REGULARIZING MODALITY CONTRIBUTION DRIFT IN MULTIMODAL CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal continual learning (MMCL) aims to acquire new knowledge from multimodal data while retaining previously learned knowledge. Existing MMCL methods primarily mitigate forgetting by aligning cross-modal representations or preserving feature-level semantic similarity. However, **different tasks may rely on different modalities, and learning new tasks can alter how modalities contribute to predictions on previously learned tasks**. It remains **underexplored how modality contributions evolve across incremental stages and how such changes relate to forgetting in MMCL**. We term such changes Modality Contribution Drift (MCD) and introduce an MCD score based on controlled modality-subset interventions. Our theoretical and empirical analyses show how **contribution drift can lead to forgetting and reveal that existing MMCL and conventional CL methods do not effectively mitigate MCD.** To address this issue, we propose Continual Modality Contribution Drift Regularization (CMCDR) to preserve the modality contribution profiles of previously learned tasks in both replay-based and replay-free settings. In the replay-based setting, CMCDR estimates modality contributions from stored old samples and regularizes their drift relative to a frozen previous model. In the replay-free setting, CMCDR uses current-task samples as probes to match old-class contribution profiles between the current and frozen models without storing old exemplars. Across six benchmarks covering multimodal class-incremental learning and continual multimodal question answering, CMCDR substantially reduces modality contribution drift, improves average accuracy by $0.88{\sim}7.66$ percentage points, and reduces average forgetting by $0.85{\sim}11.55$ percentage points over the corresponding baselines.

1 INTRODUCTION

Multimodal continual learning (MMCL) aims to learn from a sequence of multimodal tasks while retaining knowledge acquired from previous tasks. A key challenge in mitigating catastrophic forgetting in MMCL is to preserve cross-modal relationships established during earlier stages. Existing MMCL methods primarily address this challenge by stabilizing multimodal representations and preserving cross-modal semantic relationships. For example, AV-CIL (Pian et al., 2023) aligns audio-visual representations by preserving instance- and class-aware semantic similarity, while MG-CLIP (Huang et al., 2025) preserves the image–text modality gap across incremental stages. However, stabilizing multimodal representations does not necessarily preserve how individual modalities and cross-modal interactions contribute to model predictions.

This issue is particularly important in MMCL because different tasks can rely on substantially different combinations of modality-specific and cross-modal information (Hessel & Lee, 2020; Liang et al., 2023). Figure 1(a) illustrates such task-dependent modality reliance, e.g., the road-crossing task relies predominantly on visual cues, whereas the question-answering task requires joint reasoning over visual and linguistic information. Our diagnostic analyses further reveal substantial variation in modality contributions across incremental tasks within the same dataset (see Figure 1(b)) and across multimodal benchmarks (see Figure 1(c)), showing that modality reliance can vary substantially across tasks and learning settings.

In MMCL, adapting to such task-dependent modality reliance can alter the contributions of modality-specific and cross-modal information to model predictions. These changes can alter the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

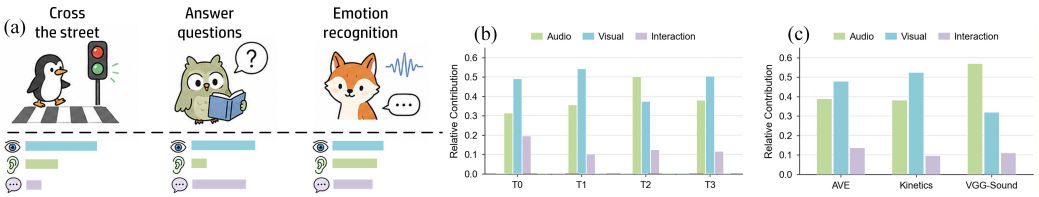


Figure 1: Task-dependent modality contributions. (a) Illustrative examples. (b) Contribution profiles obtained by partitioning AVE (Tian et al., 2018) categories into four tasks and estimating audio, visual, and cross-modal contributions. (c) Average contribution profiles computed on AVE, Kinetics-Sounds (Arandjelović & Zisserman, 2017), and VGGSound (Chen et al., 2020).

contribution profiles established for previous tasks and lead to forgetting. However, the evolution of modality contributions across incremental stages and their relation to forgetting in MMCL remain underexplored.

To characterize these contribution changes, we introduce an intervention-based estimator that quantifies the contributions of individual modalities and cross-modal interactions to model predictions. Based on these estimates, we define an MCD score to quantify changes in both the absolute magnitude and relative composition of old-task contribution profiles across incremental stages. Our theoretical and empirical analyses show how contribution drift can lead to forgetting and further show that existing MMCL and conventional CL methods do not effectively mitigate MCD. Therefore, we propose Continual Modality Contribution Drift Regularization (CMCDR), which regularizes changes in old-task contribution profiles. Our main contributions are summarized as follows:

- **Problem and Metric.** We identify Modality Contribution Drift (MCD), an underexplored phenomenon in MMCL, and introduce an intervention-based contribution estimator together with an MCD metric to quantify changes in contribution profiles across incremental stages.
- **Method.** We propose Continual Modality Contribution Drift Regularization (CMCDR), with replay-based and replay-free variants, to mitigate contribution drift and reduce forgetting.
- **Performance.** We conduct extensive experiments on multimodal class-incremental learning and continual multimodal question answering, showing that CMCDR consistently reduces MCD and improves continual-learning performance across MMCL settings and benchmarks.

2 Contribution Drift in MMCL

We use modality-subset interventions to study modality contribution drift (MCD) in MMCL (Section 2.1) and analyze how such drift can alter old-task prediction margins and induce forgetting (Section 2.2.2). Theoretical and empirical analyses further show that neither existing MMCL methods nor conventional continual learning methods effectively mitigate modality contribution drift (Sections 2.2.2 and 2.2.3).

2.1 Intervention-Based Contribution Estimation

We estimate modality contributions by intervening on modality subsets and decomposing the induced predictive benefit. Let $\mathcal{M} = \{1, \dots, M\}$ denote the set of modality indices, and $\mathbf{x}_i = (x_i^1, \dots, x_i^M)$ denote a multimodal sample. We use F_θ to denote the full predictor, comprising the modality encoders, fusion module, and classifier. For any $\mathcal{S} \subseteq \mathcal{M}$, we use $\perp^m$ to denote the absence of modality m and define the modified input and corresponding model output as

$$\widetilde{x}_i^{m,\mathcal{S}} = \begin{cases} x_i^m, & m \in \mathcal{S}, \\ \perp^m, & m \notin \mathcal{S}, \end{cases} \quad \mathbf{x}_i^{\mathcal{S}} := \mathcal{R}_{\mathcal{S}}(\mathbf{x}_i) = \big(\widetilde{x}_i^{m,\mathcal{S}}\big)_{m=1}^{M}, \quad \mathbf{z}_i(\mathcal{S}) = F_\theta\big(\mathbf{x}_i^{\mathcal{S}}\big). \tag{1}$$

Here, $\perp^m$ denotes the removal of modality m. **Appendix B.4** describes how modality removal is implemented for each architecture.

To quantify how each modality subset supports the prediction of a target class, we first define a scalar benefit from the corresponding logits. For a candidate set $\mathcal{A}$ satisfying $a \in \mathcal{A}$ and $|\mathcal{A}| \geq 2$, the target-class benefit is

$$B_i^{a,\mathcal{A}}(\mathcal{S}) = z_{i,a}(\mathcal{S}) - \log \sum_{\ell \in \mathcal{A},\, \ell \neq a} \exp\big(z_{i,\ell}(\mathcal{S})\big), \tag{2}$$

where $\mathbf{z}_i(\mathcal{S})$ is the logit vector for the modality subset, $z_{i,a}(\mathcal{S})$ is the target-class logit, and $z_{i,\ell}(\mathcal{S})$ are competing logits for classes $\ell \in \mathcal{A}$ with $\ell \neq a$. Thus, $B_i^{a,\mathcal{A}}(\mathcal{S})$ is the target-vs-competitor log-sum-exp margin when only modalities in $\mathcal{S}$ are retained.

Taking the no-modality intervention as the reference condition, $B_i^{a,\mathcal{A}}(\mathcal{S}) - B_i^{a,\mathcal{A}}(\emptyset)$ quantifies the decision-level benefit of retaining the modalities in $\mathcal{S}$ on the log-odds margin scale. Here, $B_i^{a,\mathcal{A}}(\emptyset)$ is the response when no modality is retained, while $B_i^{a,\mathcal{A}}(\mathcal{M})$ is the response under the full multimodal input.

Since the benefit of a modality subset can arise from both individual modalities and their interactions, we use the Möbius transform (Harsanyi, 1963) to decompose the subset-level benefit into coalition-wise contributions. For each nonempty modality coalition $\mathcal{T} \subseteq \mathcal{M}$, we define its contribution as:

$$C_i^{a,\mathcal{A}}(\mathcal{T}) = \sum_{\mathcal{S}\subseteq\mathcal{T}} (-1)^{|\mathcal{T}|-|\mathcal{S}|} B_i^{a,\mathcal{A}}(\mathcal{S}), \qquad \mathbf{C}_i = \left[C_i^{a,\mathcal{A}}(\mathcal{T})\right]_{\emptyset\neq\mathcal{T}\subseteq\mathcal{M}}. \tag{3}$$

By Möbius inversion, the contribution profile exactly decomposes the full-input benefit relative to the no-modality response:

$$\sum_{\emptyset\neq\mathcal{T}\subseteq\mathcal{M}} C_i^{a,\mathcal{A}}(\mathcal{T}) = B_i^{a,\mathcal{A}}(\mathcal{M}) - B_i^{a,\mathcal{A}}(\emptyset). \tag{4}$$

The Möbius decomposition separates the subset-level benefits into singleton modality contributions and higher-order cross-modal interaction contributions. Appendix A.1 provides the definition of the interventional response B and the derivation of the resulting contribution profile.

2.2 Analysis of Modality Contribution Drift

We use the estimator above to quantify changes in modality reliance across incremental tasks. We first define the MCD score using class-level contribution profiles (Section 2.2.2) and then examine whether conventional continual-learning mechanisms (Section 2.2.2) and multimodal representation alignment (Section 2.2.3) can effectively suppress this drift.

2.2.1 Exploration Setup

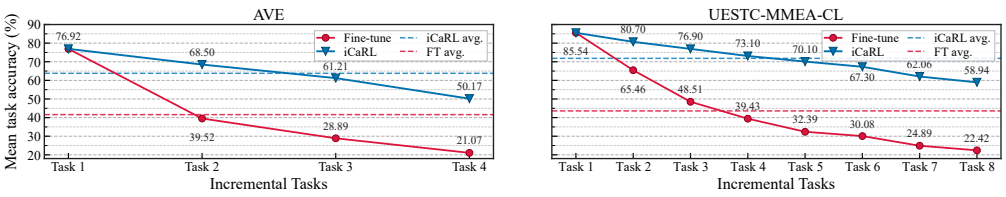


Figure 2: Mean task accuracy across incremental tasks in the exploratory diagnostic run on AVE and UESTC-MMEA-CL.

We use the fully connected classifier version of iCaRL (Rebuffi et al., 2017) (iCaRL-FC, hereafter referred to as iCaRL) as the baseline and evaluate it on two representative MMCL datasets, AVE (Tian et al., 2018) and UESTC-MMEA-CL (Xu et al., 2024). Experimental settings are provided in **Appendix B.1**. As shown in Figure 2, iCaRL's accuracy declines consistently on both datasets as incremental learning proceeds. On AVE, the mean task accuracy drops from 76.92% at Task 1 to 50.17% at Task 4, with a four-stage average of 64.20%. On UESTC-MMEA-CL, it decreases from 85.54% to 58.94%, with an average accuracy of 71.83%. Compared with Fine-tune, iCaRL achieves higher mean task accuracy at later incremental stages, indicating that it alleviates catastrophic forgetting.

2.2.2 Modality Contribution Drift in MMCL

To assess whether iCaRL preserves old-class modality contributions, we define an MCD score that combines changes in overall contribution magnitude with weighted changes in the normalized contribution profile.

Let $\tau(y)$ denote the stage at which class y is introduced, and $\mathcal{Y}_{\leq s}$ the set of classes observed up to stage s. We fix $\mathcal{A}_y = \mathcal{Y}_{\leq \tau(y)}$ so that contribution profiles are compared under a fixed class context across stages.

For each task k, we maintain a separate probe set $\mathcal{P}_k^{\text{task}}$. After learning task k, we record the contribution profiles of its probe samples as reference profiles. For each $i \in \mathcal{P}_k^{\text{task}}$ and stage $s \geq k$, let $\mathbf{c}_{i,s}$ denote its contribution vector and $\widehat{\mathbf{c}}_{i,s}$ its normalized counterpart:

$$\mathbf{c}_{i,s} = \left[C_{i,s}^{y_i,\mathcal{A}_{y_i}}(\mathcal{T})\right]_{\emptyset \neq \mathcal{T} \subseteq \mathcal{M}}, \quad \widehat{\mathbf{c}}_{i,s} = \frac{\mathbf{c}_{i,s}}{\|\mathbf{c}_{i,s}\|_1 + \epsilon}, \tag{5}$$

where each contribution term is computed using Eq. 3, and $\epsilon > 0$ stabilizes the normalization and avoids degenerate zero-magnitude cases.

For $s \in \{k, t\}$, define $m_{i,s} = \|\mathbf{c}_{i,s}\|_1$. At stage t, let $\mathcal{K}_{<t}$ denote the set of all previous tasks. The MCD score is defined as

$$\text{MCD}_t = \frac{1}{|\mathcal{K}_{<t}|} \sum_{k \in \mathcal{K}_{<t}} \frac{1}{|\mathcal{P}_k^{\text{task}}|} \sum_{i \in \mathcal{P}_k^{\text{task}}} \left[|m_{i,t} - m_{i,k}| + (\min(m_{i,t}, m_{i,k}) + \epsilon)\|\widehat{\mathbf{c}}_{i,t} - \widehat{\mathbf{c}}_{i,k}\|_1\right]. \tag{6}$$

Its two terms measure changes in the overall contribution magnitude and the normalized contribution profile, respectively.

To assess whether conventional continual learning can mitigate modality contribution drift, we compare fine-tuning with iCaRL using the MCD score. As shown in Figure 3, iCaRL only marginally reduces the average MCD compared with fine-tuning on both AVE (0.7790 vs. 0.8130) and UESTC-MMEA-CL (0.4640 vs. 0.4867). Together with the improved accuracy in Figure 2, these results show that exemplar replay and distillation can alleviate forgetting but do not effectively mitigate modality contribution drift, suggesting that improved continual learning performance does not necessarily imply more stable modality contributions.

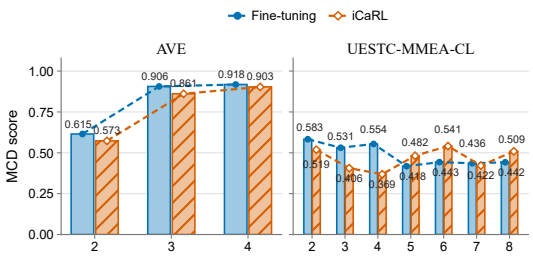


Figure 3: MCD scores across incremental stages in the exploratory diagnostic run on AVE and UESTC-MMEA-CL.

Contribution Drift and Forgetting. For an old sample i, fix a reference stage r with $\tau(y_i) \leq r < t$, and consider a candidate set $\mathcal{A}_i \subseteq \mathcal{Y}_{\leq r}$ that contains y_i and satisfies $|\mathcal{A}_i| \geq 2$. We apply the same modality-removal intervention at stages r and t. All quantities below are computed with target y_i and candidate set $\mathcal{A}_i$, which we suppress from the notation.

Define $\bar{B}_{i,s}(\mathcal{S}) = B_{i,s}(\mathcal{S}) - B_{i,s}(\emptyset)$. The retained contribution margin is

$$m_{i,r\to t}^{\text{keep}} = B_{i,t}(\emptyset) + \bar{B}_{i,r}(\mathcal{M}). \tag{7}$$

This quantity is the margin obtained by combining the stage-t no-modality response with the stage-r contribution profile. Let $\mathbf{c}_{i,s}$ denote the contribution vector at stage $s \in \{r, t\}$, and define its drift as $\Delta\mathbf{c}_{i,r\to t} = \mathbf{c}_{i,t} - \mathbf{c}_{i,r}$.

Proposition 1 (Contribution drift can cross the retention margin)**.** *Let $\gamma_{i,t}$ denote the target-class logit margin over the strongest competitor in $\mathcal{A}_i$ at stage t. Then*

$$\begin{aligned} B_{i,t}(\mathcal{M}) &= m_{i,r\to t}^{\text{keep}} + \mathbf{1}^\top \Delta\mathbf{c}_{i,r\to t}, \\ \gamma_{i,t} &\geq m_{i,r\to t}^{\text{keep}} - \|\Delta\mathbf{c}_{i,r\to t}\|_1. \end{aligned} \tag{8}$$

If $\|\Delta\mathbf{c}_{i,r\to t}\|_1 < m_{i,r\to t}^{\text{keep}}$, then y_i has a strictly higher logit than every other class in $\mathcal{A}_i$. If $m_{i,r\to t}^{\text{keep}} > 0$ and $\gamma_{i,t} \leq 0$, then $\|\Delta\mathbf{c}_{i,r\to t}\|_1 \geq m_{i,r\to t}^{\text{keep}}$. Thus, the retained margin gives a threshold below which prediction retention is guaranteed and beyond which prediction failure becomes possible.

Proposition 1 shows that contribution drift can reduce the old-task prediction margin and cause forgetting. The proof is given in **Appendix A.2**.

2.2.3 Effect of Cross-modal Representation Stability on Contribution Drift

The preceding results show that replay and distillation do not necessarily preserve modality contribution structures. Existing MMCL methods primarily stabilize cross-modal representations or semantic relationships across incremental stages (Pian et al., 2023; Mo et al., 2023; Pian et al., 2024; Huang et al., 2025). However, whether such representation-level stability also preserves modality contributions to model decisions remains unclear. To examine this question, we augment iCaRL with Cross-modal Representation Alignment (RepAlign) from AV-CIL (Pian et al., 2023) and the Cross-modal Similarity Distillation Constraint (CrossSDC) (Pian et al., 2024). At each incremental step, we quantify cumulative representation drift from the stage at which each task is introduced to the current stage using Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Over the same interval and using the same probes, we measure cumulative contribution drift using Eq. 6. The results are averaged across incremental steps on AVE and UESTC-MMEA-CL. Detailed experimental settings are provided in **Appendix B.1**.

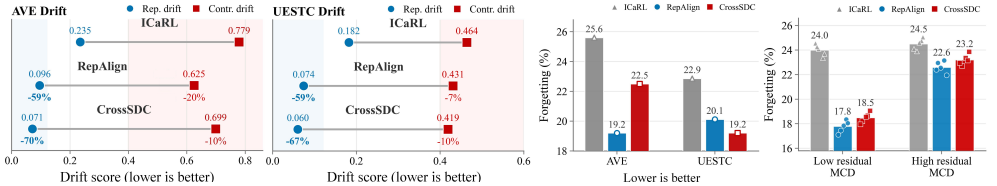


Figure 4: Representation stability and residual MCD in MMCL. Both drift metrics measure cumulative changes from each task's introduction stage to the current stage.

Figure 4(a,b) show that RepAlign and CrossSDC reduce representation drift by approximately 59%∼70% on AVE and 59%∼67% on UESTC-MMEA-CL, while achieving substantially smaller reductions in the MCD score. Figure 4(c) further shows that both methods reduce average forgetting on the two datasets. In Figure 4(d), we group old classes according to their residual MCD scores, with bars indicating group-level forgetting and points indicating class-level forgetting. Classes with low residual MCD scores exhibit clear reductions in forgetting, whereas those with high residual MCD scores retain forgetting levels comparable to iCaRL. Together, these results show that substantial reductions in representation drift can alleviate forgetting without effectively mitigating MCD, indicating that representation stability does not necessarily preserve modality contributions.

Why Representation Stability Is Insufficient. To formalize this observation, Proposition 2 bounds contribution drift by changes in modality-subset representations and the decision map, showing why representation stability alone does not guarantee contribution stability.

Proposition 2 (Representation stability does not guarantee contribution stability)**.** *For a sample i with old-class target a and stages $\tau(a) \le r < t$, fix $\mathcal{A} = \mathcal{A}_a$. Let $\mathbf{u}_{i,s}(\mathcal{S})$ be the modality-subset representation at stage s in a common normed space and $g_s^{a,\mathcal{A}}$ the corresponding target-vs-rest margin map, such that $B_{i,s}^{a,\mathcal{A}}(\mathcal{S}) = g_s^{a,\mathcal{A}}(\mathbf{u}_{i,s}(\mathcal{S}))$. Define*

$$\Delta_u(\mathcal{S}) = \|\mathbf{u}_{i,t}(\mathcal{S}) - \mathbf{u}_{i,r}(\mathcal{S})\|, \Delta_g(\mathcal{S}) = \left| g_t^{a,\mathcal{A}}(\mathbf{u}_{i,r}(\mathcal{S})) - g_r^{a,\mathcal{A}}(\mathbf{u}_{i,r}(\mathcal{S})) \right|. \tag{9}$$

If $g_t^{a,\mathcal{A}}$ is L_g-Lipschitz with respect to this norm, then for any nonempty $\mathcal{T} \subseteq \mathcal{M}$,

$$\left| C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) - C_{i,r}^{a,\mathcal{A}}(\mathcal{T}) \right| \le \sum_{\mathcal{S} \subseteq \mathcal{T}} \left(L_g \Delta_u(\mathcal{S}) + \Delta_g(\mathcal{S}) \right). \tag{10}$$

Let $\bar{B}_{i,s}^{a,\mathcal{A}}(\mathcal{S}) = B_{i,s}^{a,\mathcal{A}}(\mathcal{S}) - B_{i,s}^{a,\mathcal{A}}(\emptyset)$. If every nonempty $\mathcal{S} \subseteq \mathcal{T}$ satisfies $|\bar{B}_{i,t}^{a,\mathcal{A}}(\mathcal{S}) - \bar{B}_{i,r}^{a,\mathcal{A}}(\mathcal{S})| \le \epsilon_{\bar{B}}$, then $|C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) - C_{i,r}^{a,\mathcal{A}}(\mathcal{T})| \le (2^{|\mathcal{T}|} - 1)\epsilon_{\bar{B}}$. Moreover, there exist stage-invariant modality-subset representations and two decision maps such that $\Delta_u(\mathcal{S}) = 0$ for every $\mathcal{S}$, while $C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) \ne C_{i,r}^{a,\mathcal{A}}(\mathcal{T})$ for some nonempty $\mathcal{T}$.

Proposition 2 shows that representation stability alone does not guarantee contribution stability because changes in the decision map can induce contribution drift even when modality-subset representations are unchanged. The proof is provided in **Appendix A.4**.

3 METHOD

To control modality contribution drift, we propose Continual Modality Contribution Drift Regularization (CMCDR), which regularizes old-class contribution profiles by comparing the current model with a frozen previous model under matched modality-subset interventions. When exemplars are available, CMCDR matches contribution profiles on replayed old samples. Without exemplars, it uses current-task samples as probes to match old-class contribution profiles between the current and frozen models.

3.1 CONTRIBUTION-PROFILE COMPUTATION

At incremental step $t > 1$, let F_{θ_t}, $F_{\theta_{t-1}}$, and $\mathcal{Y}_{<t}$ be the current model, the frozen previous model, and the old-class set, respectively. For any sample $\mathbf{x}_i$, stage $s \in \{t-1, t\}$, and old-class output index $k \in \mathcal{Y}_{<t}$, CMCDR evaluates F_{θ_s} using the interventions in Eq. 1. It then computes the contribution profile $\mathbf{C}_{i,s}^{k}$ over all nonempty modality coalitions by setting $(a, \mathcal{A}) = (k, \mathcal{Y}_{<t})$ in Eqs. 2–3. For replayed old samples, $k = y_i$. In the replay-free case, k ranges over all old-class outputs in $\mathcal{Y}_{<t}$.

CMCDR compares current and frozen profiles with

$$d_{\mathrm{cp}}(\mathbf{C}, \mathbf{C}^{-}) = \rho_\delta\left(\mathbf{C}, \mathrm{sg}(\mathbf{C}^{-})\right) + \beta\, \rho_\delta\left(\frac{\mathbf{C}}{\|\mathbf{C}\|_1 + \epsilon}, \mathrm{sg}\left(\frac{\mathbf{C}^{-}}{\|\mathbf{C}^{-}\|_1 + \epsilon}\right)\right), \tag{11}$$

where ρ_δ is the mean Smooth-L_1 loss with transition $\delta > 0$ defined in Eq. 29, $\mathrm{sg}(\cdot)$ stops gradients through the frozen model, $\epsilon > 0$, and $\beta > 0$ balances absolute and relative profile matching. The first term matches the raw contribution vectors, while the second matches their normalized versions to preserve relative modality and interaction patterns.

3.2 REPLAY-BASED CONTRIBUTION CONSTRAINT

When old exemplars are available, CMCDR applies the above computation to replayed samples $\mathcal{X}_{\mathrm{old}} \subseteq \mathcal{E}_{<t}$, where $\mathcal{E}_{<t}$ denotes the exemplar buffer from previous tasks. For each $(\mathbf{x}_i, y_i) \in \mathcal{X}_{\mathrm{old}}$, the profile index is set to the ground-truth old label $k = y_i$.

The frozen previous model provides $\mathbf{C}_{i,t-1}^{y_i}$, and the current model produces $\mathbf{C}_{i,t}^{y_i}$. The replay-based objective is

$$\mathcal{L}_{\mathrm{cmcdr}}^{\mathrm{rep}} = \frac{1}{|\mathcal{X}_{\mathrm{old}}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{X}_{\mathrm{old}}} d_{\mathrm{cp}}\left(\mathbf{C}_{i,t}^{y_i}, \mathbf{C}_{i,t-1}^{y_i}\right). \tag{12}$$

Thus, replay-based CMCDR directly constrains old-sample contribution drift at both the modality-specific and interaction levels.

3.3 REPLAY-FREE CONTRIBUTION CONSTRAINT

In the replay-free setting, no old-class samples are available. CMCDR uses the $N_t = |\mathcal{X}_{\mathrm{new}}|$ current-task samples as probes (Asadi et al., 2023; Li & Hoiem, 2018) to compare the current and frozen models under the same modality-subset interventions. For each probe $\mathbf{x}_i$, the above computation yields $\mathbf{C}_{i,t}^{k}$ and $\mathbf{C}_{i,t-1}^{k}$ for every $k \in \mathcal{Y}_{<t}$.

CMCDR matches the contribution profile of each old-class output to its frozen counterpart. The replay-free objective is

$$\mathcal{L}_{\mathrm{cmcdr}}^{\mathrm{free}} = \frac{1}{N_t |\mathcal{Y}_{<t}|} \sum_{\mathbf{x}_i \in \mathcal{X}_{\mathrm{new}}} \sum_{k \in \mathcal{Y}_{<t}} d_{\mathrm{cp}}\left(\mathbf{C}_{i,t}^{k}, \mathbf{C}_{i,t-1}^{k}\right). \tag{13}$$

3.4 OVERALL TRAINING OBJECTIVE

At incremental step t, CMCDR is integrated with the objective of the underlying continual learning method as

$$\mathcal{L}_t = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{cmcdr}} \mathcal{L}_{\mathrm{cmcdr}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{base}}$ denotes the original objective of the underlying continual learning method, $\mathcal{L}_{\mathrm{cmcdr}}$ denotes the corresponding replay-based or replay-free CMCDR objective, and $\lambda_{\mathrm{cmcdr}} > 0$ controls the regularization strength.

Theoretical analysis. Proposition 1 shows how changes in modality contributions can lead to forgetting. Proposition 2 establishes that representation stability does not guarantee contribution stability. **Appendix A.5** provides theoretical guarantees that CMCDR mitigates MCD and thereby reduces forgetting.

4 EXPERIMENTS

We evaluate the proposed CMCDR on representative MMCL benchmarks by addressing three questions:

- **Q1**. Does CMCDR improve the performance of existing MMCL methods?
- **Q2**. Does CMCDR provide complementary gains to representation-preserving MMCL methods?
- **Q3**. Does CMCDR reduce modality contribution drift and alleviate forgetting?

4.1 EXPERIMENTAL SETUPS

Benchmarks. We evaluate CMCDR under two representative MMCL protocols: class-incremental recognition and continual multimodal question answering. For class-incremental learning, we use AVE (Tian et al., 2018), Kinetics-Sounds (Arandjelović & Zisserman, 2017), and UESTC-MMEA-CL (Xu et al., 2024). For continual QA, we evaluate on VQAv2 (Goyal et al., 2017) under the VQACL protocol (Zhang et al., 2023), and on Split-AVQA (Yang et al., 2022) and Split-MUSIC-AVQA (Li et al., 2022) under the AVQACL protocol (Wu et al., 2025). For VQACL and AVQACL, we instantiate CMCDR by constructing contribution profiles from answer scores, with implementation details provided in **Appendix B.3**.

Baselines. We compare CMCDR with representative continual learning and MMCL methods for audio-visual classification, continual visual question answering, and audio-visual question answering (Li & Hoiem, 2018; Kirkpatrick et al., 2017; Rebuffi et al., 2017; Pian et al., 2023; Mo et al., 2023; Yue et al., 2024; Zhang et al., 2023; Qian et al., 2023; Wu et al., 2025). We incorporate CMCDR into each baseline as an additional regularizer for preserving modality contributions. CMCDR uses stored exemplars when available. Otherwise, it uses the outputs of the frozen model on current-task probes. Apart from the additional CMCDR regularization term, all baselines retain their original architectures, loss functions, and evaluation protocols. Implementation details are provided in **Appendix B.2**. In addition, **Appendix D** discusses the distinction between CMCDR and modality-balancing regularization methods.

Evaluation. We report the replay memory budget (#Mem, measured as the total number of stored exemplars), average accuracy (Avg. Acc), and average forgetting (Avg. Forget). For contribution analysis, we additionally report the MCD score (MCD; lower is better). Values in parentheses indicate the changes obtained after adding CMCDR.

Table 1: Plug-and-play results on multimodal class-incremental learning.

Method	AVE			Kinetics-Sounds			UESTC-MMEA-CL		
	#Mem	Avg. Acc. ↑	Avg. Forget. ↓	#Mem	Avg. Acc. ↑	Avg. Forget. ↓	#Mem	Avg. Acc. ↑	Avg. Forget. ↓
LwF	None	56.61	35.11	None	65.54	16.55	None	17.13	49.37
LwF + CMCDR	None	**60.61** (+4.00)	**26.35** (-8.76)	None	**69.12** (+3.58)	**12.64** (-3.91)	None	**22.85** (+5.72)	**38.65** (-10.72)
EWC	None	41.28	68.97	None	58.42	31.84	None	19.57	29.84
EWC + CMCDR	None	**47.36** (+6.08)	**57.42** (-11.55)	None	**62.73** (+4.31)	**24.96** (-6.88)	None	**24.78** (+5.21)	**22.06** (-7.78)
iCaRL	340	64.20	25.60	500	65.54	40.57	320	71.83	22.86
iCaRL + CMCDR	340	**71.86** (+7.66)	**17.36** (-8.24)	500	**71.29** (+5.75)	**34.29** (-6.28)	320	**75.12** (+3.29)	**17.15** (-5.71)
AV-CIL	340	71.37	6.77	500	73.06	6.48	320	80.27	18.74
AV-CIL + CMCDR	340	**75.86** (+4.49)	**5.31** (-1.46)	500	**76.14** (+3.08)	**4.55** (-1.93)	320	**83.68** (+3.41)	**13.36** (-5.38)
CIGN	340	73.92	10.84	500	74.35	8.76	320	82.14	16.27
CIGN + CMCDR	340	**74.85** (+0.93)	**7.11** (-3.73)	500	**75.23** (+0.88)	**6.02** (-2.74)	320	**83.75** (+1.61)	**12.97** (-3.30)
MMAL	None	70.28	19.64	None	71.46	13.72	None	75.57	24.83
MMAL + CMCDR	None	**72.63** (+2.35)	**15.02** (-4.62)	None	**74.16** (+2.70)	**9.86** (-3.86)	None	**78.12** (+2.55)	**20.94** (-3.89)
Upper-bound	–	77.25	–	–	80.43	–	–	90.12	–

4.2 RESULTS

Plug-and-Play Effectiveness (Q1). We evaluate the plug-and-play effectiveness of CMCDR across conventional CL baselines and specialized MMCL methods using the integration procedures described in **Appendix B**. Tables 1 and 2 show consistent improvements in two distinct MMCL settings: multimodal class-incremental learning and continual multimodal question answering. Across both replay-based and replay-free methods, CMCDR improves average accuracy by 0.88∼7.66 points and reduces average forgetting by 0.85∼11.55 points. These improvements are consistently observed for conventional CL baselines, including LwF, EWC, and iCaRL, as well as specialized MMCL methods, including AV-CIL, CIGN, MMAL, VQACL, TRIPLET, and AVQACL.

Table 2: Plug-and-play results on continual multimodal question answering.

Method	VQAv2			Split-AVQA			Split-MUSIC-AVQA		
	#Mem	Avg. Acc. ↑	Avg. Forget. ↓	#Mem	Avg. Acc. ↑	Avg. Forget. ↓	#Mem	Avg. Acc. ↑	Avg. Forget. ↓
LwF	None	17.42	26.85	None	21.53	30.49	None	30.93	30.14
LwF + CMCDR	None	**22.35** (+4.93)	**21.01** (-5.84)	None	**25.81** (+4.28)	**24.11** (-6.38)	None	**35.14** (+4.21)	**24.88** (-5.26)
EWC	None	15.77	30.62	None	18.89	40.69	None	28.76	32.58
EWC + CMCDR	None	**19.15** (+3.38)	**25.03** (-5.59)	None	**22.38** (+3.49)	**33.27** (-7.42)	None	**32.67** (+3.91)	**26.41** (-6.17)
iCaRL	5000	30.62	12.48	5000	25.28	33.76	700	28.46	33.16
iCaRL + CMCDR	5000	**36.64** (+6.02)	**8.62** (-3.86)	5000	**29.31** (+4.03)	**27.45** (-6.31)	700	**32.33** (+3.87)	**27.12** (-6.04)
VQACL	5000	38.77	3.96	5000	28.91	11.76	700	31.08	28.64
VQACL + CMCDR	5000	**41.36** (+2.59)	**2.71** (-1.25)	5000	**32.47** (+3.56)	**8.34** (-3.42)	700	**33.98** (+2.90)	**23.47** (-5.17)
TRIPLET	None	34.96	7.24	None	26.72	16.83	None	31.47	23.51
TRIPLET + CMCDR	None	**38.21** (+3.25)	**4.86** (-2.38)	None	**29.86** (+3.14)	**14.26** (-2.57)	None	**34.59** (+3.12)	**20.07** (-3.44)
AVQACL	5000	36.12	6.85	5000	32.05	2.47	700	33.64	27.08
AVQACL + CMCDR	5000	**39.44** (+3.32)	**4.31** (-2.54)	5000	**34.18** (+2.13)	**1.62** (-0.85)	700	**36.12** (+2.48)	**24.95** (-2.13)
Upper-bound	–	51.64	–	–	55.42	–	–	65.57	–

Complementarity with Representation Preservation (Q2). We evaluate whether CMCDR complements representation-preservation methods. For both LwF and iCaRL, we compare the baseline, each regularizer alone, and their combinations under the same training protocol.

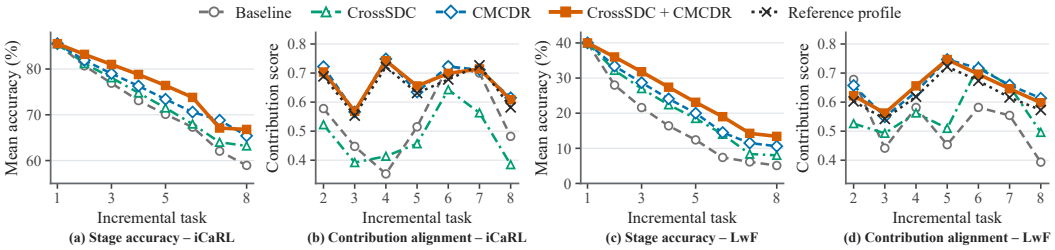


Figure 5: Complementarity of representation and contribution preservation on UESTC-MMEA-CL. Panels (a,b) use replay-based iCaRL, while (c,d) use replay-free LwF. Panels (b,d) show the visual-modality contribution score for each task.

Table 3: Complementarity with representation-preservation methods built on continual-learning baselines.

Method	AVE			UESTC-MMEA-CL		
	Avg. Acc. ↑	Avg. Forget. ↓	MCD ↓	Avg. Acc. ↑	Avg. Forget. ↓	MCD ↓
LwF	56.61	35.11	0.842	17.13	49.37	0.513
LwF + CMCDR	**60.61** (+4.00)	**26.35** (-8.76)	**0.401** (-0.441)	**22.85** (+5.72)	**38.65** (-10.72)	**0.299** (-0.214)
LwF + RepAlign	61.84	29.06	0.674	20.59	43.22	0.450
LwF + RepAlign + CMCDR	**65.22** (+3.38)	**23.41** (-5.65)	**0.251** (-0.423)	**24.88** (+4.29)	**35.13** (-8.09)	**0.286** (-0.164)
LwF + CrossSDC	60.73	30.18	0.641	21.31	41.79	0.436
LwF + CrossSDC + CMCDR	**64.04** (+3.31)	**24.72** (-5.46)	**0.278** (-0.363)	**25.62** (+4.31)	**33.87** (-7.92)	**0.243** (-0.193)
iCaRL	64.20	25.60	0.779	71.83	22.86	0.464
iCaRL + CMCDR	**71.86** (+7.66)	**17.36** (-8.24)	**0.389** (-0.390)	**75.12** (+3.29)	**17.15** (-5.71)	**0.288** (-0.176)
iCaRL + RepAlign	70.30	19.20	0.625	73.01	20.12	0.431
iCaRL + RepAlign + CMCDR	**72.65** (+2.35)	**16.39** (-2.81)	**0.203** (-0.422)	**76.96** (+3.95)	**16.77** (-3.35)	**0.235** (-0.196)
iCaRL + CrossSDC	67.80	22.50	0.699	73.29	19.21	0.419
iCaRL + CrossSDC + CMCDR	**72.89** (+5.09)	**19.96** (-2.54)	**0.214** (-0.485)	**76.69** (+3.40)	**15.38** (-3.83)	**0.186** (-0.233)

As summarized in Table 3 and Figure 5, adding CMCDR to RepAlign or CrossSDC improves accuracy by 2.35∼5.09 points, reduces forgetting by 2.54∼8.09 points, and lowers MCD scores by 0.164∼0.485. Compared with CrossSDC alone, the CrossSDC + CMCDR variants achieve higher accuracy in Figure 5(a,c) and track the reference contribution profile more closely in Figure 5(b,d). These results demonstrate that CMCDR complements representation preservation by directly constraining residual contribution drift.

MCD and Forgetting Analysis (Q3). To examine the relationship between contribution drift and forgetting across old classes, we compare classes with different baseline MCD scores. Figure 6(a) ranks old classes according to their baseline MCD scores. CMCDR consistently reduces MCD across classes, with larger reductions for classes with higher baseline drift. Figure 6(b) further groups old classes into low- and high-MCD groups and compares their accuracy across incremental tasks. Under the baseline, high-MCD classes exhibit larger accuracy decreases, whereas CMCDR reduces contribution drift, improves their classification accuracy, and narrows the performance gap between the two groups. **Appendix C.2** further analyzes the paired changes in MCD and class-wise forgetting on AVE under iCaRL and LwF.

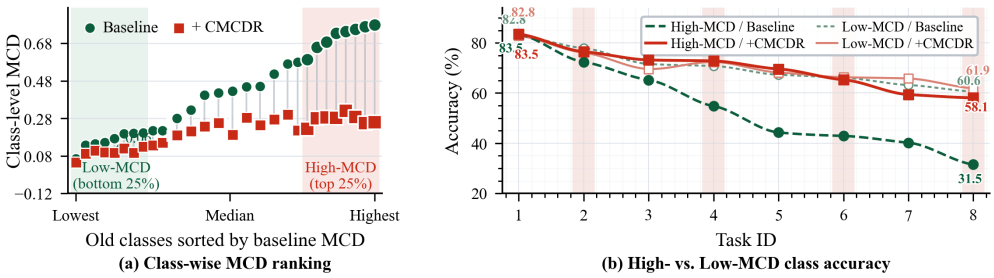


Figure 6: Class-wise effect of CMCDR on UESTC-MMEA-CL with replay-based iCaRL.

5 Limitations and Future Work

Task Similarity. CMCDR constrains deviations from previous contribution profiles and may therefore restrict contribution changes that benefit old-task predictions. When tasks share similar modality contribution patterns, changes induced by learning new tasks may also improve old-task predictions, as discussed in **Appendix C.2**. Future work will investigate how similarity in modality contribution patterns affects contribution shifts and develop adaptive regularization that preserves old-task contributions while allowing beneficial changes.

Computational Cost. Computing exact contribution profiles with CMCDR requires $2^{|\mathcal{M}|}$ modality-subset evaluations. The computational cost therefore increases exponentially with the number of modalities (**Appendix C.3**). Future work will explore sampled or adaptive coalition selection to estimate contribution profiles more efficiently.

6 Conclusion

In this work, we study modality contribution drift, an underexplored problem in MMCL. An intervention-based contribution estimator and an MCD score are introduced to quantify changes in modality contributions across incremental stages. Our theoretical and empirical analyses show how contribution drift can lead to forgetting and further show that existing MMCL and conventional CL methods do not effectively mitigate MCD. To address this problem, we propose CMCDR to preserve old-task contribution profiles in both replay-based and replay-free settings. CMCDR aligns contribution profiles between the current and frozen previous models using stored old samples or current-task samples as probes for old-class outputs, respectively. Experiments on six benchmarks spanning multimodal class-incremental learning and continual multimodal QA show that CMCDR consistently reduces MCD, improves average accuracy, and reduces forgetting. These results demonstrate that CMCDR effectively complements representation preservation in MMCL.

AI USE STATEMENT

We used generative AI tools (ChatGPT and Codex) to help improve the readability and grammar of parts of the manuscript, draft and edit implementation code, and design and refine the experimental protocol, including baselines and ablation studies.

All AI-assisted text and code were carefully reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of this submission, including all text, claims, code, and experimental results.

REPRODUCIBILITY STATEMENT

Section 2 and Section 3 specify the contribution estimator, MCD scores, and CMCDR objectives. Appendix A provides the theoretical proofs, and Appendix B describes task construction, modality-removal rules, baseline implementations, and training settings.

REFERENCES

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

Nader Asadi, Mohammadreza Davari, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Prototype-sample relation distillation: Towards replay-free continual learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1093–1106, 2023.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, 2020.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19573–19597, 2024.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in CLIP-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3777–3786, 2025.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28708–28720, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19108–19118, 2022.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22214–22224, 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393, 2023.

Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7788–7798, 2023.

Malvina Nikandrou, Georgios Pantazopoulos, Ioannis Konstas, and Alessandro Suglia. Enhancing continual learning in visual question answering with modality-aware feature distillation. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 73–85, 2024.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, 2022.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7799–7811, 2023.

Weiguo Pian, Yiyang Nan, Shijian Deng, Shentong Mo, Yunhui Guo, and Yapeng Tian. Continual audio-visual sound separation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 76058–76079, 2024.

Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2953–2962, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 247–263, 2018.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. CMoB: Modality valuation via causal effect for balanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82585–82607, 2025.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52559–52572, 2024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27338–27347, 2024.

Kaixuan Wu, Xinde Li, Xinling Li, Chuanfei Hu, and Guoliang Wu. AVQACL: A novel benchmark for audio-visual question answering continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3252–3261, 2025.

Linfeng Xu, Qingbo Wu, Lili Pan, Fanman Meng, Hongliang Li, Chiyuan He, Hanxin Wang, Shaoxu Cheng, and Yu Dai. Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning. *IEEE Transactions on Multimedia*, 26:2430–2443, 2024.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pp. 3480–3491, 2022.

Xianghu Yue, Xueyi Zhang, Yiming Chen, Chengwei Zhang, Mingrui Lao, Huiping Zhuang, Xinyuan Qian, and Haizhou Li. MMAL: Multi-modal analytic learning for exemplar-free audio-visual class incremental tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pp. 2428–2437, 2024.

Xi Zhang, Feifei Zhang, and Changsheng Xu. VQACL: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19102–19112, 2023.

APPENDIX: TABLE OF CONTENTS

A Theoretical Analysis and Proofs 13
A.1 Interventional Contribution Estimator 13
A.2 Contribution Drift and Forgetting 14
A.3 Fixed-Reference Drift and Later-Class Competition 15
A.4 Why Representation Stability Is Insufficient 15
A.5 Contribution-Drift Bounds for CMCDR 16
A.6 Möbius Matching vs. Centered-Response Matching 18

B Experimental Settings and Implementation Details 19
B.1 Experimental Settings for Exploratory Analyses 19
B.2 Main Training and Evaluation Setup 20
B.3 CMCDR for Continual Multimodal Question Answering 21
B.4 Modality-Removal Realizations 21

C Additional Experiments 23
C.1 Random-Seed Robustness 23
C.2 Modality Contribution Drift and Class-wise Forgetting 23
C.3 Computational Efficiency and Scalability 24
C.4 Hyperparameter Sensitivity Analysis 26
C.5 Replay-Free Probe Transfer 27
C.6 Ablation and Controlled Comparisons 27
C.7 Robustness to Alternative Contribution Estimators 29

D Comparison with Related Methods 29
D.1 Multimodal Continual Learning 30
D.2 Balanced Multimodal Learning 30
D.3 Comparison with Modality-Balancing Methods 30

A THEORETICAL ANALYSIS AND PROOFS

A.1 INTERVENTIONAL CONTRIBUTION ESTIMATOR

For a sample and model, we use the scalar benefit in Eq. 2 to compare retaining a modality subset $\mathcal{S}$ with the no-modality baseline $\emptyset$. For logits restricted to $\mathcal{A}$, define

$$p_{i,a}^{\mathcal{A}}(\mathcal{S}) = \frac{\exp(z_{i,a}(\mathcal{S}))}{\sum_{\ell \in \mathcal{A}} \exp(z_{i,\ell}(\mathcal{S}))}. \tag{15}$$

Here $z_{i,\ell}(\mathcal{S})$ denotes the logit for candidate class $\ell \in \mathcal{A}$. Substituting the softmax probability into Eq. 2 gives

$$B_i^{a,\mathcal{A}}(\mathcal{S}) = \log \frac{p_{i,a}^{\mathcal{A}}(\mathcal{S})}{1 - p_{i,a}^{\mathcal{A}}(\mathcal{S})}. \tag{16}$$

Thus $B_i^{a,\mathcal{A}}(\mathcal{S})$ is the one-vs-rest log-odds of target class a. The contrast $B_i^{a,\mathcal{A}}(\mathcal{S}) - B_i^{a,\mathcal{A}}(\emptyset)$ is the log-odds ratio relative to the no-modality baseline, quantifying the total decision-level effect of retaining $\mathcal{S}$. It is unchanged if all logits in $\mathcal{A}$ are shifted by the same constant, and it avoids the saturation of probabilities by working in log-odds space. Let $\gamma_i^{a,\mathcal{A}}(\mathcal{S}) = z_{i,a}(\mathcal{S}) - \max_{\ell \in \mathcal{A},\, \ell \neq a} z_{i,\ell}(\mathcal{S})$ denote the target-class logit margin. Then

$$\gamma_i^{a,\mathcal{A}}(\mathcal{S}) - \log(|\mathcal{A}| - 1) \leq B_i^{a,\mathcal{A}}(\mathcal{S}) \leq \gamma_i^{a,\mathcal{A}}(\mathcal{S}). \tag{17}$$

The benefit is therefore a smooth multiclass margin that compares the target logit with the log-sum-exp of all competing logits in $\mathcal{A}$.

Each $B_i^{a,\mathcal{A}}(\mathcal{S})$ is evaluated by one controlled forward pass as in Eq. 1, using the modality-removal rule in Appendix B.4.

Finally, we use the Möbius coefficients in Eq. 3 to separate singleton effects relative to the no-modality baseline from higher-order residual effects not explained by lower-order subsets. For example, in the two-modality case,

$$C_i^{a,\mathcal{A}}(\{m,n\}) = B_i^{a,\mathcal{A}}(\{m,n\}) - B_i^{a,\mathcal{A}}(\{m\}) - B_i^{a,\mathcal{A}}(\{n\}) + B_i^{a,\mathcal{A}}(\emptyset), \tag{18}$$

which isolates the interaction beyond the two singleton effects. By Möbius inversion,

$$B_i^{a,\mathcal{A}}(\mathcal{S}) = \sum_{\mathcal{T}\subseteq\mathcal{S}} C_i^{a,\mathcal{A}}(\mathcal{T}), \qquad C_i^{a,\mathcal{A}}(\emptyset) = B_i^{a,\mathcal{A}}(\emptyset). \tag{19}$$

Setting $\mathcal{S} = \mathcal{M}$ and subtracting the empty-set term yields the completeness identity in Eq. 4.

A.2 Contribution Drift and Forgetting

Proposition 1. *Contribution drift can cross the retention margin.* For an old sample i, let $\tau(y_i) \le r < t$, fix a candidate set $\mathcal{A}_i \subseteq \mathcal{Y}_{\le r}$ containing y_i with $|\mathcal{A}_i| \ge 2$, and use the same modality-removal rule at stages r and t. All benefits and contribution profiles use target y_i and candidate set $\mathcal{A}_i$. Define $\bar{B}_{i,s}(\mathcal{S}) = B_{i,s}(\mathcal{S}) - B_{i,s}(\emptyset)$ and

$$m^{\text{keep}}_{i,r\to t} = B_{i,t}(\emptyset) + \bar{B}_{i,r}(\mathcal{M}).$$

Let $\mathbf{c}_{i,s}$ denote the contribution vector at stage $s \in \{r,t\}$ under $(y_i, \mathcal{A}_i)$. Let $\gamma_{i,t}$ denote the target-class logit margin over its strongest competitor in $\mathcal{A}_i$ at stage t. Then

$$B_{i,t}(\mathcal{M}) = m^{\text{keep}}_{i,r\to t} + \mathbf{1}^\top(\mathbf{c}_{i,t} - \mathbf{c}_{i,r}),$$
$$\gamma_{i,t} \ge m^{\text{keep}}_{i,r\to t} - \|\mathbf{c}_{i,t} - \mathbf{c}_{i,r}\|_1.$$

If $\|\mathbf{c}_{i,t} - \mathbf{c}_{i,r}\|_1 < m^{\text{keep}}_{i,r\to t}$, then y_i has a strictly higher logit than every other class in $\mathcal{A}_i$. Conversely, if $m^{\text{keep}}_{i,r\to t} > 0$ and $\gamma_{i,t} \le 0$, then $\|\mathbf{c}_{i,t} - \mathbf{c}_{i,r}\|_1 \ge m^{\text{keep}}_{i,r\to t}$.

Proof. Let $\mathbf{c}_s$ be the contribution vector at stage $s \in \{r,t\}$ under $(y_i, \mathcal{A}_i)$. By Möbius inversion,

$$B_{i,t}(\mathcal{M}) = m^{\text{keep}}_{i,r\to t} + \mathbf{1}^\top(\mathbf{c}_t - \mathbf{c}_r). \tag{20}$$

Equation 17 and the triangle inequality give

$$\gamma_{i,t} \ge B_{i,t}(\mathcal{M}) \ge m^{\text{keep}}_{i,r\to t} - \|\mathbf{c}_t - \mathbf{c}_r\|_1. \tag{21}$$

If $\|\mathbf{c}_t - \mathbf{c}_r\|_1 < m^{\text{keep}}_{i,r\to t}$, then $\gamma_{i,t} > 0$. Conversely, $\gamma_{i,t} \le 0$ implies $\|\mathbf{c}_t - \mathbf{c}_r\|_1 \ge m^{\text{keep}}_{i,r\to t}$. □

The retained margin equals the reference full-input benefit plus the change in the no-modality response:

$$m^{\text{keep}}_{i,r\to t} = B_{i,r}(\mathcal{M}) + B_{i,t}(\emptyset) - B_{i,r}(\emptyset).$$

A prediction can flip even when the modality-subset representations remain unchanged. Consider two modalities $\mathcal{M} = \{1,2\}$ and two classes $\mathcal{A} = \{a,b\}$. Let $x = (1,1)$ and use zero-valued modality removal. For this fixed sample, we omit the sample index and keep the subset representations unchanged across stages:

$$u_r(\mathcal{S}) = u_t(\mathcal{S}) = x^{\mathcal{S}}, \qquad z_{s,a}(x^{\mathcal{S}}) = w_s^\top x^{\mathcal{S}}, \qquad z_{s,b}(x^{\mathcal{S}}) = 0.$$

For any $\gamma, \eta > 0$, choose

$$w_r = \left(\frac{\gamma}{2}, \frac{\gamma}{2}\right), \qquad w_t = \left(\frac{\gamma}{2}, -\frac{\gamma}{2} - \eta\right).$$

Since $\mathcal{A}$ is binary, the benefit equals the target-class logit margin. The resulting benefits and contribution profiles are

$$B_r(\emptyset) = B_t(\emptyset) = 0, \qquad B_r(\mathcal{M}) = \gamma, \qquad B_t(\mathcal{M}) = -\eta,$$
$$\mathbf{c}_r = \left(\frac{\gamma}{2}, \frac{\gamma}{2}, 0\right), \qquad \mathbf{c}_t = \left(\frac{\gamma}{2}, -\frac{\gamma}{2} - \eta, 0\right),$$

where $\mathbf{c}_s = (C_s(\{1\}), C_s(\{2\}), C_s(\{1,2\}))$. Here $m^{\text{keep}}_{r\to t} = \gamma$. The binary margin changes from $\gamma > 0$ to $-\eta < 0$, turning a correct prediction into an incorrect one while preserving the modality-subset representations and the no-modality response.

A.3 Fixed-Reference Drift and Later-Class Competition

For an old sample i with target $a = y_i$, let $\mathcal{A}_i$ denote its candidate set and $\mathcal{N}_t = \mathcal{Y}_{\le t} \setminus \mathcal{A}_i$ denote the later classes. Define

$$Z_{\mathcal{A}_i,t}(\mathcal{S}) = \sum_{\ell \in \mathcal{A}_i \setminus \{a\}} \exp z_{i,t,\ell}(\mathcal{S}), \qquad Z_{\mathcal{N}_t,t}(\mathcal{S}) = \sum_{\ell \in \mathcal{N}_t} \exp z_{i,t,\ell}(\mathcal{S}).$$

For the centered benefit $\bar{B}(\mathcal{S}) = B(\mathcal{S}) - B(\emptyset)$,

$$\bar{B}_{i,t}^{a,\mathcal{A}_i \cup \mathcal{N}_t}(\mathcal{S}) = \bar{B}_{i,t}^{a,\mathcal{A}_i}(\mathcal{S}) - \left[q_{i,t}(\mathcal{S}) - q_{i,t}(\emptyset)\right], \qquad q_{i,t}(\mathcal{S}) = \log\left(1 + \frac{Z_{\mathcal{N}_t,t}(\mathcal{S})}{Z_{\mathcal{A}_i,t}(\mathcal{S})}\right). \tag{22}$$

Thus, later classes change the coalition values whenever their relative logit mass varies across modality subsets. A stage-dependent candidate set would combine contribution changes within $\mathcal{A}_i$ with direct competition from newly introduced classes.

MCD measures cumulative contribution drift from class introduction under the candidate set $\mathcal{A}_y = \mathcal{Y}_{\le \tau(y)}$. Fixing $\mathcal{A}_y$ excludes changes caused solely by candidate-set expansion and keeps the competing classes unchanged across stages and methods.

In class-incremental learning, the model must continue to distinguish previously learned classes as new classes arrive. At update t, CMCDR therefore matches contributions under $\mathcal{Y}_{<t}$ to cover all previously learned classes, including those introduced after y. This set expands across updates but is shared by both models and all modality-subset interventions within each update.

Eq. 22 also yields a sufficient condition for prediction retention over the full class space. With $m_{i,r\to t}^{\text{keep}}$ and the contribution vectors computed under the same fixed $\mathcal{A}_i$ as in **Proposition 1**, Eq. 21 gives

$$B_{i,t}^{y_i,\mathcal{A}_i \cup \mathcal{N}_t}(\mathcal{M}) \ge m_{i,r\to t}^{\text{keep}} - \|\mathbf{c}_{i,t} - \mathbf{c}_{i,r}\|_1 - q_{i,t}(\mathcal{M}). \tag{23}$$

Thus, $\|\mathbf{c}_{i,t} - \mathbf{c}_{i,r}\|_1 + q_{i,t}(\mathcal{M}) < m_{i,r\to t}^{\text{keep}}$ guarantees that, at stage t, y_i has a strictly higher logit than every other class in $\mathcal{Y}_{\le t}$.

A.4 Why Representation Stability Is Insufficient

Proposition 2. *Representation stability does not guarantee contribution stability.* For a sample i with target old class a and stages $\tau(a) \le r < t$, fix $\mathcal{A} = \mathcal{A}_a$. Let $\mathbf{u}_{i,s}(\mathcal{S})$ denote the modality-subset representation at stage s in a common normed space, and let $g_s^{a,\mathcal{A}}$ denote the corresponding target-vs-rest margin map, such that $B_{i,s}^{a,\mathcal{A}}(\mathcal{S}) = g_s^{a,\mathcal{A}}(\mathbf{u}_{i,s}(\mathcal{S}))$. Define

$$\Delta_u(\mathcal{S}) = \|\mathbf{u}_{i,t}(\mathcal{S}) - \mathbf{u}_{i,r}(\mathcal{S})\|, \qquad \Delta_g(\mathcal{S}) = \left|g_t^{a,\mathcal{A}}(\mathbf{u}_{i,r}(\mathcal{S})) - g_r^{a,\mathcal{A}}(\mathbf{u}_{i,r}(\mathcal{S}))\right|.$$

If $g_t^{a,\mathcal{A}}$ is L_g-Lipschitz with respect to this norm, then for any nonempty $\mathcal{T} \subseteq \mathcal{M}$,

$$\left|C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) - C_{i,r}^{a,\mathcal{A}}(\mathcal{T})\right| \le \sum_{\mathcal{S} \subseteq \mathcal{T}} \left(L_g \Delta_u(\mathcal{S}) + \Delta_g(\mathcal{S})\right).$$

Let $\bar{B}_{i,s}^{a,\mathcal{A}}(\mathcal{S}) = B_{i,s}^{a,\mathcal{A}}(\mathcal{S}) - B_{i,s}^{a,\mathcal{A}}(\emptyset)$. If every nonempty $\mathcal{S} \subseteq \mathcal{T}$ satisfies $|\bar{B}_{i,t}^{a,\mathcal{A}}(\mathcal{S}) - \bar{B}_{i,r}^{a,\mathcal{A}}(\mathcal{S})| \le \epsilon_{\bar{B}}$, then $|C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) - C_{i,r}^{a,\mathcal{A}}(\mathcal{T})| \le (2^{|\mathcal{T}|} - 1)\epsilon_{\bar{B}}$. Moreover, there exist identical modality-subset representations at stages r and t and two decision maps such that $\Delta_u(\mathcal{S}) = 0$ for every $\mathcal{S}$, while $C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) \ne C_{i,r}^{a,\mathcal{A}}(\mathcal{T})$ for some nonempty $\mathcal{T}$.

Proof. Fix the quantities in Proposition 2. For any $\mathcal{S} \subseteq \mathcal{T}$, adding and subtracting $g_t^{a,\mathcal{A}}(\mathbf{u}_{i,r}(\mathcal{S}))$ and applying the triangle inequality and Lipschitz continuity gives

$$\left|B_{i,t}^{a,\mathcal{A}}(\mathcal{S}) - B_{i,r}^{a,\mathcal{A}}(\mathcal{S})\right| \le L_g \Delta_u(\mathcal{S}) + \Delta_g(\mathcal{S}). \tag{24}$$

Applying Eq. 3 and the triangle inequality,

$$\left|C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) - C_{i,r}^{a,\mathcal{A}}(\mathcal{T})\right| \le \sum_{\mathcal{S}\subseteq\mathcal{T}} \left|B_{i,t}^{a,\mathcal{A}}(\mathcal{S}) - B_{i,r}^{a,\mathcal{A}}(\mathcal{S})\right| \le \sum_{\mathcal{S}\subseteq\mathcal{T}} \left(L_g \Delta_u(\mathcal{S}) + \Delta_g(\mathcal{S})\right), \tag{25}$$

which proves Eq. 10.

For any nonempty coalition $\mathcal{T}$, Eq. 3 gives

$$C_{i,s}^{a,\mathcal{A}}(\mathcal{T}) = \sum_{\mathcal{S}\subseteq\mathcal{T}} (-1)^{|\mathcal{T}|-|\mathcal{S}|} \bar{B}_{i,s}^{a,\mathcal{A}}(\mathcal{S}), \tag{26}$$

because $\sum_{\mathcal{S}\subseteq\mathcal{T}}(-1)^{|\mathcal{T}|-|\mathcal{S}|} = (1-1)^{|\mathcal{T}|} = 0$. Under the centered-response condition in Proposition 2, Eq. 26 and the triangle inequality give

$$\left|C_{i,t}^{a,\mathcal{A}}(\mathcal{T}) - C_{i,r}^{a,\mathcal{A}}(\mathcal{T})\right| \le \sum_{\emptyset\neq\mathcal{S}\subseteq\mathcal{T}} \left|\bar{B}_{i,t}^{a,\mathcal{A}}(\mathcal{S}) - \bar{B}_{i,r}^{a,\mathcal{A}}(\mathcal{S})\right| \le (2^{|\mathcal{T}|}-1)\epsilon_{\bar{B}}. \tag{27}$$

To show that representation stability need not imply contribution stability, consider two modalities and a binary linear classifier operating on fixed post-fusion features. Suppressing the fixed superscript $(a, \mathcal{A})$, set $B_{i,s}(\mathcal{S}) = w_s^\top \mathbf{u}_i(\mathcal{S})$, with $\mathbf{u}_{i,t}(\mathcal{S}) = \mathbf{u}_{i,r}(\mathcal{S}) = \mathbf{u}_i(\mathcal{S})$ for every $\mathcal{S}$. Set $\mathbf{u}_i(\emptyset) = 0$, $\mathbf{u}_i(\{1\}) = (1,0)^\top$, $\mathbf{u}_i(\{2\}) = (0,1)^\top$, and $\mathbf{u}_i(\{1,2\}) = (1,1)^\top$. Choose $w_r = (0,0)^\top$ and $w_t = (1,-1)^\top$. Then $C_{i,r}(\{1\}) = C_{i,r}(\{2\}) = 0$, but $C_{i,t}(\{1\}) = 1$ and $C_{i,t}(\{2\}) = -1$, while $\bar{B}_{i,r}(\{1,2\}) = \bar{B}_{i,t}(\{1,2\}) = 0$. Thus, representation stability does not imply contribution stability. □

A.5 Contribution-Drift Bounds for CMCDR

CMCDR and contribution drift. Fix a sample i, a target class $a \in \mathcal{A}$ with $|\mathcal{A}| \ge 2$, and stages $r < t$. Use the same candidate set, modality-removal rule, and coalition order at both stages. Let $D = 2^{|\mathcal{M}|} - 1$ and $\epsilon > 0$. Write

$$\mathbf{c}_s = \left[C_{i,s}^{a,\mathcal{A}}(\mathcal{T})\right]_{\emptyset\neq\mathcal{T}\subseteq\mathcal{M}}, \qquad m_s = \|\mathbf{c}_s\|_1, \qquad \widehat{\mathbf{c}}_s = \frac{\mathbf{c}_s}{m_s + \epsilon}, \quad s \in \{r, t\}. \tag{28}$$

The raw and normalized differences are $d_{i,r\to t}^{\mathrm{abs}} = \|\mathbf{c}_t - \mathbf{c}_r\|_1$ and $d_{i,r\to t}^{\mathrm{rel}} = \|\widehat{\mathbf{c}}_t - \widehat{\mathbf{c}}_r\|_1$. The mean Smooth-$L_1$ loss in Eq. 11 is

$$\rho_\delta(\mathbf{u}, \mathbf{v}) = \frac{1}{D}\sum_{j=1}^{D} h_\delta(u_j - v_j), \qquad h_\delta(x) = \begin{cases} x^2/(2\delta), & |x| \le \delta, \\ |x| - \delta/2, & |x| > \delta, \end{cases} \tag{29}$$

where $\delta > 0$. The per-profile loss is

$$\ell_{i,r\to t} = \rho_\delta(\mathbf{c}_t, \mathbf{c}_r) + \beta\rho_\delta(\widehat{\mathbf{c}}_t, \widehat{\mathbf{c}}_r), \qquad \beta > 0. \tag{30}$$

Each training comparison uses $r = t - 1$ and $\mathcal{A} = \mathcal{Y}_{<t}$.

Lemma 3 (CMCDR bounds contribution drift)**.** *Let $\Psi_\delta(u) = \sqrt{2\delta u} + u$ for $u \ge 0$. Then*

$$d_{i,r\to t}^{\mathrm{abs}} \le D\Psi_\delta(\ell_{i,r\to t}), \qquad d_{i,r\to t}^{\mathrm{rel}} \le D\Psi_\delta(\ell_{i,r\to t}/\beta). \tag{31}$$

For any probability weighting π over finitely many profile comparisons with the same D, δ, and β, let $\bar{\ell} = \mathbb{E}_\pi[\ell_{i,r\to t}]$. Then

$$\mathbb{E}_\pi[d_{i,r\to t}^{\mathrm{abs}}] \le D\Psi_\delta(\bar{\ell}), \qquad \mathbb{E}_\pi[d_{i,r\to t}^{\mathrm{rel}}] \le D\Psi_\delta(\bar{\ell}/\beta). \tag{32}$$

Proof. For every $x \in \mathbb{R}$, $|x| \le \sqrt{2\delta h_\delta(x)} + h_\delta(x)$. When $|x| \le \delta$, the square-root term equals $|x|$. When $|x| > \delta$, $h_\delta(x) \ge \delta/2$, so the square-root term is at least δ and the inequality follows. Summing over coordinates and applying Cauchy–Schwarz gives

$$\|\mathbf{u} - \mathbf{v}\|_1 \le D\Psi_\delta(\rho_\delta(\mathbf{u}, \mathbf{v})). \tag{33}$$

The two loss terms satisfy $\rho_\delta(\mathbf{c}_t, \mathbf{c}_r) \le \ell_{i,r\to t}$ and $\rho_\delta(\widehat{\mathbf{c}}_t, \widehat{\mathbf{c}}_r) \le \ell_{i,r\to t}/\beta$. Monotonicity of Ψ_δ proves Eq. 31. Its concavity and Jensen's inequality prove Eq. 32. □

Prediction retention. For a replayed old sample, set $a = y_i$, $r = t-1$, and $\mathcal{A} = \mathcal{Y}_{<t}$. Write $\Gamma(u) = D\Psi_\delta(u)$ and let $\gamma_{i,t}$ be the logit margin over the strongest competitor in $\mathcal{A}$. Proposition 1 and Lemma 3 give

$$\begin{aligned} \gamma_{i,t} &\geq m^{\text{keep}}_{i,t-1\to t} - \Gamma(\ell_{i,t-1\to t}), \\ m^{\text{keep}}_{i,t-1\to t} &= B_{i,t-1}(\mathcal{M}) + B_{i,t}(\emptyset) - B_{i,t-1}(\emptyset). \end{aligned} \tag{34}$$

Thus, $\Gamma(\ell_{i,t-1\to t}) < m^{\text{keep}}_{i,t-1\to t}$ ensures $\gamma_{i,t} > 0$ within $\mathcal{Y}_{<t}$. For prediction over $\mathcal{Y}_{\leq t}$, Eq. 23 gives the sufficient condition

$$\Gamma(\ell_{i,t-1\to t}) + q_{i,t}(\mathcal{M}) < m^{\text{keep}}_{i,t-1\to t}, \tag{35}$$

where $q_{i,t}$ is defined in Eq. 22 with $\mathcal{A}_i = \mathcal{Y}_{<t}$ and $\mathcal{N}_t = \mathcal{Y}_{\leq t} \setminus \mathcal{Y}_{<t}$.

Replay-free transfer. Fix the current and frozen models. Let $\mathcal{D}_t$ be the uniform empirical distribution over $\mathcal{X}_{\text{new}}$ and let $U(\mathcal{Y}_{<t})$ be uniform over old classes. The probe distribution is

$$\nu_t = \mathcal{D}_t \otimes U(\mathcal{Y}_{<t}), \qquad \mathcal{L}^{\text{free}}_{\text{cmcdr}} = \mathbb{E}_{(\mathbf{x},a)\sim\nu_t}[\ell_{t-1\to t}(\mathbf{x}, a)], \tag{36}$$

where $\ell_{t-1\to t}(\mathbf{x}, a)$ is the per-profile loss under $\mathcal{A} = \mathcal{Y}_{<t}$. Let μ_t be the old-task evaluation distribution over input–label pairs and define

$$f_t(\mathbf{x}, a) = \|\mathbf{C}^a_t(\mathbf{x}) - \mathbf{C}^a_{t-1}(\mathbf{x})\|_1. \tag{37}$$

Equip $\mathcal{X} \times \mathcal{Y}_{<t}$ with a fixed metric d_t. Let $W_1^{d_t}$ be the induced Wasserstein-1 distance (Peyré & Cuturi, 2019).

Proposition 4 (Replay-free contribution drift)**.** *Assume that μ_t and ν_t have finite first moments and that f_t is L_t-Lipschitz with respect to d_t. Then*

$$\mathbb{E}_{\mu_t}[f_t] \leq \Gamma\big(\mathcal{L}^{\text{free}}_{\text{cmcdr}}\big) + L_t W_1^{d_t}(\mu_t, \nu_t). \tag{38}$$

Proof. Lemma 3 with $\pi = \nu_t$ gives $\mathbb{E}_{\nu_t}[f_t] \leq \Gamma(\mathcal{L}^{\text{free}}_{\text{cmcdr}})$. For any coupling ζ of μ_t and ν_t,

$$|\mathbb{E}_{\mu_t}[f_t] - \mathbb{E}_{\nu_t}[f_t]| \leq \mathbb{E}_{(Z,Z')\sim\zeta}|f_t(Z) - f_t(Z')| \leq L_t \mathbb{E}_{(Z,Z')\sim\zeta} d_t(Z, Z').$$

Taking the infimum over couplings proves the result. □

Appendix C.5 compares MCD scores on held-out current-task and old-task probes to assess how closely their contribution drift agrees.

Corollary 5 (Replay-free old-class forgetting)**.** *Under the assumptions of Proposition 4, let $\gamma_{i,s}$ be the target-class logit margin over $\mathcal{A} = \mathcal{Y}_{<t}$ at $s \in \{t-1, t\}$ for $(\mathbf{x}_i, y_i) \sim \mu_t$. Define $\mathcal{F}_t = \{\gamma_{i,t-1} > 0,\ \gamma_{i,t} \leq 0\}$. For any $m_0 > 0$,*

$$\Pr_{\mu_t}(\mathcal{F}_t) \leq \Pr_{\mu_t}\Big(m^{\text{keep}}_{i,t-1\to t} \leq m_0\Big) + \frac{\Gamma(\mathcal{L}^{\text{free}}_{\text{cmcdr}}) + L_t W_1^{d_t}(\mu_t, \nu_t)}{m_0}. \tag{39}$$

Proof. On $\mathcal{F}_t \cap \{m^{\text{keep}}_{i,t-1\to t} > m_0\}$, Proposition 1 implies $f_t(\mathbf{x}_i, y_i) > m_0$. The union bound and Markov's inequality give $\Pr_{\mu_t}(\mathcal{F}_t) \leq \Pr_{\mu_t}(m^{\text{keep}}_{i,t-1\to t} \leq m_0) + \mathbb{E}_{\mu_t}[f_t]/m_0$. □

Relation to cumulative MCD. For a fixed sample, target class, candidate set, and stage pair, MCD and raw contribution drift differ by at most a factor of three.

Lemma 6 (MCD and contribution drift)**.** *Fix i, a, $\mathcal{A}$, r, and t as in Eq. 28, and let $d^{\text{MCD}}_{i,r\to t} = |m_t - m_r| + (\min(m_t, m_r) + \epsilon) d^{\text{rel}}_{i,r\to t}$. Then*

$$d^{\text{abs}}_{i,r\to t} \leq d^{\text{MCD}}_{i,r\to t} \leq 3 d^{\text{abs}}_{i,r\to t} \leq 3\Gamma(\ell_{i,r\to t}). \tag{40}$$

The centered responses $\bar{B}^{a,\mathcal{A}}_{i,s}(\mathcal{S}) = B^{a,\mathcal{A}}_{i,s}(\mathcal{S}) - B^{a,\mathcal{A}}_{i,s}(\emptyset)$ satisfy

$$\begin{aligned} \bar{B}^{a,\mathcal{A}}_{i,t}(\mathcal{M}) - \bar{B}^{a,\mathcal{A}}_{i,r}(\mathcal{M}) &= \mathbf{1}^\top(\mathbf{c}_t - \mathbf{c}_r), \\ \max_{\mathcal{S}\subseteq\mathcal{M}} \Big|\bar{B}^{a,\mathcal{A}}_{i,t}(\mathcal{S}) - \bar{B}^{a,\mathcal{A}}_{i,r}(\mathcal{S})\Big| &\leq d^{\text{abs}}_{i,r\to t}. \end{aligned} \tag{41}$$

The profiles are equal if and only if all centered subset responses are equal.

Proof. Assume $m_t \geq m_r$ without loss of generality and let $\Delta m = m_t - m_r$. Since $\|\widehat{\mathbf{c}}_t\|_1 \leq 1$, the identity

$$\mathbf{c}_t - \mathbf{c}_r = (m_r + \epsilon)(\widehat{\mathbf{c}}_t - \widehat{\mathbf{c}}_r) + \Delta m \widehat{\mathbf{c}}_t$$

gives $d^{\mathrm{abs}}_{i,r\to t} \leq d^{\mathrm{MCD}}_{i,r\to t}$. Rearranging the same identity gives $(m_r + \epsilon) d^{\mathrm{rel}}_{i,r\to t} \leq d^{\mathrm{abs}}_{i,r\to t} + \Delta m$. The reverse triangle inequality yields $\Delta m \leq d^{\mathrm{abs}}_{i,r\to t}$, hence $d^{\mathrm{MCD}}_{i,r\to t} \leq 3 d^{\mathrm{abs}}_{i,r\to t}$. Lemma 3 completes Eq. 40. Möbius inversion gives

$$\bar{B}^{a,\mathcal{A}}_{i,t}(\mathcal{S}) - \bar{B}^{a,\mathcal{A}}_{i,r}(\mathcal{S}) = \sum_{\emptyset \neq \mathcal{T} \subseteq \mathcal{S}} \left[C^{a,\mathcal{A}}_{i,t}(\mathcal{T}) - C^{a,\mathcal{A}}_{i,r}(\mathcal{T}) \right].$$

Taking $\mathcal{S} = \mathcal{M}$ and bounding each subset sum proves Eq. 41. Invertibility of the transform gives the equivalence of profile and centered-response equality. □

Equation 6 measures drift from task introduction under $\mathcal{A}_y$, whereas each training update uses $\mathcal{Y}_{<t}$. Cumulative drift is bounded by the sum of per-update drift and candidate-set residuals.

Corollary 7 (MCD after multiple task updates)**.** *Fix an old-task probe i with label y introduced at stage k and let $A = \mathcal{A}_y$, with $|A| \geq 2$. Use the same modality-removal rule at stages $k, k+1, \ldots, t$. For each update $u \in \{k+1, \ldots, t\}$, let $B_u = \mathcal{Y}_{<u} \supseteq A$ and define, for $v \in \{u-1, u\}$,*

$$\begin{aligned} q^{A,B_u}_{i,v}(\mathcal{S}) &= \log\left(1 + \frac{\sum_{c \in B_u \setminus A} \exp z_{i,v,c}(\mathcal{S})}{\sum_{c \in A \setminus \{y\}} \exp z_{i,v,c}(\mathcal{S})}\right), \\ \widetilde{q}^{A,B_u}_{i,v}(\mathcal{S}) &= q^{A,B_u}_{i,v}(\mathcal{S}) - q^{A,B_u}_{i,v}(\emptyset). \end{aligned} \tag{42}$$

Let Mob *be the Möbius transform over nonempty coalitions and set $R^A_{i,u} = \|\mathsf{Mob}(\widetilde{q}^{A,B_u}_{i,u} - \widetilde{q}^{A,B_u}_{i,u-1})\|_1$. Let $\ell^{B_u}_{i,u}$ be the loss in Eq. 30 evaluated on this probe with target y, candidate set B_u, and stages $(u-1, u)$. Then*

$$d^{\mathrm{MCD},A}_{i,k\to t} \leq 3 \sum_{u=k+1}^{t} \left[\Gamma(\ell^{B_u}_{i,u}) + R^A_{i,u} \right], \tag{43}$$

where $d^{\mathrm{MCD},A}_{i,k\to t}$ is the per-probe score in Eq. 6.

Proof. Write $\mathbf{c}^A_{i,v}$ and $\mathbf{c}^{B_u}_{i,v}$ for the profiles with target y under the respective candidate sets. Equation 22 and linearity of the Möbius transform give

$$\mathbf{c}^A_{i,u} - \mathbf{c}^A_{i,u-1} = \mathbf{c}^{B_u}_{i,u} - \mathbf{c}^{B_u}_{i,u-1} + \mathsf{Mob}(\widetilde{q}^{A,B_u}_{i,u} - \widetilde{q}^{A,B_u}_{i,u-1}). \tag{44}$$

Thus, $\|\mathbf{c}^A_{i,u} - \mathbf{c}^A_{i,u-1}\|_1 \leq \Gamma(\ell^{B_u}_{i,u}) + R^A_{i,u}$ by Lemma 3. Summing over updates bounds $\|\mathbf{c}^A_{i,t} - \mathbf{c}^A_{i,k}\|_1$. □

The residual $R^A_{i,u}$ measures changes in competition from classes in $B_u \setminus A$. It vanishes when $q^{A,B_u}_{i,u}(\mathcal{S}) - q^{A,B_u}_{i,u-1}(\mathcal{S})$ is constant across modality subsets. Averaging Eq. 43 over the fixed old-task probes with the task weights in Eq. 6 gives a bound on MCD_t. Each $\ell^{B_u}_{i,u}$ is evaluated on the corresponding old-task probe.

A.6 Möbius Matching vs. Centered-Response Matching

The absolute term of CMCDR penalizes changes in each modality contribution and interaction separately. Centered-response matching penalizes their sum within each modality subset, where opposing changes can cancel. Although the two complete profiles contain the same information, their matching losses assign different costs to these changes.

Fix the sample, target class, candidate set, and removal rule as in Appendix A.5. For stages $s \in \{r, t\}$, write $b_s(\mathcal{S}) = B_s(\mathcal{S}) - B_s(\emptyset)$, $c_s(\mathcal{T}) = C_s(\mathcal{T})$, and $d_{\mathcal{T}} = c_t(\mathcal{T}) - c_r(\mathcal{T})$.

Proposition 8 (Penalty for offsetting interaction changes)**.** *Let $D = 2^{|\mathcal{M}|} - 1$ and let h_δ be the Smooth-L_1 function in Eq. 29. The absolute matching losses are*

$$\mathcal{L}^{\text{abs}}_{\text{center}} = \frac{1}{D} \sum_{\emptyset \neq \mathcal{S} \subseteq \mathcal{M}} h_\delta \left(\sum_{\emptyset \neq \mathcal{T} \subseteq \mathcal{S}} d_{\mathcal{T}} \right), \qquad \mathcal{L}^{\text{abs}}_{\text{Mob}} = \frac{1}{D} \sum_{\emptyset \neq \mathcal{T} \subseteq \mathcal{M}} h_\delta(d_{\mathcal{T}}). \tag{45}$$

Both losses vanish exactly when $\mathbf{c}_t = \mathbf{c}_r$. For two modalities, let $\Delta b_{12} = d_1 + d_2 + d_{12}$ be the change in the joint centered response. Then

$$\mathcal{L}^{\text{abs}}_{\text{Mob}} - \mathcal{L}^{\text{abs}}_{\text{center}} = \frac{h_\delta(d_{12}) - h_\delta(\Delta b_{12})}{3}. \tag{46}$$

Thus, Möbius matching assigns a strictly larger absolute loss whenever $|d_{12}| > |\Delta b_{12}|$.

Proof. Möbius inversion gives $b_t(\mathcal{S}) - b_r(\mathcal{S}) = \sum_{\emptyset \neq \mathcal{T} \subseteq \mathcal{S}} d_{\mathcal{T}}$. Substitution gives Eq. 45. Invertibility of the Möbius transform and $h_\delta(x) = 0$ if and only if $x = 0$ give the common zero-loss condition. For two modalities, $\Delta b_1 = d_1$, $\Delta b_2 = d_2$, and $\Delta b_{12} = d_1 + d_2 + d_{12}$. The singleton terms in the two losses coincide, leaving Eq. 46. The strict inequality follows because $h_\delta(x)$ is strictly increasing in $|x|$. □

For example, $\Delta\mathbf{c} = (\eta, \eta, -2\eta)$ with $\eta > 0$ gives $\Delta\mathbf{b} = (\eta, \eta, 0)$. Centered-response matching penalizes the two singleton changes, while Möbius matching adds $h_\delta(2\eta)/3$ for the interaction decrease. For $0 < \eta \leq \delta/2$, the absolute losses are $\eta^2/(3\delta)$ and η^2/δ, respectively. Möbius matching therefore imposes three times the penalty on this update, whose joint centered response remains unchanged. Table 11 compares the two parameterizations empirically.

B Experimental Settings and Implementation Details

B.1 Experimental Settings for Exploratory Analyses

We detail the settings for the exploratory analyses in Section 2.

Modality-contribution drift analysis. We analyze whether iCaRL preserves old-task modality contributions on AVE and UESTC-MMEA-CL. AVE contains audio and visual modalities, while UESTC-MMEA-CL contains video, acceleration, and gyroscope streams.

For AVE, visual and audio features are extracted using pretrained VideoMAE (Tong et al., 2022) and AudioMAE (Huang et al., 2022) models. For UESTC-MMEA-CL, the video stream follows the same visual pipeline, while acceleration and gyroscope signals are normalized and encoded by a CNN-LSTM. Modality features are projected into a shared space before fusion and classification.

Both datasets follow fixed class orders throughout the experiment. The model is trained sequentially from the previous-stage checkpoint using the same iCaRL configuration at each stage. After each update, we report the mean accuracy over observed tasks and compute the MCD score over old classes using the contribution estimator in Section 2.1. These stage-wise results are reported separately from the aggregate results in Tables 1 and 3.

The exploratory analyses use four tasks on AVE and eight tasks on UESTC-MMEA-CL. At each stage, all test samples from previously learned classes serve as diagnostic probes. For each probe from task k, we compare its current contribution profile with the profile recorded immediately after learning task k, using the fixed candidate set $\mathcal{A}_{y_i} = \mathcal{Y}_{\leq \tau(y_i)}$ for both profiles. We aggregate these cumulative changes according to Eq. 6.

Cross-modal stability analysis. We compare iCaRL with RepAlign (Pian et al., 2023) and CrossSDC (Pian et al., 2024) under the same backbone, task order, memory budget, optimizer, and training schedule. RepAlign preserves instance- and class-level audio–visual similarity, while CrossSDC further preserves cross-modal similarities between the current and previous models. Both constraints are added to the original iCaRL objective.

We measure representation drift using linear CKA (Kornblith et al., 2019) on the same old-class probes used for the MCD score. For task k and modality m, let $\mathbf{H}^m_{k,s} \in \mathbb{R}^{|\mathcal{P}^{\text{task}}_k| \times d}$ contain the

column-centered projected representations of the fixed probes in $\mathcal{P}_k^{\text{task}}$ after learning stage s, in the same row order. We compare the representations recorded immediately after learning task k with those at stage t:

$$\begin{aligned} \mathrm{CKA}(\mathbf{H}_{k,k}^m, \mathbf{H}_{k,t}^m) &= \frac{\|{\mathbf{H}_{k,k}^m}^\top \mathbf{H}_{k,t}^m\|_F^2}{\|{\mathbf{H}_{k,k}^m}^\top \mathbf{H}_{k,k}^m\|_F \|{\mathbf{H}_{k,t}^m}^\top \mathbf{H}_{k,t}^m\|_F}, \\ D_t^{\text{rep}} &= 1 - \frac{1}{M|\mathcal{K}_{<t}|} \sum_{k \in \mathcal{K}_{<t}} \sum_{m=1}^{M} \mathrm{CKA}(\mathbf{H}_{k,k}^m, \mathbf{H}_{k,t}^m). \end{aligned} \tag{47}$$

Lower D_t^{rep} indicates smaller cumulative representation drift from task introduction to stage t. The MCD score measures cumulative contribution drift over the same interval on the same probes under each task's fixed candidate set, as defined in Eq. 6. Both metrics are evaluated after each incremental update and averaged within each dataset.

For Figure 4(d), the residual MCD score of each class is averaged across incremental stages and across RepAlign and CrossSDC. Classes are divided into low- and high-MCD groups using the within-dataset median. The same groups are used for all methods.

B.2 Main Training and Evaluation Setup

Task construction. All methods follow the published train/test splits, task construction, and task order for each benchmark. For multimodal class-incremental learning, each benchmark is divided into a fixed sequence of disjoint class groups, introduced one per stage. AVE uses four tasks of seven classes, and Kinetics-Sounds uses five tasks of six classes, following AV-CIL (Pian et al., 2023). UESTC-MMEA-CL uses eight tasks of four classes (Xu et al., 2024). For continual multimodal question answering, VQAv2 follows the VQACL protocol (Zhang et al., 2023), while Split-AVQA and Split-MUSIC-AVQA follow the AVQACL protocol (Wu et al., 2025).

Model and baseline implementation. We use each method's released codebase when available, retaining its architecture and baseline-specific components. For class-incremental tasks, we implement classical continual-learning methods and methods without publicly released code in the AV-CIL (Pian et al., 2023) codebase. This framework extracts visual and audio features using pretrained VideoMAE (Tong et al., 2022) and AudioMAE (Huang et al., 2022), respectively. On UESTC-MMEA-CL, acceleration and gyroscope signals are normalized and encoded by CNN-LSTM. Within each baseline, the original method and its CMCDR variant use the same architecture, memory budget, task order, and inference procedure.

Training. All methods are trained sequentially, starting with the initial task at $t = 1$. At subsequent stages, replay-based methods use current-task data and their exemplar buffer, while replay-free methods use only current-task data. CMCDR compares the current model with the frozen previous model, using buffered old exemplars in the replay-based setting and current-task samples in the replay-free setting. All contribution profiles use the zero-representation removal in Appendix B.4.

Gradient-trained baselines initialize each subsequent stage from the previous checkpoint. We retain the original training loss, optimizer, learning-rate schedule, batch size, number of epochs, and exemplar-selection rule, with the same settings across incremental stages. CMCDR adds the contribution-preservation regularizer in Section 3 to the original objective.

MMAL updates the fusion and modality-specific compensation branches analytically at each incremental stage. We then initialize separate prediction-head copies from the resulting solutions. With the feature extractors frozen, we optimize these copies on current-task data using the normalized ridge objectives and replay-free CMCDR. Contribution profiles are computed from the combined branch scores. The refined heads are used for inference and as the frozen reference at the next stage.

Hyperparameters. We use $\lambda_{\text{cmcdr}} = \beta = 1.0$ in the main experiments, with $\epsilon = 10^{-6}$ for contribution-profile normalization and $\delta = 1.0$ as the Smooth-L_1 transition threshold. Appendix C.4 reports sensitivity analyses for $\lambda_{\text{cmcdr}} \in \{0.1, 0.5, 1.0, 2.0\}$ and $\beta \in \{0.25, 0.5, 1.0, 2.0\}$.

Evaluation. Class-incremental recognition is evaluated over all classes observed so far, without task identity at inference. MCD measures cumulative contribution drift using old-task test probes and their contribution profiles recorded at task introduction. These probes and recorded profiles are used only for evaluation. Both CMCDR variants and the baselines use the fixed-reference MCD protocol in Eq. 6.

B.3 CMCDR FOR CONTINUAL MULTIMODAL QUESTION ANSWERING

We implement LwF (Li & Hoiem, 2018), EWC (Kirkpatrick et al., 2017), and iCaRL (Rebuffi et al., 2017) in the official AVQACL codebase (Wu et al., 2025). On each benchmark, these three baselines share the modality encoders, fusion module, and classification-based answer predictor. They differ in their continual-learning objectives and memory mechanisms. Each baseline and its CMCDR variant use the same architecture, answer vocabulary, optimizer, training schedule, memory budget, and evaluation protocol.

For Split-AVQA and Split-MUSIC-AVQA, we follow the AVQACL protocol and retain audio, visual, and question inputs. Their answer vocabularies contain 2,135 and 40 classes, respectively. CMCDR applies the modality-removal rule in Appendix B.4.

Let $\mathbf{x}_i = (x_i^m)_{m\in\mathcal{M}}$ denote an input and a_i its reference answer. At stage $s \in \{t-1, t\}$, $z_{i,s,a}(\mathcal{S})$ denotes the score of candidate answer a when modalities $\mathcal{S} \subseteq \mathcal{M}$ are retained. The candidate set $\mathcal{A}_{<t}$ is shared by the current and frozen models and remains fixed across all modality subsets within each stage.

In classification-based QA, $z_{i,s,a}(\mathcal{S})$ is the logit of answer class a, and $\mathcal{A}_{<t}$ contains the answer classes represented by the previous-stage prediction head, with $|\mathcal{A}_{<t}| \geq 2$. For a fixed answer vocabulary, $\mathcal{A}_{<t}$ is the full vocabulary; for an expanding head, it contains the classes included before stage t.

For generative VQACL, we retain the original decoder and score each answer sequence $a = (a_1, \ldots, a_{L_a})$ by teacher forcing:

$$z_{i,s,a}(\mathcal{S}) = \frac{1}{L_a} \sum_{j=1}^{L_a} \log p_{\theta_s}\left(a_j \mid a_{<j}, \mathbf{x}_i^{\mathcal{S}}\right), \tag{48}$$

where L_a is the number of scored answer tokens. The candidate set $\mathcal{A}_{<t}$ contains the distinct normalized reference answers in the old-task replay buffer.

For both prediction formulations, we compute the answer benefit using Eq. 2 with target answer $a \in \mathcal{A}_{<t}$ and candidate set $\mathcal{A} = \mathcal{A}_{<t}$ at each stage s. The Möbius transform in Eq. 3 then yields $C_{i,s}^a(\mathcal{T})$ for each nonempty coalition $\mathcal{T} \subseteq \mathcal{M}$. We collect these terms into $\mathbf{C}_{i,s}^a = [C_{i,s}^a(\mathcal{T})]_{\emptyset\neq\mathcal{T}\subseteq\mathcal{M}}$, with a fixed coalition ordering.

Replay-based regularization. Let $\mathcal{X}_{\text{old}}$ contain the replayed input–answer pairs. CMCDR matches the contribution profiles of their ground-truth answers using Eq. 12, with y_i replaced by $a_i \in \mathcal{A}_{<t}$ and profiles computed over $\mathcal{A}_{<t}$. The distance d_{cp} is defined in Eq. 11.

Replay-free regularization. For replay-free answer classifiers, current-stage inputs $\mathcal{X}_{\text{new}}$ serve as intervention probes. For each input, CMCDR matches the contribution profiles of all answer classes in $\mathcal{A}_{<t}$ using Eq. 13, with $\mathcal{Y}_{<t}$ replaced by $\mathcal{A}_{<t}$ and the class index k by the answer index a. Here $N_t = |\mathcal{X}_{\text{new}}|$.

Both variants evaluate all $2^{|\mathcal{M}|}$ modality subsets using the same removal rule for the current and frozen models. Gradients are stopped through the frozen model. We use the training objective in Eq. 14, with $\mathcal{L}_{\text{cmcdr}}$ set to the replay-based or replay-free regularizer above. The baseline answer prediction and evaluation procedures remain unchanged.

B.4 MODALITY-REMOVAL REALIZATIONS

Equation 1 removes excluded modalities through an absence symbol. For fixed-interface models, we realize this intervention at the fusion input. Let $\mathbf{h}_{i,s}^m = E_{m,s}(x_i^m)$ denote the representation of

modality m at stage s. Given a replacement vector $\mathbf{q}_s^m$, we define

$$\widetilde{\mathbf{h}}_{i,s}^{m,\mathbf{q}}(\mathcal{S}) = \begin{cases} \mathbf{h}_{i,s}^m, & m \in \mathcal{S}, \\ \mathbf{q}_s^m, & m \notin \mathcal{S}, \end{cases} \qquad \mathbf{z}_{i,s}^{\mathbf{q}}(\mathcal{S}) = G_{\theta_s}\left(\{\widetilde{\mathbf{h}}_{i,s}^{m,\mathbf{q}}(\mathcal{S})\}_{m=1}^{M}\right). \tag{49}$$

The same removal rule is applied to the current and frozen models.

Zero-representation removal. The default realization in the main experiments uses $\mathbf{q}_s^m = \mathbf{0}$ for each excluded modality. This replacement is deterministic, sample-independent, reference-free, and identical across incremental stages.

Reference-mean removal. We also consider the mean representation of a fixed unlabeled reference set $\mathcal{R}^m$:

$$\mathbf{q}_s^m = \boldsymbol{\mu}_s^m = \frac{1}{|\mathcal{R}^m|} \sum_{x_j^m \in \mathcal{R}^m} E_{m,s}(x_j^m). \tag{50}$$

Because the encoder changes across stages, the resulting reference representation may also change.

Reference-anchor removal. Alternatively, for each modality, we select a fixed reference sample $x_{r_m}^m \in \mathcal{R}^m$ and set $\mathbf{q}_s^m = E_{m,s}(x_{r_m}^m)$. The reference sample is fixed across stages, while its encoded representation changes with the encoder.

Reference-mean and reference-anchor removal require a fixed auxiliary unlabeled reference set and are used only in the robustness experiments below.

Table 4: Robustness to different modality-removal realizations.

AVE					
Protocol	**Method**	**Removal**	**Avg. Acc.** ↑	**Avg. Forget.** ↓	**MCD** ↓
Replay	iCaRL	Zero	64.20	25.60	0.779
Replay	**iCaRL + CMCDR**	Zero	71.86	17.36	0.389
Replay	iCaRL	Reference mean	64.20	25.60	0.812
Replay	**iCaRL + CMCDR**	Reference mean	71.63	17.82	0.417
Replay	iCaRL	Reference anchor	64.20	25.60	0.864
Replay	**iCaRL + CMCDR**	Reference anchor	71.34	18.07	0.436
Replay-free	LwF	Zero	56.61	35.11	0.842
Replay-free	**LwF + CMCDR**	Zero	60.61	26.35	0.401
Replay-free	LwF	Reference mean	56.61	35.11	0.879
Replay-free	**LwF + CMCDR**	Reference mean	60.28	26.84	0.423
Replay-free	LwF	Reference anchor	56.61	35.11	0.817
Replay-free	**LwF + CMCDR**	Reference anchor	60.03	27.16	0.451
UESTC-MMEA-CL					
Protocol	**Method**	**Removal**	**Avg. Acc.** ↑	**Avg. Forget.** ↓	**MCD** ↓
Replay	iCaRL	Zero	71.83	22.86	0.464
Replay	**iCaRL + CMCDR**	Zero	75.12	17.15	0.288
Replay	iCaRL	Reference mean	71.83	22.86	0.493
Replay	**iCaRL + CMCDR**	Reference mean	74.96	17.58	0.307
Replay	iCaRL	Reference anchor	71.83	22.86	0.538
Replay	**iCaRL + CMCDR**	Reference anchor	74.73	17.94	0.332
Replay-free	LwF	Zero	17.13	49.37	0.513
Replay-free	**LwF + CMCDR**	Zero	22.85	38.65	0.299
Replay-free	LwF	Reference mean	17.13	49.37	0.557
Replay-free	**LwF + CMCDR**	Reference mean	22.47	39.14	0.319
Replay-free	LwF	Reference anchor	17.13	49.37	0.526
Replay-free	**LwF + CMCDR**	Reference anchor	22.16	39.63	0.347

Modality-removal robustness. We compare zero, reference-mean, and reference-anchor removal using iCaRL and LwF, with and without CMCDR. For each CMCDR variant, the same removal rule

is used for contribution-profile construction, regularization, and evaluation, while all other training settings remain fixed. For the corresponding baselines, modality removal is applied only when computing the MCD score. Their accuracy and forgetting values are therefore identical across removal rules.

Table 4 reports average accuracy, average forgetting, and the MCD score. Across both datasets and learning protocols, CMCDR improves accuracy and reduces forgetting and the MCD score under all three removal rules.

C Additional Experiments

C.1 Random-Seed Robustness

Table 5: Five-seed robustness results on AVE and UESTC-MMEA-CL (mean $\pm$ sample standard deviation).

Protocol	Method	AVE			UESTC-MMEA-CL		
		Avg. Acc. ↑	Avg. Forget. ↓	MCD ↓	Avg. Acc. ↑	Avg. Forget. ↓	MCD ↓
Replay-free	LwF	56.61 $\pm$ 0.31	35.11 $\pm$ 0.28	0.842 $\pm$ 0.042	17.13 $\pm$ 0.44	49.37 $\pm$ 0.37	0.513 $\pm$ 0.029
Replay-free	**LwF + CMCDR**	**60.61** $\pm$ 0.56	**26.35** $\pm$ 0.52	**0.401** $\pm$ 0.032	**22.85** $\pm$ 0.70	**38.65** $\pm$ 0.60	**0.299** $\pm$ 0.024
Replay-free	EWC	41.28 $\pm$ 0.27	68.97 $\pm$ 0.42	0.962 $\pm$ 0.057	19.57 $\pm$ 0.48	29.84 $\pm$ 0.22	0.482 $\pm$ 0.030
Replay-free	**EWC + CMCDR**	**47.36** $\pm$ 0.79	**57.42** $\pm$ 0.88	**0.375** $\pm$ 0.037	**24.78** $\pm$ 0.68	**22.06** $\pm$ 0.35	**0.308** $\pm$ 0.026
Replay-free	MMAL	70.28 $\pm$ 0.35	19.64 $\pm$ 0.24	0.724 $\pm$ 0.038	75.57 $\pm$ 0.47	24.83 $\pm$ 0.51	0.458 $\pm$ 0.017
Replay-free	**MMAL + CMCDR**	**72.63** $\pm$ 0.50	**15.02** $\pm$ 0.36	**0.321** $\pm$ 0.030	**78.12** $\pm$ 0.59	**20.94** $\pm$ 0.63	**0.319** $\pm$ 0.018
Replay	iCaRL	64.20 $\pm$ 0.41	25.60 $\pm$ 0.37	0.779 $\pm$ 0.044	71.83 $\pm$ 0.52	22.86 $\pm$ 0.45	0.464 $\pm$ 0.028
Replay	**iCaRL + CMCDR**	**71.86** $\pm$ 0.82	**17.36** $\pm$ 0.58	**0.389** $\pm$ 0.026	**75.12** $\pm$ 0.67	**17.15** $\pm$ 0.58	**0.288** $\pm$ 0.025
Replay	AV-CIL	71.37 $\pm$ 0.40	6.77 $\pm$ 0.28	0.649 $\pm$ 0.030	80.27 $\pm$ 0.58	18.74 $\pm$ 0.34	0.449 $\pm$ 0.029
Replay	**AV-CIL + CMCDR**	**75.86** $\pm$ 0.60	**5.31** $\pm$ 0.37	**0.227** $\pm$ 0.025	**83.68** $\pm$ 0.72	**13.36** $\pm$ 0.45	**0.275** $\pm$ 0.026
Replay	CIGN	73.92 $\pm$ 0.45	10.84 $\pm$ 0.18	0.691 $\pm$ 0.039	82.14 $\pm$ 0.53	16.27 $\pm$ 0.31	0.461 $\pm$ 0.026
Replay	**CIGN + CMCDR**	**74.85** $\pm$ 0.42	**7.11** $\pm$ 0.27	**0.240** $\pm$ 0.026	**83.75** $\pm$ 0.62	**12.97** $\pm$ 0.38	**0.254** $\pm$ 0.023

We assess whether the gains of CMCDR persist under stochastic training on AVE and UESTC-MMEA-CL for the class-incremental methods reported in Table 1. For each setting, we use the same five random seeds for a baseline and its CMCDR counterpart. Model initialization, data order, and exemplar sampling are paired within a seed, whereas the benchmark split, class order, memory budget, and the coefficients selected for the main evaluation remain fixed. Table 5 reports the mean and sample standard deviation of average accuracy, average forgetting, and the MCD score over the five runs.

C.2 Modality Contribution Drift and Class-wise Forgetting

Class-wise effects of CMCDR. We compare class-wise MCD and forgetting on AVE for iCaRL and LwF, with and without CMCDR. Class-wise MCD averages the per-probe score in Eq. 6 over each old class. Figure 7 shows 21 class pairs for each baseline.

CMCDR reduces MCD for all displayed classes and reduces forgetting for 19 of 21 iCaRL classes and 20 of 21 LwF classes. The remaining classes have lower MCD but higher forgetting. For one iCaRL class, MCD decreases from 0.391 to 0.180, while forgetting increases from 17.01% to 21.97%. For the LwF exception, MCD decreases from 0.494 to 0.241, while forgetting increases from 24.41% to 28.24%. CMCDR improves retention for most classes, with a few class-specific exceptions.

These exceptions indicate that reducing MCD does not necessarily reduce forgetting for every class. One possible explanation is task similarity: when new and old tasks share modality contribution patterns, changes in modality-specific contributions and cross-modal interactions may benefit old-class discrimination. Regularizing these changes may also restrict beneficial adaptation. This explanation remains a hypothesis to be examined in future work.

MCD and forgetting under comparable full-input benefit drift. The class-wise comparisons above suggest an association between MCD and forgetting, with some exceptions. We next examine whether this association persists when changes in the overall full-input benefit are comparable.

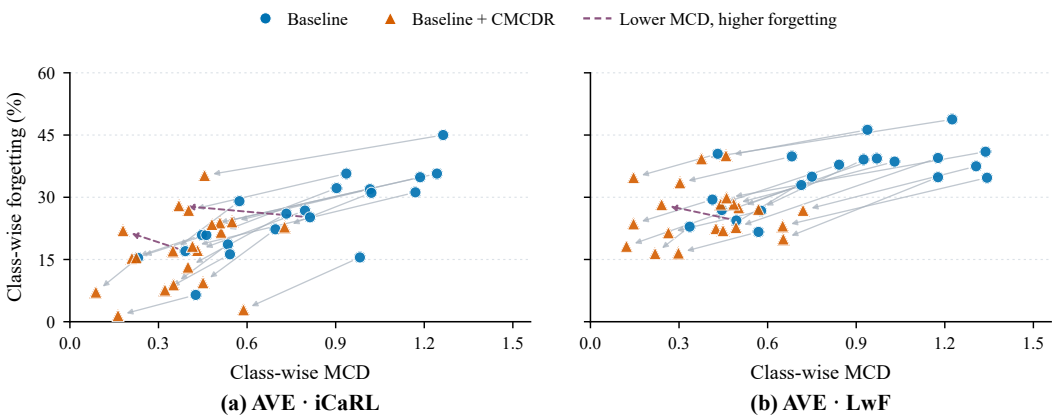


Figure 7: Class-wise MCD and forgetting on AVE. Arrows connect each class under the baseline and its CMCDR variant. Dashed purple arrows mark classes with lower MCD but higher forgetting.

We conduct this analysis separately for iCaRL and LwF on AVE, treating each old class at each incremental stage as an observation.

For each observation, we use the same probe samples and fixed candidate set as in the MCD evaluation. We first subtract the no-modality benefit from the full-input benefit. For each probe, we take the absolute change in this centered benefit from the class's introduction stage to the current stage, and then average these changes over the class. This gives the cumulative full-input benefit drift for that observation. For each method, we divide the observed drift range into three equal-width bins and split observations within each bin into low- and high-MCD groups using that bin's median MCD. We compute the mean class-wise forgetting for each group within each bin, and then average the three bin-level means with equal weights.

Figure 8 reports these equal-bin averages. For iCaRL, mean forgetting increases from 63.8% in the low-MCD group to 74.5% in the high-MCD group, a gap of 10.7 percentage points. For LwF, the corresponding values are 53.1% and 65.1%, with a gap of 12.0 percentage points. These positive gaps support an association between higher MCD and greater forgetting after stratification by centered full-input benefit drift.

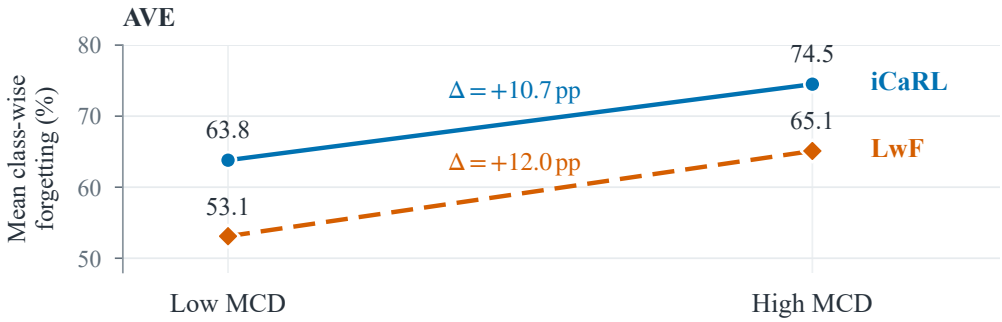


Figure 8: MCD and forgetting after stratification by full-input benefit drift on AVE. Class–stage observations are divided into three equal-width bins of centered full-input benefit drift and split at the median MCD within each bin. Each point is the equally weighted average of the three bin-level forgetting means for the indicated MCD group.

C.3 Computational Efficiency and Scalability

We report the mean wall-clock training time per incremental stage, measured using four NVIDIA GeForce RTX 3090 GPUs. Each baseline and its CMCDR variant use the same architecture,

Table 6: Mean training time per incremental stage on UESTC-MMEA-CL. Overhead is relative to each baseline. Δ_M is the CMCDR training-time increase from two to three modalities. Acc. and Gyro. denote acceleration and gyroscope signals.

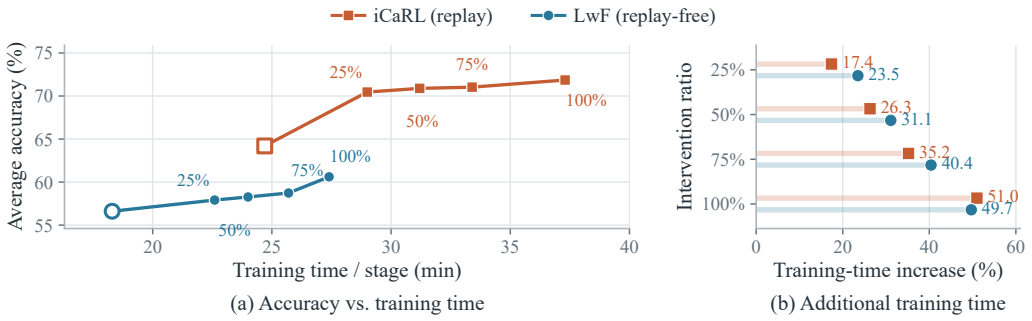
Protocol	Method	Modalities	Time (min) ↓	Overhead (%) ↓	Δ_M (%) ↓
Replay	iCaRL	Video + Acc.	46.8	–	–
Replay	**iCaRL + CMCDR**	Video + Acc.	62.9	+34.4	–
Replay	iCaRL	Video + Acc. + Gyro.	57.6	–	–
Replay	**iCaRL + CMCDR**	Video + Acc. + Gyro.	85.4	+48.3	+35.8
Replay-free	LwF	Video + Acc.	31.9	–	–
Replay-free	**LwF + CMCDR**	Video + Acc.	43.4	+36.1	–
Replay-free	LwF	Video + Acc. + Gyro.	45.7	–	–
Replay-free	**LwF + CMCDR**	Video + Acc. + Gyro.	69.8	+52.7	+60.8

batch size, training schedule, replay budget, and hardware. The relative overhead is computed as $100(t_{\text{CMCDR}}/t_{\text{base}} - 1)\%$ under the same modality setting.

We further evaluate how CMCDR's training time scales with the number of modalities on UESTC-MMEA-CL using iCaRL and LwF. The two-modality setting uses video and acceleration signals. The three-modality setting additionally includes gyroscope signals.

On UESTC-MMEA-CL, CMCDR increases training time by $34.4\%\sim52.7\%$ for iCaRL and LwF across the two- and three-modality classification settings (Table 6). Increasing the number of modalities from two to three increases the total training time of iCaRL + CMCDR and LwF + CMCDR by 35.8% and 60.8%, respectively.

Effect of intervention sample ratio. We study the efficiency–performance trade-off of CMCDR by varying the fraction of samples used for modality-subset interventions from 25% to 100%, while keeping all other training settings fixed. A lower intervention ratio reduces the additional forward passes required to estimate modality contributions. We evaluate replay-based iCaRL and replay-free LwF in terms of average accuracy, average forgetting, the MCD score, and average training time per incremental stage.



(a) Accuracy vs. training time (b) Additional training time

Figure 9: Training-time–accuracy trade-off on AVE using the measurements in Table 7. (a) Average accuracy versus mean training time per incremental stage. (b) Relative training-time increase over each baseline.

As shown in Table 7 and Figure 9, CMCDR remains effective with partial interventions. For iCaRL, using only 25% of samples reduces the MCD score from 0.779 to 0.452 and improves average accuracy from 64.20% to 70.46%, while limiting the training overhead to 17.4%. Increasing the intervention ratio further reduces the MCD score and improves accuracy, with the lowest forgetting achieved at a 100% ratio.

A similar trend is observed for LwF. At a 25% ratio, CMCDR reduces the MCD score from 0.842 to 0.474 and forgetting from 35.11% to 28.99%, with a 23.5% training overhead. Full intervention further improves performance but increases the overhead to 49.7%. Overall, these results show

that CMCDR can reduce its additional overhead through partial interventions, providing a favorable efficiency–performance trade-off and improving its scalability to larger continual-learning settings.

Table 7: Effect of the intervention sample ratio on AVE.

Protocol	Method	Ratio	Avg. Acc. (%) ↑	Avg. Forget. (%) ↓	MCD ↓	Time / stage (min) ↓	Overhead (%) ↓
Replay	iCaRL	–	64.20	25.60	0.779	24.7	–
Replay	iCaRL + CMCDR	25%	70.46	18.14	0.452	**29.0**	+17.4
Replay	iCaRL + CMCDR	50%	70.89	17.41	0.421	31.2	+26.3
Replay	iCaRL + CMCDR	75%	71.03	17.72	0.402	33.4	+35.2
Replay	iCaRL + CMCDR	100%	**71.86**	**17.36**	**0.389**	37.3	+51.0
Replay-free	LwF	–	56.61	35.11	0.842	18.3	–
Replay-free	LwF + CMCDR	25%	57.91	28.99	0.474	**22.6**	+23.5
Replay-free	LwF + CMCDR	50%	58.27	28.01	0.468	24.0	+31.1
Replay-free	LwF + CMCDR	75%	58.73	27.88	0.433	25.7	+40.4
Replay-free	LwF + CMCDR	100%	**60.61**	**26.35**	**0.401**	27.4	+49.7

C.4 Hyperparameter Sensitivity Analysis

We vary one hyperparameter at a time while keeping all other training and evaluation settings fixed. We evaluate $\lambda_{\mathrm{cmcdr}} \in \{0.1, 0.5, 1.0, 2.0\}$ and $\beta \in \{0.25, 0.5, 1.0, 2.0\}$ on AVE and UESTC-MMEA-CL, using iCaRL for replay-based training and LwF for replay-free training. For λ_{cmcdr}, we report average accuracy, average forgetting, the MCD score, and new-task accuracy. For β, we report average accuracy, average forgetting, and the raw and normalized contribution drifts defined in Appendix A.5, denoted Abs-MCD and Rel-MCD.

Table 8: Sensitivity to λ_{cmcdr} on AVE and UESTC-MMEA-CL.

Protocol	λ_{cmcdr}	AVE				UESTC-MMEA-CL			
		Avg. Acc. ↑	Avg. Forget. ↓	MCD ↓	New-task Acc. ↑	Avg. Acc. ↑	Avg. Forget. ↓	MCD ↓	New-task Acc. ↑
Replay (iCaRL)	0.1	68.742	21.835	0.547	**76.923**	73.418	20.464	0.387	**82.369**
Replay (iCaRL)	0.5	**71.937**	18.106	0.442	76.258	74.834	17.577	0.335	81.692
Replay (iCaRL)	1.0	71.860	17.360	0.389	75.783	**75.120**	17.152	0.288	81.257
Replay (iCaRL)	2.0	71.416	**16.934**	**0.342**	74.605	74.707	**16.663**	**0.254**	79.976
Replay-free (LwF)	0.1	58.743	31.276	0.587	**67.425**	20.158	44.724	0.428	**32.179**
Replay-free (LwF)	0.5	60.167	26.774	0.472	66.829	**22.913**	39.116	0.353	31.568
Replay-free (LwF)	1.0	**60.609**	26.352	0.401	65.957	22.851	38.648	0.299	30.912
Replay-free (LwF)	2.0	60.076	**25.713**	**0.361**	64.618	22.167	**37.979**	**0.262**	29.715

Effect of λ_{cmcdr}. Increasing λ_{cmcdr} from 0.1 to 1.0 reduces the MCD score and forgetting in all settings, while average accuracy generally improves (Table 8). New-task accuracy decreases gradually as the regularization becomes stronger. At $\lambda_{\mathrm{cmcdr}} = 2.0$, the MCD score and forgetting decrease further, but both average accuracy and new-task accuracy drop across all settings. Thus, $\lambda_{\mathrm{cmcdr}} = 0.5{\sim}1.0$ provides the best trade-off between retaining old knowledge and learning new tasks.

Table 9: Sensitivity to β on AVE and UESTC-MMEA-CL.

Protocol	β	AVE				UESTC-MMEA-CL			
		Avg. Acc. ↑	Avg. Forget. ↓	Abs-MCD ↓	Rel-MCD ↓	Avg. Acc. ↑	Avg. Forget. ↓	Abs-MCD ↓	Rel-MCD ↓
Replay (iCaRL)	0.25	70.512	18.731	**0.223**	0.201	74.326	18.427	**0.145**	0.176
Replay (iCaRL)	0.50	71.724	17.582	0.251	0.158	**75.342**	17.618	0.174	0.135
Replay (iCaRL)	1.00	**71.860**	**17.360**	0.281	0.129	75.120	**17.152**	0.201	0.112
Replay (iCaRL)	2.00	71.203	18.127	0.318	**0.105**	74.581	17.993	0.237	**0.087**
Replay-free (LwF)	0.25	59.604	28.246	**0.232**	0.235	21.527	41.653	**0.154**	0.197
Replay-free (LwF)	0.50	**60.795**	26.827	0.264	0.176	22.381	39.412	0.194	0.162
Replay-free (LwF)	1.00	60.609	**26.352**	0.289	0.138	**22.851**	**38.648**	0.209	0.118
Replay-free (LwF)	2.00	59.872	27.261	0.339	**0.112**	22.054	39.786	0.257	**0.079**

Effect of β. Increasing β consistently reduces Rel-MCD but increases Abs-MCD across both datasets and protocols (Table 9). Small β values better preserve contribution values, whereas large values better preserve the normalized contribution profile. The default setting, $\beta = 1.0$, yields the lowest average forgetting in all four settings, while average accuracy peaks at $\beta = 0.5$ or 1.0.

C.5 Replay-Free Probe Transfer

Proposition 4 bounds old-task contribution drift using the current-task probe loss and a distribution-shift term. We examine whether held-out current-task probes track old-task drift by comparing class-wise MCD scores using the replay-free LwF. Lemma 6 shows that each MCD score bounds the ℓ_1 change in the corresponding contribution profile.

After each incremental stage $t \geq 2$, we compare the frozen model $F_{\theta_{t-1}}$ and the updated model F_{θ_t} under the same modality-subset interventions. For two profiles defined as in Eq. 28, the per-profile MCD score is

$$d^{\mathrm{MCD}}_{i,r\to t} = |m_t - m_r| + (\min(m_t, m_r) + \epsilon)\|\widehat{\mathbf{c}}_t - \widehat{\mathbf{c}}_r\|_1. \tag{51}$$

For the comparison from $t-1$ to t, we denote this score for sample $\mathbf{x}_i$ and target class c by $d^{\mathrm{MCD}}_{t-1\to t}(\mathbf{x}_i, c)$. For each old class $c \in \mathcal{Y}_{<t}$, we compute

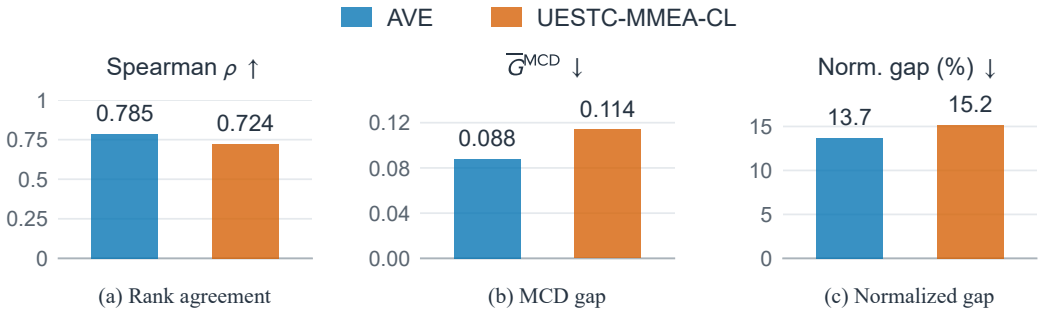
$$D^{t,c}_{\mathrm{cur}} = \frac{1}{|\mathcal{Q}_t|}\sum_{\mathbf{x}_i\in\mathcal{Q}_t} d^{\mathrm{MCD}}_{t-1\to t}(\mathbf{x}_i, c), \qquad D^{t,c}_{\mathrm{old}} = \frac{1}{|\mathcal{P}^{\mathrm{class}}_c|}\sum_{\mathbf{x}_i\in\mathcal{P}^{\mathrm{class}}_c} d^{\mathrm{MCD}}_{t-1\to t}(\mathbf{x}_i, c), \tag{52}$$

where $\mathcal{Q}_t$ contains held-out current-task probes and $\mathcal{P}^{\mathrm{class}}_c$ contains held-out probes from old class c. Both scores use c as the target class.

At each stage, we compute the Spearman correlation between $D^{t,c}_{\mathrm{cur}}$ and $D^{t,c}_{\mathrm{old}}$ across old classes. We report the arithmetic mean of these within-stage correlations across incremental stages. We also measure the mean absolute MCD gap across old classes:

$$G^{\mathrm{MCD}}_t = \frac{1}{|\mathcal{Y}_{<t}|}\sum_{c\in\mathcal{Y}_{<t}} \left|D^{t,c}_{\mathrm{old}} - D^{t,c}_{\mathrm{cur}}\right|. \tag{53}$$

We report the stage-averaged gap $\overline{G}^{\mathrm{MCD}}$. At each stage, we normalize G^{MCD}_t by the mean of $D^{t,c}_{\mathrm{old}}$ over old classes. We report the average normalized gap across stages as a percentage.



(a) Rank agreement (b) MCD gap (c) Normalized gap

Figure 10: Agreement between class-wise MCD scores on held-out current-task and old-class probes using the replay-free LwF: (a) mean within-stage Spearman correlation. (b) stage-averaged mean absolute MCD gap $\overline{G}^{\mathrm{MCD}}$. (c) average normalized gap.

Figure 10 reports mean within-stage Spearman correlations of 0.785 and 0.724 on AVE and UESTC-MMEA-CL, respectively. The stage-averaged mean absolute MCD gaps are 0.088 and 0.114, with normalized gaps of 13.7% and 15.2%. These results support using current-task probes to estimate old-class MCD in replay-free learning.

C.6 Ablation and Controlled Comparisons

Loss components. We first ablate the two terms in the CMCDR objective. Let $\widehat{\mathbf{C}} = \mathbf{C}/(\|\mathbf{C}\|_1 + \epsilon)$ and define $\widehat{\mathbf{C}}^-$ analogously. The two components are

$$\mathcal{L}_{\mathrm{abs}} = \rho_\delta\left(\mathbf{C}, \mathrm{sg}(\mathbf{C}^-)\right), \qquad \mathcal{L}_{\mathrm{rel}} = \rho_\delta\left(\widehat{\mathbf{C}}, \mathrm{sg}(\widehat{\mathbf{C}}^-)\right).$$

Equation 11 therefore gives $d_{\mathrm{cp}} = \mathcal{L}_{\mathrm{abs}} + \beta\mathcal{L}_{\mathrm{rel}}$. The relative term retains its weight β when used alone. All variants use the full Möbius contribution profile with identical training settings.

Table 10: Ablation of the CMCDR loss components.

Protocol	Components $\mathcal{L}_{\text{abs}}$	Components $\mathcal{L}_{\text{rel}}$	AVE Avg. Acc. ↑	AVE Avg. Forget. ↓	UESTC-MMEA-CL Avg. Acc. ↑	UESTC-MMEA-CL Avg. Forget. ↓
Replay (iCaRL)	×	×	64.20	25.60	71.83	22.86
Replay (iCaRL)	✓	×	69.27	20.38	74.07	19.17
Replay (iCaRL)	×	✓	68.11	21.57	73.48	20.05
Replay (iCaRL)	✓	✓	**71.86**	**17.36**	**75.12**	**17.15**
Replay-free (LwF)	×	×	56.61	35.11	17.13	49.37
Replay-free (LwF)	✓	×	59.46	29.18	21.36	42.09
Replay-free (LwF)	×	✓	58.91	30.43	20.83	43.20
Replay-free (LwF)	✓	✓	**60.61**	**26.35**	**22.85**	**38.65**

Both components improve accuracy and reduce forgetting when used independently (Table 10). When used alone, absolute matching achieves higher accuracy and lower forgetting than relative matching. Combining both terms yields the best performance across both datasets and learning protocols.

Contribution-profile parameterization. We next examine whether retaining modality interactions and explicitly representing them with Möbius coordinates are beneficial. For each old-class index k, we define the centered subset-response vector as

$$\overline{\mathbf{B}}_{i,s}^{k} = \left[B_{i,s}^{k,\mathcal{Y}_{<t}}(\mathcal{S}) - B_{i,s}^{k,\mathcal{Y}_{<t}}(\emptyset) \right]_{\emptyset \neq \mathcal{S} \subseteq \mathcal{M}}. \tag{54}$$

The centered subset-response vector and the full Möbius profile are related by an invertible linear transform and therefore retain the same subset-level information. We compare three parameterizations: singleton-only contributions, which exclude interaction terms; centered subset responses, which retain interactions implicitly; and the full Möbius profile, which represents interaction terms explicitly. All variants evaluate the same 2^M modality subsets and use both terms in Eq. 11. The regularization coefficients are selected using the same search grid and tuning budget.

Table 11: Ablation of the contribution-profile parameterization.

Protocol	Parameterization	Interactions	AVE Avg. Acc. ↑	AVE Avg. Forget. ↓	UESTC-MMEA-CL Avg. Acc. ↑	UESTC-MMEA-CL Avg. Forget. ↓
Replay (iCaRL)	Singleton-only	Excluded	68.52	21.36	73.41	20.02
Replay (iCaRL)	Centered subsets	Implicit	70.13	18.89	73.96	18.84
Replay (iCaRL)	Möbius	Explicit	**71.86**	**17.36**	**75.12**	**17.15**
Replay-free (LwF)	Singleton-only	Excluded	58.74	31.02	20.11	44.27
Replay-free (LwF)	Centered subsets	Implicit	59.46	28.41	21.53	40.66
Replay-free (LwF)	Möbius	Explicit	**60.61**	**26.35**	**22.85**	**38.65**

Both parameterizations that retain interaction information achieve higher accuracy and lower forgetting than singleton-only matching. The full Möbius profile improves average accuracy by 1.15∼1.73 percentage points and reduces forgetting by 1.53∼2.06 points over centered-response matching across both datasets and learning protocols. Möbius matching penalizes each contribution change separately. For two modalities, its absolute loss includes an additional interaction penalty whenever singleton changes offset interaction drift and preserve the joint centered response (Proposition 8).

Controlled comparison with subset-wise distillation. To distinguish contribution-profile preservation from generic subset-wise regularization, we compare CMCDR with subset-wise distillation (SubsetKD). SubsetKD uses the same intervention samples, all 2^M modality subsets, and the same forward passes through the current and frozen models as CMCDR. The two methods differ in their preservation targets: SubsetKD matches old-class output distributions for each modality subset, whereas CMCDR matches the corresponding modality contribution profiles.

For stage $s \in \{t-1, t\}$, the old-class distribution under modality subset $\mathcal{S}$ is

$$\mathbf{p}_{i,s}^{<t}(\mathcal{S}; T) = \mathrm{softmax}\left(\frac{[z_{i,s,k}(\mathcal{S})]_{k \in \mathcal{Y}_{<t}}}{T}\right), \tag{55}$$

where T is the distillation temperature. Let $\mathcal{I}_t$ denote the intervention inputs, with $\mathcal{I}_t = \{\mathbf{x}_i : (\mathbf{x}_i, y_i) \in \mathcal{X}_{\mathrm{old}}\}$ for replay-based training and $\mathcal{I}_t = \mathcal{X}_{\mathrm{new}}$ for replay-free training. SubsetKD is defined as

$$\mathcal{L}_{\mathrm{subKD}} = \frac{T^2}{|\mathcal{I}_t|\, 2^M} \sum_{\mathbf{x}_i \in \mathcal{I}_t} \sum_{\mathcal{S} \subseteq \mathcal{M}} \mathrm{KL}\left(\mathrm{sg}\left[\mathbf{p}_{i,t-1}^{<t}(\mathcal{S}; T)\right] \middle\| \mathbf{p}_{i,t}^{<t}(\mathcal{S}; T)\right). \tag{56}$$

We tune $T \in \{1, 2, 4, 8\}$ and $\lambda_{\mathrm{subKD}} \in \{0.1, 0.5, 1.0, 2.0\}$, using the same weight grid as for λ_{cmcdr}. We use $T = 2$ and $\lambda_{\mathrm{subKD}} = 1$, which achieve the best performance under both protocols.

The two objectives impose different constraints even when evaluated on the same modality subsets. A subset-independent shift in class benefits is removed by centering and therefore leaves the Möbius contribution profile unchanged, while it can still alter the corresponding output distribution. Thus, SubsetKD preserves subset-wise predictions, whereas CMCDR preserves the contribution structure induced jointly by all subset responses.

Table 12: Controlled comparison between CMCDR and subset-wise distillation (SubsetKD).

Protocol	Method	Preservation target	# Subsets	AVE		UESTC-MMEA-CL	
				Avg. Acc. ↑	Avg. Forget. ↓	Avg. Acc. ↑	Avg. Forget. ↓
Replay	iCaRL	Full-input logits	1	64.20	25.60	71.83	22.86
Replay	iCaRL + SubsetKD	Subset-wise distributions	2^M	66.74	23.63	72.46	21.61
Replay	iCaRL + CMCDR	Möbius contribution profile	2^M	**71.86**	**17.36**	**75.12**	**17.15**
Replay-free	LwF	Full-input distillation	1	56.61	35.11	17.13	49.37
Replay-free	LwF + SubsetKD	Subset-wise distributions	2^M	57.89	33.78	19.86	46.57
Replay-free	LwF + CMCDR	Möbius contribution profile	2^M	**60.61**	**26.35**	**22.85**	**38.65**

SubsetKD consistently improves accuracy and reduces forgetting over the underlying continual-learning baselines, showing that subset-wise distillation is beneficial (Table 12). CMCDR further improves average accuracy by $2.66{\sim}5.12$ percentage points and reduces average forgetting by $4.46{\sim}7.92$ percentage points across both datasets and protocols. With matched intervention inputs, subset coverage, and forward-pass budget, these results support preserving modality contribution profiles rather than subset-wise predictions alone. SubsetKD directly constrains the output distribution of every modality subset and may therefore penalize changes that leave the underlying contribution structure intact, potentially restricting adaptation to new tasks. CMCDR instead preserves contribution structure while allowing greater flexibility in subset-level predictions.

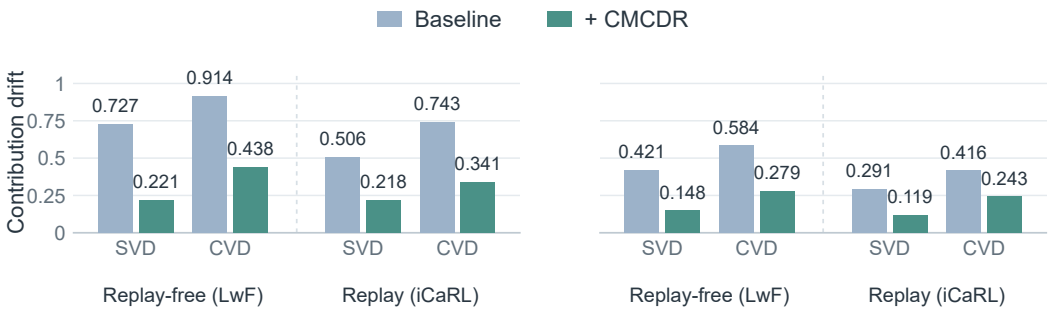
C.7 Robustness to Alternative Contribution Estimators

We finally examine whether the observed contribution stabilization depends on the Möbius-based estimator used for the MCD score. We re-evaluate the same trained models using CMoB (Wang et al., 2025) and Shapley-based modality valuation (Wei et al., 2024). Both estimators are used only for post-hoc evaluation. For each estimator, we construct contribution vectors at each stage and compute their drift using the distance in Eq. 6. We call the resulting measures CMoB-value drift (CVD) and Shapley-value drift (SVD).

As shown in Figure 11, CMCDR reduces contribution drift across both datasets and learning protocols under the Shapley- and CMoB-based estimators. Together with the MCD results, these findings demonstrate consistent contribution stabilization across three estimators.

D Comparison with Related Methods

We discuss how CMCDR differs from existing methods in multimodal continual learning and balanced multimodal learning, and compare it with representative modality-balancing methods under MMCL settings.



(a) AVE (b) UESTC-MMEA-CL

Figure 11: Contribution drift under Shapley-based (SVD) and CMoB-based (CVD) estimators on (a) AVE and (b) UESTC-MMEA-CL.

D.1 Multimodal Continual Learning

Multimodal continual learning (MMCL) learns sequential tasks from multiple modalities while retaining previously acquired knowledge. Existing methods mainly reduce forgetting by preserving multimodal representations. AV-CIL (Pian et al., 2023) maintains instance- and class-level audio–visual semantic similarity and distills audio-guided visual attention. CIGN (Mo et al., 2023) distills audio–visual class tokens to preserve class-aware cross-modal representations. MAFED (Nikandrou et al., 2024) applies modality-aware distillation to visual and question representations. MG-CLIP (Huang et al., 2025) preserves the image–text modality gap in CLIP-based continual learning. These methods regularize feature similarity, token representations, attention, or cross-modal relations. CMCDR instead constrains changes in old-task modality contribution profiles, including both individual modalities and their interactions.

D.2 Balanced Multimodal Learning

Balanced multimodal learning addresses modality dominance and under-optimization during joint multimodal training (Peng et al., 2022; Fan et al., 2023; Li et al., 2023; Wei & Hu, 2024; Hua et al., 2024). OGM-GE (Peng et al., 2022) adjusts modality-specific gradients according to their optimization states. Shapley-based modality valuation estimates sample-level modality contributions and strengthens low-contributing modalities (Wei et al., 2024). CMoB (Wang et al., 2025) estimates modality contributions through causal effects and selectively improves under-optimized modalities.

In MMCL, modality contributions are task-dependent. As illustrated in Figure 1, different tasks can rely on different modalities and modality interactions. A low modality contribution therefore does not necessarily indicate under-optimization. Modality-balancing methods adjust how modalities are optimized for the current task. CMCDR instead constrains changes in old-task contribution profiles for individual modalities and their interactions across incremental stages, rather than encouraging a common contribution pattern across tasks.

D.3 Comparison with Modality-Balancing Methods

We integrate OGM-GE (Peng et al., 2022) and CMoB (Wang et al., 2025) into replay-free and replay-based MMCL pipelines. Under the replay-free protocol, both methods operate on the current-task multimodal objective while the original LwF (Li & Hoiem, 2018) distillation objective remains unchanged. Replay-free CMCDR uses current-task probes and the frozen previous model to preserve old-class contribution profiles. Under the replay-based protocol, OGM-GE and CMoB are integrated into the same iCaRL training pipeline used by CMCDR. Within each protocol, all methods use the same backbone, classifier, task order, data augmentation, optimizer, memory budget, and number of updates. CMCDR uses the same contribution-profile construction and coefficient as in the main experiments.

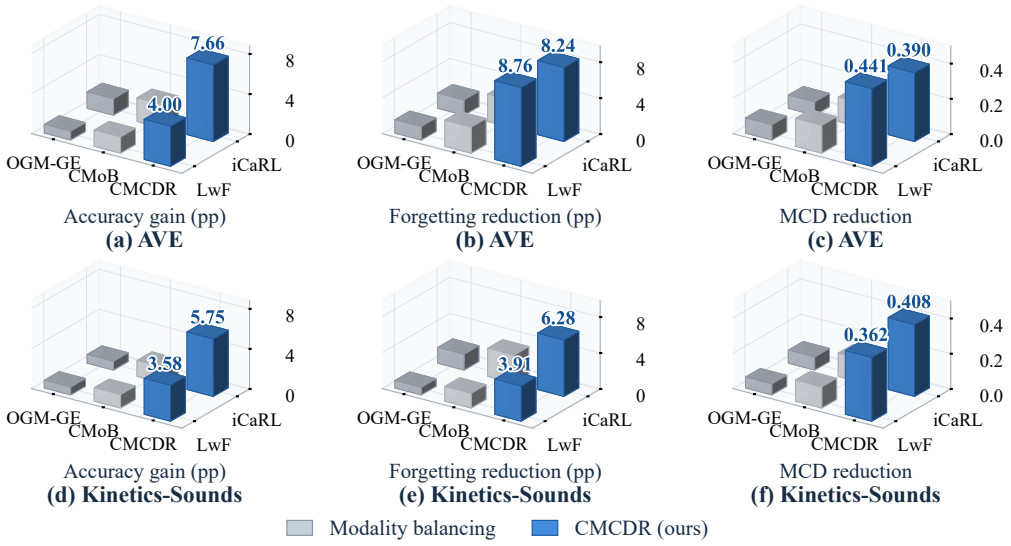


Figure 12: Comparison with modality-balancing methods on AVE and Kinetics-Sounds. Bars show accuracy gains (pp), forgetting reductions (pp), and MCD reductions relative to the corresponding LwF or iCaRL baseline. Higher is better in all panels.

Figure 12 shows that OGM-GE and CMoB improve accuracy and reduce forgetting over the corresponding LwF and iCaRL baselines. They also reduce the MCD score, indicating that more balanced multimodal optimization can partially alleviate contribution drift. However, multimodal balancing methods generally encourage under-optimized modalities to contribute more during the current task, whereas different tasks can legitimately exhibit different modality dependencies. As a result, enforcing modality balance across tasks can distort the modality dependence learned for earlier tasks. CMCDR instead preserves the task-specific modality dependence of old tasks, leading to lower MCD scores, higher accuracy, and less forgetting across both datasets and protocols.

Overall, modality-balancing methods adjust modality optimization for the current task without explicitly preserving old-task contribution profiles. When old and new tasks rely on different modalities, these updates may disrupt the contributions supporting old-task predictions. CMCDR instead constrains these changes by matching old-class contribution profiles to those of the frozen previous model.