Audio-Visual Continual Test-Time Adaptation *without* Forgetting

Sarthak Kumar Maharana[1], Akshay Mehra[2], Bhavya Ramakrishna[2], Yunhui Guo[1], and Guan-Ming Su[2]

[1] The University of Texas at Dallas, Richardson, TX 75080, USA
[2] Dolby Laboratories, Inc., San Francisco, CA 94103, USA
sarthak.maharana@utdallas.edu, akshay.mehra@dolby.com
Project page: https://sarthaxxxxx.github.io/AVReCAP/index.html

Abstract. Audio-visual continual test-time adaptation involves continually adapting a source audio-visual model at test-time, to unlabeled non-stationary domains, where either or both modalities can be distributionally shifted, which hampers online cross-modal learning and eventually leads to poor accuracy. While previous works have tackled this problem, we find that SOTA methods suffer from *catastrophic forgetting* where the model's performance drops well below even the source model due to continual parameter updates at test-time. In this work, we first show that adapting only the modality fusion layer to a target domain not only improves performance on that domain but can also enhance performance on subsequent domains. Based on this strong cross-task transferability of the fusion layer's parameters, we propose a method, AVReCAP, that improves test-time performance of the models without access to any source data. Our approach works by using a selective parameter retrieval mechanism that dynamically retrieves the best fusion layer parameters from a buffer using only a small batch of test data. These parameters are then integrated into the model, adapted to the current test distribution, and saved back for future use. Extensive experiments on benchmark datasets involving unimodal and bimodal corruptions show our proposed AVReCAP significantly outperforms existing methods while minimizing catastrophic forgetting.

1 Introduction

The development of audio-visual models has accelerated rapidly in recent years [1, 2, 15, 16, 21, 24, 29]. However, most prior work evaluates these models under in-distribution settings that closely mirror their pre-training data. In real-world deployment, models operate under evolving test-time distribution shifts that significantly hinder generalization [31, 44]. Consider safety-critical scenarios such as self-driving cars equipped with audio-visual sensors. As the vehicle encounters changing visual conditions (*e.g. snow*, *fog*) or diverse acoustic environments, cross-modal associations learned during training may no longer hold, leading to degraded perception. The challenge becomes even more severe when both modalities shift simultaneously, such as in a robot deployed for scene understanding in a *crowded* street [32].

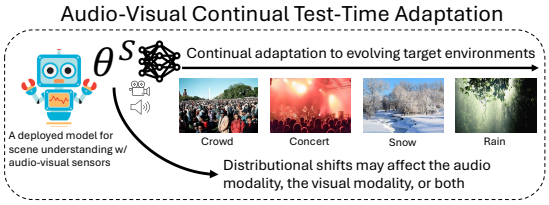


Fig. 1: We illustrate **source-free audio-visual continual test-time adaptation** using an example of a deployed model (say, in a robot) with audio-visual sensors for scene understanding. Starting from a source model parameterized by θ^S, it encounters a sequence of evolving target environments where distributional shifts may affect the audio modality, the visual modality, or both, motivating continual adaptation at test-time. The goal is to maintain robust performance at test-time without access to the source data and task boundaries.

Test-time adaptation (TTA) has emerged as an effective paradigm for improving model robustness under distribution shifts [28, 45]. In realistic scenarios with privacy and real-time constraints [36], source data is unavailable, and adaptation must be efficient. Consequently, updates are typically limited to a *single forward pass* per test sample [11, 38], while ensuring that source knowledge is not catastrophically forgotten [17]. Although there is a long TTA literature on vision models [8, 10, 38, 39, 51], it has been recently gaining traction for audio-visual models too [19, 26, 32, 46, 50]. We name such methods as AV-TTA.

But, this single-domain assumption is unrealistic in real-world deployment, where models face dynamic and unpredictable domain changes without explicit task boundaries [48]. This setting, termed continual TTA or CTTA, introduces two key challenges: ***catastrophic forgetting*** of the source knowledge due to continual parameter updates across domains, and ***error accumulation***, where miscalibrated [18] pseudo-labels compound over time. Although CTTA has been widely studied in vision-only models [34, 43, 49, 54], it is substantially harder in such multimodal settings. Shifts may affect audio and visual modalities differently, creating a modality gap [32] that hinders cross-modal alignment [12]. Continual updates under such bimodal shifts further amplify error accumulation, leading to progressive performance degradation (see Figure 2). We use the terms distribution, domain, and corruption interchangeably.

In this work, we study *source-free* audio-visual continual test-time adaptation under both unimodal and challenging bimodal corruptions, as illustrated in Figure 1. Although existing AV-TTA methods can be extended to the continual setting, they exhibit limitations. Prior work [32] shows that, even in single-domain AV-TTA, READ [50] suffers from increasing attention imbalance between modality tokens under strong bimodal corruptions, causing performance degradation. We observe a consequence of this in the CTTA setting: when extending READ to VGGSound-2C [32], performance initially improves but progressively declines as new domains arrive (see Figure 2). Furthermore, BriMPR [26], an AV-TTA method which relies on source data, degrades significantly when source access is removed in our continual setting; we denote this source-free variant as BriMPR*.

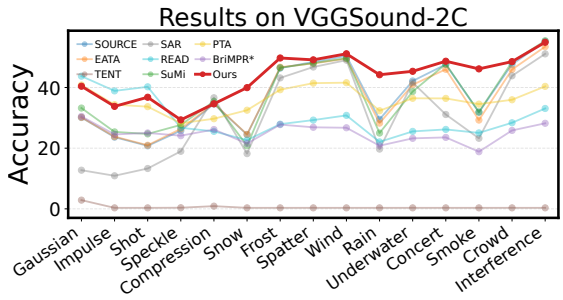


Fig. 2: AVReCAP achieves SOTA performance on audio-visual CTTA. We report task-wise accuracy on VGGSound-2C [32] at a severity level of 5 under correlated bimodal corruptions in the continual setting. CAV-MAE [16] is used as the **SOURCE** model. We extend TTA (**TENT**, **EATA**, **SAR**) and AV-TTA (**READ**, **PTA**, **SuMi**, **BriMPR***) methods to the continual setting. Existing AV-TTA methods struggle under severe correlated bimodal corruptions.

With the dual goals of minimizing source knowledge forgetting and enhancing test-time performance, we propose a new framework AVReCAP. In §4.1, we present a key observation: adapting only the attention weights of the fusion layer to a single target domain not only improves performance over the source model on related domains within the same category, but also achieves better or competitive improvements on unseen domains of other categories. This reveals both intra- and cross-task transferability of the adapted fusion parameters. For the continual setting, however, this suggests that, at any time-step, we can retrieve *similar* parameters from past time-steps and reuse them for further adaptation, instead of the continual overwriting of shared parameters that could result in poor accuracy (see Figure 2). In §4.2, we leverage this core insight and propose a selective parameter retrieval scheme to obtain the most relevant parameter state for adaptation. We introduce a shared buffer to store "snapshots" of the attention weights, inspired by replay-based methods [6, 7, 40], but adhering to strict data privacy rules [41] by storing only modality-specific input-level statistics and the attention weights, without retaining or revisiting past domain data. At every time-step, we compute current modality-specific statistics to guide the selection. To bound buffer growth (§4.3), we merge statistically similar buffer elements. This effectively avoids overfitting and largely minimizes *catastrophic forgetting* (see Figure 7).

Our **contributions** are as follows: 1) We comprehensively study source-free audio-visual CTTA under unimodal and bimodal corruptions. 2) We show that the attention fusion layer shows intra- and cross-task transferability, suggesting that parameters from previous time-steps can be selectively retrieved and used for adaptation (see §4.1). Our method AVReCAP then proposes maintaining a buffer, under a memory budget, to selectively retrieve such parameters based on a proposed criterion. Crucially, AVReCAP substantially mitigates **catastrophic forgetting**, incurring only a **2.9%** performance drop compared to **27.9%** with READ on a difficult test set like VGGSound-2C, involving 15 bimodal corruptions. 3) Our method outperforms all baselines for audio-visual CTTA across datasets.

2 Related Work

Audio-Visual Test-Time Adaptation (AV-TTA). TTA aims to adapt a source model to unlabeled target data without access to source samples, mitigating source–target domain gaps under privacy and real-time constraints [8, 37–39, 45]. Typically, adaptation is performed independently per domain with a single forward pass per sample [36]. TENT [45] minimizes the Shannon entropy of model predictions by fine-tuning the affine parameters of normalization layers. EATA [38] penalizes high entropy samples while minimizing the entropy. SAR [39] identified that model adaptation is affected by samples with large gradients.

TTA has also been widely studied in audio-visual settings *i.e.* AV-TTA. The first work, READ [50], proposed adapting the attention weights of the joint encoder to ensure robust cross-modal fusion. SuMi [19] employs a mutual information sharing loss to enhance the alignment between modalities. PTA [46] argues that under bimodal domain shifts, the source model results in overlapping class representations, leading to biased predictions. PTA partitions samples by prediction bias and jointly reweighs bias and confidence. BriMPR [26] introduces modality-specific prompts to enhance re-alignment. While prompts are optimized for a specific domain, a layer-wise discrepancy loss is minimized between the statistics of the intermediate features and their corresponding source features.

Continual TTA (CTTA). The research community has extensively studied CTTA in the visual domain [13, 14, 30, 33, 34, 43, 48, 49, 54]. However, CTTA for audio-visual data has been relatively underexplored. We notice that only PTA [46] and BriMPR [26] provide limited extensions of their respective AV-TTA methods to the continual setting (see their Appendices). [53] also studies audio-visual CTTA assuming complete access to the source data. A comprehensive study of source-free audio–visual CTTA is still missing, which is the primary focus of this paper.

Remarks. While promising in performance, prior vision-based CTTA methods [30, 43, 54] use a small amount of source data to warm-start meta-networks/adapters or guide visual prompt tuning [23] at test-time, respectively. Though it might be a common practice, under *strict* real-time and privacy-constrained settings, even limited source access may be unavailable. Similarly, BriMPR reports using only 32 source samples. Interestingly, we notice that when the unlabeled source data is absent, the performance drastically drops, as in Figure 2. Moving forward, we abide by strict test-time settings and provide a realistic source-free audio-visual CTTA method, and so, do not compare against [53].

3 Preliminaries

Notations. Let f_a and f_v denote two unimodal transformer encoders that process audio and visual inputs, respectively. These are followed by a joint transformer encoder f_j, which operates on the concatenated audio-visual output tokens, and a final prediction head h for classification. The total parameter set is $\theta^S = \{\theta_a, \theta_v, \theta_j, \theta_h\}$, where θ_a, θ_v, θ_j, and θ_h are the parameters of the mentioned

modules, respectively. In essence, let $z_a = f_a(x_a; \theta_a)$ where $z_a \in \mathbb{R}^{T_a \times D}$ be the audio tokens and $z_v = f_v(x_v; \theta_v)$ where $z_v \in \mathbb{R}^{T_v \times D}$ be the visual tokens. T_a and T_v represent the respective number of output tokens and D refers to the embedding dimension (usually $D = 768$). For cross-modal fusion, these features are concatenated along the token dimension to form $z_c = [z_a; z_v]$, resulting in a joint representation $z_c \in \mathbb{R}^{(T_a+T_v)\times D}$. The final logits, mapped to the concept space, are $l = h(f_j(z_c; \theta_j); \theta_h)$. x_a may represent raw audio or its corresponding spectrogram, while x_v denotes the corresponding video frames. Specific to CAV-MAE [16], there are 11 attention blocks in each of f_a and f_v and 1 attention block in f_j. We denote the weights of the query, key, and value projection matrices in f_j as $\mathcal{W}_\mathcal{Q}$, $\mathcal{W}_\mathcal{K}$, and $\mathcal{W}_\mathcal{V} \in \mathbb{R}^{D\times D}$, respectively.

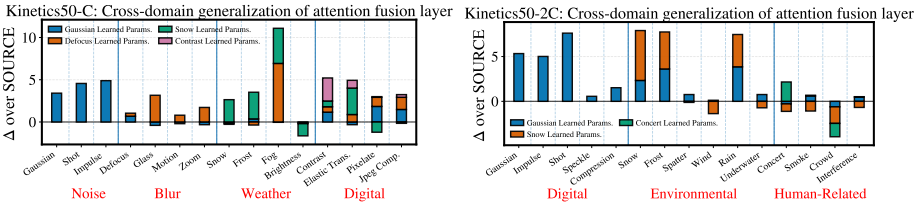
Problem Setting. We begin with a model pre-trained on a source domain $\mathcal{T}^S = \{(x^S_{a,i}, x^S_{v,i}, y^S_i)\}^n_{i=1}$, where $x^S_i = (x^S_{a,i}, x^S_{v,i})$ represents a pair of source audio and visual inputs and y^S_i denotes the corresponding ground-truth labels. In the CTTA setting [48], this source model encounters a sequence of U target tasks/domains at test-time, $\mathcal{T} = \{\mathcal{T}_u\}^U_{u=1}$. The objective is to continually adapt the model to each incoming domain in an online fashion, without any parameter reset. For a given task $\mathcal{T}_u$, the model receives a stream of B unlabeled batches. At each time step t, the input is a multimodal batch $x^t = \{x^t_a, x^t_v\}$, with labels unknown. We assume a strictly online setting where each batch is available only once for a *single forward pass*. Crucially, the distribution of the source domain differs from the target domain, *i.e.* $p(x^S_i) \neq p(x^t_i)$ but $p(y^S_i) = p(y^t_i)$ for an i^{th} sample. Target domains also differ from one another. Throughout, the task boundaries are unknown. We assume that the sequence of tasks at test-time are disjoint, *i.e.* $\mathcal{T}_u \cap \left(\bigcup^{u-1}_{i=1} \mathcal{T}_i\right) = \emptyset$. Each $\mathcal{T}_u$ introduces a unique domain that has not been seen earlier, but the concept or label spaces remain the same. To be noted, *either the audio, visual, or both modalities can be shifted/corrupted,* i.e. *unimodal or bimodal corruptions, respectively.*

4 Method

In this section, we introduce our proposed method AVReCAP for audio-visual continual test-time adaptation. Before diving in, we provide a motivating rationale.

4.1 Motivation

Cross-domain generalization [22, 25, 47, 52] is a well-studied problem in computer vision that aims to improve performance on unseen and unlabeled test domains by training models on diverse labeled source domains. Although most previous works focus on improving generalization through training [4, 55], we slightly abuse the setting here as the source data is absent. In this section, we present an interesting observation by studying cross-domain generalization at test-time. *Specifically, we show that when adapting the attention fusion layer,* i.e. $(\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V)$ *of a source model on an unlabeled test/target domain of a corruption category, the learned model parameters exhibit improved/competitive transfer to other unseen*



(a) Kinetics50-C with visual corruptions. **(b)** Kinetics50-2C with bimodal corruptions.

Fig. 3: Attention fusion layer adapted on a single domain successfully transfers, achieving performance exceeding or comparable to the source model, motivating us to store parameter snapshots in a buffer that can be reused during audio-visual CTTA. We adapt the projection matrices $\{\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V\}$ of the joint encoder f_j of a pre-trained CAV-MAE (SOURCE) [16] on the first unseen domain of each corruption category, following READ [50]. This adapted state is then frozen and inferred on the remaining sequence of unseen domains. We report the accuracy change Δ (in %) over SOURCE.

test/target domains within the same category. We illustrate the results in Figure 3.

We evaluate on Kinetics50-C [50] with visual-only corruptions [20], grouped into *Noise*, *Blur*, *Weather*, and *Digital* categories [20]. Starting from a CAV-MAE source model [16], pre-trained on Kinetics50, we adapt only the attention fusion layer ($\mathcal{W}_\mathcal{Q}$, $\mathcal{W}_\mathcal{K}$, and $\mathcal{W}_\mathcal{V}$ of f_j) following READ [50], with pseudo-labels based on the model's predictions. For each category, we perform single-domain TTA by adapting on the first domain and evaluating on the remaining domains. We see that across nearly all corruption categories, the adapted $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ consistently outperform the SOURCE baseline (pre-trained). We see strong intra-category generalization; $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ learned on *Gaussian Noise* generalize effectively to *Shot* and *Impulse Noise*. This is consistent in other categories, too. We also notice cross-category transferability. In *Blur*, the parameters learned on *Gaussian Noise* perform slightly better than SOURCE, but are outperformed by the *Blur*-specific parameters. Similar trends are in *Digital*, too. This *suggests* that the attention fusion layer captures domain-invariant correlations, enabling a *reuse* of past parameters across related corruption categories.

We perform a similar evaluation on Kinetics50-2C, constructed by applying challenging bimodal audio-visual corruptions from [32] to the Kinetics50 test set. Following their taxonomy, corruptions are grouped into *Digital*, *Environmental*, and *Human-Related* categories. We draw similar conclusions as earlier. Now, with the audio also being shifted, causing more data challenges, $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ learned on *Gaussian Noise* consistently outperform SOURCE throughout, while also generalizing really well to other domains within *Digital*. We also observe patterns of cross-category transferability, as earlier.

Major takeaway. As observed, the adapted $\mathcal{W}_\mathcal{Q}$, $\mathcal{W}_\mathcal{K}$, and $\mathcal{W}_\mathcal{V}$ of the joint-encoder f_j on a single domain (for instance, *Gaussian Noise*) yields strong intra-category and cross-category transfers, beating the SOURCE baseline. To exploit this transfer to unseen tasks/domains in a continual audio-visual setting, specifically,

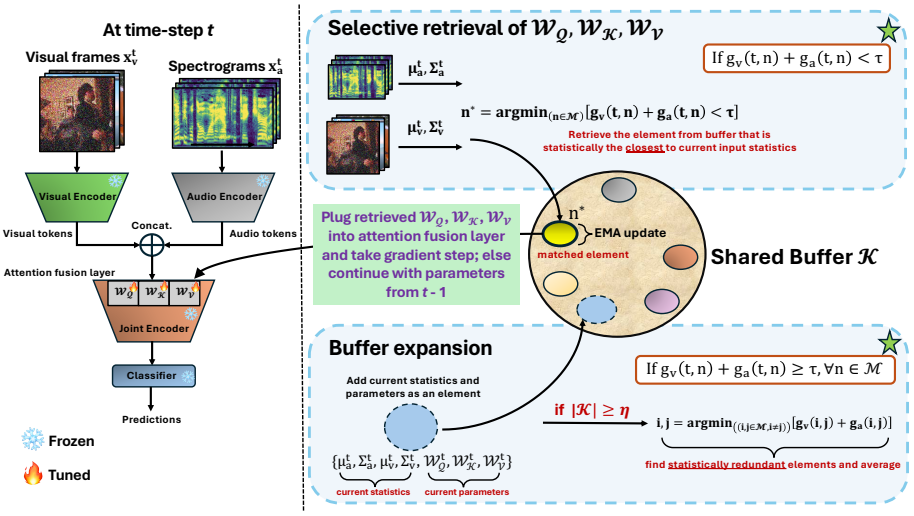


Fig. 4: Illustration of AVReCAP. At time-step t, audio-visual inputs via $\mu_a^t, \Sigma_a^t, \mu_v^t, \Sigma_v^t$ are modeled as Gaussian distributions. The **selection retrieval** stage uses KL divergence $g(\cdot)$ to compare current statistics against all elements in the shared buffer $\mathcal{K}$ ($\mathcal{M}$ is the current set of indices). For the best match within threshold τ, stored parameters $(\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V)$ are retrieved for adaptation at time-step t. If not, the **buffer expansion** stage involves adding current statistics and parameters in $\mathcal{K}$. Redundant elements are merged to maintain a memory budget η. Continual adaptation proceeds with current parameters, *i.e.* from time-step $t-1$.

we propose using a shared buffer that can save model "snapshots". This way, we can *selectively* retrieve a previously optimized state and regain high performance. This leverages past successful adaptations to enable further, ongoing adaptation. When an input-level shift is detected, the model queries the buffer for the most "relevant" parameters rather than continuing from a drifted state, as seen in the results of Figure 2. Consequently, two critical design questions remain unanswered: 1) **Selective retrieval**: How do we identify and retrieve the optimal $\{\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V\}$ for adaptation at time-step t? 2) **Buffer expansion**: What criterion determines the addition of a new element to the shared buffer? Can we constrain the buffer size due to memory requirements? We answer these questions below. Figure 4 illustrates our approach.

4.2 Selective retrieval of $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$

In Figure 2, we see that continual adaptation of $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ in joint-encoder f_j to bimodal shifts, following READ [50], leads to eventual overfitting. This is due to poor and degraded online cross-modal learning and fusion, with unrestricted continual parameter updates. In §4.1, we observed robust cross-domain generalization of the adapted attention fusion layer, outperforming the source model. To leverage this transferability during continual parameter updates, we propose

utilizing a shared buffer that can save parameter snapshots, *i.e.* $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$, to enable selective retrieval during audio-visual CTTA.

At $t = 0$, *i.e.* before adaptation begins with source model parameters θ^S, the buffer $\mathcal{K} = \phi$. At time-step $t = 1$ when the first set of inputs arrive (x^1), we follow READ and only online adapt $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ where the loss function is described in 4.4 and $\theta^S \rightarrow \theta^1$. For $t > 1$ and to enable selective retrieval of these parameters that can be plugged back and adapted, we characterize each x^t by its underlying distribution. To note, the true distribution $p(x^t, y^t)$ is unavailable. Our selection criterion, instead, depends on the statistics (mean and covariance) of the raw data x^t to capture modality-specific information via Gaussian distributions, owing to their simplicity and compactness. We assume a diagonal covariance structure for independence, *i.e.* $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_r^2)$, where r is the dimension. Let the visual frames be $x_v^t \in \mathbb{R}^{B \times H \times W \times C}$, where H, W, C denote the height, width, and channel dimension, respectively. The corresponding spectrograms are $x_a^t \in \mathbb{R}^{B \times T \times F}$, where T and F denote the time and frequency axes, respectively. For the visual modality, we compute the channel-wise mean and covariance as,

$$\mu_v^t = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} x_v^t(b, h, w), \tag{1}$$

$$\Sigma_v^t = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(x_v^t(b, h, w) - \mu_v^t \right)^2 \tag{2}$$

Similarly, for the audio modality, we compute the mean and variance along the time axis T to capture the frequency information of the spectrogram. That is,

$$\mu_a^t = \frac{1}{BT} \sum_{b=1}^{B} \sum_{e=1}^{T} x_a^t(b, e), \tag{3}$$

$$\Sigma_a^t = \frac{1}{BT} \sum_{b=1}^{B} \sum_{e=1}^{T} \left(x_a^t(b, e) - \mu_a^t \right)^2 \tag{4}$$

Clearly, μ_v^t, $\Sigma_v^t \in \mathbb{R}^C$ and μ_a^t, $\Sigma_a^t \in \mathbb{R}^F$. At $t = 1$, the buffer $\mathcal{K}$ is first initialized with an element consisting of $\{\mu_v^1, \Sigma_v^1, \mu_a^2, \Sigma_a^2, \mathcal{W}_Q^1, \mathcal{W}_K^1, \mathcal{W}_V^1\}$, where $\mathcal{W}_Q^1, \mathcal{W}_K^1, \mathcal{W}_V^1$ refer to the parameters *after adaptation*. We assume that, throughout, the buffer takes the form $\mathcal{K} = \left[\{d_1, \widehat{\theta}_1\}, \{d_2, \widehat{\theta}_2\}, \{d_3, \widehat{\theta}_3\}, \dots\right]$. Any element in $\mathcal{K}$ with index n is $\mathcal{K}[n] = \{d_n, \widehat{\theta}_n\}$ where $d_n = (\mu_v^n, \Sigma_v^n, \mu_a^n, \Sigma_a^n)$ and $\widehat{\theta}_n = (\mathcal{W}_Q^n, \mathcal{W}_K^n, \mathcal{W}_V^n)$. $\mathcal{K}[n]$ could refer to any pair in $\mathcal{K}$ from a previous time-step since the domain boundaries are unknown in the continual setting, unlike §4.1. For clarification, an element $\{d_n, \widehat{\theta}_n\}$ is only added as per our proposed §4.3. We will discuss this shortly.

For $t > 1$ with multimodal inputs x_a^t and x_v^t and corresponding input-statistics $(\mu_v^t, \Sigma_v^t, \mu_a^t, \Sigma_a^t)$, we perform a selective retrieval to identify the most compatible

parameter state within the buffer. Our goal is to select the buffer element that is closest to the current input-statistics, retrieve its corresponding parameter state, and plug it back into the model for further adaptation. This way, as illustrated by Figure 3, we can possibly exploit transferability at any time-step when inputs arrive continually. This directly operationalizes the strong intra-category and cross-category generalizations, as observed earlier. For this, we model two modality-specific Gaussian distributions as $P_v^t = \mathcal{N}(\mu_v^t, \Sigma_v^t)$ and $P_a^t = \mathcal{N}(\mu_a^t, \Sigma_a^t)$. We define a distance metric $g_u(t,n)_{u\in\{a,v\}}$ (say, KL divergence), which is the distance between modality u's distribution at time-step t and a stored buffer element n (modeled as a corresponding distribution). Ideally, $g_u(t,n) = \mathcal{D}_{KL}(\mathcal{N}(\mu_u^t, \Sigma_u^t) \parallel \mathcal{N}(\mu_u^n, \Sigma_u^n))$ or $\mathcal{D}_{KL}^u$ as a short-hand. This degenerates to a closed-form solution. Now, we sum the modality-specific distances $g^*(t,n) = g_a(t,n) + g_v(t,n)$ to best capture the complete audio-visual input distance at time-step t. In the case of a unimodal corruption setting, let's say with visual-only corruptions, $g_a(t,n)$ might be negligible and $g_v(t,n)$ drives the retrieval. In a bimodal corruption setting, however, $g_v(t,n)$ and $g_a(t,n)$ both contribute.

Now, from here, two cases arise. For a threshold τ, $g^*(t,n) < \tau$ or $g^*(t,n) \geq \tau$. $g^*(t,n) < \tau$ indicates that the n^{th} element is distributionally closer to x_a^t and x_v^t, *i.e.* in the raw input space. As seen earlier in Figure 3, for example, attention fusion parameters adapted to *Gaussian Noise* transferred to *Shot Noise* and *Impulse Noise* with improved performances over the source model. Hence, at any time-step t in the continual setting, we are interested in finding the best or matched buffer element that is the closest to x_a^t and x_v^t. To select, we solve the following optimization problem,

$$n^* = \underset{n\in\mathcal{M}}{\operatorname{argmin}} \left[g^*(t,n) < \tau\right] \tag{5}$$

where $\mathcal{M}$ is the current set of indices in $\mathcal{K}$. The selected buffer element is $\mathcal{K}[n^*]$. Based on this, we retrieve the corresponding parameters $(\mathcal{W}_Q^{n^*}, \mathcal{W}_K^{n^*}, \mathcal{W}_V^{n^*})$. These selected parameters represent the configuration of the attention fusion layer projection matrices that most closely align with the current inputs x^t. We plug these specific parameters into f_j, indicating that the model had already seen a statistically similar audio-visual shift. In this way, from the raw input-space with domain shift, we exploit the intra-category and cross-category transferability. Following losses as proposed by READ (see 4.4), we adapt the parameters once, *i.e.* $\theta^{t-1} \rightarrow \theta^t$ where θ^t refers to $(\mathcal{W}_Q^{n^*}, \mathcal{W}_K^{n^*}, \mathcal{W}_V^{n^*})$ in joint-encoder f_j.

In addition to this, we perform a moment-preserving exponential moving average (EMA) update of the retrieved input-statistics $(\mu_v^{n^*}, \Sigma_v^{n^*}, \mu_a^{n^*}, \Sigma_a^{n^*})$ and parameter state $(\mathcal{W}_Q^{n^*}, \mathcal{W}_K^{n^*}, \mathcal{W}_V^{n^*})$. This ensures a smooth transition and encourages transferability, *i.e.* intra-task and cross-task. Let β (usually 0.99) be the EMA smoothing factor. For brevity, we denote $O \in \{Q, K, V\}$. Each parameter in the selected element of $\mathcal{K}$ is then updated as,

$$\mathcal{W}'^{n^*}_O \leftarrow \beta\mathcal{W}_O^{n^*} + (1-\beta)\mathcal{W}_O^t \tag{6}$$

An EMA update of the input-statistics is done as follows,

$$\mu_u'^{n^*} \leftarrow \beta \mu_u^{n^*} + (1-\beta)\mu_u^t \tag{7}$$

$$\Sigma_u'^{n^*} \leftarrow \beta \left[\Sigma_u^{n^*} + (\mu_u^{n^*} - \mu_u'^{n^*})^2\right] + (1-\beta)\left[\Sigma_u^t + (\mu_u^t - \mu_u'^{n^*})^2\right] \tag{8}$$

where $\mu_u'^{n^*}$ and $\Sigma_u'^{n^*}$ denote updated modality-specific mean and diagonal covariance, and $u \in \{a, v\}$. Overall, the different components of the selected element are updated to $\mathcal{K}[n^*] \leftarrow \{\mu_v'^{n^*}, \Sigma_v'^{n^*}, \mu_a'^{n^*}, \Sigma_a'^{n^*}, \mathcal{W}'^{n^*}_Q, \mathcal{W}'^{n^*}_K, \mathcal{W}'^{n^*}_V\}$ via Eqns. (6), (7), and (8).

4.3 Buffer expansion

In our method, an element is added to $\mathcal{K}$ *only when* $g^*(t,n) \geq \tau$, $\forall n \in \mathcal{M}$. This means that all the current elements in $\mathcal{K}$ are distributionally distinct from x_a^t and x_v^t. Since no parameters were retrieved, we optimize the current model parameters as-is, following 4.4. Once a gradient step is taken, we add a new element to $\mathcal{K}$ as $\mathcal{K} \cup \{\mu_v^t, \Sigma_v^t, \mu_a^t, \Sigma_a^t, \mathcal{W}_Q^t, \mathcal{W}_K^t, \mathcal{W}_V^t\}$. This new element will serve as an anchor for future occurrences. A possibility in such a scenario is the unconstrained growth of the buffer, a key consideration in memory-constrained continual learning [42]. To address this, we explore a variant of *element merging*, *i.e.* for a fixed buffer budget/size η, we perform a pairwise comparison (when $|\mathcal{K}| \geq \eta$) to find out the two nearest elements based on their statistics and merge them. That is, for indices n_1 and n_2, we solve,

$$(n_1, n_2) = \underset{i,j \in \mathcal{M}, i \neq j}{\operatorname{argmin}} \; g^*(i,j) \tag{9}$$

Once the most similar pair (n_1, n_2) is identified, we merge them to maintain the fixed budget η. As an approximation, we average the corresponding elements in $\mathcal{K}[n_1]$ and $\mathcal{K}[n_2]$ and add the element back in $\mathcal{K}$. Now, $|\mathcal{K}| = \eta - 1$. We show an ablation in 5.2.

4.4 Losses

At every time-step t, we optimize following the loss functions proposed in READ [50]. For a notational purpose, let σ denote the softmax operation. Let the prediction confidence of the i^{th} input be $p_{\max} = \max(\sigma(l_i))$, where l_i refers to the logits from the classifier. For a batch of B inputs, the confidence-based loss function $\mathcal{L}_{conf}$ is,

$$\mathcal{L}_{conf} = \mathbb{E}_{i \sim B}[-p_{i,max} \log(p_{i,max})] \tag{10}$$

In addition, a negative entropy loss [27] as below is also used,

$$\mathcal{L}_{ne} = \sum_{a=1}^{A} \sigma(k_a) \log(\sigma(k_a)) \tag{11}$$

where, $k_a = \sum_{i=1}^{B} \sigma(l_i)$. A refers to the specific class. The complete loss function is, $\mathcal{L} = \mathcal{L}_{conf} + \mathcal{L}_{ne}$, where only $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ of the the joint-encoder f_j are **continually** adapted.

4.5 Motivation of this design choice

In [50], the authors identified the attention fusion layer, with frozen encoders, as critical for cross-modal TTA. As a follow-up, [35] showed that bimodal corruptions induce attention imbalance, i.e, in online TTA, there is a widening gap in self and cross-attention between the visual and audio tokens. Motivated by these observations, our method focuses on adapting only the fusion parameters and retrieves similar parameters from the shared buffer instead of continually overwriting them. Throughout, both encoders are frozen. We later show that this design choice, combined with our proposed method AVReCAP, minimizes catastrophic forgetting (Figure 7).

5 Experiments

Settings. We evaluate in two main corruption settings - unimodal and bimodal. Corruptions are applied to either one modality or both, respectively. Here, tasks arrive sequentially (see §3), and parameter updates happen continually without any reset. Each corruption spans the entire test set and defines a task.
Datasets. We evaluate on the test sets of Kinetics50 [5] and VGGSound [9], following READ [50]. For unimodal corruptions, we use Kinetics50-C and VGGSound-C, which include 15 visual [20] and 6 audio corruptions at severity level 5. For bimodal corruptions, we adopt 15 corruptions from AVRobustBench [32], applied to both modalities, resulting in Kinetics50-2C and VGGSound-2C. Kinetics50 is *visually dominant*, while VGGSound is *audio dominant*, meaning task-relevant cues primarily reside in the visual and audio modalities, respectively.
Baselines. We compare our proposed AVReCAP to popular TTA baselines (extensible to continual) like TENT [45], EATA [38], and SAR [39]. The prior audio-visual baselines include READ [50], SuMi [19], PTA [46], and BriMPR* [26]. As discussed in §2, we eliminate access to the source data in BriMPR to make it more deployment-friendly in a real-world setting. We also do not compare against [53] as it requires full access to source data.
Implementation Details. Following previous works, we use CAV-MAE [16] as the source model, trained on Kinetics50 or VGGSound. We use a batch size of 32 across all experiments. We optimize with Adam using a learning rate of 1×10^{-4}. For unimodal corruptions, we set τ to 0.005 and 0.01 for bimodal corruptions. For all the baselines, we adopt their recommended hyperparameters. All experiments are conducted on an NVIDIA RTX A5000 GPU.

5.1 Main Results

Results on the unimodal corruption setting. In Tables 1 and 2, we present results for unimodal video and audio corruptions, respectively. Under visual

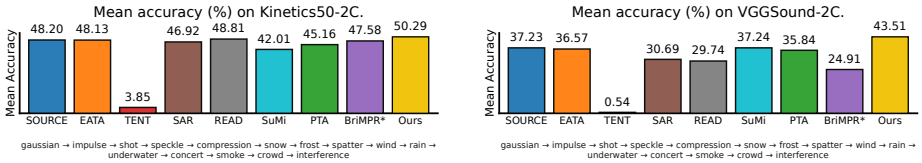
Table 1: AVReCAP achieves SOTA results on Kinetics50-C (top) **and VGGSound-C** (below) with **video corruptions** (unimodal). We report the task-wise accuracy (%) at a severity level of 5, in the **continual** setting.

	Method	Noise			Blur				Weather				Digital				Mean
		Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic Transform	Pixelate	JPEG	
Kinetics50-C	SOURCE	46.55	47.36	46.51	67.23	61.70	70.67	66.39	62.22	60.54	47.32	74.88	51.36	64.14	66.39	61.98	59.68
	TENT	45.55	42.51	27.84	22.76	8.33	3.64	2.24	2.40	2.08	2.00	2.04	2.28	2.04	2.00	2.00	11.31
	EATA	46.80	47.62	46.81	67.62	63.02	70.77	66.94	60.65	60.65	48.23	75.48	52.42	65.40	66.45	62.22	60.07
	SAR	46.83	47.00	46.64	65.71	62.30	69.99	66.35	60.54	60.22	52.60	73.96	50.76	65.99	63.42	59.90	59.48
	READ	49.96	51.80	51.56	68.07	65.63	65.38	61.50	52.20	49.48	42.19	54.93	28.93	39.82	33.17	27.32	49.46
	SuMi	45.75	45.43	44.95	64.30	65.38	69.23	67.91	62.86	66.67	63.22	69.23	50.16	71.15	66.31	63.94	61.10
	PTA	46.84	45.95	45.91	56.17	55.73	55.49	54.81	51.80	52.60	50.24	54.73	49.80	53.32	52.24	51.12	51.78
	BriMPR*	49.12	49.48	48.64	58.25	59.82	58.57	58.53	53.69	55.49	54.33	59.13	50.32	59.98	57.77	57.61	55.38
	AVReCAP ($\eta = 50$)	48.88	51.00	48.72	68.23	64.78	70.99	68.47	65.02	66.35	61.58	72.52	53.81	66.79	68.19	65.99	**62.75**
	AVReCAP ($\eta = 100$)	48.52	50.52	48.08	67.95	63.86	71.59	67.75	65.14	67.19	62.14	72.80	54.17	66.43	67.59	65.18	**62.59**
	AVReCAP ($\eta = \infty$)	48.52	50.24	48.32	67.91	63.18	71.60	68.03	64.86	67.15	62.10	72.80	53.57	65.43	67.87	64.02	**62.37**
VGGSound-C	SOURCE	52.78	52.69	52.73	57.19	57.21	58.52	57.61	56.25	56.58	55.31	58.95	53.66	56.95	55.81	56.89	55.94
	TENT	52.58	51.82	51.20	53.85	53.90	54.49	54.39	51.80	52.39	52.59	52.14	51.08	52.49	51.70	51.87	52.55
	EATA	52.79	52.81	52.46	56.60	55.90	57.43	56.16	54.36	54.54	53.35	57.06	51.94	54.56	53.23	54.51	54.51
	SAR	52.92	52.86	52.92	57.03	56.66	58.65	57.61	55.95	56.74	56.22	58.25	54.21	57.13	55.42	56.46	55.94
	READ	52.43	52.65	52.44	55.85	55.22	56.07	55.69	54.51	54.96	54.69	55.38	54.34	54.79	54.65	54.65	54.55
	SuMi	53.14	53.42	53.43	57.01	56.69	57.67	56.91	55.60	56.22	51.73	56.91	53.67	56.37	55.16	55.25	55.28
	PTA	51.06	49.86	49.46	52.88	52.12	52.31	51.94	50.96	51.21	51.17	51.92	50.46	51.59	51.03	51.05	51.27
	BriMPR*	43.24	39.11	38.86	39.99	41.22	39.41	37.97	33.12	30.76	31.47	30.71	26.08	31.95	34.30	36.15	35.62
	AVReCAP ($\eta = 300$)	52.87	52.87	52.86	57.16	56.97	58.65	58.02	56.66	57.13	56.67	58.18	54.66	57.42	56.22	56.71	**56.20**
	AVReCAP ($\eta = 400$)	52.86	52.83	52.91	57.30	57.25	58.66	57.95	56.50	56.76	56.47	57.89	54.80	57.40	55.82	56.54	**56.13**
	AVReCAP ($\eta = \infty$)	52.87	52.83	52.91	57.42	57.29	58.68	57.77	56.38	56.80	55.50	59.13	53.88	57.07	55.96	57.01	**56.10**

Table 2: AVReCAP achieves SOTA results on Kinetics50-C (top) **and VGGSound-C** (below) with **audio corruptions** (unimodal). We report the task-wise accuracy (%) at a severity level of 5, in the **continual** setting.

	Method	Noise			Weather			Mean
		Gaussian	Traffic	Crowd	Rain	Thunder	Wind	
Kinetics50-C	SOURCE	73.48	65.30	67.59	70.11	67.67	70.11	69.04
	TENT	73.84	68.55	70.51	69.59	73.12	70.15	70.96
	EATA	73.56	65.30	67.71	70.11	68.07	70.15	69.15
	SAR	73.36	65.95	68.11	69.71	69.15	69.71	69.33
	READ	74.28	69.63	70.63	70.35	71.76	69.39	71.01
	SuMi	73.76	68.19	70.59	69.27	73.24	69.43	70.75
	PTA	72.68	69.03	69.79	68.83	71.79	69.75	70.31
	BriMPR*	72.84	67.67	67.43	65.95	69.07	63.70	67.78
	AVReCAP ($\eta = 50$)	73.72	68.35	70.27	70.11	72.88	69.91	70.87
	AVReCAP ($\eta = 100$)	73.72	68.47	69.79	70.27	72.56	70.39	70.87
	AVReCAP ($\eta = \infty$)	74.15	68.90	70.30	70.62	72.06	70.57	**71.10**
VGGSound-C	SOURCE	37.36	21.12	16.80	21.64	27.29	25.54	24.96
	TENT	6.20	0.46	0.29	0.28	0.28	0.28	1.30
	EATA	37.51	21.64	17.58	22.82	27.98	25.34	25.48
	SAR	36.53	8.01	4.12	4.49	13.43	3.34	11.65
	READ	28.01	15.10	17.35	13.69	20.24	14.37	18.13
	SuMi	37.66	19.28	14.79	20.73	28.82	28.05	24.89
	PTA	36.30	28.79	28.42	25.35	30.60	26.09	29.26
	BriMPR*	23.12	14.95	15.15	13.85	20.08	15.24	17.07
	AVReCAP ($\eta = 300$)	39.52	28.43	24.36	29.81	33.58	20.85	**29.42**
	AVReCAP ($\eta = 400$)	39.52	28.43	24.32	29.37	35.65	26.38	**30.61**
	AVReCAP ($\eta = \infty$)	39.52	28.43	24.31	29.44	35.52	27.05	**30.71**

corruptions in Table 1, particularly on Kinetics50-C, where task information is heavily degraded, TENT exhibits severe overfitting, as entropy minimization updates all LayerNorm [3] affine parameters across attention blocks in all encoders. This update is independent of any modality interaction or cross-modal fusion. In contrast, EATA and SAR also update norm parameters but maintain performance due to their respective loss functions, achieving accuracy comparable to the source model. In READ, continual updates to $\mathcal{W}_{\mathcal{Q}}, \mathcal{W}_{\mathcal{V}}, \mathcal{W}_{\mathcal{V}}$ lead to an eventual decline, as continual adaptation amplifies modality imbalance [32] and error accumulation, ultimately biasing the attention mechanism. Without access to source data, the randomly initialized visual prompts in BriMPR* struggle to maintain semantic alignment and fail to effectively adapt to the target domains. At a small budget $\eta = 50$, AVReCAP achieves a relative improvement of 2.7% over the next best method, SuMi, on average. On a larger test set like VGGSound-C with audio corruptions, we have similar observations and AVReCAP obtains a relative



(a) Kinetics50-2C **(b)** VGGSound-2C

Fig. 5: AVReCAP achieves SOTA results on Kinetics50-2C (left) **and VGGSound-2C** (right). We report mean accuracy (%) at a severity level of 5 in the **continual** setting. Here, buffer size $\eta = \infty$.

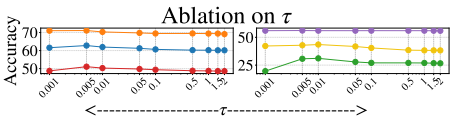


Fig. 6: Ablation on τ. Results on Kinetics50 (left) under **audio**, **visual**, and **bimodal** corruptions, and on VGGSound (right) under **visual**, **audio**, and **bimodal** corruptions.

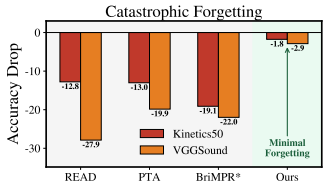


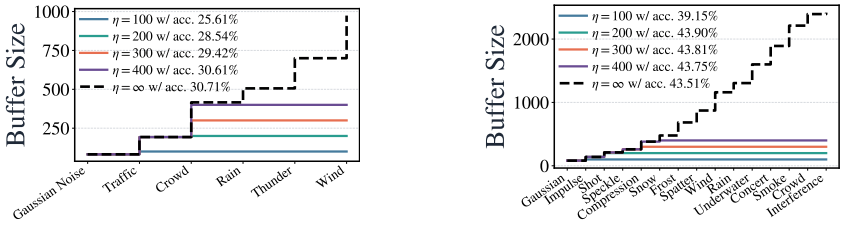
Fig. 7: AVReCAP minimizes *catastrophic forgetting*. We indicate the accuracy drop after continual adaptation to Kinetics50-2C and VGGSound-2C compared to the respective source data performance of CAV-MAE.

improvement of 4.95% over PTA. On VGGSound-C with visual corruptions and Kinetics50-C with audio corruptions, *i.e.* where task-dominant information is clean, AVReCAP achieves comparable to better mean accuracies. We show more results with varying η in 5.2 (ablation).

Results on the bimodal corruption setting. In Figures 5a and 5b, we illustrate the mean results in the challenging bimodal corruptions. First off, TENT overfits drastically due to continual norm updates across domains. On Kinetics50-2C, the proposed AVReCAP obtains an improvement of 1.25% over READ (next best). Interestingly, most methods fall short of a pre-trained SOURCE model under continual parameter updates with bimodal corruptions since cross-modal learning is hampered. A similar trend is observed on the more challenging VGGSound-2C benchmark with 309 classes and 14,046 samples in each task, where AVReCAP achieves a substantial 6.28% improvement over SOURCE. In contrast, most methods progressively overfit and perform worse than SOURCE.

5.2 Ablation Studies

Effect of threshold τ. In Figure 6, we ablate the distance threshold τ and report results on both datasets. We sweep $\tau \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 1.5, 2\}$. Performance gradually degrades as τ increases, with a more pronounced drop under bimodal corruptions. Larger τ leads to larger buffers that allow minor distributional fluctuations to be added, as elements that reduce cross-domain transferability. In contrast, a smaller τ value strikes a favorable balance.



(a) VGGSound-C w/ audio corruptions. **(b)** VGGSound-2C w/ bimodal corruptions.

Fig. 8: Effect of buffer size η. Mean accuracy (%) on VGGSound-C (left) and VGGSound-2C (right).

Catastrophic forgetting. We evaluate the final adapted parameters back on the source test set. We continually adapt on Kinetics50-2C and VGGSound-2C, and illustrate the results in Figure 7. For Kinetics50-2C, this corresponds to 15×2466 updates and 15×14046 updates for VGGSound-2C. As seen, in the presence of challenging bimodal corruptions, selective parameter retrieval, as done in our work, significantly preserves the source knowledge. In prior works, long-term continual parameter adaptation can lead to major source knowledge forgetting, as in Figure 7.

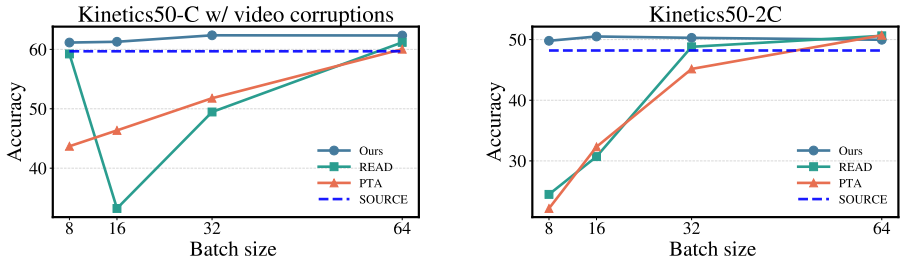
Buffer budget η. In Figure 8, we study the scalability of AVReCAP with a fixed buffer size η on VGGSound-C (w/ audio corruptions) and VGGSound-2C. Each task comprises approximately 439 batches ($\frac{14{,}046}{32}$) across 6/15 tasks, respectively. For VGGSound-C, we notice that at a fixed budget $\eta \geq 300$ throughout, AVReCAP still achieves better or competitive accuracies compared to baselines. On VGGSound-2C, with a tighter budget of $\eta = 200$, we notice improved performances. Element merging indeed consolidates task-specific knowledge and remains stable in long-term continual adaptation.

Distance metric $g_u(t, n)$. In Table 3, we experiment with a norm-based distance as $g_u(t, n) = \|\mu_u^t - \mu_u^n\|_2 + \|\sigma_u^t - \sigma_u^n\|_2$, where $u \in \{a, v\}$. Then, $g^*(t, n) = g_a(t, n) + g_v(t, n)$. As expected, with only $\mathcal{D}_{KL}^a$ and visual corruptions in Kinetics50-C, we see a drop in performance since retrieval is independent of the corrupted video modality. In general, there is a need to capture both audio and visual distances via $\mathcal{D}_{KL}^{a+v}$.

Table 3: Alternative distance metric $g_u(t, n)$. Results on Kinetics50-C (visual corruptions) and Kinetics50-2C.

Metric $g_u(t,n)$	K50-C	K50-2C
$\|\mu_u^t - \mu_u^n\|_2 + \|\sigma_u^t - \sigma_u^n\|_2$	49.47	48.79
$\mathcal{D}_{KL}^v$	61.53	49.28
$\mathcal{D}_{KL}^a$	60.52	49.70
$\mathcal{D}_{KL}^{a+v}$ (Ours)	**62.37**	**50.29**

Effect of batch size. Here, we investigate the sensitivity of AVReCAP to the choice of batch size B. We evaluate a range of values $B \in \{8, 16, 32, 64\}$ to assess the framework's robustness across varying computational constraints. The effect of batch size can be critical in our setting. In Eqns. (1), (2), (3), and (4), the mean and covariance are computed along the batch axis, where a smaller batch size can *potentially* introduce noise into these statistics. We demonstrate results on Kinetics50-C (w/ video corruptions) and Kinetics50-2C. Our results in Figures 9a and 9b demonstrate that AVReCAP maintains competitive performance even



(a) Kinetics50-C (video corruptions) **(b)** Kinetics50-2C (bimodal corruptions)

Fig. 9: AVReCAP maintains stable performance under small batch sizes. The x-axis shows batch size, and the y-axis reports mean accuracy for AVReCAP and baselines.

at $B = 8$. This highlights the suitability of our proposed method under memory limitations, especially on small data streams.

6 Conclusion

In this paper, we comprehensively study audio-visual continual test-time adaptation, taking a major step forward in the real-time deployment of audio-visual models in continually changing target environments. We propose AVReCAP, a novel method to address the same. Our motivation stems from the fact that the attention fusion layer exhibits intra- and cross-task transferability, suggesting that attention parameters from previous domains can be retrieved, plugged, and reused for continual adaptation. Inspired by the replay-based continual learning literature [40], we introduce a shared buffer across tasks to store and selectively retrieve these generalizable attention parameters based on low-level input statistics. This effectively mitigates catastrophic forgetting and offers a scalable, memory-efficient framework. Through extensive evaluations on benchmark datasets involving unimodal and bimodal audio-visual corruptions, we establish SOTA baselines for robust audio-visual continual adaptation.

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in neural information processing systems **34**, 24206–24221 (2021)
2. Araujo, E., Rouditchenko, A., Gong, Y., Bhati, S., Thomas, S., Kingsbury, B., Karlinsky, L., Feris, R., Glass, J.R., Kuehne, H.: Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18794–18803 (2025)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2229–2238 (2019)

5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. In: ICLR (2019)
7. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H., Ranzato, M.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
8. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 295–305 (2022)
9. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020)
10. Choi, S., Yang, S., Choi, S., Yun, S.: Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In: European Conference on Computer Vision. pp. 440–458. Springer (2022)
11. Danilowski, M., Chatterjee, S., Ghosh, A.: Botta: Benchmarking on-device test time adaptation. arXiv preprint arXiv:2504.10149 (2025)
12. Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O.: Simmmdg: A simple and effective framework for multi-modal domain generalization. Advances in Neural Information Processing Systems **36**, 78674–78695 (2023)
13. Gan, Y., Bai, Y., Lou, Y., Ma, X., Zhang, R., Shi, N., Luo, L.: Decorate the newcomers: Visual domain prompt for continual test time adaptation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 7595–7603 (2023)
14. Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., Lee, S.J.: Note: Robust continual test-time adaptation against temporal correlation. Advances in Neural Information Processing Systems **35**, 27253–27266 (2022)
15. Gong, Y., Liu, A.H., Rouditchenko, A., Glass, J.: Uavm: Towards unifying audio and visual models. IEEE Signal Processing Letters **29**, 2437–2441 (2022)
16. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.R.: Contrastive audio-visual masked autoencoder. In: The Eleventh International Conference on Learning Representations (2023)
17. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211 (2013)
18. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
19. Guo, Z., Jin, T.: Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In: The Thirteenth International Conference on Learning Representations (2025)
20. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
21. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C., et al.: Mavil: Masked audio-video learners. Advances in Neural Information Processing Systems **36**, 20371–20393 (2023)
22. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: European conference on computer vision. pp. 124–140. Springer (2020)

23. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
24. Kim, J., Lee, H., Rho, K., Kim, J., Chung, J.S.: Equiav: Leveraging equivariance for audio-visual contrastive learning. In: International Conference on Machine Learning. PMLR (2024)
25. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5542–5550 (2017)
26. Li, J., Feng, S.: Bridging modalities via progressive re-alignment for multimodal test-time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 22931–22939 (2026)
27. Li, J., Yu, Z., Du, Z., Zhu, L., Shen, H.T.: A comprehensive survey on source-free domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5743–5762 (2024)
28. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
29. Likhosherstov, V., Arnab, A., Choromanski, K., Lucic, M., Tay, Y., Weller, A., Dehghani, M.: Polyvit: Co-training vision transformers on images, videos and audio. arXiv preprint arXiv:2111.12993 (2021)
30. Liu, J., Yang, S., Jia, P., Zhang, R., Lu, M., Guo, Y., Xue, W., Zhang, S.: Vida: Homeostatic visual domain adapter for continual test time adaptation. In: International Conference on Learning Representations. vol. 2024, pp. 48396–48417 (2024)
31. Liu, Z., Xu, Y., Xu, Y., Qian, Q., Li, H., Jin, R., Ji, X., Chan, A.B.: An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation. arXiv preprint arXiv:2205.12753 (2022)
32. Maharana, S.K., Kushwaha, S.S., Zhang, B., Rodriguez, A., Wei, S., Tian, Y., Guo, Y.: Avrobustbench: Benchmarking the robustness of audio-visual recognition models at test-time. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
33. Maharana, S.K., Mishra, S., Zhang, Y., Niu, S., Rafi, T.H., Hamm, J., Pedersoli, M., Dolz, J., Guo, Y.: Continual test-time adaptation in computer vision: Methods, benchmarks, and future directions. arXiv preprint arXiv:2607.08164 (2026)
34. Maharana, S.K., Zhang, B., Guo, Y.: Palm: Pushing adaptive learning rate mechanisms for continual test-time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 19378–19386 (2025)
35. Maharana, S.K., Zhang, B., Karlinsky, L., Feris, R., Guo, Y.: Batclip: Bimodal online test-time adaptation for clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
36. Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., Sanner, S.: Online continual learning in image classification: An empirical survey. Neurocomputing **469**, 28–51 (2022)
37. Niu, S., Miao, C., Chen, G., Wu, P., Zhao, P.: Test-time model adaptation with only forward passes. In: The International Conference on Machine Learning (2024)
38. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International Conference on Machine Learning. pp. 16888–16905. PMLR (2022)
39. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: Internetional Conference on Learning Representations (2023)

40. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
41. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1310–1321 (2015)
42. Smith, J., Balloch, J., Hsu, Y.C., Kira, Z.: Memory-efficient semi-supervised continual learning: The world is its own replay buffer. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
43. Song, J., Lee, J., Kweon, I.S., Choi, S.: Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11920–11929 (2023)
44. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems **33**, 18583–18599 (2020)
45. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021)
46. Wang, G., Lyu, F., Ding, C.: Partition-then-adapt: Combating prediction bias for reliable multi-modal test-time adaptation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
47. Wang, H., He, Z., Lipton, Z.L., Xing, E.P.: Learning robust representations by projecting superficial statistics out. In: International Conference on Learning Representations (2019)
48. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
49. Wang, Y., Hong, J., Cheraghian, A., Rahman, S., Ahmedt-Aristizabal, D., Petersson, L., Harandi, M.: Continual test-time domain adaptation via dynamic sample selection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1701–1710 (2024)
50. Yang, M., Li, Y., Zhang, C., Hu, P., Peng, X.: Test-time adaptation against multi-modal reliability bias. In: The twelfth international conference on learning representations (2024)
51. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in neural information processing systems **35**, 38629–38642 (2022)
52. Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., Liu, H.: Towards unsupervised domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4910–4920 (2022)
53. Zhang, Y., Xu, Y., Wei, H., Lin, Z., Zou, X., Chen, C., Zhuang, H.: Analytic continual test-time adaptation for multi-modality corruption. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 1929–1937 (2025)
54. Zhang, Y., Mehra, A., Niu, S., Hamm, J.: Dpcore: Dynamic prompt coreset for continual test-time adaptation. In: Forty-second International Conference on Machine Learning (2025)
55. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: European conference on computer vision. pp. 561–578. Springer (2020)