respectively. The highlighted layers differ across modalities: the audio branch selects layers 2, 3, and 10, while the visual branch selects layers 3, 4, and 9. This indicates that audio and visual encoders respond to the same deployment shift at different depths, supporting branch-wise layer selection rather than a unified update rule. The right panel compares different update-layer choices under the same ZOO query budget. Updating high-LSS layers achieves the fastest convergence and the highest final accuracy. In contrast, updating all layers introduces unnecessary perturbation dimensions, while low-LSS, random, audio-only, or visual-only selections provide weaker adaptation. These results support the key principle of SSLS: effective forward-only adaptation requires an update space that is not only compact, but also shift-relevant and modality-aware.

**Analysis of BARC** Figure 4 examines BARC's reliability-calibrated learning signal. Correct high-confidence samples receive positive weights, wrong high-confidence samples receive negative weights, and low-confidence samples remain near zero. Compared with entropy-only ZOO, BARC also reduces error concentration on over-amplified classes: wrong predictions to Classes 1 and 18 fall

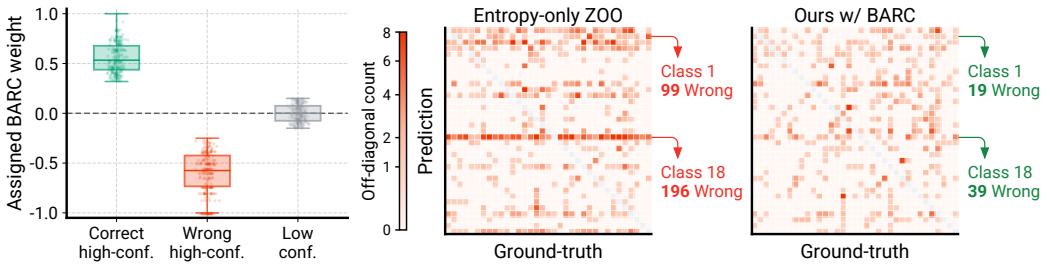



Figure 4: Analysis of BARC on Kinetics50-C PairTrack-Async. Left: BARC assigns positive weights to correct high-confidence samples and negative weights to wrong high-confidence samples. Middle and right: compared with entropy-only ZOO, BARC reduces prediction concentration on over-amplified classes, e.g., wrong predictions to Classes 1 and 18 drop from 99/196 to 19/39.

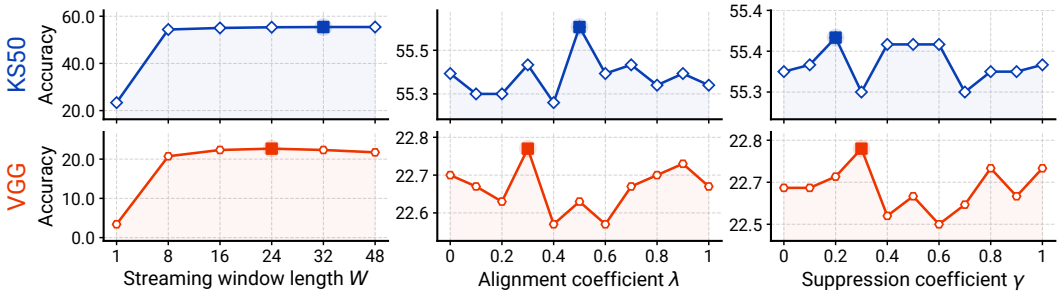



Figure 5: Hyperparameter sensitivity on the two datasets under the PairTrack-Async setting.

Table 3: Accuracy (%) with quantized CAV-MAE backbones. ICS and PACS denote interleaved corruption streams and PairTrack-Async corruption streams, respectively.

| Method | Kinetics50-C | | | | VGGSound-C | | | |
|---|---|---|---|---|---|---|---|---|
| | ICS | | PACS | | ICS | | PACS | |
| | W8A8 | W4A4 | W8A8 | W4A4 | W8A8 | W4A4 | W8A8 | W4A4 |
| FOA | 65.25 | 45.98 | 47.40 | 28.07 | 46.32 | 17.89 | 15.13 | 4.63 |
| T3A | 49.85 | 24.44 | 28.80 | 13.63 | 17.67 | 5.88 | 3.37 | 1.33 |
| FOZO | 54.50 | 35.16 | 42.97 | 24.33 | 29.76 | 10.97 | 14.70 | 3.57 |
| **ZORA** | **66.93** | **46.84** | **54.90** | **33.60** | **47.32** | **18.12** | **22.17** | **6.00** |

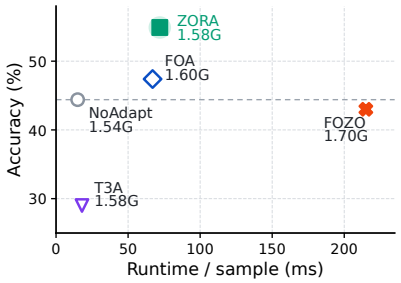



Figure 6: Accuracy–runtime trade-off with peak-memory annotations.

from 99/196 to 19/39. This confirms that BARC promotes reliable entropy signals while suppressing misleading confident predictions.

**Hyperparameter Sensitivity** We evaluate the sensitivity of ZORA to three key hyperparameters: the local window length $W$, the source-anchored alignment coefficient $\lambda$, and the biased-confidence suppression coefficient $\gamma$. As shown in Figure 5, ZORA remains stable across a broad range of values, indicating that its gains do not rely on a narrow hyperparameter configuration.

**Quantized Backbone Results** To evaluate the practicality of ZORA under resource-constrained deployment, we further conduct experiments with quantized CAV-MAE backbones. Specifically, the audio and visual branches are quantized under W8A8 and W4A4 settings, while the fusion attention remains in full precision. As shown in Table 3, ZORA achieves the best performance across all datasets, stream protocols, and quantization settings. Compared with strongest backpropagation-free baseline FOA, ZORA obtains an average relative improvement of 9.31%, with 9.89% under W8A8 and 8.27% under W4A4. These results indicate that ZORA remains effective even when the backbone is quantized, supporting its applicability to forward-only adaptation in resource-constrained scenarios.

**Efficiency Comparison** Figure 6 compares per-sample runtime, peak memory, and adaptation accuracy. ZORA achieves a favorable trade-off because SSLS restricts the update space and BARC calibrates forward objectives without backpropagation. Deployment and transfer analyses are provided in Appendices E.2–E.3.

## 5 Conclusion

We study forward-only AV-CTTA for resource-constrained deployment and show that naïve ZOO is hindered by shift-insensitive perturbations and WHLE-reinforcing entropy minimization. ZORA combines SSLS for shift-sensitive update-space selection with BARC for reliability-calibrated forward objectives. Across Kinetics50-C and VGGSound-C, ZORA consistently outperforms CTTA, multimodal adaptation, and forward-only baselines under multiple continual shifts. Quantized-backbone and efficiency results further demonstrate its practicality, while future work will extend ZORA to larger multimodal models and realistic streams.

## AI USE STATEMENT

Generative AI tools were used to assist with experimental design and implementation, code editing, result analysis and interpretation, literature and citation checking, figure preparation, and language and formatting revisions. AI-assisted outputs were treated as drafts and reviewed by the authors before inclusion. The authors take full responsibility for the final content of this work, including all text, claims, code, experimental results, figures, and references produced or revised with AI assistance.

## REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2236–2246, 2018.

Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, and Lihua Xie. Multi-modal continual test-time adaptation for 3D semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 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 (ICASSP)*, pp. 721–725, 2020.

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 multi-modal data. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 9711–9727. PMLR, 2025.

Zeshuai Deng, Guohao Chen, Shuaicheng Niu, Hui Luo, Shuhai Zhang, Yifan Yang, Renjie Chen, Wei Luo, and Mingkui Tan. Test-time model adaptation for quantized neural networks. In *ACM International Conference on Multimedia (ACM MM)*, pp. 7258–7267, 2025. doi: 10.1145/3746027.3754730.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *International Conference on Learning Representations (ICLR)*, pp. 2215–2232, 2025.

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

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 2427–2440, 2021.

Hong Jia, Young D. Kwon, Alessio Orsino, Ting Dang, Domenico Talia, and Cecilia Mascolo. TinyTTA: Efficient test-time adaptation via early-exit ensembles on edge devices. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 43274–43299, 2024.

Will Kay, Joao 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. *arXiv preprint arXiv:1705.06950*, 2017.

Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. BECoTTA: Input-dependent online blending of experts for continual test-time adaptation. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 27072–27093. PMLR, 2024a.

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 (ICLR)*, pp. 46977–47002, 2024b.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision (IJCV)*, 133(1):31–64, 2025. doi: 10.1007/s11263-024-02181-w.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, and Shanghang Zhang. Continual-MAE: Adaptive distribution masked autoencoders for continual test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28653–28663, 2024.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O. Hero, and Pramod K. Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Anqi Lu, Youbing Hu, Yun Cheng, Dawei Wei, Zhiqiang Cao, Jie Liu, and Zhijun Li. TeamTTA: Efficient multi-device collaboration for open-set test-time adaptation via cloud integration. *Journal of Artificial Intelligence Research*, 85, 2026.

Shaocong Ma and Heng Huang. Revisiting zeroth-order optimization: Minimum-variance two-point estimators and directionally aligned perturbations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, pp. 13267–13293, 2025.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv:2006.10963*, 2020.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

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 (ICLR)*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 38298–38315. PMLR, 2024.

Weichuang Shao, Iman Yi Liao, Tomas Henrique Bode Maul, and Tissa Chandesa. An investigation of test-time adaptation for audio classification under background noise. *arXiv:2507.15523*, 2025.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

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 (ICLR)*, 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 (NeurIPS)*, volume 38, pp. 2575–2601, 2025.

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

Xingyu Wang and Tao Wang. FOZO: Forward-only zeroth-order prompt optimization for test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7936–7945, 2026.

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

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of LLMs in random subspaces. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4475–4485, 2025.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

Yufei Zhang, Yicheng Xu, Hongxin Wei, Zhiping Lin, Xiaofeng Zou, Cen Chen, and Huiping Zhuang. Analytic continual test-time adaptation for multi-modality corruption. In *ACM International Conference on Multimedia (ACM MM)*, pp. 1929–1937, 2025.

Jichun Zhao, Haiyong Jiang, Haoxuan Song, Jun Xiao, and Dong Gong. $D^3$CTTA: Domain-dependent decorrelation for continual test-time adaption of 3D lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11864–11874, 2025a.

Yusheng Zhao, Junyu Luo, Xiao Luo, Jinsheng Huang, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. Attention bootstrapping for multi-modal test-time adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22849–22857, 2025b.

# A BENCHMARK AND STREAM CONSTRUCTION

This appendix provides benchmark construction, pretrained-model and implementation details, extended analyses of SSLS and BARC, statistical and deployment evaluations, and complete benchmark tables.

We construct two audio-visual corruption benchmarks based on Kinetics50 Kay et al. (2017) and VGGSound Chen et al. (2020) to evaluate online adaptation under continual multimodal domain shifts. Unlike conventional static evaluation, all protocols are evaluated in a single-sample streaming setting with batch size 1. We consider two continual evaluation setups: interleaved single-modality corruption and paired multimodal co-shift corruption. In the main paper, we abbreviate these two protocols as ICS and PACS, respectively; in this appendix, we keep the expanded names for readability. The interleaved setup alternates corruptions between audio and video over time, while keeping the other modality clean. The paired co-shift setup corrupts audio and video simultaneously according to fixed audio-video corruption pairs, providing a more challenging scenario where both modalities shift together.

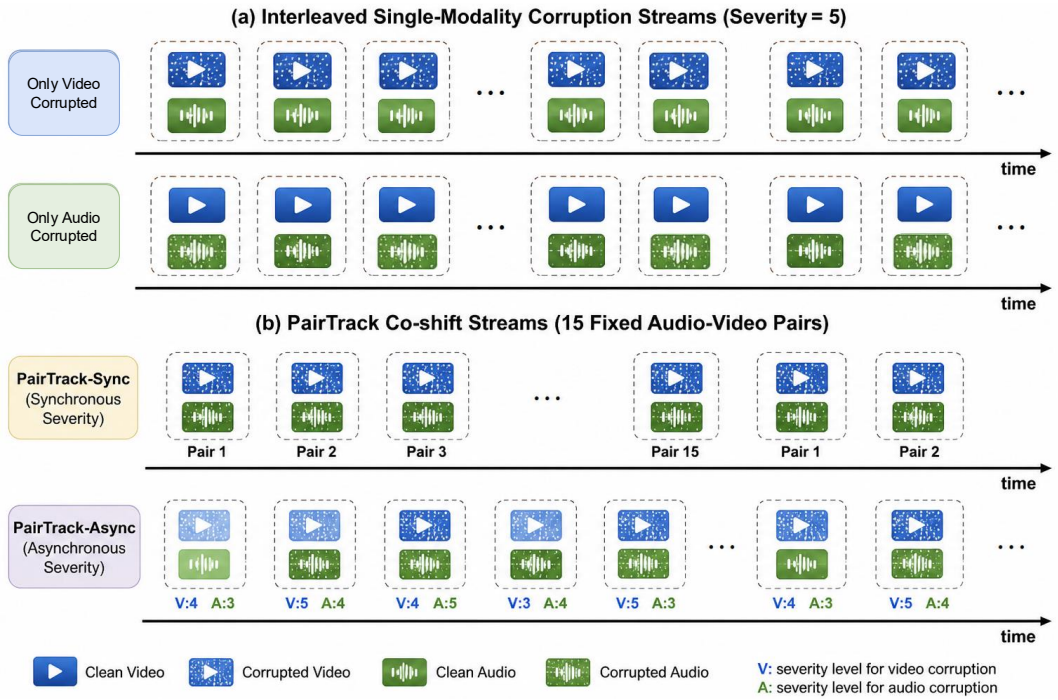



Figure 7: Illustration of the audio-visual corruption benchmark construction. We reorganize static Kinetics50-C and VGGSound-C corruptions into single-sample streaming protocols, including interleaved single-modality shifts and paired multimodal co-shift streams.

**Kinetics50** Kinetics is a large-scale action recognition dataset collected from YouTube videos, consisting of short clips annotated with human action categories. Following prior multimodal TTA work Yang et al. (2024); Zhao et al. (2025b), we use the Kinetics50 subset for audio-video classification, which contains 50 action classes with paired audio-visual samples.

**VGGSound** VGGSound is a large-scale in-the-wild audio-visual dataset collected from YouTube videos. It is designed for audio-visual correspondence, where the sound source is visually identifiable in the video. Each sample contains paired audio and visual clips, making it suitable for evaluating multimodal robustness under both audio and video corruptions.

**Kinetics50-C and VGGSound-C.** Based on the clean Kinetics50 and VGGSound test sets, we construct corrupted benchmarks by applying modality-specific corruptions while preserving the

original labels. Following prior work Yang et al. (2024); Zhao et al. (2025b), for the video modality, we use 15 corruption types: Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate, and JPEG compression. For the audio modality, we use 6 corruption types: Gaussian noise, traffic noise, crowd noise, rain, thunder, and wind. Each corruption type is applied at five severity levels.

**Streaming protocols.** For interleaved single-modality corruption, columns denoted by `V:*` indicate that only the video modality is corrupted while the audio input remains clean; columns denoted by `A:*` indicate that only the audio modality is corrupted while the video input remains clean. This protocol simulates alternating modality shifts between audio and video, and evaluates whether a method can continuously switch between different shifted modalities while maintaining adaptation stability. All interleaved-stream results reported in the main paper use severity level 5.

For PairTrack co-shift streams, we construct 15 fixed audio-video corruption pairs by matching audio and video corruptions according to the order reported in the tables. This protocol simulates scenarios where audio and video shift simultaneously, and evaluates robustness under joint multimodal degradation. We further consider two variants: **PairTrack-Sync** and **PairTrack-Async**. In PairTrack-Sync, both modalities are corrupted synchronously at a fixed severity level. In PairTrack-Async, asynchronous severity changes are introduced within each corruption pair, using the audio severity schedule $A = [3, 4, 5, 4, 3]$ and the video severity schedule $V = [4, 5, 4, 3, 5]$. This setting further evaluates whether a method remains stable when the shift intensities of audio and video evolve asynchronously.

**CMU-MOSI.** CMU-MOSI Zadeh et al. (2016) is a widely used English multimodal sentiment analysis dataset collected from YouTube videos. Each sample contains aligned text, audio, and visual modalities, with continuous sentiment intensity annotations and categorical sentiment labels.

**CMU-MOSEI.** CMU-MOSEI Bagher Zadeh et al. (2018) is a larger-scale English multimodal sentiment and emotion analysis dataset collected from in-the-wild YouTube videos. It contains synchronized text, audio, and visual modalities, and provides continuous sentiment scores and emotion annotations.

**CH-SIMS.** CH-SIMS Yu et al. (2020) is a Chinese multimodal sentiment analysis dataset collected from Chinese video content. It provides aligned text, audio, and visual modalities, together with continuous sentiment scores and categorical sentiment labels.

**Cross-dataset sentiment streams.** Since the original annotation spaces of these datasets are not fully identical, we unify them into a three-class sentiment classification protocol: negative, neutral, and positive according to whether the continuous sentiment score is smaller than, equal to, or larger than $0$. We use one dataset as the source domain and the remaining two datasets as alternating target domains to construct continual cross-dataset test streams. This protocol evaluates online adaptation under natural multimodal distribution shifts across language, speaking style, emotional expression, and data collection conditions.

## B Pretrained Model Details

In our synthetic multimodal domain-shift experiments, we adopt the pretrained CAV-MAE as the base audio-visual model, following the experimental setting of prior multimodal TTA works Yang et al. (2024); Guo & Jin (2025). Specifically, the model is pretrained on the training sets of Kinetics50 and VGGSound, respectively. Therefore, the clean Kinetics50 and VGGSound test sets are regarded as the source distributions, while Kinetics50-C and VGGSound-C constructed with audio or video corruptions serve as target test streams. All test samples are evaluated in an online manner to assess continual adaptation under the single-sample AV-CTTA setting.

To further evaluate edge-side deployment, we conduct experiments on quantized CAV-MAE. Following common settings in test-time adaptation for quantized neural networks Deng et al. (2025), we use W8A8 and W4A4 weight/activation quantization to simulate low-bit inference. In implementation, we adopt a lightweight post-training quantization scheme with symmetric fake quantization and percentile calibration, without introducing complex Hessian-based scale search. For the quantization scope, we quantize only the audio and video encoder branches while keeping the fusion attention

module in full precision, reducing the main computational cost while preserving cross-modal fusion stability.

# C Implementation Details

All baselines are evaluated under the same online single-sample streaming protocols on Kinetics50-C and VGGSound-C. For backpropagation-based methods, we use the Adam optimizer with a learning rate of $1 \times 10^{-4}$. For single-modal TTA/CTTA baselines, we adapt them to the audio-visual setting by computing their original objectives from the final fused prediction of CAV-MAE. Unless otherwise specified, the tunable parameters are set to the LayerNorm parameters in the audio and video encoders, while the fusion module and classifier are frozen.

- **T3A** Iwasawa & Matsuo (2021). We freeze the backbone and update classifier prototypes using confident fused test-time features.
- **SAR** Niu et al. (2023). We compute SAR's reliable sharpness-aware entropy loss on fused predictions and update audio/video LayerNorm parameters.
- **TSA** Chen et al. (2025). We apply its test-time self-adaptation objective to fused predictions and update audio/video LayerNorm parameters.
- **PTA** Wang et al. (2025). We follow its partition-then-adaptation design and use a memory window of length 8 for batch-level statistics.
- **BECoTTA** Lee et al. (2024a). We use its official CTTA configuration and keep method-specific hyperparameters unchanged.
- **Tent** Wang et al. (2021). We minimize fused prediction entropy and update audio/video LayerNorm parameters.
- **READ** Yang et al. (2024). We follow its multimodal TTA setting and update the query/key/value matrices in the fusion attention layer.
- **FOZO** Wang & Wang (2026). We freeze the CAV-MAE backbone and optimize lightweight modality-specific prompt tokens with forward-only zeroth-order updates.
- **CMAE** Liu et al. (2024). We follow the CAV-MAE adaptation setting with the same pretrained backbone and original hyperparameters.
- **DeYO** Lee et al. (2024b). We apply its reliable sample selection and deformed entropy loss to fused predictions and update audio/video LayerNorm parameters.
- **DASP** He et al. (2026). We follow its asymmetric multimodal adaptation setting with the original tunable modules and hyperparameters.
- **FOA** Niu et al. (2024). We freeze CAV-MAE and optimize modality-specific prompt tokens using FOA's derivative-free forward-only update.
- **ZORA**. We update only SSLS-selected audio/video LayerNorm parameters, keeping the fusion module, classifier, and unselected layers frozen. For BARC, we use bias and confidence quantile levels $q_b = 0.5$ and $q_c = 0.7$, respectively; the corresponding thresholds $\tau_b$ and $\tau_c$ are recomputed from each local window. The negative-weight coefficient is fixed to $\gamma = 0.5$.

# D Extended Method Analysis

## D.1 Offline-to-Online Procedure

ZORA consists of source-side caching, a one-time deployment-start initialization stage, and sample-wise online adaptation. Source statistics are cached before deployment, whereas SSLS is executed once after collecting the early unlabeled target subset; the selected update space then remains fixed throughout the stream. The online stage is then executed on a batch-size-one target stream, where ZORA dynamically constructs a reliability-calibrated objective from the current sample and a local history window, and updates only the selected parameters through forward-only ZOO.

**Offline preparation.** In the offline preparation stage, ZORA first extracts layer-wise features from the audio and visual branches using a small set of source reference samples and early unlabeled target samples. Since the early target samples are used without labels, this step does not introduce supervised target-domain information. ZORA then computes the Layer Shift Sensitivity (LSS) score for each candidate normalization layer to measure its sensitivity to the current domain shift. Based on LSS, SSLS selects the top-$K$ shift-sensitive normalization layers in each modality branch, yielding the updatable layer sets $\mathcal{S}^a$ and $\mathcal{S}^v$. The normalization parameters in these layers form the low-dimensional ZOO update space. Finally, ZORA pre-computes the source feature statistics $\{\mu_l^{S,m}, \sigma_l^{S,m}\}$ on the selected layers, which serve as stable anchors for the source-anchored feature alignment in BARC. This offline stage does not train the model or require target labels; it only prepares a compact update space and stable source statistics for online adaptation.

**Online adaptation.** During online adaptation, target samples arrive sequentially with batch size one. Given the current sample $\mathbf{x}_t$, the model first performs a forward pass to obtain the fusion prediction $p(\mathbf{x}_t)$ and updates the local history window $\mathcal{W}_t$. BARC estimates the prediction bias of the current sample from the predicted-label frequency within the window and combines it with prediction confidence to obtain the sample weight $I(\mathbf{x}_t)$. Using this weight, BARC constructs a fusion-level bias-aware weighted entropy objective, which promotes entropy minimization on reliable low-bias samples while suppressing misleading confident or unreliable samples. Meanwhile, BARC computes target-window feature statistics on the selected branch layers and aligns them with the pre-computed source statistics, forming a branch-level source-anchored feature alignment constraint. The weighted entropy term and the feature-statistics alignment term are then combined as the overall online objective.

**Forward-only update.** ZOO is applied only to the selected branch normalization parameters, $\Theta_{\mathrm{br}} = \Theta^a_{\mathcal{S}^a} \cup \Theta^v_{\mathcal{S}^v}$. Specifically, ZORA applies bidirectional random perturbations to $\Theta_{\mathrm{br}}$, queries the objective values with two forward passes, estimates the finite-difference update direction, and updates the selected parameters. The fusion module, classifier, and all unselected branch layers remain frozen throughout adaptation. Therefore, ZORA does not retain computation graphs or activations for backward propagation and does not compute gradients.

Through this offline-to-online procedure, ZORA improves ZOO-based AV-CTTA from two aspects: SSLS constrains the update space to low-dimensional shift-sensitive branch layers, while BARC provides a more reliable and smoother learning signal through bias-aware reweighting and source-anchored alignment.

**Algorithm 1** Zeroth-Order Reliability Adaptation (ZORA)

**Require:** Model $f_\Theta$, source set $\mathcal{D}_s^{\mathrm{ref}}$, target stream $\{\mathbf{x}_t\}_{t=1}^T$, window $W$, layer number $K$, $q_b, q_c, \gamma, \lambda, \eta, \delta$
**Ensure:** Adapted branch parameters $\Theta_{\mathrm{br}}$

**Offline preparation.**
Collect early unlabeled target samples $\mathcal{D}_t^{\mathrm{early}}$
**for** each branch $m \in \{a, v\}$ **do**
  Extract layer features on $\mathcal{D}_s^{\mathrm{ref}}$ and $\mathcal{D}_t^{\mathrm{early}}$; compute $\mathrm{LSS}_l^m$
  Select $\mathcal{S}^m = \mathrm{TopK}\{\mathrm{LSS}_l^m\}_{l=1}^{L_m}$ and set tunable parameters $\Theta_{\mathcal{S}^m}^m$
  Compute source-reference statistics $\{\mu_l^{S,m}, \sigma_l^{S,m}\}_{l\in\mathcal{S}^m}$ on $\mathcal{D}_s^{\mathrm{ref}}$
**end for**
Set $\Theta_{\mathrm{br}} = \Theta_{\mathcal{S}^a}^a \cup \Theta_{\mathcal{S}^v}^v$, freeze other parameters, and initialize $\mathcal{W}_0 \leftarrow \emptyset$
**Online adaptation.**
**for** $t = 1$ to $T$ **do**
  Receive $\mathbf{x}_t$, compute $p(\mathbf{x}_t)$, and update $\mathcal{W}_t$
  Compute $\{B_i, C_i\}_{\mathbf{x}_i\in\mathcal{W}_t}$ from the local prediction history
  Set $\tau_b = \mathrm{Quantile}_{q_b}(\{B_i\})$ and $\tau_c = \mathrm{Quantile}_{q_c}(\{C_i\})$
  Compute BARC weight $I(\mathbf{x}_t; \tau_b, \tau_c, \gamma)$ and entropy loss $\mathcal{L}_t^{\mathrm{ent}}$
  **for** each branch $m \in \{a, v\}$ **do**
    Compute target-window statistics and alignment loss $\mathcal{L}_t^{\mathrm{align},m}$ on $\mathcal{S}^m$
  **end for**
  Set $\mathcal{L}_t = \mathcal{L}_t^{\mathrm{ent}} + \lambda \sum_{m\in\{a,v\}} \mathcal{L}_t^{\mathrm{align},m}$
  Sample $\mathbf{u}_t$, estimate $\hat{\mathbf{g}}_t$ with two forward queries, and update $\Theta_{\mathrm{br}} \leftarrow \Theta_{\mathrm{br}} - \eta\hat{\mathbf{g}}_t$
**end for**
**return** $\Theta_{\mathrm{br}}$

### D.2 SSLS Robustness and Initialization Cost

**Fixed selection protocol.** SSLS is computed once from $N_s$ cached source-reference samples and $N_t$ early unlabeled target samples. It selects the Top-3 LayerNorm layers independently in the audio and visual branches, and the selected sets remain fixed for the rest of the stream. The initialization samples are used only for layer scoring, not for parameter adaptation.

**Subset size.** Table 4 independently varies $N_s$ and $N_t$ on Kinetics50-C PACS while fixing the other subset at 64. The overlap is measured against the default sets, audio $\{2, 3, 10\}$ and video $\{3, 4, 9\}$. Layer identities vary when only 16–32 samples are available, but the largest accuracy reduction is 0.61 points; performance saturates at $N_s = N_t = 64$.

**Representativeness of the early target subset.** We retain the same 64 sample identities and change only their shift condition. After selection, every condition restarts from the same checkpoint and adapts on the identical remaining stream. Table 5 shows that mild or single-modality early shifts alter mainly the clean or weakly shifted branch, while suffix accuracy drops by at most 0.76 points.

**Cross-domain stability.** We independently recompute the Top-3 sets for each of the 15 PACS corruption pairs while keeping the sample budget and candidate layers fixed. Across the $\binom{15}{2} = 105$ domain pairs, a recurrent modality-dependent core remains, although the visual branch is less stable. We therefore claim cross-domain robustness rather than strict domain invariance.

**Complexity and simpler scores.** All candidate-layer representations are collected by shared forward passes. With $L_m$ candidate layers, $M_m$ retained tokens, feature dimension $d_m$, and $I$ two-means iterations in branch $m$, the one-time scoring cost is

$$\mathcal{O}\left(\sum_{m\in\{a,v\}} L_m M_m (N_s + N_t) d_m I\right), \tag{9}$$

Table 4: Sensitivity of SSLS to source-reference and early-target subset sizes on Kinetics50-C PACS.

| Varied subset | Size | Accuracy | Audio / Video Top-3 | Avg. overlap |
|---|---|---|---|---|
| $N_s$ ($N_t = 64$) | 16 | 54.96 | 3,5,6 / 4,5,6 | 33.3% |
| | 32 | 55.01 | 2,4,8 / 3,5,9 | 50.0% |
| | 64 | **55.57** | 2,3,10 / 3,4,9 | 100.0% |
| | 128 | **55.57** | 2,3,10 / 3,4,9 | 100.0% |
| | 256 | **55.57** | 2,3,10 / 3,4,9 | 100.0% |
| $N_t$ ($N_s = 64$) | 16 | 55.07 | 8,9,10 / 3,4,6 | 50.0% |
| | 32 | 55.10 | 3,4,10 / 3,4,6 | 66.7% |
| | 64 | **55.57** | 2,3,10 / 3,4,9 | 100.0% |
| | 128 | **55.57** | 2,3,9 / 3,4,9 | 83.3% |
| | 256 | **55.57** | 2,3,10 / 3,4,9 | 100.0% |

Table 5: SSLS sensitivity to the representativeness of the early unlabeled target subset.

| Condition | Audio Top-3 | Video Top-3 | Avg. overlap | Suffix Acc. | Drop |
|---|---|---|---|---|---|
| Reference | 2,3,10 | 3,4,9 | 100.0% | **55.68** | – |
| Mild A/V | 2,3,10 | 3,5,9 | 83.3% | 55.28 | 0.40 |
| Audio-only | 2,3,10 | 4,5,6 | 66.7% | 54.92 | 0.76 |
| Video-only | 0,3,9 | 3,4,9 | 66.7% | 55.11 | 0.57 |

in addition to the shared feature-extraction cost $\mathcal{O}((N_s + N_t)C_{\text{fwd}})$. Table 7 compares LSS with non-iterative alternatives under the same features, Top-3 budget, stream, and adaptation configuration. LSS adds 307.21 ms once, or approximately 0.10 ms when amortized over the 3,000-sample stream, and improves accuracy by 0.92–1.09 points.

### D.3 BARC Weighting Mechanism and Variants

We further analyze why the bias-aware weighting in BARC can stabilize forward-only online adaptation from the perspective of ZOO direction estimation. Unlike backpropagation-based TTA methods, ZOO does not explicitly compute gradients with respect to model parameters. Instead, it estimates the update direction from objective-value differences under parameter perturbations. For the current test sample $\mathbf{x}_t$, if the raw entropy objective is directly used, the entropy-driven ZOO direction can be written as

$$\hat{\mathbf{g}}_t^{\text{ent}} = \frac{H(\mathbf{x}_t; \Theta + \delta \mathbf{u}_t) - H(\mathbf{x}_t; \Theta - \delta \mathbf{u}_t)}{2\delta} \mathbf{u}_t, \tag{10}$$

where $H(\mathbf{x}_t; \Theta) = -\sum_c p_c(\mathbf{x}_t; \Theta) \log p_c(\mathbf{x}_t; \Theta)$ denotes the prediction entropy under model parameters $\Theta$, $\mathbf{u}_t$ is the random perturbation direction, and $\delta$ is the perturbation radius. Eq. (10) shows that the ZOO update direction is directly determined by the perturbed entropy difference. Thus, if the entropy objective provides unreliable feedback, the estimated direction can also become biased.

This issue becomes critical in AV-CTTA. Under continual audio-visual shifts, the relative reliability of audio and visual modalities may change over time, causing the fused prediction to over-rely on one modality and become biased toward a few over-predicted classes. As a result, some incorrectly predicted samples may still exhibit high confidence and low entropy. Directly minimizing entropy on such samples can further reinforce their erroneous predictions and push the model toward over-confident biased outputs. For ZOO, this misleading effect is reflected in the perturbed entropy difference, causing biased confident samples to dominate the finite-difference feedback and leading to error accumulation in the streaming process.

To mitigate this issue, BARC replaces the raw entropy objective with a bias-aware weighted entropy objective:

$$\mathcal{L}_t^{\text{barc}} = I(\mathbf{x}_t) H(\mathbf{x}_t; \Theta), \tag{11}$$

where $I(\mathbf{x}_t)$ is determined by the prediction bias and confidence of the current sample. The corresponding ZOO direction becomes

$$\hat{\mathbf{g}}_t^{\text{barc}} = I(\mathbf{x}_t) \frac{H(\mathbf{x}_t; \Theta + \delta \mathbf{u}_t) - H(\mathbf{x}_t; \Theta - \delta \mathbf{u}_t)}{2\delta} \mathbf{u}_t. \tag{12}$$

Table 6: Pairwise overlap of independently selected SSLS Top-3 sets across 15 corruption pairs.

| Branch | Most frequent layers | Avg. shared Top-3 | Share $\geq 1$ | Share $\geq 2$ |
|---|---|---|---|---|
| Audio | L10: 86.7%, L3: 80.0%, L2: 40.0% | 1.71/3 | 94.3% | 64.8% |
| Video | L9: 60.0%, L3: 46.7%, L4: 46.7% | 1.04/3 | 72.4% | 30.5% |

Table 7: One-time SSLS initialization cost and simpler layer-scoring alternatives.

| Score | Shared extraction | Scoring | Selected layers (A/V) | Accuracy |
|---|---|---|---|---|
| Random | 1449.74 ms | 0.00 ms | 5,9,10 / 3,4,7 | $54.55 \pm 0.31$ |
| Mean Gap | 1449.74 ms | 0.82 ms | 8,9,10 / 0,9,10 | $54.72 \pm 0.38$ |
| Moment Gap | 1449.74 ms | 1.29 ms | 8,9,10 / 0,9,10 | $54.71 \pm 0.38$ |
| **LSS** | 1449.74 ms | 307.21 ms | 2,3,10 / 3,4,9 | $\mathbf{55.64 \pm 0.38}$ |

Eq. (12) indicates that $I(\mathbf{x}_t)$ directly controls the contribution of the current sample to the finite-difference feedback. For low-bias high-confidence samples, BARC assigns positive weights, allowing reliable predictions to provide entropy-minimization signals. For high-bias high-confidence samples, BARC assigns negative weights, reversing their entropy-driven feedback to suppress over-confident biased predictions. For the remaining samples, BARC assigns zero weights, preventing ambiguous or unreliable samples from injecting noisy finite-difference signals.

Therefore, BARC does not merely perform sample filtering. Instead, it calibrates the test-time objective used by ZOO, converting raw entropy minimization into a bias-aware reliability-calibrated signal. This reduces the influence of misleading confident samples on finite-difference direction estimation and provides more stable forward-only updates under continual multimodal shifts. Table 8 further compares several BARC weighting variants on the Kinetics50-C PairTrack-Async setting. BARC-Full denotes our default design, which promotes low-bias high-confidence samples, suppresses high-bias high-confidence samples, and neutralizes the remaining samples, where $\rho_c$ and $\rho_b$ denote confidence and bias ranks, respectively. BARC-Hard uses ranked positive and negative weights, BARC-Suppress only suppresses high-bias high-confidence samples, and BARC-Continuous assigns continuous signed weights to all samples. The results show that BARC-Full achieves the best performance, indicating that a hard reliability partition with explicit promotion, suppression, and neutralization provides a more stable learning signal for ZOO.

Table 8: BARC weighting variants.

| Variant | Positive | Negative | Other | Accuracy |
|---|---|---|---|---|
| BARC-Hard | $+\rho_c\rho_b$ | $-\gamma\rho_c\rho_b^-$ | 0 | 54.61 |
| BARC-Suppress | $H$ | $-\gamma$ | $H$ | 53.72 |
| BARC-Continuous | signed | signed | signed | 54.27 |
| **BARC-Full** | $+\rho_c\rho_b$ | $-\gamma$ | 0 | **55.37** |

### D.4 BARC Scope under Non-IID Class Priors

BARC treats local predicted-class frequency as an over-amplification risk signal rather than direct evidence of an incorrect prediction. To test its boundary, we construct two single-seed controlled stress tests from the same 3,000-sample Kinetics50-C PACS stream. Target labels are used only to construct the streams and compute offline diagnostics; online adaptation remains label-free. False-suppression rate (FSR) is the fraction of genuinely correct, high-confidence head- or burst-class samples assigned a negative weight. Negative-weight precision (NWP) is the fraction of negatively weighted samples that are misclassified. Recovery is the sample offset at which a 24-sample sliding window first returns to within one percentage point of its pre-burst accuracy.

**Persistent long-tailed prior.** At imbalance ratio (IR) 1, every class contains 60 samples. At IR 10, class counts follow an exponential long tail from 152 to 15 samples, while consecutive instances of the same class are limited to two. Table 9 shows that BARC remains beneficial, but genuine

head-class prevalence increases FSR and lowers NWP. Consequently, weaker suppression ($\gamma = 0.25$) is preferable under the imbalanced stream.

Table 9: BARC under persistent class imbalance on Kinetics50-C PACS. These controlled diagnostics use one seed.

| IR | Method | Acc. | Macro-F1 | Head FSR ↓ | NWP ↑ |
|---|---|---|---|---|---|
| 1 | Entropy-only | 53.80 | 54.38 | – | – |
| 1 | BARC-Positive | 54.71 | 55.14 | – | – |
| 1 | Full BARC, $\gamma = 0.25$ | 55.25 | 55.41 | 11.44 | 61.13 |
| 1 | Full BARC, $\gamma = 0.50$ | **55.61** | **55.82** | **9.08** | **63.45** |
| 10 | Entropy-only | 52.90 | 48.85 | – | – |
| 10 | BARC-Positive | 54.03 | 49.97 | – | – |
| 10 | Full BARC, $\gamma = 0.25$ | **54.87** | **50.88** | **18.98** | **55.60** |
| 10 | Full BARC, $\gamma = 0.50$ | 54.51 | 50.58 | 19.12 | 55.19 |

**Transient same-class bursts.** In 25 of the 75 segments, 24 samples from one genuine class and 16 samples from other classes are arranged as 8 other $\rightarrow$ 24 same-class $\rightarrow$ 8 other. The dispersed control contains exactly the same samples, corruptions, and severities, but limits consecutive same-class runs to two. As shown in Table 10, all BARC variants improve overall accuracy over entropy-only adaptation. However, contiguous genuine bursts remain harder than the dispersed control, confirming that short-term class concentration is a boundary of the local frequency estimate. Increasing $W$ improves end-to-end accuracy and recovery, although FSR and NWP do not improve monotonically.

Table 10: BARC under transient same-class concentration on Kinetics50-C PACS. These controlled diagnostics use one seed.

| Method / stream | Overall Acc. | Macro-F1 | Burst Acc. | FSR ↓ | NWP ↑ | Recovery ↓ |
|---|---|---|---|---|---|---|
| Entropy-only / contiguous | 50.87 | 50.60 | 36.67 | – | – | 63.50 |
| BARC-Positive / contiguous | 51.80 | 51.53 | 36.67 | 0.00 | – | **44.42** |
| Full BARC, $W = 24$ / contiguous | 51.51 | 50.92 | 36.07 | 32.71 | 39.04 | 45.13 |
| Full BARC, $W = 48$ / contiguous | **52.13** | **51.56** | **36.67** | 34.84 | 31.31 | **44.42** |
| Full BARC, $W = 24$ / dispersed | **53.10** | **52.59** | **43.33** | 29.50 | 44.31 | – |

These tests support a bounded interpretation: BARC is designed for corruption-induced reliability bias, not as a general label-shift correction mechanism. Persistent prior skew favors conservative negative weighting, whereas transient bursts benefit from a longer history window.

# E Statistical Reliability, Deployment, and Generalization

## E.1 Five-Seed Statistics

Table 11 reports the mean, standard deviation, and two-sided 95% Student-$t$ confidence interval over five independent runs. The ranking in Table 1 is preserved in every setting, and ZORA's standard deviation remains at or below 0.74 points.

Table 11: Five-seed accuracy statistics. Each setting reports mean ± standard deviation and the two-sided 95% confidence interval.

| | ICS | | | | PACS | | | |
|---|---|---|---|---|---|---|---|---|
| | Kinetics50-C | | VGGSound-C | | Kinetics50-C | | VGGSound-C | |
| Method | Mean ± Std | 95% CI | Mean ± Std | 95% CI | Mean ± Std | 95% CI | Mean ± Std | 95% CI |
| Source | 61.47 ± 0.47 | [60.89,62.05] | 41.54 ± 0.09 | [41.43,41.65] | 44.47 ± 0.02 | [44.45,44.49] | 12.62 ± 0.45 | [12.06,13.18] |
| T3A | 49.70 ± 1.06 | [48.38,51.02] | 17.67 ± 0.59 | [16.94,18.40] | 30.02 ± 0.47 | [29.44,30.60] | 3.58 ± 0.16 | [3.38,3.78] |
| SAR | 41.59 ± 1.48 | [39.75,43.43] | 24.85 ± 0.94 | [23.68,26.02] | 27.45 ± 0.98 | [26.23,28.67] | 7.19 ± 0.10 | [7.07,7.31] |
| TSA | 55.51 ± 0.91 | [54.38,56.64] | 24.93 ± 0.65 | [24.12,25.74] | 43.28 ± 1.36 | [41.59,44.97] | 12.68 ± 0.50 | [12.06,13.30] |
| PTA | 63.58 ± 0.90 | [62.46,64.70] | 43.75 ± 0.32 | [43.35,44.15] | 35.23 ± 2.73 | [31.84,38.62] | 4.83 ± 0.10 | [4.71,4.95] |
| BECoTTA | 63.54 ± 0.63 | [62.76,64.32] | 44.08 ± 0.28 | [43.73,44.43] | 44.85 ± 0.21 | [44.59,45.11] | 7.57 ± 1.32 | [5.93,9.21] |
| Tent | 63.62 ± 0.60 | [62.88,64.36] | 45.72 ± 1.07 | [44.39,47.05] | 42.13 ± 0.17 | [41.92,42.34] | 8.76 ± 0.24 | [8.46,9.06] |
| READ | 64.10 ± 0.24 | [63.80,64.40] | 44.98 ± 1.01 | [43.73,46.23] | 46.71 ± 0.13 | [46.55,46.87] | 5.27 ± 0.22 | [5.00,5.54] |
| FOZO | 63.73 ± 0.16 | [63.53,63.93] | 44.91 ± 0.84 | [43.87,45.95] | 44.12 ± 0.25 | [43.81,44.43] | 10.70 ± 0.35 | [10.27,11.13] |
| CMAE | 63.62 ± 0.44 | [63.07,64.17] | 45.54 ± 1.10 | [44.17,46.91] | 44.39 ± 0.04 | [44.34,44.44] | 12.50 ± 0.40 | [12.00,13.00] |
| DeYO | 63.65 ± 0.74 | [62.73,64.57] | 44.80 ± 0.42 | [44.28,45.32] | 45.51 ± 0.19 | [45.27,45.75] | 12.56 ± 0.34 | [12.14,12.98] |
| DASP | 63.54 ± 0.47 | [62.96,64.12] | 44.11 ± 0.17 | [43.90,44.32] | 44.40 ± 0.02 | [44.38,44.42] | 12.63 ± 0.42 | [12.11,13.15] |
| FOA | 63.36 ± 0.51 | [62.73,63.99] | 44.21 ± 0.99 | [42.98,45.44] | 50.68 ± 0.36 | [50.23,51.13] | 17.93 ± 0.44 | [17.38,18.48] |
| **ZORA** | **66.59 ± 0.74** | [65.67,67.51] | **47.58 ± 0.54** | [46.91,48.25] | **55.64 ± 0.67** | [54.81,56.47] | **22.06 ± 0.45** | [21.50,22.62] |

## E.2 Computation and Deployment Analysis

We distinguish adaptation queries, which directly construct an update, from total model evaluations, which also include non-perturbed prediction and reliability-estimation passes. Table 12 reports the official configurations. ZORA uses two adaptation queries and four total evaluations per sample; it updates only SSLS-selected LayerNorm affine parameters and uses no target labels.

Table 12: Implementation-level accounting for forward-only baselines and ZORA on CAV-MAE.

| Method | Updated state | Params | Adapt. queries | Total evals | Selection protocol | Labels |
|---|---|---|---|---|---|---|
| T3A | Prototype bank | 0 | 0 | 1 | Official, $K = 5$ | No |
| FOA | Fusion prompts | 2,304 | 28 | 28 | 3 prompts, population 27 | No |
| FOZO | A/V input prompts | 4,608 | 2 | 2 | Official multimodal extension | No |
| ZORA | Selected LN affine | 27,648 | 2 | 4 | Label-free fixed setting | No |

We also vary the number of sequential ZOO updates performed on each incoming sample. The best multi-step result changes accuracy by only 0.06 points while every multi-step setting is slower than the one-step default. Since latency was measured in a single timing run and is subject to GPU variation, these values characterize the operating point rather than precise linear scaling.

Table 13: Within-sample ZOO iterations on Kinetics50-C PACS.

| Steps | Perturbed queries | Accuracy | Gain | Latency/sample | Throughput |
|---|---|---|---|---|---|
| 1 | 2 | 55.57 | – | **77.16 ms** | **12.96/s** |
| 2 | 4 | 55.57 | 0.00 | 97.41 ms | 10.27/s |
| 3 | 6 | **55.63** | +0.06 | 84.44 ms | 11.84/s |
| 5 | 10 | 55.53 | −0.04 | 117.68 ms | 8.50/s |

The low-bit experiments in Table 3 use symmetric fake quantization with percentile calibration. They verify numerical robustness under W8A8 and W4A4 perturbations, but should not be interpreted as native integer-kernel latency or memory measurements.

## E.3 Cross-Architecture and Task Extension

**UAVM architecture.** We evaluate a second audio-visual architecture with modality-specific encoders followed by shared cross-modal attention on the full VGGSound-C PACS stream. SSLS selects three LayerNorm layers per branch, while the cross-modal Transformer and classifier remain frozen. Table 14 is a one-seed controlled architecture study, not a five-seed benchmark comparison.

Table 14: Transfer to the UAVM cross-modal attention architecture on VGGSound-C PACS.

| Method | Acc. | Macro-F1 | Params | Obj. evals | Total evals | Backward |
|---|---|---|---|---|---|---|
| Source | 20.53 | 23.94 | 0 | 0 | 1 | No |
| Tent | 20.30 | 23.78 | 12,288 | 1 | 1 | Yes |
| READ | 23.80 | 22.90 | 3,145,728 | 1 | 2 | Yes |
| FOZO | 18.57 | 22.03 | 6,144 | 2 | 2 | No |
| FOA | 26.73 | 24.66 | 3,072 | 28 | 28 | No |
| **ZORA** | **28.30** | **29.45** | 12,288 | 2 | 4 | No |

**Unimodal and gradient-based transfer.** On a unimodal ImageNet-C interleaved stream, ZORA reaches 58.65% average accuracy, compared with 55.59%, 55.45%, 55.38%, and 56.08% for Source, Tent, FOA, and FOZO, respectively. We further replace finite differences with a standard one-step backpropagation update on Kinetics50-C PACS. Table 15 shows that SSLS and BARC remain useful when exact gradients are available, supporting their roles as update-space selection and objective calibration rather than ZOO-specific heuristics.

Table 15: Transfer of SSLS and BARC to backpropagation-based online adaptation. This controlled experiment uses one seed.

| Optimizer | Layer selection | BARC | Updated params | Accuracy |
|---|---|---|---|---|
| BP | Random Top-3 | Off | 18,432 | 53.23 |
| BP | SSLS Top-3 | Off | 18,432 | 54.47 |
| BP | Random Top-3 | On | 18,432 | 55.31 |
| BP | SSLS Top-3 | On | 18,432 | **56.23** |
| ZOO | SSLS Top-3 | On | 18,432 | 55.74 |

Finally, Table 26 evaluates a different tri-modal sentiment task under natural cross-dataset shifts among SIMS, MOSI, and MOSEI. Together, the UAVM, ImageNet-C, optimizer-transfer, and sentiment experiments test changes in backbone, fusion design, modality composition, task, and optimization mechanism. The applicable interface is explicit: SSLS requires accessible modality-specific intermediate representations and lightweight normalization parameters, while BARC requires a fused predictive distribution.

## F Additional Benchmark Results

**Detailed ICS and PairTrack-Async results.** We provide complete interleaved corruption stream results on Kinetics50-C and VGGSound-C in Table 16 and Table 17. These tables report the accuracy of each method at every local corruption position, where audio and video corruptions appear alternately while the other modality remains clean. We further provide complete PairTrack-Async results on Kinetics50-C and VGGSound-C in Table 18 and Table 19. These detailed results complement the averaged ICS and PACS results reported in the main paper and show that ZORA consistently improves the overall streaming performance.

**PairTrack-Sync co-shift.** Tables 20 and 21 report synchronized multimodal co-shifts with fixed severity levels. ZORA reaches $56.83\%$ and $23.77\%$ on Kinetics50-C and VGGSound-C, exceeding FOA by $4.23/3.90$ points and the source model by $9.50/10.04$ points, respectively.

**Results on single-modality corruption streams.** Tables 22–25 evaluate localized audio-only and video-only shifts. ZORA obtains average accuracies of $69.72\%$, $23.05\%$, $61.10\%$, and $52.43\%$ across the four settings, achieving the best or near-best performance and consistently improving over the corresponding source models. This confirms that SSLS and BARC remain effective beyond paired multimodal shifts.

Table 16: Accuracy comparison with SOTAs on Kinetics50-C for interleaved single-modality corruption streams (severity level = 5). Here, $*$ denotes the clean modality.

| Method | A:* | | V:* | | A:* | | V:* | | A:* | | | | | V:* | A:* | | V:* | | A:* | | V:* | | A:* | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | V:Gaus. | V:Shot | A:Gaus. | V:Imp. | V:Def. | A:Traf. | V:Glas. | V:Mot. | A:Crowd | V:Zoom | V:Snow | V:Frost | A:Rain | V:Fog | V:Brit. | A:Thun. | V:Contr. | V:Elas. | A:Wind | V:Pix. | V:JPEG | Avg. |
| | $t$ ⟶ | | | | | | | | | | | | | | | | | | | | | |
| Source | 48.88 | 46.44 | 66.75 | 44.88 | 63.62 | 62.84 | 70.66 | 78.47 | 65.97 | 55.03 | 59.72 | 59.72 | 65.97 | 47.22 | 73.78 | 72.22 | 44.78 | 59.72 | 73.78 | 66.75 | 63.62 | 61.47 |
| FOA | 46.44 | 48.78 | 66.75 | 40.19 | 65.97 | 62.84 | 72.22 | 78.47 | 68.31 | 55.03 | 61.28 | 64.41 | 67.53 | 56.94 | 73.00 | 73.00 | 51.91 | 63.62 | 76.12 | 68.31 | 67.53 | 63.36 |
| T3A | 30.47 | 37.50 | 50.00 | 35.16 | 55.47 | 47.66 | 58.59 | 67.19 | 59.38 | 49.22 | 39.84 | 45.31 | 46.88 | 37.50 | 57.03 | 57.81 | 33.59 | 58.59 | 64.06 | 56.25 | 56.25 | 49.70 |
| TSA | 46.88 | 45.31 | 66.41 | 39.84 | 55.47 | 56.25 | 67.97 | 74.22 | 56.25 | 51.56 | 53.12 | 57.81 | 47.66 | 55.47 | 61.72 | 53.91 | 42.97 | 53.12 | 55.47 | 60.16 | 64.06 | 55.51 |
| Tent | 51.66 | 50.44 | 67.53 | 46.88 | 65.62 | 66.41 | 71.88 | 80.47 | 70.31 | 57.03 | 61.72 | 62.72 | 67.75 | 49.22 | 75.00 | 74.22 | 46.78 | 59.38 | 78.12 | 67.19 | 65.62 | 63.62 |
| SAR | 42.97 | 34.38 | 46.88 | 32.03 | 34.38 | 35.16 | 59.38 | 62.50 | 27.34 | 47.66 | 44.53 | 35.16 | 47.66 | 50.00 | 57.81 | 26.56 | 28.91 | 31.25 | 51.56 | 39.06 | 38.28 | 41.59 |
| READ | 51.88 | 49.22 | 68.75 | 46.88 | 65.62 | 66.41 | 70.31 | 79.69 | 70.31 | 58.59 | 60.94 | 65.62 | 63.72 | 54.91 | 75.00 | 73.44 | 46.56 | 61.72 | 76.56 | 68.75 | 67.19 | 64.10 |
| CMAE | 50.66 | 48.44 | 68.75 | 46.88 | 65.62 | 64.84 | 72.66 | 80.47 | 69.53 | 57.03 | 61.72 | 61.72 | 67.97 | 49.22 | 75.78 | 74.22 | 47.78 | 61.72 | 76.56 | 68.75 | 65.62 | 63.62 |
| DeYO | 46.88 | 48.44 | 68.75 | 46.88 | 65.62 | 66.41 | 72.66 | 80.47 | 70.31 | 58.59 | 60.94 | 60.16 | 65.53 | 50.09 | 75.78 | 75.00 | 48.44 | 60.16 | 78.91 | 70.31 | 66.41 | 63.65 |
| FOZO | 51.66 | 48.44 | 68.75 | 46.88 | 65.62 | 65.62 | 72.66 | 80.47 | 69.53 | 57.03 | 61.72 | 62.50 | 67.97 | 51.56 | 75.00 | 75.00 | 46.78 | 60.94 | 78.12 | 66.41 | 65.62 | 63.73 |
| PTA | 46.88 | 49.22 | 68.75 | 46.09 | 66.41 | 66.41 | 71.88 | 79.69 | 70.31 | 57.03 | 61.72 | 64.84 | 67.53 | 53.31 | 75.00 | 74.22 | 48.00 | 60.94 | 78.91 | 68.75 | 63.28 | 63.58 |
| BECoTTA | 45.31 | 49.66 | 69.09 | 46.09 | 66.41 | 66.41 | 73.44 | 80.47 | 67.97 | 54.69 | 60.16 | 57.81 | 68.71 | 46.09 | 74.22 | 74.88 | 47.78 | 60.16 | 74.22 | 66.41 | 65.62 | 63.54 |
| DASP | 47.66 | 48.44 | 68.75 | 46.88 | 65.62 | 64.84 | 72.66 | 80.47 | 67.97 | 57.03 | 61.74 | 61.70 | 67.97 | 49.22 | 75.78 | 74.22 | 50.78 | 61.72 | 76.56 | 68.75 | 65.62 | 63.54 |
| **ZORA** | **53.91** | **54.25** | **69.53** | **47.09** | **67.19** | **69.53** | **74.56** | **81.25** | **71.44** | **58.81** | **62.94** | **66.06** | **68.75** | **64.41** | **75.88** | **75.34** | **52.44** | **65.62** | **79.12** | **70.78** | **70.53** | **66.59** |

Table 17: Accuracy comparison with SOTAs on VGGSound-C for interleaved single-modality corruption streams (severity level = 5). Here, $*$ denotes the clean modality.

| Method | A:* | | V:* | | A:* | | V:* | | A:* | | | | | V:* | A:* | | V:* | | A:* | | V:* | | A:* | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | V:Gaus. | V:Shot | A:Gaus. | V:Imp. | V:Def. | A:Traf. | V:Glas. | V:Mot. | A:Crowd | V:Zoom | V:Snow | V:Frost | A:Rain | V:Fog | V:Brit. | A:Thun. | V:Contr. | V:Elas. | A:Wind | V:Pix. | V:JPEG | Avg. |
| | $t$ ⟶ | | | | | | | | | | | | | | | | | | | | | |
| Source | 47.22 | 41.75 | 22.22 | 44.09 | 49.56 | 17.53 | 54.25 | 62.06 | 19.09 | 55.25 | 56.38 | 51.12 | 21.44 | 44.09 | 60.50 | 31.66 | 44.69 | 56.59 | 20.66 | 52.81 | 61.28 | 41.54 |
| FOA | 48.78 | 41.22 | 23.16 | 46.78 | 50.69 | 17.09 | 54.59 | 60.72 | 20.12 | 55.47 | 53.03 | 49.91 | 25.47 | 48.56 | 60.84 | 32.91 | 44.25 | 56.59 | 21.78 | 52.69 | 60.84 | 44.21 |
| T3A | 6.25 | 15.62 | 4.69 | 14.06 | 27.34 | 3.91 | 34.38 | 27.34 | 5.47 | 32.81 | 22.66 | 13.28 | 6.25 | 8.59 | 35.16 | 3.91 | 25.00 | 26.56 | 7.03 | 25.00 | 25.78 | 17.67 |
| TSA | 42.97 | 26.56 | 5.47 | 23.44 | 35.16 | 14.84 | 28.12 | 33.59 | 14.06 | 25.00 | 34.38 | 34.38 | 10.16 | 24.22 | 35.16 | 10.94 | 21.88 | 33.59 | 14.84 | 25.78 | 28.91 | 24.93 |
| Tent | 50.54 | 43.43 | 23.66 | 47.66 | 50.78 | 19.75 | 57.03 | 64.62 | 22.09 | 58.38 | 57.03 | 54.78 | 25.44 | 52.14 | 63.04 | 32.97 | 45.81 | 58.59 | 22.88 | 54.91 | 62.62 | 45.72 |
| SAR | 31.25 | 26.56 | 6.25 | 25.78 | 32.81 | 1.56 | 35.94 | 39.84 | 3.91 | 32.81 | 39.06 | 26.56 | 2.34 | 24.22 | 35.16 | 9.38 | 34.38 | 42.19 | 8.59 | 32.03 | 31.25 | 24.85 |
| READ | 50.78 | 42.97 | 22.66 | 48.44 | 53.89 | 15.62 | 57.03 | 64.84 | 21.88 | 57.81 | 56.25 | 53.12 | 23.44 | 48.44 | 61.72 | 27.97 | 43.91 | 58.59 | 19.53 | 52.34 | 63.28 | 44.98 |
| CMAE | 49.22 | 43.75 | 24.22 | 46.09 | 51.56 | 19.53 | 56.25 | 64.06 | 22.09 | 56.25 | 58.38 | 53.12 | 23.44 | 46.09 | 62.50 | 33.66 | 45.69 | 58.59 | 23.66 | 54.81 | 63.28 | 45.54 |
| DeYO | 48.22 | 42.75 | 24.22 | 45.09 | 50.56 | 17.75 | 56.81 | 63.06 | 20.88 | 56.03 | 57.59 | 52.12 | 24.78 | 46.66 | 63.06 | 32.88 | 44.03 | 57.59 | 22.44 | 53.69 | 61.50 | 44.80 |
| FOZO | 48.22 | 43.53 | 22.44 | 45.88 | 50.56 | 16.97 | 56.03 | 62.84 | 20.88 | 59.16 | 57.59 | 51.34 | 24.00 | 50.56 | 63.06 | 28.53 | 42.12 | 59.16 | 20.88 | 53.69 | 64.62 | 44.91 |
| PTA | 50.00 | 43.75 | 21.88 | 48.56 | 53.12 | 20.97 | 55.47 | 64.06 | 15.62 | 57.81 | 57.03 | 53.91 | 22.31 | 46.44 | 63.28 | 15.62 | 50.78 | 56.25 | 19.53 | 46.09 | 56.25 | 43.75 |
| BECoTTA | 51.00 | 43.52 | 10.94 | 43.75 | 53.12 | 18.75 | 57.81 | 64.06 | 19.53 | 56.25 | 55.47 | 53.91 | 24.22 | 45.31 | 60.16 | 10.59 | 46.44 | 57.03 | 21.09 | 56.25 | 60.94 | 44.08 |
| DASP | 51.56 | 42.31 | 24.22 | 42.56 | 51.31 | 19.53 | 54.09 | 57.22 | 21.09 | 54.31 | 52.88 | 54.66 | 25.44 | 51.56 | 64.01 | 32.66 | 41.44 | 58.22 | 24.66 | 45.31 | 57.34 | 44.11 |
| **ZORA** | **52.34** | **44.53** | **24.94** | **49.62** | **53.91** | **21.00** | **58.59** | **64.94** | **23.47** | **60.47** | **58.69** | **55.81** | **27.91** | **52.34** | **64.06** | **33.72** | **46.88** | **59.38** | **25.00** | **56.69** | **64.94** | **47.58** |

Table 18: Accuracy comparison with SOTAs on Kinetics50-C under the PairTrack-Async stream. The best performances are highlighted.

| Method | A:Gaus. | | | V:Def. | V:Glas. | A:Traf. | | | A:Crowd. | | | A:Rain | | A:Thun. | | | A:Wind | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | V:Gaus. | V:Shot | V:Imp. | | | V:Mot. | V:Zoom | V:Snow | V:Frost | V:Fog | V:Brit. | V:Contr. | V:Elas. | V:Pix. | V:JPEG | Avg. |
| Source | 30.90 | 29.89 | 27.39 | 33.72 | 39.63 | 54.63 | 65.81 | 48.12 | 42.45 | 23.69 | 66.76 | 40.35 | 49.01 | 60.74 | 53.97 | 44.47 |
| FOA | 33.83 | 36.68 | 30.65 | 37.35 | 43.88 | 59.21 | 67.92 | 53.75 | 54.88 | 39.95 | 70.89 | 52.11 | 57.77 | 62.62 | 58.72 | 50.68 |
| T3A | 12.27 | 21.37 | 19.34 | 22.67 | 27.65 | 34.54 | 47.79 | 27.21 | 20.95 | 19.76 | 51.11 | 26.43 | 32.36 | 45.55 | 41.33 | 30.02 |
| TSA | 32.42 | 33.76 | 28.73 | 36.91 | 44.11 | 47.34 | 55.55 | 47.84 | 36.61 | 32.96 | 58.87 | 45.77 | 51.34 | 48.28 | 48.74 | 43.28 |
| Tent | 30.43 | 29.95 | 26.73 | 32.93 | 35.55 | 57.10 | 65.85 | 48.41 | 38.91 | 14.63 | 63.87 | 38.26 | 44.26 | 55.28 | 49.82 | 42.13 |
| SAR | 26.77 | 20.05 | 16.15 | 17.65 | 24.87 | 31.74 | 42.35 | 32.84 | 21.32 | 10.23 | 36.74 | 26.77 | 33.85 | 39.28 | 31.21 | 27.45 |
| READ | 30.31 | 32.74 | 28.93 | 35.45 | 43.96 | 55.66 | 65.61 | 48.27 | 47.45 | 27.59 | 66.54 | 46.95 | 55.48 | 60.86 | 54.88 | 46.71 |
| CMAE | 30.26 | 29.54 | 27.32 | 33.90 | 39.75 | 55.07 | 65.09 | 48.51 | 42.74 | 22.94 | 66.35 | 40.59 | 49.39 | 60.05 | 54.41 | 44.39 |
| DeYO | 30.86 | 30.23 | 26.68 | 33.35 | 39.62 | 55.06 | 66.86 | 49.38 | 43.29 | 24.69 | 67.31 | 44.31 | 53.08 | 59.95 | 57.91 | 45.51 |
| FOZO | 30.83 | 29.12 | 26.90 | 33.43 | 37.65 | 55.96 | 66.71 | 47.21 | 41.20 | 22.31 | 66.39 | 42.36 | 48.75 | 59.73 | 53.21 | 44.12 |
| PTA | 28.56 | 27.48 | 24.06 | 32.04 | 24.55 | 57.97 | 68.94 | 47.87 | 38.52 | 12.84 | 58.93 | 22.32 | 19.22 | 38.15 | 27.07 | 35.23 |
| BECoTTA | 30.12 | 33.28 | 28.49 | 32.11 | 37.89 | 55.33 | 66.54 | 48.90 | 44.09 | 22.66 | 63.83 | 40.44 | 52.52 | 61.31 | 55.18 | 44.85 |
| DASP | 30.50 | 30.00 | 27.00 | 33.50 | 39.50 | 55.00 | 65.50 | 48.50 | 42.50 | 23.50 | 66.50 | 40.50 | 49.00 | 60.50 | 54.00 | 44.40 |
| **ZORA** | **36.20** | **44.12** | **34.83** | **48.55** | **51.87** | **64.11** | **70.10** | **58.56** | **57.38** | **48.33** | **73.56** | **59.07** | **62.59** | **64.10** | **61.27** | **55.64** |

Table 19: Accuracy comparison with SOTAs on VGGSound-C under the PairTrack-Async stream.

| Method | A:Gaus. | | | V:Def. | V:Glas. | A:Traf. | | | A:Crowd. | | | A:Rain | | A:Thun. | | | A:Wind | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | V:Gaus. | V:Shot | V:Imp. | | | V:Mot. | V:Zoom | V:Snow | V:Frost | V:Fog | V:Brit. | V:Contr. | V:Elas. | V:Pix. | V:JPEG | Avg. |
| Source | 4.17 | 8.69 | 6.17 | 7.26 | 7.78 | 21.54 | 16.22 | 17.13 | 5.14 | 3.59 | 22.84 | 13.72 | 15.70 | 20.16 | 19.14 | 12.62 |
| FOA | 10.61 | 11.38 | 11.06 | 11.80 | 11.28 | 26.40 | 19.83 | 19.28 | 10.98 | 9.72 | 31.35 | 21.81 | 22.16 | 28.96 | 22.30 | 17.93 |
| T3A | 1.02 | 0.21 | 0.22 | 2.52 | 0.49 | 3.48 | 6.83 | 6.80 | 0.82 | 3.18 | 9.85 | 1.23 | 4.73 | 5.97 | 6.30 | 3.58 |
| TSA | 8.49 | 9.25 | 9.41 | 7.84 | 11.03 | 14.94 | 15.92 | 12.21 | 5.22 | 2.92 | 26.80 | 14.66 | 19.54 | 16.61 | 15.33 | 12.68 |
| Tent | 3.87 | 4.69 | 1.90 | 5.82 | 6.79 | 22.86 | 15.67 | 13.34 | 4.16 | 0.91 | 17.12 | 7.69 | 9.76 | 10.34 | 6.51 | 8.76 |
| SAR | 2.02 | 5.81 | 3.82 | 4.10 | 6.24 | 10.14 | 9.98 | 7.99 | 1.59 | 1.37 | 11.45 | 10.74 | 12.61 | 9.31 | 10.64 | 7.19 |
| READ | 2.47 | 2.09 | 0.20 | 4.68 | 7.01 | 20.81 | 17.68 | 14.31 | 0.69 | 0.19 | 6.22 | 0.47 | 0.73 | 1.21 | 0.26 | 5.27 |
| CMAE | 4.42 | 7.41 | 5.69 | 7.04 | 7.78 | 21.71 | 16.80 | 16.88 | 5.31 | 3.48 | 21.20 | 14.32 | 15.71 | 20.22 | 19.49 | 12.50 |
| DeYO | 4.21 | 8.54 | 6.26 | 7.88 | 7.61 | 21.79 | 16.59 | 17.70 | 4.97 | 3.67 | 22.10 | 12.32 | 14.91 | 19.67 | 20.15 | 12.56 |
| FOZO | 4.15 | 4.71 | 2.60 | 6.76 | 6.51 | 23.10 | 16.83 | 15.81 | 4.74 | 1.40 | 18.86 | 10.82 | 12.12 | 18.27 | 13.76 | 10.70 |
| PTA | 4.09 | 2.97 | 0.73 | 3.15 | 3.43 | 19.32 | 13.29 | 6.78 | 1.20 | 0.67 | 9.58 | 2.40 | 2.26 | 1.85 | 0.77 | 4.83 |
| BECoTTA | 1.25 | 1.11 | 0.26 | 6.55 | 6.47 | 22.16 | 16.91 | 15.09 | 5.26 | 3.69 | 23.46 | 7.74 | 1.06 | 2.03 | 0.56 | 7.57 |
| DASP | 4.50 | 8.50 | 6.50 | 7.00 | 7.50 | 21.50 | 16.50 | 17.00 | 5.00 | 3.50 | 22.50 | 14.00 | 15.50 | 20.50 | 19.50 | 12.63 |
| **ZORA** | **13.69** | **18.20** | **12.81** | **14.11** | **20.08** | **29.26** | **25.05** | **23.09** | **15.09** | **13.76** | **36.15** | **26.69** | **28.77** | **29.20** | **24.93** | **22.06** |

Table 20: Complete accuracy comparison on Kinetics50-C under the PairTrack-Sync stream. The proposed method is highlighted.

| Method | A:Gaus. | | | A:Traf. | | A:Crowd. | | | A:Rain | | A:Thun. | | | A:Wind | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | V:Gaus. | V:Shot | V:Imp. | V:Def. | V:Glas. | V:Mot. | V:Zoom | V:Snow | V:Frost | V:Fog | V:Brit. | V:Contr. | V:Elas. | V:Pix. | V:JPEG | Avg. |
| Source | 31.69 | 38.56 | 35.39 | 37.33 | 37.92 | 55.70 | 67.33 | 47.33 | 43.74 | 29.14 | 68.17 | 44.16 | 49.20 | 63.40 | 60.79 | 47.32 |
| FOA | 36.05 | 43.88 | 36.42 | 40.82 | 43.55 | 57.86 | 69.11 | 53.15 | 51.77 | 47.23 | 70.35 | 56.10 | 57.66 | 62.13 | 63.80 | 52.66 |
| T3A | 14.71 | 22.26 | 22.90 | 23.65 | 28.70 | 40.18 | 45.82 | 28.55 | 22.47 | 22.67 | 50.73 | 34.85 | 33.72 | 51.34 | 45.40 | 32.53 |
| TSA | 33.50 | 38.00 | 33.50 | 39.50 | 44.00 | 48.00 | 56.00 | 50.00 | 34.00 | 39.00 | 57.50 | 49.50 | 52.00 | 49.00 | 49.00 | 44.83 |
| Tent | 31.41 | 37.82 | 35.31 | 36.30 | 38.40 | 57.53 | 69.16 | 45.60 | 39.87 | 26.86 | 66.71 | 42.21 | 46.44 | 63.18 | 58.73 | 46.37 |
| SAR | 24.85 | 25.46 | 20.20 | 20.84 | 25.44 | 27.72 | 38.76 | 35.05 | 22.01 | 12.27 | 39.22 | 32.22 | 35.70 | 45.68 | 33.78 | 29.28 |
| READ | 31.22 | 38.71 | 35.28 | 38.18 | 40.71 | 56.62 | 68.74 | 48.18 | 48.34 | 36.29 | 69.06 | 51.59 | 54.48 | 64.19 | 60.72 | 49.49 |
| CMAE | 31.52 | 38.55 | 35.45 | 37.48 | 38.31 | 55.97 | 67.79 | 46.79 | 42.68 | 28.36 | 68.48 | 44.56 | 48.24 | 63.21 | 60.85 | 47.22 |
| DeYO | 31.76 | 38.73 | 35.75 | 37.11 | 40.65 | 56.24 | 70.46 | 47.25 | 41.04 | 29.53 | 68.99 | 48.28 | 50.95 | 65.08 | 60.26 | 48.14 |
| FOZO | 31.64 | 38.34 | 35.57 | 36.77 | 38.30 | 57.31 | 68.63 | 46.49 | 41.01 | 28.42 | 67.73 | 46.29 | 46.90 | 63.71 | 60.65 | 47.18 |
| PTA | 30.83 | 33.68 | 29.17 | 39.33 | 39.33 | 58.74 | 69.27 | 48.70 | 42.66 | 26.37 | 68.06 | 42.37 | 47.46 | 61.72 | 56.90 | 46.31 |
| BECoTTA | 31.66 | 37.49 | 34.27 | 37.86 | 38.72 | 56.79 | 68.81 | 46.59 | 41.49 | 25.70 | 66.22 | 43.37 | 46.12 | 60.09 | 61.30 | 46.43 |
| DASP | 31.50 | 38.50 | 35.50 | 37.50 | 38.00 | 55.50 | 67.50 | 47.00 | 43.50 | 29.00 | 68.50 | 44.50 | 49.00 | 63.50 | 61.00 | 47.33 |
| **ZORA** | **37.44** | **45.17** | **39.95** | **50.71** | **49.73** | **65.31** | **72.12** | **57.91** | **52.72** | **55.72** | **74.15** | **60.84** | **59.54** | **65.22** | **64.56** | **56.74** |

Table 21: Complete accuracy comparison on VGGSound-C under the PairTrack-Sync stream. The proposed method is highlighted.

| Method | A:Gaus. | | | A:Traf. | | A:Crowd. | | | A:Rain | | A:Thun. | | | A:Wind | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | V:Gaus. | V:Shot | V:Imp. | V:Def. | V:Glas. | V:Mot. | V:Zoom | V:Snow | V:Frost | V:Fog | V:Brit. | V:Contr. | V:Elas. | V:Pix. | V:JPEG | Avg. |
| Source | 3.65 | 10.04 | 8.30 | 8.99 | 7.83 | 19.20 | 17.31 | 18.82 | 7.16 | 4.73 | 26.60 | 16.89 | 15.27 | 22.28 | 17.78 | 13.66 |
| FOA | 12.75 | 15.29 | 15.45 | 13.82 | 12.38 | 23.08 | 22.38 | 20.79 | 13.09 | 9.72 | 35.20 | 26.12 | 25.23 | 26.71 | 25.16 | 19.81 |
| T3A | 1.30 | 0.83 | 0.10 | 2.32 | 0.97 | 4.48 | 5.49 | 5.69 | 1.31 | 1.64 | 11.18 | 4.44 | 4.74 | 6.84 | 8.93 | 4.02 |
| TSA | 10.50 | 12.00 | 10.00 | 11.50 | 11.00 | 14.00 | 18.00 | 13.50 | 4.50 | 4.50 | 24.50 | 18.00 | 17.50 | 15.00 | 14.00 | 13.23 |
| Tent | 3.84 | 7.40 | 3.24 | 7.20 | 8.51 | 18.23 | 14.92 | 18.71 | 4.18 | 2.10 | 18.69 | 9.30 | 9.28 | 10.35 | 7.25 | 9.55 |
| SAR | 0.20 | 5.17 | 4.68 | 3.70 | 6.28 | 8.91 | 9.65 | 7.25 | 2.43 | 2.28 | 13.93 | 12.13 | 11.68 | 10.48 | 9.67 | 7.23 |
| READ | 2.87 | 2.67 | 0.10 | 5.77 | 8.82 | 19.26 | 17.52 | 16.93 | 1.23 | 0.25 | 5.16 | 0.38 | 0.56 | 1.34 | 0.22 | 5.54 |
| CMAE | 3.95 | 10.35 | 8.21 | 8.66 | 8.28 | 19.79 | 17.36 | 18.74 | 7.34 | 4.60 | 26.39 | 16.81 | 14.66 | 22.18 | 18.31 | 13.71 |
| DeYO | 3.95 | 10.27 | 9.39 | 8.89 | 8.22 | 20.22 | 17.53 | 19.13 | 7.64 | 4.39 | 24.75 | 17.02 | 13.81 | 22.41 | 19.28 | 13.79 |
| FOZO | 3.26 | 8.29 | 5.17 | 8.12 | 8.44 | 18.67 | 16.25 | 18.99 | 5.77 | 3.56 | 20.81 | 12.89 | 11.72 | 19.15 | 14.17 | 11.68 |
| PTA | 3.19 | 3.47 | 1.78 | 4.79 | 3.20 | 17.84 | 15.17 | 12.20 | 1.73 | 0.05 | 9.81 | 3.31 | 2.43 | 1.95 | 0.62 | 5.44 |
| BECoTTA | 2.41 | 1.32 | 0.22 | 1.27 | 0.76 | 5.15 | 0.29 | 1.14 | 1.05 | 0.23 | 4.23 | 0.12 | 1.63 | 1.38 | 0.71 | 1.46 |
| DASP | 4.00 | 10.00 | 8.50 | 9.00 | 8.00 | 19.50 | 17.50 | 18.50 | 7.50 | 4.50 | 26.50 | 17.00 | 15.00 | 22.50 | 18.00 | 13.73 |
| **ZORA** | **15.76** | **18.19** | **16.21** | **15.72** | **19.22** | **27.37** | **24.18** | **21.46** | **17.70** | **19.71** | **40.18** | **32.39** | **31.60** | **28.48** | **28.37** | **23.77** |

Table 22: Kinetics50-C audio-only corruption streams.

| Method | Crowd | Gaus. | Rain | Thun. | Traf. | Wind | Avg. |
|---|---|---|---|---|---|---|---|
| Source | 64.56 | 70.25 | 68.62 | 68.69 | 62.44 | 67.19 | 67.12 |
| FOA | 68.38 | 71.38 | 71.25 | **70.81** | 64.94 | 69.69 | 69.41 |
| T3A | 57.00 | 58.31 | 58.69 | 57.62 | 50.56 | 61.12 | 57.22 |
| TSA | 51.88 | 53.12 | 48.50 | 54.12 | 47.00 | 50.81 | 50.91 |
| Tent | 66.00 | 70.06 | 68.19 | 69.06 | 66.50 | 68.31 | 68.02 |
| SAR | 34.69 | 49.75 | 40.81 | 38.75 | 35.62 | 44.00 | 40.60 |
| READ | 66.62 | 70.25 | 68.50 | 70.31 | 65.88 | 67.50 | 68.18 |
| CMAE | 64.56 | 70.38 | 69.62 | 68.69 | 62.44 | 67.25 | 67.16 |
| DeYO | 66.44 | 68.62 | 68.12 | 70.62 | 66.62 | 68.50 | 68.16 |
| FOZO | 64.25 | 68.25 | 67.06 | 67.25 | 63.81 | 66.56 | 66.03 |
| PTA | 65.81 | 70.50 | 69.50 | 69.19 | **67.00** | 67.38 | 68.56 |
| BECoTTA | 67.94 | 70.69 | 71.00 | 70.06 | 63.69 | 69.81 | 68.86 |
| DASP | 66.56 | 70.25 | 70.75 | 69.69 | 63.44 | 68.25 | 68.16 |
| **ZORA** | **69.38** | **72.19** | **72.13** | 70.25 | 63.81 | **70.56** | **69.72** |

Table 23: VGGSound-C audio-only corruption streams.

| Method | Crowd | Gaus. | Rain | Thun. | Traf. | Wind | Avg. |
|---|---|---|---|---|---|---|---|
| Source | 22.25 | 17.69 | 20.19 | 20.12 | 18.56 | 27.62 | 21.07 |
| FOA | 23.12 | 22.19 | 21.88 | 21.44 | 18.12 | 27.31 | 22.34 |
| T3A | 8.12 | 10.56 | 8.75 | 6.31 | 7.62 | 10.00 | 8.56 |
| TSA | 7.00 | 0.31 | 5.81 | 7.81 | 6.25 | 7.00 | 5.70 |
| Tent | 5.69 | 4.62 | 6.81 | 8.38 | 4.81 | 15.31 | 7.60 |
| SAR | 5.69 | 5.75 | 5.19 | 5.38 | 5.00 | 7.38 | 5.73 |
| READ | 7.38 | 2.25 | 2.81 | 5.69 | 5.31 | 6.50 | 4.99 |
| CMAE | 23.19 | 20.50 | 21.19 | 21.00 | 18.56 | 27.50 | 21.99 |
| DeYO | 23.50 | 20.38 | 22.50 | **22.62** | 16.75 | 27.06 | 22.14 |
| FOZO | 12.69 | 10.19 | 10.62 | 12.62 | 12.06 | 26.12 | 14.05 |
| PTA | 2.94 | 2.69 | 3.12 | 4.06 | 2.06 | 6.56 | 3.57 |
| BECoTTA | 7.69 | 3.12 | 1.06 | 6.31 | 7.44 | 10.25 | 5.98 |
| DASP | 23.31 | 7.62 | 21.19 | 21.12 | 17.56 | 25.69 | 19.42 |
| **ZORA** | **25.25** | **22.44** | **24.19** | 19.81 | **18.81** | **27.81** | **23.05** |

**Batch-size analysis.** Figure 8 evaluates $B \in \{4, 8, 16, 32\}$ in addition to our batch-size-one protocol. ZORA performs best at every batch size; at $B = 4$, it reaches $53.6\%$ and exceeds DASP, FOA, PTA, and Tent by $3.3$, $5.8$, $9.4$, and $9.6$ points, respectively. The margin narrows with larger batches but remains positive throughout.

Table 24: Kinetics50-C video-only corruption streams.

| Method | Brit. | Contr. | Defoc. | Elas. | Fog | Frost | Gaus. | Glass | Imp. | JPEG | Motion | Pix. | Shot | Snow | Zoom | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | 70.62 | 44.79 | 62.71 | 59.58 | 42.50 | 62.29 | 44.17 | 56.04 | 46.67 | 60.62 | 67.08 | 61.67 | 44.38 | 54.38 | 62.50 | 56.00 |
| FOA | 74.04 | 48.42 | 66.75 | 63.21 | 51.33 | 68.79 | 48.88 | 60.92 | 52.58 | 64.46 | 70.92 | 65.29 | 49.67 | 58.42 | 65.29 | 60.60 |
| T3A | 59.79 | 35.62 | 53.96 | 53.54 | 39.38 | 39.17 | 42.08 | 52.71 | 44.38 | 55.62 | 58.12 | 56.04 | 43.54 | 40.83 | 53.75 | 48.57 |
| TSA | 66.67 | 42.92 | 57.08 | 57.08 | 45.83 | 63.33 | 46.67 | 54.38 | 49.17 | 57.08 | 62.71 | 56.67 | 46.04 | 50.62 | 58.96 | 54.35 |
| Tent | 73.00 | 47.58 | 65.08 | 63.42 | 45.92 | 65.92 | 47.38 | 59.25 | 51.75 | 62.79 | 68.62 | 64.67 | 49.25 | 57.58 | 66.54 | 59.25 |
| SAR | 50.21 | 35.21 | 48.75 | 51.04 | 41.25 | 48.96 | 37.71 | 50.83 | 35.62 | 43.96 | 53.96 | 47.92 | 35.62 | 40.00 | 51.67 | 44.85 |
| READ | 74.04 | 47.58 | 65.08 | 64.04 | 47.38 | 66.75 | 46.96 | 60.08 | 51.54 | 63.62 | 69.46 | 65.71 | 47.58 | 58.00 | 64.67 | 59.50 |
| CMAE | 72.83 | 46.79 | 64.50 | 61.58 | 43.50 | 64.29 | 46.17 | 58.04 | 48.67 | 62.62 | 69.08 | 63.67 | 46.38 | 56.38 | 64.50 | 58.00 |
| DeYO | 75.25 | 49.25 | 67.71 | 63.58 | 37.75 | 64.83 | 47.96 | 58.38 | 52.12 | 66.08 | 71.25 | 67.17 | 50.62 | 57.54 | 63.42 | 59.53 |
| FOZO | 70.17 | 45.38 | 62.88 | 64.83 | **48.25** | 65.33 | 46.21 | 59.33 | 49.12 | 61.83 | 66.25 | 61.88 | 45.62 | 56.21 | 65.38 | 57.78 |
| PTA | 74.17 | 46.04 | 65.96 | 64.25 | 40.83 | 66.25 | 43.54 | 55.62 | 43.33 | 57.29 | 70.62 | 61.25 | 43.54 | 57.08 | 62.92 | 56.85 |
| BECoTTA | 75.41 | 49.38 | 68.12 | 64.75 | 46.67 | 67.71 | 49.58 | 59.38 | 53.29 | 65.88 | 71.29 | 66.25 | 49.38 | 59.36 | 66.45 | 60.97 |
| DASP | 75.62 | 49.79 | 67.71 | 64.58 | 47.50 | 67.29 | 49.38 | 61.04 | 51.67 | 65.62 | 72.08 | 66.67 | 49.38 | **59.38** | **67.50** | 61.01 |
| **ZORA** | **75.42** | **50.01** | **68.75** | **65.42** | 47.29 | **69.38** | **49.38** | **61.88** | **54.38** | **66.04** | **71.88** | **67.50** | **52.08** | 55.18 | 64.17 | **61.10** |

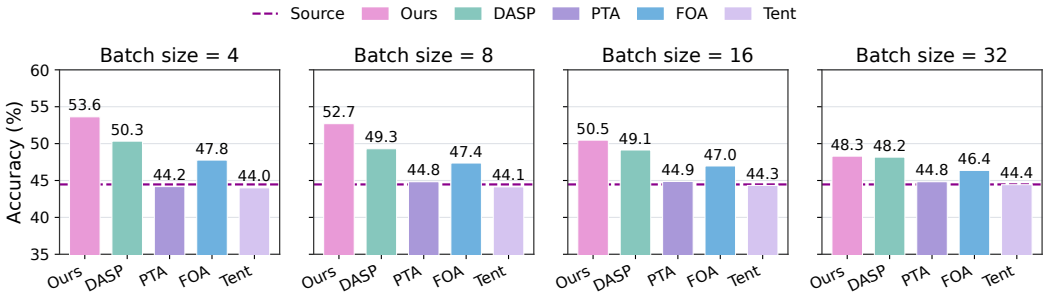



Figure 8: Results on different batch sizes under the Kinetics50-C PairTrack-Async protocol.

Table 25: VGGSound-C video-only corruption streams.

| Method | Brit. | Contr. | Defoc. | Elas. | Fog | Frost | Gaus. | Glass | Imp. | JPEG | Motion | Pix. | Shot | Snow | Zoom | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | 55.04 | 39.42 | 57.96 | 47.54 | 48.17 | 50.04 | 39.00 | 49.83 | 42.54 | 51.50 | 58.79 | 50.25 | 43.17 | 44.00 | 51.29 | 49.57 |
| FOA | 57.67 | 44.54 | 61.62 | 51.21 | 52.46 | 53.50 | 41.62 | 51.21 | 46.42 | 53.50 | 61.62 | 54.92 | 49.00 | 43.25 | 53.71 | 51.75 |
| T3A | 28.75 | 15.21 | 27.92 | 19.38 | 11.25 | 14.58 | 15.62 | 13.96 | 16.46 | 26.67 | 24.79 | 21.04 | 19.38 | 16.25 | 21.25 | 19.50 |
| TSA | 37.50 | 26.88 | 37.29 | 31.46 | 27.08 | 32.29 | 25.42 | 26.88 | 27.08 | 34.58 | 40.00 | 32.71 | 28.54 | 28.33 | 32.50 | 31.24 |
| Tent | 56.04 | 40.62 | 60.00 | 50.42 | 51.88 | 52.71 | 40.21 | 54.17 | 43.96 | 52.08 | 59.17 | 51.04 | 45.00 | 44.17 | 55.21 | 50.44 |
| SAR | 29.58 | 30.83 | 38.96 | 29.38 | 39.38 | 43.54 | 23.54 | 31.25 | 27.08 | 31.46 | 37.08 | 28.33 | 27.29 | 33.75 | 40.62 | 32.81 |
| READ | 57.71 | 41.25 | 60.62 | 53.12 | **57.29** | 53.12 | 43.54 | 54.17 | 44.38 | 51.67 | 58.75 | 52.50 | 46.46 | 46.88 | **56.25** | 51.85 |
| CMAE | 55.25 | 39.42 | 47.96 | 47.54 | 48.38 | 50.04 | 39.00 | 49.83 | 42.54 | 51.50 | 58.79 | 50.25 | 43.17 | 44.00 | 51.50 | 48.61 |
| DeYO | 58.04 | 42.83 | 63.04 | 52.21 | 52.00 | 53.04 | 43.79 | 52.71 | 47.62 | 53.88 | 61.58 | 54.29 | 46.79 | 45.33 | 56.17 | 52.22 |
| FOZO | 54.83 | 38.83 | 57.96 | 49.42 | 49.83 | 53.38 | 38.79 | 52.33 | 43.58 | 49.62 | 57.96 | 49.83 | 42.54 | 44.00 | 54.21 | 49.21 |
| PTA | 55.42 | 40.42 | 59.17 | 51.46 | 56.46 | 54.79 | 42.92 | 53.33 | 44.38 | 50.42 | 54.79 | 50.62 | 46.04 | 45.00 | 55.00 | 50.68 |
| BECoTTA | 58.21 | 42.58 | 61.33 | 51.75 | 54.46 | 53.83 | 41.96 | 53.21 | 48.08 | 51.04 | 61.17 | 54.42 | 46.33 | 45.92 | 53.83 | 51.88 |
| DASP | 58.04 | 45.42 | 62.96 | **53.54** | 54.67 | 54.54 | 39.50 | 50.33 | 43.04 | 52.00 | 59.29 | 50.75 | 43.67 | 44.50 | 55.00 | 52.08 |
| **ZORA** | **58.92** | **46.42** | **63.08** | 51.00 | 51.21 | **55.17** | **43.92** | **53.50** | **48.08** | **52.04** | **62.04** | **55.38** | **50.79** | **44.96** | 49.96 | **52.43** |

**Real-world cross-dataset transfer.** Table 26 evaluates natural continual shifts among MOSI, MOSEI, and SIMS. For SIMS $\rightarrow$ MOSI+MOSEI, ZORA improves Acc/F1 over the source model by $6.88/11.86$ points on MOSI and $1.26/10.55$ points on MOSEI; for MOSI $\rightarrow$ SIMS+MOSEI, the gains are $2.67/1.59$ points on SIMS and $2.62/2.85$ points on MOSEI. These results confirm generalization beyond synthetic corruptions and indicate that BARC reduces biased confident feedback under cross-dataset shifts.

Table 26: Real-world cross-dataset continual transfer results on multimodal sentiment streams. Each source dataset is adapted to the remaining target datasets in sequence.

| Method | Source: SIMS | | | | Source: MOSI | | | |
|---|---|---|---|---|---|---|---|---|
| | SIMS→MOSI | | SIMS→MOSEI | | MOSI→SIMS | | MOSI→MOSEI | |
| | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 |
| `Source` | 36.43 | 27.45 | 48.87 | 33.10 | 38.82 | 40.74 | 66.65 | 65.20 |
| `SuMi` | 40.02 | 24.39 | 48.44 | 31.97 | 39.93 | 40.82 | 68.33 | 67.83 |
| `FOA` | 36.73 | 28.00 | 46.56 | 30.33 | 39.82 | 40.76 | 67.40 | 67.12 |
| `PTA` | 38.54 | 37.78 | 41.98 | 41.98 | 38.84 | 40.03 | 66.98 | 66.60 |
| **ZORA** | **43.31** | **39.31** | **50.13** | **43.65** | **41.49** | **42.33** | **69.27** | **68.05** |

## G LIMITATIONS

ZORA focuses on forward-only online adaptation for edge-side AV-CTTA, where efficiency is critical. Although it avoids backpropagation, activation storage, and gradient computation, ZOO still requires additional forward queries to estimate update directions, which may introduce extra latency compared with standard inference. Moreover, SSLS depends on a small number of early target samples for layer selection, and BARC relies on a local history window for prediction-bias estimation. While our experiments show that these estimates are stable with limited samples, extremely short or rapidly changing streams may still affect their reliability. In future work, we will explore more query-efficient ZOO updates and extend ZORA to broader multimodal tasks beyond audio-visual classification.