AV-CTTA: Audio-Visual Continual Test-Time Adaptation *without* Forgetting

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

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 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, AV-CTTA, 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 AV-CTTA significantly outperforms existing methods while minimizing catastrophic forgetting.

1 Introduction

The development of audio-visual models has accelerated rapidly in recent years (Gong et al., 2022b;c; Huang et al., 2023; Kim et al., 2024; Araujo et al., 2025). 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 or continual test-time distribution shifts that significantly hinder generalization (Taori et al., 2020; Liu et al., 2022). Consider safety-critical scenarios such as self-driving cars equipped with audio-visual sensors.

Test-time adaptation (TTA) improves robustness under distribution shifts (Wang et al., 2020; Liang et al., 2025). In privacy and latency-constrained settings (Mai et al., 2022), adaptation must be efficient while avoiding forgetting (Niu et al., 2022; Danilowski et al., 2025; Goodfellow et al., 2013). While well-studied in vision, it is now extending to audio-visual models (AV-TTA) (Yang et al., 2024; Guo & Jin, 2025; Li & Feng, 2025; Wang et al., 2025; Maharana et al., 2025a). But, this single-domain assumption is unrealistic in real-world deployment, where models face dynamic and unpredictable domain changes without explicit task boundaries (Wang et al., 2022). This setting, termed continual TTA or CTTA, introduces two key challenges: *catastrophic forgetting* of the source knowledge and *error accumulation*, where miscalibrated (Guo et al., 2017) pseudo-labels compound over time. Although CTTA has been widely studied in vision-only models (Song et al., 2023; Wang et al., 2024; Maharana et al., 2025b; Zhang et al., 2024), it is substantially harder in multimodal settings. *Shifts may affect audio and visual modalities differently.*

In this work, we study **source-free** audio-visual CTTA under both unimodal and challenging bimodal corruptions. We extend existing single-domain AV-TTA works to the continual setting. For fairness, BriMPR (Li & Feng, 2025), an AV-TTA method, has source data access removed; we denote this as BriMPR*.

With the dual goals of minimizing source knowledge *forgetting* and enhancing test-time performance, we propose a new framework **AV-CTTA**. In Section B.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

[1]The University of Texas at Dallas, USA [2]Dolby Laboratories, USA. Correspondence to: Sarthak Kumar Maharana <SKM200005@utdallas.edu>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Copyright 2026 by the author(s).

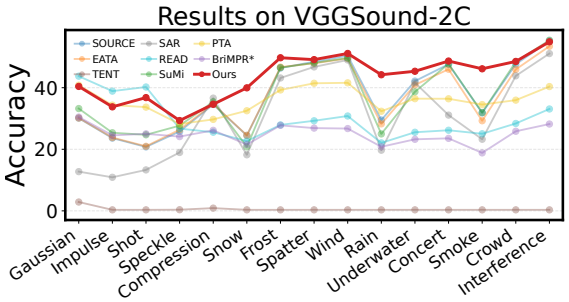


Figure 1. **AV-CTTA achieves SOTA performance on audio-visual CTTA.** We report task-wise accuracy on VGGSound-2C (Maharana et al., 2025a) at severity level 5 under correlated bimodal corruptions in the continual setting. We use a batch size of 32.

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 1). In Section 2.1, 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 (Rebuffi et al., 2017; Chaudhry et al., 2018; 2019), but adhering to strict data privacy rules (Shokri & Shmatikov, 2015) 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 (Section 2.2), we merge statistically similar buffer elements. This effectively avoids overfitting and largely minimizes *catastrophic forgetting* (see Figure 5).

Our **contributions** are as follows: 1) We comprehensively study 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 Section B.1). Our method AV-CTTA then proposes maintaining a buffer, under a memory budget, to selectively retrieve such parameters based on a proposed criterion. Crucially, AV-CTTA 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. 3) Our method outperforms all baselines for audio-visual CTTA across datasets.

2 Proposed Method

In this section, we introduce our proposed method AV-CTTA. **We urge the reader to read Section B, especially B.1 in the Appendix, for notations and the core motivation.**

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

In Section B.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 $t = 1$ 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 2.3 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, $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_r^2)$. 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 $\mathcal{K} = \left[\{d_1, \widehat{\theta}_1\}, \{d_2, \widehat{\theta}_2\}, \{d_3, \widehat{\theta}_3\}, \dots\right]$. An element in 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 $\widehat{\theta}_n = (\mathcal{W}_Q^n, \mathcal{W}_K^n, \mathcal{W}_V^n)$.

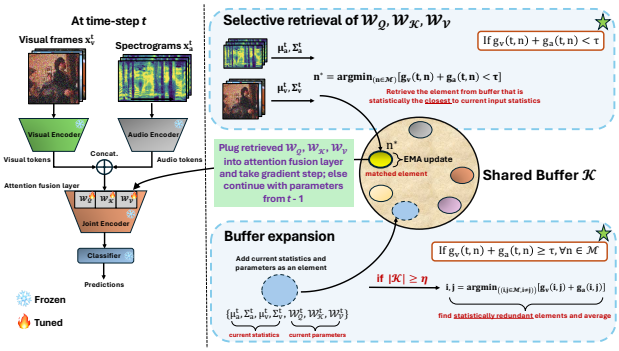


Figure 2. **Illustration of our proposed AV-CTTA.**

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 7, 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 (see Appendix C.1). 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 nth element is distributionally closer to x_a^t and x_v^t, i.e, in the raw input space. As seen earlier in Figure 7, 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 t, 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 2.3), 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,

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

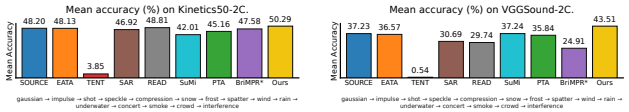
$$\Sigma_u'^{n^*} \leftarrow \beta\left[\Sigma_u^{n^*} + (\mu_u^{n^*} - \mu_u'^{n^*})^2\right] + (1-\beta)\left[\Sigma_u^t + (\mu_u^t - \mu_u'^{n^*})^2\right] \tag{8}$$

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).

2.2 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 2.3. 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 (Smith et al., 2021). 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,

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



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

Figure 3. **AV-CTTA 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. We let $\eta = \infty$, but performances are still comparable/better with η as low as 50.

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 E.2 and explore another strategy in Appendix E.6.

2.3 Losses

At every time-step t, we optimize the loss functions proposed in READ (Yang et al., 2024).

3 Experiments and Results

Please refer to Section D for all the details on datasets, baselines, and implementation. In Section E, we present more results and ablation studies.

Table 1. **AV-CTTA 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
	AV-CTTA ($\eta = 50$)	73.72	68.35	70.27	70.11	72.88	69.91	70.87
	AV-CTTA ($\eta = 100$)	73.72	68.47	69.79	70.27	72.56	70.39	70.87
	AV-CTTA ($\eta = \infty$)	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
	AV-CTTA ($\eta = 300$)	39.52	28.43	24.36	29.81	33.58	20.85	**29.42**
	AV-CTTA ($\eta = 400$)	39.52	28.43	24.32	29.37	35.65	26.38	**30.61**
	AV-CTTA ($\eta = \infty$)	39.52	28.43	24.31	29.44	35.52	27.05	**30.71**

3.1 Main Results

Results on the unimodal corruption setting. In Tables 2 (see Appendix) and 1, we present results for unimodal video and audio corruptions, respectively. More results are in the Appendix due to a lack of space.

Results on the bimodal corruption setting. In Figures 3a and 3b, we illustrate the mean results in the challenging bimodal corruptions. In this, the task-wise accuracies are in Appendix E.1.

3.2 Ablation Studies

Effect of threshold τ. In Figure 4, performance gradually degrades as τ increases, with a more pronounced drop under bimodal corruptions. A smaller τ value strikes a favorable balance.

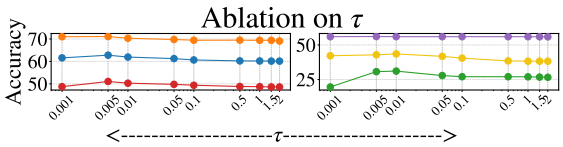


Figure 4. **Ablation on τ.** Results on Kinetics50 (left) under **audio**, **visual**, and **bimodal** corruptions, and on VGGSound (right) under **visual**, **audio**, and **bimodal** corruptions. The source accuracies are 83.33% and 65.5%, respectively.

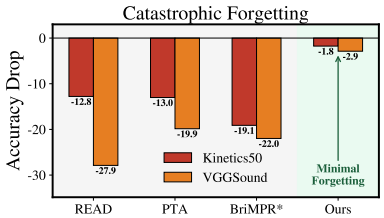


Figure 5. **AV-CTTA minimizes *catastrophic forgetting* to a large extent.**

Catastrophic forgetting. We evaluate the final adapted parameters back on the source test set and illustrate the results in Figure 5. 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 5.

Additional ablation studies. In the Appendix (see Section E), we present additional ablations.

4 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 AV-CTTA, 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. Through extensive evaluations, we establish SOTA baselines for robust audio-visual continual test-time adaptation.

References

Araujo, E., Rouditchenko, A., Gong, Y., Bhati, S., Thomas, S., Kingsbury, B., Karlinsky, L., Feris, R., Glass, J. R., and 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.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and 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.

Carreira, J. and 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.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., and Ranzato, M. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Chen, D., Wang, D., Darrell, T., and Ebrahimi, S. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vg-gsound: 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.

Danilowski, M., Chatterjee, S., and Ghosh, A. Botta: Benchmarking on-device test time adaptation. *arXiv preprint arXiv:2504.10149*, 2025.

Gan, Y., Bai, Y., Lou, Y., Ma, X., Zhang, R., Shi, N., and Luo, L. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7595–7603, 2023.

Georgescu, M.-I., Fonseca, E., Ionescu, R. T., Lucic, M., Schmid, C., and Arnab, A. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16144–16154, 2023.

Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., and Lee, S.-J. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022a.

Gong, Y., Liu, A. H., Rouditchenko, A., and Glass, J. Uavm: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, 29:2437–2441, 2022b.

Gong, Y., Rouditchenko, A., Liu, A. H., Harwath, D., Karlinsky, L., Kuehne, H., and Glass, J. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022c.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Guo, Z. and Jin, T. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. *arXiv preprint arXiv:2503.02616*, 2025.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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.

Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In *European conference on computer vision*, pp. 124–140. Springer, 2020.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Kim, J., Lee, H., Rho, K., Kim, J., and Chung, J. S. Equiav: Leveraging equivariance for audio-visual contrastive learning. *arXiv preprint arXiv:2403.09502*, 2024.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, J. and Feng, S. Bridging modalities via progressive re-alignment for multimodal test-time adaptation. *arXiv preprint arXiv:2511.22862*, 2025.

Liang, J., He, R., and Tan, T. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Liu, J., Yang, S., Jia, P., Zhang, R., Lu, M., Guo, Y., Xue, W., and Zhang, S. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023.

Liu, Z., Xu, Y., Xu, Y., Qian, Q., Li, H., Jin, R., Ji, X., and 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.

Maharana, S. K., Kushwaha, S. S., Zhang, B., Rodriguez, A., Wei, S., Tian, Y., and 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*, 2025a.

Maharana, S. K., Zhang, B., and Guo, Y. Palm: Pushing adaptive learning rate mechanisms for continual test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19378–19386, 2025b.

Maharana, S. K., Zhang, B., Karlinsky, L., Feris, R., and Guo, Y. Batclip: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c.

Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., and Sanner, S. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., and Tan, M. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

Niu, S., Miao, C., Chen, G., Wu, P., and Zhao, P. Test-time model adaptation with only forward passes. *arXiv preprint arXiv:2404.01650*, 2024.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and 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.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Smith, J., Balloch, J., Hsu, Y.-C., and 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.

Song, J., Lee, J., Kweon, I. S., and 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.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Wang, G., Lyu, F., and 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.

Wang, H., He, Z., Lipton, Z. C., and Xing, E. P. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

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

Wang, Y., Hong, J., Cheraghian, A., Rahman, S., Ahmedt-Aristizabal, D., Petersson, L., and 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.

Yang, M., Li, Y., Zhang, C., Hu, P., and Peng, X. Test-time adaptation against multi-modal reliability bias. In *The twelfth international conference on learning representations*, 2024.

Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., and Liu, H. Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4910–4920, 2022.

Zhang, Y., Mehra, A., Niu, S., and Hamm, J. Dpcore: Dynamic prompt coreset for continual test-time adaptation. *arXiv preprint arXiv:2406.10737*, 2024.

Zhou, K., Yang, Y., Hospedales, T., and Xiang, T. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pp. 561–578. Springer, 2020.

Appendix

A 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 (Wang et al., 2020; Chen et al., 2022; Niu et al., 2022; 2023; 2024). Typically, adaptation is performed independently per domain with a single forward pass per sample (Mai et al., 2022). TENT (Wang et al., 2020) minimizes the Shannon entropy of model predictions by fine-tuning the affine parameters of normalization layers. EATA (Niu et al., 2022) penalizes high entropy samples while minimizing the entropy. SAR (Niu et al., 2023) identified that model adaptation is affected by samples with large gradients.

TTA has also been widely studied in audio-visual settings. The first work, READ (Yang et al., 2024), proposed adapting the attention weights of the joint encoder to ensure robust cross-modal fusion. SuMi (Guo & Jin, 2025) employs a mutual information sharing loss to enhance the alignment between modalities. PTA (Wang et al., 2025) 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 (Li & Feng, 2025) 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 (Wang et al., 2022; Gong et al., 2022a; Song et al., 2023; Gan et al., 2023; Liu et al., 2023; Wang et al., 2024; Maharana et al., 2025b; Zhang et al., 2024). However, CTTA for audio-visual data has been relatively underexplored. We notice that only PTA (Wang et al., 2025) and BriMPR (Li & Feng, 2025) provide limited extensions of their respective AV-TTA methods to the continual setting (see their Appendices). A comprehensive study of audio–visual CTTA is still missing, which is the primary focus of this paper.

Remarks. While promising in performance, prior vision-based CTTA methods (Song et al., 2023; Liu et al., 2023; Zhang et al., 2024) use a small amount of source data to warm-start meta-networks/adapters or guide visual prompt tuning (Jia et al., 2022) 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 1. Moving forward, we abide by strict test-time settings and provide a realistic source-free audio-visual CTTA method.

B 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 (Gong et al., 2022c), 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 (Wang et al., 2022), 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

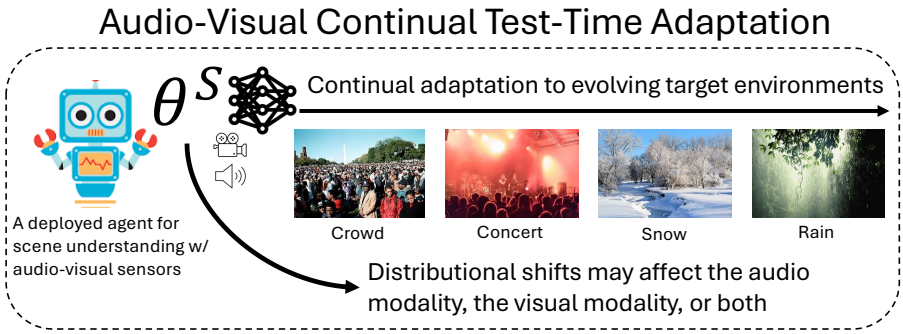
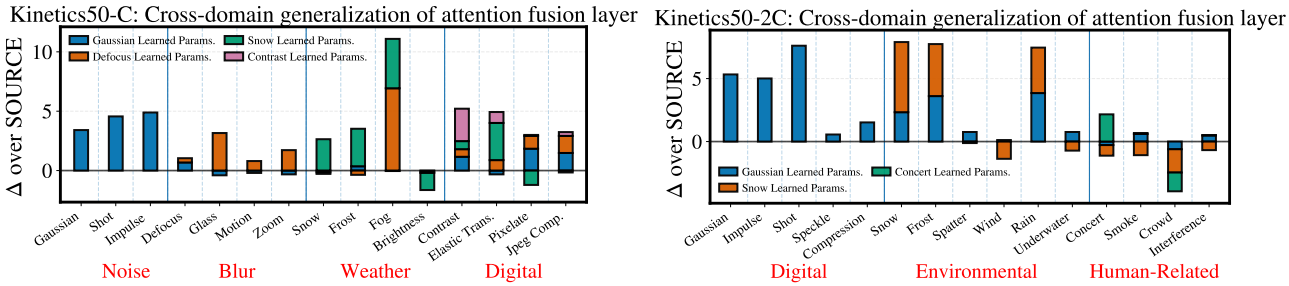


Figure 6. We illustrate **audio-visual continual test-time adaptation** using an example of a deployed "agent" with audio-visual sensors for scene understanding. Starting from a source model parameterized by θ^S, the agent encounters a sequence of evolving target environments where distributional shifts may affect the audio modality, the visual modality, or both, motivating continual adaptation at test-time. The goal is to maintain robust performance without access to the source data and task boundaries.



(a) Kinetics50-C with visual corruptions.

(b) Kinetics50-2C with bimodal corruptions.

Figure 7. **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) (Gong et al., 2022c) on the first unseen domain of each corruption category, following READ (Yang et al., 2024). This adapted state is then frozen and inferred on the remaining sequence of unseen domains. We report the accuracy change Δ (in %) over SOURCE.

multimodal batch $x^t = \{x_a^t, x_v^t\}$, 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., $p(x_i^S) \neq p(x_i^t)$ but $p(y_i^S) = p(y_i^t)$ for an ith 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., $\mathcal{T}_u \cap \left(\bigcup_{i=1}^{u-1} \mathcal{T}_i\right) = \emptyset$. Each $\mathcal{T}_u$ introduces a unique domain that has not been seen earlier, but the concept 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.*

Before diving in, we provide a motivating rationale to our proposed method `AV−CTTA` and illustrate the idea of audio-visual CTTA in Figure 6.

B.1 Motivation

Cross-domain generalization (Huang et al., 2020; Wang et al., 2019; Zhang et al., 2022; Li et al., 2017) 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 (Carlucci et al., 2019; Zhou et al., 2020), 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 test/target domains within the same category.* We illustrate the results in Figure 7.

We evaluate on Kinetics50-C (Yang et al., 2024) with visual-only corruptions (Hendrycks & Dietterich, 2019), grouped into

Noise, *Blur*, *Weather*, and *Digital* categories (Hendrycks & Dietterich, 2019). Starting from a CAV-MAE source model (Gong et al., 2022c), 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 (Yang et al., 2024). 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 Maharana et al. (2025a) 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, 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 1. 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. In Appendix C.2, we provide a neat algorithm of our proposed method. Figure 2 illustrates our approach.

C Additional Details of AV-CTTA

C.1 KL Divergence Between Modality-Specific Gaussian Distributions

In Section 2.1, we utilize the KL divergence as a distance metric between the current modality-specific statistics and the stored buffer elements. We define the distributions at time-step t and for the nth buffer element as $P_u^t = \mathcal{N}(\mu_u^t, \Sigma_u^t)$ and $P_u^n = \mathcal{N}(\mu_u^n, \Sigma_u^n)$, respectively. For simplicity and to ensure real-time retrieval efficiency as the buffer grows, we assume a diagonal covariance structure, $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_r^2)$. The PDF of a multivariate Gaussian is defined as $p(x) = (2\pi)^{-r/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$, where r is the feature dimension. The closed-form solution for the KL divergence is derived as follows:

$$g_u(t,n) = \mathcal{D}_{KL}(P_u^t \parallel P_u^n) = \mathbb{E}_{P_u^t}\left[\frac{\ln p_u^t(x)}{\ln p_u^n(x)}\right] \tag{10}$$

Now, upon substituting the PDF of a multivariate Gaussian in Eqn. (10),

$$\ln \frac{p_u^t(x)}{p_u^n(x)} = \frac{1}{2}\ln\frac{|\Sigma_u^n|}{|\Sigma_u^t|} - \frac{1}{2}(x-\mu_u^t)^\top(\Sigma_u^t)^{-1}(x-\mu_u^t) + \frac{1}{2}(x-\mu_u^n)^\top(\Sigma_u^n)^{-1}(x-\mu_u^n) \tag{11}$$

This simplifies to,

$$g_u(t,n) = \frac{1}{2}\ln\frac{|\Sigma_u^n|}{|\Sigma_u^t|} - \frac{1}{2}\mathbb{E}_{P_u^t}\left[(x-\mu_u^t)^\top(\Sigma_u^t)^{-1}(x-\mu_u^t)\right] + \frac{1}{2}\mathbb{E}_{P_u^t}\left[(x-\mu_u^n)^\top(\Sigma_u^n)^{-1}(x-\mu_u^n)\right] \tag{12}$$

Using the identity $\mathbb{E}_P[x^\top Ax] = \text{tr}(A\Sigma) + \mu^\top A\mu$ for a distribution with mean μ and covariance Σ, we evaluate the quadratic forms as, $\mathbb{E}_{P_u^t}[(x-\mu_u^t)^\top(\Sigma_u^t)^{-1}(x-\mu_u^t)] = \text{tr}((\Sigma_u^t)^{-1}\Sigma_u^t) = \text{tr}(\mathbf{I}_r) = r$. Similarly, $\mathbb{E}_{P_u^t}[(x-\mu_u^n)^\top(\Sigma_u^n)^{-1}(x-\mu_u^n)] = \text{tr}((\Sigma_u^n)^{-1}\Sigma_u^t) + (\mu_u^t-\mu_u^n)^\top(\Sigma_u^n)^{-1}(\mu_u^t-\mu_u^n)$. Upon substituting these into Eqn. (12), we obtain

$$g_u(t,n) = \frac{1}{2}\left[\underbrace{\text{tr}((\Sigma_u^n)^{-1}\Sigma_u^t)}_{\text{A}} + \underbrace{(\mu_u^t - \mu_u^n)^\top(\Sigma_u^n)^{-1}(\mu_u^t - \mu_u^n)}_{\text{B}} - r + \underbrace{\ln\frac{|\Sigma_u^n|}{|\Sigma_u^t|}}_{\text{C}}\right] \tag{13}$$

In our formulation, we assume a diagonal covariance matrix to maintain computational efficiency. Under this assumption, the individual terms simplify substantially. Term A in Eqn. (13) reduces to $\sum_{i=1}^{r}\frac{(\sigma_{u,i}^t)^2}{(\sigma_{u,i}^n)^2}$, i.e, a sum over diagonal entries only. Term B reduces to a weighted Euclidean distance as $\sum_{i=1}^{r}\frac{(\mu_{u,i}^t-\mu_{u,i}^n)^2}{(\sigma_{u,i}^n)^2}$. In addition, log-det term C simplifies as the log-product of diagonal elements as $\ln\frac{\prod(\sigma_{u,i}^n)^2}{\prod(\sigma_{u,i}^t)^2} = \sum_{i=1}^{r}\ln\frac{(\sigma_{u,i}^n)^2}{(\sigma_{u,i}^t)^2} = 2\sum_{i=1}^{r}\ln\frac{\sigma_{u,i}^n}{\sigma_{u,i}^t}$. Putting all of them back in Eqn. (12),

$$\begin{aligned} g_u(t,n) &= \frac{1}{2}\sum_{i=1}^{r}\left[\frac{(\sigma_{u,i}^t)^2}{(\sigma_{u,i}^n)^2} + \frac{(\mu_{u,i}^t-\mu_{u,i}^n)^2}{(\sigma_{u,i}^n)^2} - 1 + \ln\frac{\sigma_{u,i}^n}{\sigma_{u,i}^t}\right] \\ &= \sum_{i=1}^{r}\left[\frac{(\sigma_{u,i}^t)^2 + (\mu_{u,i}^t-\mu_{u,i}^n)^2}{2(\sigma_{u,i}^n)^2} - \frac{1}{2} + \ln\frac{\sigma_{u,i}^n}{\sigma_{u,i}^t}\right] \end{aligned} \tag{14}$$

For audio, $r = 128$ and for visual frames, $r = 3$.

C.2 AV-CTTA Algorithm

In Alg. 1, we present the full pipeline of our proposed AV-CTTA for audio-visual CTTA.

Algorithm 1 AV-CTTA: Selective Parameter Retrieval and Buffer Expansion

Inputs: Multimodal batch at time-step t, $x^t = \{(x_a^{t,i}, x_v^{t,i})\}_{i=1}^{B}$, model parameters θ^{t-1} (where $\theta \in \{\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V\}$), threshold τ, buffer size η, EMA smoothing factor β, learning rate α
Outputs: Predictions $y_t = \{y_i\}_{i=1}^{B}$
Initialize empty buffer $\mathcal{K} \leftarrow \emptyset$
Compute input-level statistics (μ_u^t, Σ_u^t) for $u \in \{a, v\}$ # Eqns. (1), (2), (3), (4)
$n^* \leftarrow \varnothing$, dist $\leftarrow \infty$
if $\mathcal{K}$ is empty **then**
  $\mathcal{K} \leftarrow \mathcal{K} \cup \{(\mu_u^t, \Sigma_u^t, \theta)\}$ for $u \in \{a, v\}$
  continue
end if
if $\mathcal{K}$ is not empty **then**
  n^*, dist $\leftarrow \arg\min_{n\in\mathcal{K}}(\mathcal{D}_{KL}^a(\mathcal{N}_a^t \parallel \mathcal{N}_a^n) + \mathcal{D}_{KL}^v(\mathcal{N}_v^t \parallel \mathcal{N}_v^n))$
end if
if dist $< \tau$ **then**
  $\theta \leftarrow \mathcal{K}[n^*]$.params # Selective Parameter Retrieval of $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$
  EMA update $\mathcal{K}[n^*]$.params # Eqn. (6)
  EMA update $\mathcal{K}[n^*]$.stats # Eqns. (7), (8)
else
  if $|\mathcal{K}| \geq \eta$ **then**
    $\mathcal{K} \leftarrow$ MergeClosestElements$(\mathcal{K})$
  end if
  $\mathcal{K} \leftarrow \mathcal{K} \cup \{(\mu_u^t, \Sigma_u^t, \theta)\}$ for $u \in \{a, v\}$
end if
$\mathcal{L} \leftarrow \mathcal{L}_{conf}(f_\theta(x_a^t, x_v^t)) - \mathcal{L}_{ne}(f_\theta(x_a^t, x_v^t))$
$\theta^t \leftarrow \theta^{t-1} - \alpha\nabla_\theta\mathcal{L}$ # Adaptation step of $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$
$y_t \leftarrow f_\theta(x_a^t, x_v^t)$

D Experimental Details

D.1 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 Section B), and parameter updates happen continually without any reset. Each corruption spans the entire test set and defines a task.

D.2 Datasets

Our experiments are conducted on the test sets of popular audio-visual datasets like Kinetics50 (Carreira & Zisserman, 2017), and VGGSound (Chen et al., 2020), following READ (Yang et al., 2024) and other baselines. Specifically, we adopt Kinetics50-C and VGGSound-C from READ, which contain unimodal corruptions and include 50 and 309 classes, respectively. Kinetics50-C includes 2,466, while VGGSound-C contains 14,046 audio–visual test pairs. We apply 15 visual (Hendrycks & Dietterich, 2019) and 6 audio corruptions at a severity level of 5, following READ. For the bimodal corruption setting, we borrow the 15 corruptions from AVRobustBench (Maharana et al., 2025a), i.e., *Gaussian, Impulse, Shot, Speckle* ... *Interference*, and apply them to the Kinetics50 and VGGSound source test sets. We name them Kinetics50-2C and VGGSound-2C, respectively, following their nomenclature. **These reflect *realistic*, *co-occurring*, and *correlated* bimodal corruptions, giving the feel of a real-world setting.** *It is worth mentioning that Kinetics50 (Carreira & Zisserman, 2017) and VGGSound (Chen et al., 2020) are visual-dominant and audio-dominant, respectively.* This means that task-specific information is in the visual and audio modalities, respectively.

D.3 Baselines

We compare our proposed AV-CTTA to popular TTA baselines (extensible to continual) like TENT (Wang et al., 2020), EATA (Niu et al., 2022), and SAR (Niu et al., 2023). The prior audio-visual baselines include READ (Yang et al., 2024), SuMi (Guo & Jin, 2025), PTA (Wang et al., 2025), and BriMPR* (Li & Feng, 2025). As discussed in Section A, we eliminate access to the source data in BriMPR to make it more deployment-friendly in a real-world setting.

D.4 Implementation Details

Following previous works, we use CAV-MAE (Gong et al., 2022c) as the source model, trained on Kinetics50 or VGGSound. We use a batch size of 32 across all experiments (see Appendix E.4 for an ablation). 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. We expand on the details in Appendix D.6, D.7. All experiments are conducted on an NVIDIA RTX A5000 GPU.

D.5 Source Model Architecture

Following all prior AV-TTA methods, we use CAV-MAE (Gong et al., 2022c) as the default audio-visual recognition model. CAV-MAE consists of 11-layer transformer encoders for both audio and visual modalities. Each video clip is sampled at 10 frames, with one randomly selected frame used as input to the visual encoder, while the corresponding 10-second audio waveform is converted into a spectrogram and processed by the audio encoder. The joint-encoder has only 1 transformer layer. Following prior work (Yang et al., 2024), spectrogram inputs are normalized using a dataset mean of -5.081 and a standard deviation of 4.4849.

D.6 Implementation of the Baselines

Here, we outline the implementation schemes of all the baselines used in our study. Throughout, we use a batch size of 32.

SOURCE[1] **(Gong et al., 2022c)**: We follow (Yang et al., 2024) and adopt CAV-MAE (Gong et al., 2022c) as the source model, which is used for inference. We use pre-trained weights of Kinetics50 and VGGSound, as released by (Yang et al., 2024).

TENT[2] **(Wang et al., 2020)**: We adapt all the affine parameters of all LayerNorms and minimize the entropy of model

[1]https://github.com/YuanGongND/cav-mae
[2]https://github.com/DequanWang/tent

predictions with a learning rate of 1×10^{-4} and an Adam optimizer.

EATA[3] **(Niu et al., 2022)**: With LayerNorm affine parameters being continually adapted, we optimize using Adam with a learning rate of 1×10^{-4}. The entropy threshold is set to 0.4×log(C), where C denotes the number of classes. As source data is not accessible during adaptation, we do not apply Fisher regularization to preserve source-domain knowledge.

SAR[4] **(Niu et al., 2023)**: We continually adapt only the LayerNorm affine parameters, optimized with Adam at a constant learning rate of 1×10^{-4}. To stabilize entropy minimization, we employ the same confidence filtering strategy used in EATA (Niu et al., 2022), and trigger model recovery when the confidence drops below 0.2. Model predictions are smoothed using ENA with a momentum coefficient of 0.9.

READ[5] **(Yang et al., 2024)**: Following their core implementation, we continually adapt $\mathcal{W}_Q$, $\mathcal{W}_K$, and $\mathcal{W}_V$ of the joint-encoder with a learning rate of 1×10^{-4} and an Adam optimizer.

SuMi[6] **(Guo & Jin, 2025)**: We use learning rates of 1×10^{-4} and 1×10^{-5} on the derivatives of Kinetics50 and VGGSound, respectively, with an Adam optimizer. Throughout, LayerNorm affine parameters are continually adapted. For bimodal corruptions, the mutual information loss is applied once every two iterations. All remaining hyperparameters follow the original paper.

PTA[7] **(Wang et al., 2025)**: For the test sets from Kinetics50, we use a recommended learning rate of 2×10^{-4} and 1×10^{-4} for VGGSound derivatives. $\mathcal{W}_Q$, $\mathcal{W}_K$, and $\mathcal{W}_V$ of the joint-encoder are continually adapted using the loss-specific hyperparameters recommended in the original method.

BriMPR*[8] **(Li & Feng, 2025)**: As discussed earlier, we remove access to the 32 source-domain samples. We use 10 prompt tokens for both the audio and visual encoders. During adaptation, the visual prompts are continually adapted using a cross-modal masked embedding loss together with an instance-wise contrastive loss. The mask ratio is set to 0.5, and the contrastive temperature is fixed to 0.7. Additional implementation details follow the original paper.

D.7 Implementation Details of AV-CTTA

At the onset of adaptation, we initialize a shared rehearsal buffer, $\mathcal{K} = \emptyset$. Each element in $\mathcal{K}$ is a tuple of modality-specific statistics and their corresponding adapted parameters: $(\mu_v^i, \Sigma_v^i, \mu_a^i, \Sigma_a^i, \mathcal{W}_Q^i, \mathcal{W}_K^i, \mathcal{W}_V^i)$. Here, μ and Σ denote the mean and covariance for the visual (v) and audio (a) modalities, while $\{\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V\}$ are the Query, Key, and Value projection weights of the attention fusion layer. We use a batch size of 32 and optimize with a learning rate of 1×10^{-4} using an Adam optimizer, for all the benchmark datasets. On datasets involving unimodal corruptions, i.e., Kinetics50-C and VGGSound-C (Yang et al., 2024), we set τ to be 0.005. For Kinetics50-C and VGGSound-2C (Maharana et al., 2025a) that involve bimodal corruptions, τ is set to be 0.01.

E Additional Results and Ablation Studies

E.1 Task-wise results of bimodal corruptions

In Table 4, we report the task-wise accuracies on Kinetics50-2C (top) and VGGSound-2C (bottom). For our proposed AV-CTTA, we also report the performances for various buffer sizes. Despite the increased complexity of the bimodal shifts, AV-CTTA consistently achieves superior performance compared to existing baselines. Our rationale and findings remain the same as in Section E.5. Over a wide range of buffer budgets, we observe consistent performances on Kinetics50-2C, which has about $2,466 \times 15$ test samples. On a larger dataset like VGGSound-2C, $\eta \geq 100$ achieves SOTA results.

E.2 Buffer budget η

In Figure 8, we study the scalability of AV-CTTA 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,

[3]https://github.com/mr-eggplant/EATA
[4]https://github.com/mr-eggplant/SAR
[5]https://github.com/XLearning-SCU/2024-ICLR-READ
[6]https://github.com/zrguo/SuMi
[7]https://github.com/MPI-Lab/PTA
[8]https://github.com/Luchicken/BriMPR

Table 2. **AV−CTTA achieves SOTA results on Kinetics50-C** with **video corruptions** (unimodal). We report the task-wise accuracy (%) at a severity level of 5, in the **continual** setting. We also report results with varying buffer sizes η.

	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
	AV−CTTA ($\eta = 50$)	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
	AV−CTTA ($\eta = 100$)	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
	AV−CTTA ($\eta = 200$)	48.52	50.24	48.44	67.91	63.18	71.68	68.03	64.94	66.99	61.58	72.80	53.97	66.11	67.87	64.42	62.45
	AV−CTTA ($\eta = 300$)	48.52	50.24	48.48	67.91	63.18	71.60	68.07	64.94	67.15	62.10	72.96	53.57	65.67	67.87	64.54	62.45
	AV−CTTA ($\eta = 400$)	48.52	50.24	48.16	67.91	63.22	71.60	67.99	64.90	67.15	62.06	72.76	53.57	65.55	67.83	64.66	62.41
	AV−CTTA ($\eta = \infty$)	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
	AV−CTTA ($\eta = 50$)	52.83	52.66	52.63	57.18	57.04	58.52	58.02	55.85	56.39	56.24	57.44	54.50	57.22	55.23	55.79	55.84
	AV−CTTA ($\eta = 100$)	52.81	52.76	52.81	57.13	56.74	58.33	58.02	55.87	56.35	56.17	57.52	54.40	56.17	55.32	55.95	55.76
	AV−CTTA ($\eta = 200$)	52.84	52.83	52.83	57.18	56.91	58.68	58.20	56.33	56.83	56.45	57.82	54.76	57.47	55.99	56.78	56.13
	AV−CTTA ($\eta = 300$)	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
	AV−CTTA ($\eta = 400$)	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
	AV−CTTA ($\eta = \infty$)	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

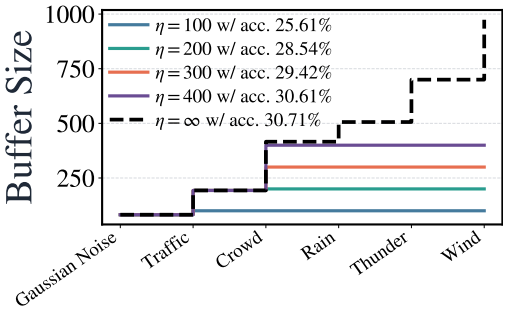
Table 3. **AV−CTTA achieves SOTA results on Kinetics-50C and VGGSound-C** with **audio corruptions** (unimodal). We report the task-wise accuracy (%) at a severity level of 5, in the **continual** setting. We also report results with varying buffer sizes η.

	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
	AV−CTTA ($\eta = 50$)	73.72	68.35	70.27	70.11	72.88	69.91	70.87
	AV−CTTA ($\eta = 100$)	73.72	68.47	69.79	70.27	72.56	70.39	70.87
	AV−CTTA ($\eta = 200$)	73.72	68.47	69.87	70.19	71.63	70.51	70.73
	AV−CTTA ($\eta = 300$)	73.72	68.47	69.87	70.19	71.63	70.15	70.67
	AV−CTTA ($\eta = 400$)	73.72	68.47	69.87	70.19	71.63	70.15	70.67
	AV−CTTA ($\eta = \infty$)	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
	AV−CTTA ($\eta = 50$)	39.67	27.99	17.31	16.99	19.05	14.12	22.52
	AV−CTTA ($\eta = 100$)	39.55	28.63	23.15	25.50	21.49	15.34	25.61
	AV−CTTA ($\eta = 200$)	39.55	28.42	24.45	30.14	30.90	17.81	28.54
	AV−CTTA ($\eta = 300$)	39.52	28.43	24.36	29.81	33.58	20.85	29.42
	AV−CTTA ($\eta = 400$)	39.52	28.43	24.32	29.37	35.65	26.38	30.61
	AV−CTTA ($\eta = \infty$)	39.52	28.43	24.31	29.44	35.52	27.05	30.71

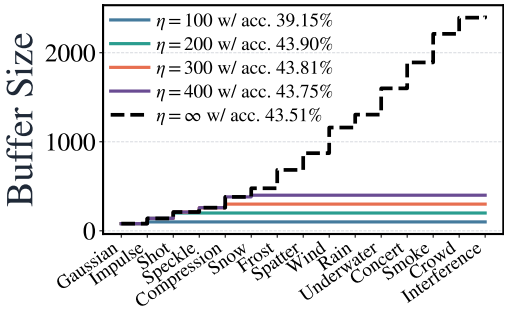
we notice that at a fixed budget $\eta \geq 300$ throughout, AV−CTTA still achieves better or competitive accuracies compared to baselines. On VGGSound-2C, with a tighter budget of $\eta = 200$, we notice improved performances. Element merging indeed consolidates task-specific knowledge and remains stable in long-term continual adaptation.

Table 4. **AV-CTTA achieves SOTA results on Kinetics50-2C and VGGSound-2C** (bimodal). We report the task-wise accuracy (%) at a severity level of 5, in the **continual** setting. We also report results with varying buffer sizes η.

	Method	Digital					Environmental						Human-Related				Mean
		Gaussian	Impulse	Shot	Speckle	Compression	Snow	Frost	Spatter	Wind	Rain	Underwater	Concert	Smoke	Crowd	Interference	
Kinetics50-2C	SOURCE	32.29	29.69	28.97	36.46	54.85	46.43	50.12	52.40	66.83	45.95	16.43	67.07	49.96	72.68	72.88	48.20
	TENT	22.76	3.93	2.00	2.12	6.81	1.96	2.00	2.00	2.36	1.96	1.96	1.96	1.96	2.00	1.96	3.85
	EATA	32.29	29.33	29.01	36.42	54.85	46.39	50.12	52.40	66.79	45.35	16.35	67.03	50.00	72.68	73.00	48.13
	SAR	32.05	29.53	28.12	35.54	53.40	45.67	49.12	50.44	64.50	44.15	17.43	64.38	47.32	69.99	72.20	46.92
	READ	37.62	36.74	38.22	42.83	59.66	48.64	55.25	52.48	59.02	47.52	37.86	57.33	40.87	58.53	59.58	48.81
	SuMi	35.34	31.85	26.76	42.47	56.97	42.79	40.70	44.79	61.02	34.13	7.69	52.72	28.49	61.82	62.54	42.01
	PTA	33.13	30.81	31.37	31.97	46.59	40.70	46.87	45.27	56.29	45.91	39.18	58.41	44.71	62.78	63.46	45.16
	BriMPR*	38.18	37.02	37.94	45.31	57.93	46.59	51.24	48.36	54.13	46.31	39.54	54.77	43.15	57.25	56.01	47.58
	AV-CTTA ($\eta = 50$)	34.13	30.77	33.73	37.62	58.97	50.00	59.02	57.01	65.38	56.17	19.35	66.79	51.16	70.39	72.32	50.85
	AV-CTTA ($\eta = 100$)	34.13	31.01	33.69	36.94	59.05	49.36	57.25	56.85	66.07	56.21	13.30	67.15	51.08	72.07	73.20	50.49
	AV-CTTA ($\eta = 200$)	34.13	30.73	33.69	36.86	59.02	48.36	55.13	55.73	66.83	55.61	13.34	67.23	51.44	72.36	73.28	50.25
	AV-CTTA ($\eta = 300$)	34.13	31.01	33.77	36.86	59.02	48.32	55.09	56.29	66.95	55.21	12.74	67.11	51.88	72.40	73.04	50.25
	AV-CTTA ($\eta = 400$)	34.13	30.81	33.77	36.86	59.02	48.32	55.13	56.61	66.87	55.09	13.90	67.15	51.60	72.44	73.32	50.33
	AV-CTTA ($\eta = \infty$)	34.13	31.09	33.73	36.82	59.02	48.32	55.21	56.57	66.87	55.13	14.50	66.83	51.12	71.87	73.12	50.29
VGGSound-2C	SOURCE	30.11	23.63	20.77	25.77	34.57	24.48	46.39	48.43	50.40	29.50	42.15	47.60	31.83	47.64	55.21	37.23
	TENT	2.87	0.33	0.33	0.37	0.89	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.54
	EATA	30.16	23.81	20.97	26.36	34.73	24.55	46.69	47.94	49.50	28.26	40.90	46.00	29.27	45.95	53.46	36.57
	SAR	12.74	10.90	13.30	18.97	36.62	18.21	43.16	46.75	49.01	19.70	41.63	31.10	23.26	43.85	51.09	30.69
	READ	43.74	38.89	40.23	26.72	25.54	22.65	27.92	29.26	30.81	22.08	25.53	26.18	25.04	28.38	33.09	29.74
	SuMi	33.23	25.37	24.70	27.67	35.44	20.79	46.54	48.04	49.66	24.99	38.67	47.76	31.93	48.43	55.42	37.24
	PTA	40.82	34.35	33.69	28.21	29.71	32.52	39.27	41.44	41.59	32.35	36.44	36.41	34.52	35.96	40.37	35.84
	BriMPR*	30.47	24.53	25.02	24.12	26.15	21.54	27.77	26.88	26.73	20.81	23.24	23.55	18.83	25.85	28.20	24.91
	AV-CTTA ($\eta = 50$)	40.63	34.77	37.70	31.37	34.95	41.07	50.02	49.03	52.51	38.79	29.37	29.85	24.06	28.40	35.58	37.21
	AV-CTTA ($\eta = 100$)	40.42	33.95	37.66	30.05	34.80	40.81	50.07	49.86	52.58	44.52	39.74	34.44	28.85	31.85	37.68	39.15
	AV-CTTA ($\eta = 200$)	40.42	33.58	37.02	29.60	34.63	40.33	49.81	49.49	51.96	44.63	45.90	49.47	46.10	51.00	54.55	43.90
	AV-CTTA ($\eta = 300$)	40.42	33.90	36.86	29.57	34.61	40.22	49.84	49.35	51.54	44.31	45.71	49.34	46.04	50.69	54.79	43.81
	AV-CTTA ($\eta = 400$)	40.42	33.75	37.03	29.55	34.59	40.07	49.83	49.26	51.45	44.29	45.65	49.20	46.04	50.18	54.89	43.75
	AV-CTTA ($\eta = \infty$)	40.42	33.78	36.79	29.33	34.62	39.96	49.73	49.12	51.12	44.21	45.34	48.66	46.14	48.53	54.91	43.51



(a) VGGSound-C w/ audio corruptions.



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

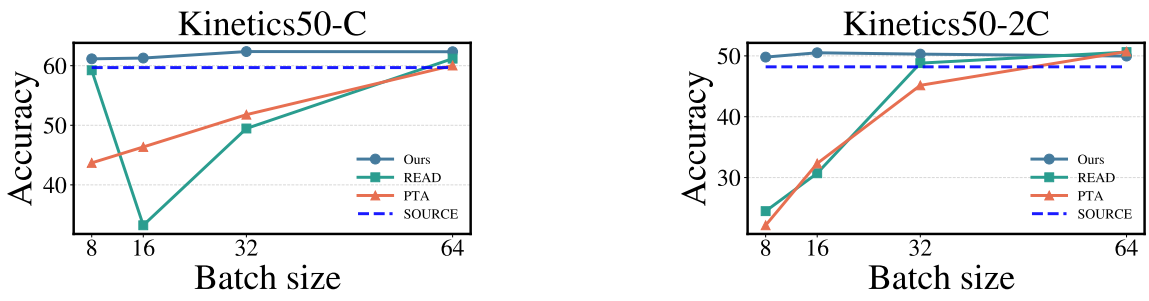
Figure 8. **Effect of buffer size η.** Mean accuracy (%) on VGGSound-C (left) and VGGSound-2C (right).

Table 5. **Alternative distance metric $g_u(t, n)$.** Results are reported on Kinetics50-C (w/ visual corruptions) and Kinetics50-2C.

Metric $g_u(t, n)$	Kinetics50-C	Kinetics50-2C
$\|\mu_u^t - \mu_u^n\|_2 + \|\sigma_u^t - \sigma_u^n\|_2$	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

E.3 Distance metric $g_u(t, n)$

In Table 5, 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 $\mathcal{D}_{KL}^{a+v}$.



(a) Kinetics50-C (video corruptions)

(b) Kinetics50-2C (bimodal corruptions)

Figure 9. **AV-CTTA maintains stable performance under small batch sizes.** The x-axis shows batch size, and the y-axis reports mean accuracy for AV-CTTA and baselines.

E.4 Effect of batch size

Here, we investigate the sensitivity of AV-CTTA 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 AV-CTTA maintains competitive performance even at $B = 8$. This highlights the suitability of our proposed method under memory limitations, especially on small data streams.

E.5 Fixed buffer budget η

In Tables 2 and 3, we report additional results on Kinetics50-C and VGGSound-C, respectively. In Kinetics50-C, the video is corrupted, whereas in VGGSound-C, the audio is corrupted. As a reminder, the task-specific information lies in the visual and audio modalities for Kinetics50 and VGGSound, respectively. The results indicate that AV-CTTA maintains superior performance even when the "primary" modality is corrupted.

We observe distinct scaling behaviors across these datasets. On Kinetics50-C, AV-CTTA achieves stable and competitive performance with a buffer size as small as $\eta = 50$. Given that the test set comprises $36,990$ samples ($2,466 \times 15$ tasks), a budget of $\eta = 50$ represents a compression ratio of nearly $740 : 1$, yet it is sufficient. Conversely, on VGGSound-C, comparable or superior performance is achieved at $\eta \geq 200$. We attribute this to the size. VGGSound-C has a significantly larger amount of data (14046×6 tasks), necessitating a slightly larger set of buffer elements to adequately cover the expanded distribution. Despite these differences in scale, the overall results confirm that AV-CTTA effectively bounds memory growth without sacrificing adaptation quality, making it ideal for large-scale, long-term deployment.

E.6 Alternatives to maintaining a buffer budget

In our proposed method, to maintain a strict budget η of buffer $\mathcal{K}$, we perform statistical merging of the most redundant elements. We perform pairwise comparisons of all stored elements in $\mathcal{K}$ by solving Eqn. (9). Here, we explore another strategy following (Zhang et al., 2024), where the oldest element in $\mathcal{K}$ is removed when $|\mathcal{K}| \geq \eta$. We conduct experiments on Kinetics50-2C and VGGSound-2C. In Table 6, we observe that removing the oldest element maintains robust performance for both datasets involving bimodal corruptions. On VGGSound-2C, with η as low as 50, AV-CTTA maintains a stronger performance compared to merging by averaging elements (see Table 4). However, we do posit that this strategy is highly dependent on the order of tasks. In settings where tasks arrive in a random order, discarding the oldest element may prove detrimental. This risks prematurely removing statistically relevant elements that belong to corruption categories, which may recur later in the stream, as evidenced by the cross-task transferability observed in Section B.1.

E.7 Sensitivity to Task Orders

In real-world continual learning, the sequence of tasks arriving is often unpredictable. To assess the robustness of AV-CTTA and competing baselines against this, we evaluate their performance across three alternative task orderings in the challenging

Table 6. **AV−CTTA remains stable in performance when the oldest element is removed from the buffer to maintain the budget.** We report the mean accuracies on Kinetics50-2C and VGGSound-2C with varying buffer sizes η.

Buffer budget η	Kinetics50-2C	VGGSound-2C
$\eta = 50$	50.38	44.10
$\eta = 100$	50.43	43.81
$\eta = 200$	50.28	43.51

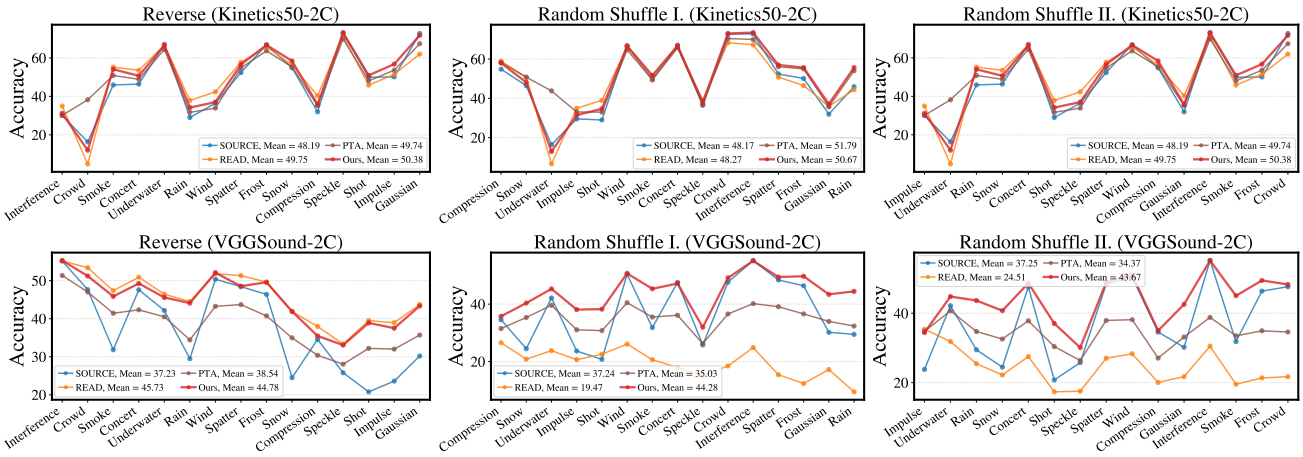


Figure 10. **AV−CTTA is robust to variations in task ordering.** The top row reports results on Kinetics50-2C, while the bottom row shows results on VGGSound-2C. In each plot, the x-axis denotes the task order in the continual setting, and the y-axis reports task-wise accuracy. The legend summarizes the mean accuracy achieved by each baseline and by our method.

bimodal corruption setting. While the standard sequence follows a fixed progression (e.g., *Gaussian* $\rightarrow$ *Impulse* $\rightarrow \cdots \rightarrow$ *Interference*), we introduce: 1) Reverse: The original sequence is inverted. 2) Random Shuffle I & II: Two distinct permutations of the 15 bimodal tasks. The objective of this study is to study whether the selective retrieval mechanism in AV−CTTA is dependent on a specific curriculum or not. We conduct experiments on Kinetics50-2C and VGGSound-2C and illustrate the results in Figure 10. On average, AV−CTTA demonstrates strong robustness to task order variations compared to baselines, particularly in the difficult bimodal setting. Although a performance drop is observed during the initial tasks, continual adaptation allows the model to stabilize over time and converge to comparable accuracy across different task orderings. Overall, the selective retrieval mechanism remains effective without long-term degradation and remains resilient to variations in task ordering.

F Limitations and Discussions

While AV−CTTA demonstrates significant advancements in audio-visual CTTA, several avenues remain for further exploration.

Compute overhead. The selective parameter retrieval mechanism described in Section 2.1 performs an $O(N)$ linear search over the buffer $\mathcal{K}$ at each time-step t. Although effective, this design introduces non-trivial computational overhead that may become prohibitive as the number of encountered target environments grows. Improving retrieval efficiency, or eliminating the need for an explicit buffer, remains an open challenge.

Model architectural assumptions. Our method AV−CTTA is developed under the assumption that audio-visual recognition models employ modality-specific encoders followed by a joint encoder for late cross-modal fusion, as commonly adopted in prior architectures (Gong et al., 2022c;b; Georgescu et al., 2023; Huang et al., 2023). Consequently, its applicability to alternative fusion paradigms or more heterogeneous architectures is not guaranteed, limiting its generality.

Threshold τ. The method relies on a single hyperparameter τ to control selective parameter retrieval. In practical test-time deployment settings, hyperparameter tuning is often infeasible or costly (Maharana et al., 2025c).

Correlation between raw input-level statistics and optimality of the retrieved fusion parameters. Modality-level input

statistics (μ, Σ) capture covariate shift, and since cross-modal attention is conditioned on these inputs, similar shifts induce similar optimal fusion parameters, consistent with prior findings (Zhang et al., 2024). We store adapted fusion parameters in a buffer and retrieve them using the KL-divergence similarity. Tracking retrievals over the first 6 tasks of Kinetics50-C (78 batches per task) shows clear alignment: Shot Noise retrieves from Gaussian 4/78 batches (and the rest are continually adapted), Impulse Noise retrieves from Gaussian (77/78). Defocus Blur uses Gaussian for (1/78) batches, while Glass, Motion, and Zoom Blur always retrieve from Defocus Blur (78/78). This confirms that similar domains retrieve parameters from each other, validating that the previous parameters retrieved by our method correspond to similar domains.