RAM-CIL: SEPARATING PREDICTIVE RELIABILITY FROM REPLAY UTILITY IN MULTIMODAL CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Experience replay preserves historical training examples, but their observed modalities need not remain reliable learning signals. We study audio–text class-incremental intent recognition under controlled changes to current and replay observations. Holding current inputs clean, replacing historical audio and transcripts with paired shifted versions reduces old-task retention by 15.90 percentage points. We introduce Reliability-Aware Multimodal Class-Incremental Learning (RAM-CIL), which combines a base retrieval score with rank-normalized, label-conditioned predictive support from independent modality classifiers, without changing replay features or the learning objective. The results reveal a protocol-dependent separation between predictive support, retention, and new-task acquisition. Over a complete six-task run, additive reliability correction raises final Macro-F1 while reducing old-task retention. From a shared task-three checkpoint, the same correction instead improves retention. A separate, privileged diagnostic that filters MIR candidates using paired degradation scores improves new-task acquisition by 6.91 points with a 0.21-point retention decrease. These findings establish neither uniform forgetting reduction nor a universal plasticity benefit. They show that predictive reliability is insufficient as a proxy for replay utility, and motivate evaluating stability and plasticity separately under explicitly matched replay protocols.

1 INTRODUCTION

Experience replay exposes a continual learner to historical examples while it acquires new classes. Retrieval methods prioritize examples using signals such as anticipated interference, whereas other approaches modify memory construction or the replay objective (Aljundi et al., 2019a;b; Rebuffi et al., 2017; Buzzega et al., 2020). These mechanisms usually treat an available historical observation as usable evidence. In multimodal applications, however, the observation itself can change: degraded audio can obscure an intent-bearing phrase, and automatic speech recognition (ASR) can propagate that change into the text modality. A difficult historical example may therefore reflect forgetting, altered evidence, or both.

These explanations have different implications for retrieval. A high-interference example may need rehearsal, but high predictive support does not establish that rehearsing it best protects historical knowledge. Conversely, an example with weak modality support can still contain useful constraints. We ask how observation changes affect replay, and whether adding predictive support to an interference-based selector improves retention, new-task acquisition, or neither. Figure 1 summarizes the setting.

Our controlled study yields three contributions. First, a current/replay factorial intervention isolates a substantial replay-observation effect that stream-level comparisons can conceal. Second, we formalize RAM-CIL as a retrieval-only correction using independent modality heads, temperature scaling, candidate-wise ranks, and a fixed replay budget. Third, we report contrasting stability–plasticity outcomes, including unfavorable and protocol-dependent results. In particular, full-sequence additive correction and a two-stage paired-reference diagnostic improve different acquisition measures without consistently reducing forgetting. A favorable retention result in checkpoint continuation

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: Study overview. Paired clean and shifted observations retain sample identities and labels, allowing current-stream and replay effects to be separated. The retrieval block depicts RAM-CIL's additive score. The inference panel illustrates prediction over all seen intents; evaluation does not supply a task identity. Arrows express the experimental workflow, not an established causal mechanism for gradient interference.

prevents a stronger claim that reliability can never help stability. The central finding is consequently a distinction between predictive reliability and replay utility, rather than a general anti-forgetting guarantee.

2 Related Work

Replay in continual learning. Maximally Interfered Retrieval (MIR) prioritizes memories whose loss increases after a virtual current-data update (Aljundi et al., 2019a). Gradient-based sample selection controls memory contents (Aljundi et al., 2019b); iCaRL combines exemplar memory with incremental recognition (Rebuffi et al., 2017); and Dark Experience Replay retains historical predictions in its objective (Buzzega et al., 2020). Our intervention concerns retrieval. A matched comparison therefore holds the candidate generator, memory identities, replay count, and learner objective fixed within each experiment. Results obtained with different objectives or checkpoints are reported separately.

Multimodal confidence and predictive support. Trusted multi-view classification models view-specific evidence (Han et al., 2021), and dynamic fusion studies learning from modalities of differing quality (Zhang et al., 2023). Such signals address prediction and fusion; they do not by themselves identify which rehearsal examples preserve past knowledge. True-class probability is an established label-conditioned confidence quantity (Corbière et al., 2019). We adopt this quantity with temperature scaling (Guo et al., 2017), without claiming novelty for either component. Because historical labels are available during replay, support can be evaluated directly. It remains distinct from an independently validated measure of corruption severity.

3 Setting and Controlled Observation

Task $t \in \{1, \ldots, T\}$ introduces disjoint classes $\mathcal{Y}_t^{\text{new}}$, and $\mathcal{Y}_t = \bigcup_{\tau \le t} \mathcal{Y}_\tau^{\text{new}}$ is the seen-class set. A labeled observation is $x_i = (x_i^a, x_i^s, y_i)$, with audio a and transcript s. The learner predicts over

Table 1: Current/replay factorial intervention from a shared clean task-three checkpoint (seed 42). The entire historical buffer is assigned its specified observation version in tasks 4–6. F1 is final pooled Macro-F1; P_{new}, Ret_{old}, and F_{old} use Eqs. (1)–(2).

	Current	Replay	F1 ↑	P_{new} ↑	Ret_{old} ↑	F_{old} ↓
A	Clean	Clean	31.67	74.28	32.42	40.36
B	Shift	Clean	33.80	48.61	47.17	25.62
C	Clean	Shift	20.76	72.73	16.53	56.26
D	Shift	Shift	30.99	70.36	35.66	37.13

$\mathcal{Y}_t$. At a training update, $\mathcal{B}_t$ denotes current examples, $\mathcal{M}_{t-1}$ historical memory, and $\mathcal{C}_t \subseteq \mathcal{M}_{t-1}$ a candidate set of size n_t. The stage index t is reused across its training updates.

Definition 1 (Replay-observation intervention)**.** *A paired replay intervention replaces* $(x_i^{a,0}, x_i^{s,0}, y_i)$ *by* $(x_i^{a,c}, x_i^{s,c}, y_i)$ *for the same historical sample identity* i*, while holding the current batch, historical labels, memory identities, initialization, and training budget fixed. Audio-only shift changes* x_i^a*; coupled shift also changes its associated ASR transcript.*

Experimental scope. We use the 30 in-scope intents of MIntRec2.0 (Zhang et al., 2024), arranged into six tasks of five classes under a fixed order (order 0). The study uses fixed audio/text feature caches and a concatenation classifier, 10 epochs per task, current batch size 32, and a memory allocation of 20 examples per historical class. Evaluation uses clean observations, including a pooled test set of 2,033 examples. The shifted condition is the recorded Gaussian10 audio condition and, for coupled shift, its paired ASR text. We distinguish acquisition-time stream shift from replacement of the entire historical buffer; they are different interventions. Full details and per-seed results appear in Appendix A.

Separate stability and plasticity metrics. Let $Q_{t,j}$ be the reported task-wise Macro-F1 on task j after learning task t, on a 0–100 scale. Tasks 1–3 constitute the old-task group; tasks 4–6 constitute the acquisition group. We report

$$P_{\text{new}} = \tfrac{1}{3}\sum_{j=4}^{6} Q_{j,j}, \qquad \text{Ret}_{\text{old}} = \tfrac{1}{3}\sum_{j=1}^{3} Q_{6,j}, \tag{1}$$

$$F_{\text{old}} = \tfrac{1}{3}\sum_{j=1}^{3}\left(\max_{k=j,\ldots,6} Q_{k,j} - Q_{6,j}\right), \qquad \text{BWT}_{\text{old}} = \tfrac{1}{3}\sum_{j=1}^{3}(Q_{6,j} - Q_{j,j}). \tag{2}$$

P_{new} measures performance immediately after each new task, not learning speed or the optimizer learning rate. Final Macro-F1 is computed separately from pooled predictions over all 30 classes; it is not the arithmetic mean of task-wise F1. Forgetting and retention also differ: forgetting depends on the preceding performance trajectory. All changes below are percentage points (pp), unless specified otherwise.

Observation changes have distinct effects. Condition C versus A in Table 1 isolates replay shift: old-task retention decreases by 15.90 pp, whereas new-task acquisition decreases by 1.55 pp. This establishes a substantial effect of historical observation quality within this controlled run. It does not establish that every form of shift increases forgetting. Condition B instead lowers acquisition and raises retention. The retention interaction $(D - B) - (C - A)$ is $+4.39$ pp, further showing that current and historical shifts cannot be treated as independent, interchangeable perturbations. Contrasts use the precision retained in the experiment summaries; displayed endpoints are rounded.

4 RAM-CIL: Reliability-Aware Retrieval

RAM-CIL denotes *Reliability-Aware Multimodal Class-Incremental Learning*. Its core operation combines an existing retrieval priority with predictive support, while leaving the selected observations and learner objective unchanged (Figure 2). The base scorer is an explicit input, not an additional latent risk variable.

Algorithm 1 RAM-CIL update at task stage t

Inputs: learner (θ, ω); current batch $\mathcal{B}_t$; historical memory $\mathcal{M}_{t-1}$; candidate generator; base scorer $\mathcal{I}$; independent heads and temperatures; budgets n, b; coefficients γ, λ.
Output: updated learner and memory under the base storage policy.

1: If memory is empty, apply $\mathrm{OptStep}(\theta, \omega, \nabla_\theta \mathcal{L}_{\mathrm{cur}})$ and invoke the base storage policy; return.
2: Obtain $\mathcal{C}_t$ using the base candidate generator; set $b_t = \min(b, |\mathcal{C}_t|)$.
3: Compute $I_{i,t}$ for all $i \in \mathcal{C}_t$ using $\mathcal{I}$, including its internal virtual update when required; retain the original learner state.
4: In evaluation mode, compute $r^a_{i,t}, r^s_{i,t}, \widehat{R}_{i,t}$ by Eq. (3), without a gradient graph.
5: Compute $\bar{I}_{i,t}, \bar{R}_{i,t}$ and $\mathrm{stopgrad}(S_{i,t})$ by Eqs. (4)–(5). Retrieve the top b_t candidates, resolving ties by sample identity.
6: In training mode, evaluate Eq. (6) with b_t observed replay examples, without feature weighting. Set $(\theta^+, \omega^+) = \mathrm{OptStep}(\theta, \omega, \nabla_\theta \mathcal{L})$.
7: Invoke the base memory-update policy at its configured update point; return the updated state.

Definition 2 (Label-conditioned modality predictive support)**.** *For modality $m \in \{a, s\}$, let $z^m_i = E_m(x^m_i)$ and $u^m_{i,t} = h_{\phi_{m,t}}(z^m_i) \in \mathbb{R}^{|\mathcal{Y}_t|}$ be logits from an independently supervised modality head. With temperature $T_{m,t} > 0$, define*

$$q^m_{i,t} = \mathrm{softmax}(u^m_{i,t}/T_{m,t}), \qquad r^m_{i,t} = q^m_{i,t}(y_i), \qquad \widehat{R}_{i,t} = \tfrac{1}{2}(r^a_{i,t} + r^s_{i,t}). \tag{3}$$

The temperature minimizes cross-entropy on a clean, seen-class development set at the configured calibration points and remains fixed between them (Appendix B). The heads are not a fusion classifier evaluated with one modality zeroed out. The identity $r^m_{i,t} = \exp(-\ell^{m,\mathrm{cal}}_{i,t})$ follows directly from cross-entropy; it is not independent evidence of quality estimation. Labels are used only for training-time retrieval, never supplied at inference.

A correction to an explicit base score. Let $I_{i,t} = \mathcal{I}(\theta_t, \mathcal{B}_t, \mathcal{C}_t; i)$, with larger values indicating greater baseline priority. MIR instantiates $\mathcal{I}$ using the increase in historical loss after its virtual current-batch update (Aljundi et al., 2019a). We retain that scorer and its virtual-update configuration in the matched baseline and corrected method. For a candidate vector v, let $\mathcal{N}_{\mathcal{C}_t}(v_i)$ be its ascending midrank, shifted and scaled to $[0, 1]$; when $n_t = 1$, set it to $1/2$. The complete tie-aware definition appears in Appendix B. Set

$$\bar{I}_{i,t} = \mathcal{N}_{\mathcal{C}_t}(I_{i,t}), \qquad \bar{R}_{i,t} = \mathcal{N}_{\mathcal{C}_t}(\widehat{R}_{i,t}), \tag{4}$$

$$S_{i,t} = \bar{I}_{i,t} - \gamma(1 - \bar{R}_{i,t}), \qquad \mathcal{R}_t = \mathrm{Top}_{b,\, i \in \mathcal{C}_t} S_{i,t}, \quad \gamma \geq 0. \tag{5}$$

Retrieval uses descending score and ascending sample identity for ties. This ranks predictive support jointly with interference; it does not estimate a candidate's causal contribution to retention.

Proposition 1 (Baseline recovery)**.** *Using the same tie rule, $\gamma = 0$ recovers the base selector exactly. For any fixed γ, selecting by $S_{i,t}$ is equivalent to selecting by $\bar{I}_{i,t} + \gamma \bar{R}_{i,t}$.*

The first claim follows because midranks preserve the ordering and ties of I; the second follows by removing the candidate-independent constant $-\gamma$.

Unchanged learner. With fusion model f_θ and baseline replay coefficient λ, the objective is

$$\mathcal{L}(\theta) = \underbrace{\frac{1}{|\mathcal{B}_t|} \sum_{j \in \mathcal{B}_t} \mathrm{CE}(f_\theta(x^a_j, x^s_j), y_j)}_{\mathcal{L}_{\mathrm{cur}}} + \lambda \underbrace{\frac{1}{b} \sum_{i \in \mathcal{R}_t} \mathrm{CE}(f_\theta(x^a_i, x^s_i), y_i)}_{\mathcal{L}_{\mathrm{rep}}}. \tag{6}$$

All selection scores and indices are detached from learner backpropagation. This does not prohibit gradients inside MIR's own virtual update. Modality support is evaluated without constructing a gradient graph, and the original optimizer state is retained for the actual learner step.

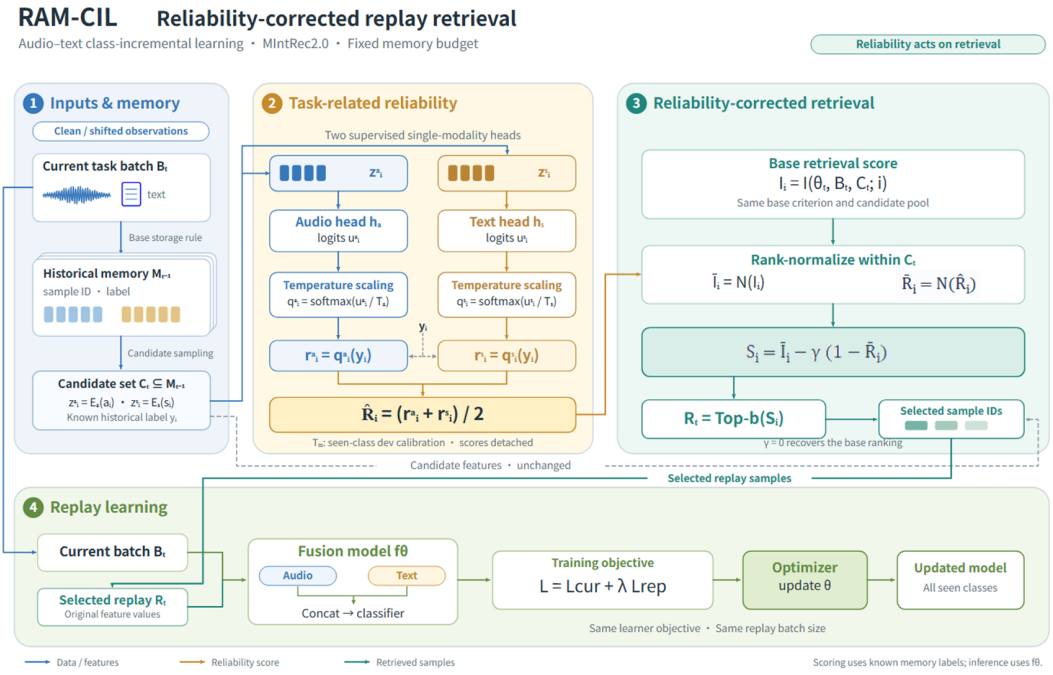

Figure 2: RAM-CIL retrieval and learner update. Independent modality heads provide temperature-scaled true-class support; candidate-wise rank normalization makes its correction comparable to the base priority. Scores determine discrete retrieval only. The fusion learner consumes the selected observed audio/text features without reliability weighting.

A separate privileged diagnostic. To distinguish predictive support from information about paired degradation, we also evaluate a reference-assisted score R_i^{pair} computed from paired clean/shifted observations. It is an oracle diagnostic, not the deployable score in Eq. (3). Besides additive oracle retrieval, a two-stage variant first retains the k highest MIR priorities and then chooses b by R^{pair}:

$$\mathcal{H}_t = \text{Top}_{k,\, i\in\mathcal{C}_t} I_{i,t}, \qquad \mathcal{R}_t^{\text{two}} = \text{Top}_{b,\, i\in\mathcal{H}_t} R_i^{\text{pair}}, \qquad (n,k,b) = (80,16,8). \tag{7}$$

Clean references affect scoring only; they are not substituted into learner inputs. The predictive-support method and this privileged diagnostic are kept distinct in every result table.

5 Results

Matched comparisons and selection. The method comparisons use seed 42, candidate count 80, replay count 8, and $\lambda = 1$. For additive predictive support, $\gamma = 0.5$ was chosen over 1.0 using the continuation development results and then fixed for the six-task run. Continuation starts from the identical clean task-three checkpoint; full-sequence runs train from task one and therefore acquire different histories. Both use clean current inputs and shifted replay during the intervention stages. Table 2 separates these protocols. Random ER in Table 1 used a mixed-batch mean objective and a variable replay count; its absolute values are not an equal-budget baseline for Table 2.

Predictive support does not give a uniform retention benefit. Over all six tasks, additive support improves final Macro-F1 by 3.51 pp and acquisition by 1.75 pp, but reduces retention by 3.47 pp and increases forgetting by 9.59 pp. A higher final pooled score therefore does not establish successful old-task preservation. The continuation result is a necessary qualification: from the shared task-three state, additive support raises retention by 9.65 pp and final F1 by 1.96 pp while slightly reducing acquisition (-0.25 pp). Predictive support can aid retention in this controlled continuation; it does not reliably identify the best stability–plasticity trade-off across protocols.

Table 2: Retrieval results under two distinct protocols (seed 42). Additive support is RAM-CIL with $\gamma = 0.5$. Paired oracle methods use privileged clean references for scoring. The common MIR row within each panel is the matched baseline. No method uniformly dominates stability and plasticity.

Selector	Final F1 ↑	P_{new} ↑	Ret_{old} ↑	F_{old} ↓
Full sequence: tasks 1–6				
MIR	18.66	68.89	16.56	50.40
Additive support	22.18	70.64	13.09	59.98
Additive paired oracle	18.59	67.08	20.64	54.01
Continuation: shared task-three state, tasks 4–6				
MIR	17.75	68.88	17.76	55.02
Additive support	19.71	68.63	27.41	45.38
MIR top-16 → paired-oracle top-8	19.11	75.79	17.56	55.23

Paired-reference filtering separates acquisition from retention. The two-stage oracle diagnostic increases P_{new} from 68.88 to 75.79 ($+6.91$ pp), while retention changes from 17.76 to 17.56 (-0.21 pp). Final F1 rises by 1.36 pp. Thus, restricting MIR's high-interference pool using paired-reference information can improve acquisition with a small observed retention change in this run. This is not a statistical equivalence result, and it does not demonstrate the same gain for predictive-support filtering. The full-sequence additive oracle further illustrates why retention and forgetting must both be reported: retention increases by 4.08 pp, but forgetting also increases by 3.62 pp because earlier attained performance differs. The oracle does not establish uniform forgetting reduction.

Predictive support is not an input-quality meter. Across 1,200 historical stage–sample pairs, clean observations have *lower* mean combined support than shifted ones by 0.00288, and only 35.92% have higher clean support. Across 6,165 training pairs at task six, the mean clean advantage is 0.00253, but the median is negative and only 43.78% favor clean inputs. These audits support the label-conditioned interpretation of Eq. (3); they do not validate it as a monotone corruption measure. Appendix F provides modality-specific results and the limited held-out audit.

Negative controls and broader observations. Direct modality feature weighting and its norm-preserving variant reduce retention from 16.53 to 10.52 and 7.27, respectively, while improving acquisition (Appendix E). Preserving the concatenated feature norm therefore does not rescue that intervention. In a separate three-seed stream-shift study, acquisition declines while retention increases; sequential training without replay exhibits a retention floor. These results reinforce the need to separate current-input difficulty, replay versioning, and retrieval behavior, rather than pooling unlike experiments into one claim (Appendix C).

6 Discussion and Limitations

The evidence distinguishes three quantities: label-conditioned predictive support, interference under a virtual current update, and the eventual utility of replay for a chosen objective. Neither the support definition nor the rank correction equates them. The two-stage result is consistent with reducing a replay-related impediment to acquisition, but gradient conflict and faster convergence were not directly measured; the mechanism remains an interpretation. In particular, endpoint P_{new} cannot establish a higher learning rate or a steeper learning curve.

The method comparisons cover one seed, one class order, one dataset, and one shifted condition. Three-seed replication applies to the stream-level observations, not to the retrieval variants. Those small-sample comparisons do not establish statistical significance. The paired-reference selector also requires privileged observations unavailable to ordinary deployment. We therefore draw conclusions about the reported controlled setting and preserve the positive retention counterexample, rather than claiming universal failure or success of reliability-aware replay.

7 CONCLUSION

Changing historical multimodal observations can substantially alter replay even when current inputs remain clean. RAM-CIL makes the role of predictive support explicit, but its outcomes depend on the training protocol. Full-sequence correction, checkpoint continuation, and paired-reference filtering affect acquisition and retention differently. Reliability information should consequently be evaluated against separate stability and plasticity endpoints; it cannot be assumed to measure forgetting vulnerability or anti-forgetting replay utility.

AI USE DISCLOSURE

Generative AI tools were used in this work for literature discovery and retrieval, research ideation, refinement of hypotheses, discussion of experimental design, assistance with implementation and debugging, interpretation of experimental results, and drafting and language polishing of parts of the manuscript. AI-assisted suggestions were treated as provisional and were independently checked against experiments, source literature, and the authors' own analysis before being incorporated into the work.

All experimental configurations, model implementations, reported numerical results, statistical analyses, citations, scientific claims, and final manuscript content were reviewed and verified by the authors. Generative AI tools were not used to generate synthetic experimental datasets or as an authority for mathematical or empirical claims. The authors take full responsibility for the content of this submission.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/15825aee15eb335cc13f9b559f166ee8-Abstract.html.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019b. URL https://proceedings.neurips.cc/paper/2019/hash/e562cd9c0768d5464b64cf61da7fc6bb-Abstract.html.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/b704ea2c39778f07c617f6b7ce480e9e-Abstract.html.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2102.02051.

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*, pp. 2001–2010, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Rebuffi_iCaRL_Incremental_Classifier_CVPR_2017_paper.html.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ca97c1a8c52889c49e16497912244c3b-Abstract-Conference.html.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. arXiv preprint arXiv:2306.02050, 2023. URL https://arxiv.org/abs/2306.02050.

A Experimental Protocols and Metric Conventions

A.1 Three protocols that should not be pooled

Table 3 summarizes the experimental units. All use the same six-task, 30-intent setting and clean evaluation, but intervention timing and training histories differ. The per-class memory allocation is fixed at 20; the total memory grows as classes are introduced. Thus, historical memory contains 300, 400, and 500 examples before tasks 4, 5, and 6, respectively. A fixed replay count should not be confused with a fixed total memory capacity.

Table 3: Protocol distinctions used throughout the paper. Coupled shift denotes the paired shifted audio and ASR observations.

Protocol	Observation intervention	Shared state and replication
Stream shift	Tasks 1–3 clean; tasks 4–6 use clean, audio-only, or coupled current observations. Memory retains the version observed when a sample was acquired.	Clean prefix shared within each seed; seeds 42, 43, 44; Random ER and sequential training.
Factorial continuation	Current and replay versions crossed independently during tasks 4–6. The entire historical buffer is assigned its clean or coupled-shift version at each stage.	Identical clean task-three checkpoint, memory identities, batch order, and update counts; seed 42.
Retrieval comparison	Matched MIR candidates and replay budget. Clean current inputs, shifted replay in intervention stages.	Either shared task-three continuation or independent full six-task histories; seed 42.

The factorial replacement applies to all currently historical samples, including task-four examples once task five begins and task-five examples once task six begins. It therefore differs from the acquisition-time stream protocol, even when both are informally described as "current shifted/replay shifted." Clean and shifted observations are linked by sample identity and retain the original intent label. This construction controls label identity; it does not by itself establish that every corrupted observation preserves human-recognizable semantics.

The learner operates on fixed audio and text feature caches with modality projections and concatenation-based classification. Within a comparison, feature caches, learner architecture, initialization protocol, training schedule, and evaluation code are held fixed. The study is an audio–text subset of the multimodal benchmark; it does not evaluate its visual modality or out-of-scope detection task. Gaussian10 refers to the archived noise condition used to construct the shifted feature and ASR caches, rather than a sweep over noise severities.

A.2 Task-wise and pooled metrics

The matrix Q contains the task-wise scores returned by the experiment evaluator. Its diagonal quantifies immediate task acquisition; its final row describes task-wise performance after task six. The pooled Macro-F1 is computed from pooled predictions across all 30 labels. Restricting the evaluation examples to a task changes the confusion statistics for F1; a mean of separately evaluated task scores need not equal the pooled score. No task identity is an input to the fusion classifier.

Alongside Eqs. (1)–(2), standard backward transfer averages over all five historical tasks at the final stage:

$$\mathrm{BWT} = \frac{1}{5}\sum_{j=1}^{5}(Q_{6,j} - Q_{j,j}). \tag{8}$$

This differs from $\mathrm{BWT}_{\mathrm{old}}$, which covers only tasks 1–3. When each old task attains its maximum on its diagonal, $\mathrm{BWT}_{\mathrm{old}} = -F_{\mathrm{old}}$. This equality holds for the reported old-task summaries, but is not imposed by definition. Across independently trained methods, different historical peaks can produce higher final retention *and* higher forgetting.

All F1, retention, acquisition, and backward-transfer values are reported on a 0–100 scale. Changes in these metrics are pp. Support scores and their paired differences remain on their original probability scale. Tables reproduce two-decimal experiment summaries; source-reported contrasts can

differ by 0.01 from subtraction of rounded entries. The machine-readable tables accompanying the manuscript preserve this distinction. No missing seed is imputed and no standard deviation is assigned to a single run.

B Complete Mathematical Specification

B.1 Calibration and candidate normalization

For each independently trained modality head, the clean calibration set $\mathcal{V}_t$ contains development examples from the seen labels $\mathcal{Y}_t$. The temperature is

$$T_{m,t} \in \arg\min_{T>0} \frac{1}{|\mathcal{V}_t|} \sum_{j \in \mathcal{V}_t} -\log \left[\mathrm{softmax}\left(u^m_{j,t}/T\right)\right]_{y_j}. \tag{9}$$

Heads and temperatures are fixed between the configured stage-wise scoring checkpoints and are reused across matched selector comparisons. Test labels are not used for calibration or coefficient selection. The normalized probability for class $k \in \mathcal{Y}_t$ is

$$q^m_{i,t}(k) = \frac{\exp(u^m_{i,t,k}/T_{m,t})}{\sum_{c \in \mathcal{Y}_t} \exp(u^m_{i,t,c}/T_{m,t})}, \qquad \ell^{m,\mathrm{cal}}_{i,t} = -\log q^m_{i,t}(y_i). \tag{10}$$

Thus $r^m_{i,t} = \exp(-\ell^{m,\mathrm{cal}}_{i,t})$ is high when that modality's model supports the known replay label. It is not a label-free probability of correctness, a noise severity estimate, or a measured forgetting risk.

For any candidate score vector $v = (v_j)_{j \in \mathcal{C}_t}$, the ascending midrank is

$$\mathrm{Rank}_{\mathcal{C}_t}(v_i) = 1 + \sum_{j \in \mathcal{C}_t} \mathbf{1}[v_j < v_i] + \frac{\sum_{j \in \mathcal{C}_t} \mathbf{1}[v_j = v_i] - 1}{2}. \tag{11}$$

Define

$$\mathcal{N}_{\mathcal{C}_t}(v_i) = \begin{cases} \big(\mathrm{Rank}_{\mathcal{C}_t}(v_i) - 1\big)/(n_t - 1), & n_t > 1, \\ 1/2, & n_t = 1. \end{cases} \tag{12}$$

Normalization is performed within the same candidate set for both signals. It preserves order, handles ties symmetrically, and avoids identifying numerical loss increments with probabilities. It does not make the resulting combined score a calibrated probability. If all support scores are equal, their correction is constant and the base ordering is recovered for any γ.

B.2 Selection, gradients, and state

The selection computations are

$$I^{\mathrm{sel}}_{i,t} = \mathrm{stopgrad}(I_{i,t}), \qquad r^{m,\mathrm{sel}}_{i,t} = \mathrm{stopgrad}(r^m_{i,t}), \qquad S^{\mathrm{sel}}_{i,t} = \mathrm{stopgrad}(S_{i,t}). \tag{13}$$

The discrete top-b indices do not receive learner gradients. In contrast, the base scorer can internally differentiate its loss or construct a virtual update as required by MIR. That temporary state is not substituted for the actual optimizer state. With optimizer state ω, the learner update is

$$(\theta^+, \omega^+) = \mathrm{OptStep}\big(\theta, \omega, \nabla_\theta[\mathcal{L}_{\mathrm{cur}} + \lambda \mathcal{L}_{\mathrm{rep}}]\big). \tag{14}$$

Only observed selected inputs enter the replay loss. RAM-CIL does not multiply the features or loss by $\widehat{R}$. An empty memory produces a current-only step. For a nonempty candidate set smaller than the requested replay budget, Algorithm 1 uses $b_t = \min(b, n_t)$. Memory construction and its update timing remain those of the base experiment.

Two candidate-wise sorts for rank normalization and top-b selection add $O(n_t \log n_t)$ ranking work. Support additionally requires two modality-head evaluations per candidate. This accounting excludes the unchanged base scorer and does not constitute a measured wall-clock comparison.

C THREE-SEED STREAM-SHIFT RESULTS

C.1 AGGREGATE EFFECTS

Table 4 reports sample means and sample standard deviations across seeds 42, 43, and 44. Both stream shifts reduce immediate acquisition but increase old-task retention. This is not evidence that corrupted historical inputs intrinsically protect memory: current-input changes alter learning, and acquisition-time memory versioning differs from the factorial replacement. The joint intervention changes the stability–plasticity balance.

Table 4: Random ER under stream shift, mean ± sample standard deviation over three seeds. Standard BWT uses all five historical tasks.

Stream	$P_{\text{new}} \uparrow$	$\text{Ret}_{\text{old}} \uparrow$	$F_{\text{old}} \downarrow$	BWT ↑
Clean	75.00 ± 1.55	30.92 ± 2.11	41.29 ± 2.74	-37.64 ± 0.53
Audio shift	53.66 ± 2.40	54.31 ± 0.42	17.89 ± 0.67	-22.55 ± 3.08
Coupled shift	51.08 ± 1.77	55.15 ± 2.05	17.05 ± 1.27	-22.31 ± 1.71

The mean final task-wise score over tasks 4–6 is 53.54 for clean, 33.97 for audio shift, and 30.94 for coupled shift. These are final acquisition-group scores, distinct from the diagonal-based P_{new}. Pooled final Macro-F1 changes relative to clean are -1.9502 pp for audio shift and -2.9489 pp for coupled shift. These pooled reductions coexist with improved task-wise retention.

With three paired seeds, exact two-sided sign-flip tests for the reported acquisition, retention, forgetting, and BWT contrasts give $p = 0.25$; the reported Holm-adjusted values are 1.00. We report consistent directions and effect sizes without claiming statistical significance.

Table 5: Per-seed Random ER stream results. In these runs $\text{BWT}_{\text{old}} = -F_{\text{old}}$.

Seed	Stream	P_{new}	Ret_{old}	F_{old}	BWT
42	Clean	74.28	32.42	40.36	−37.09
42	Audio shift	55.91	54.58	18.21	−26.11
42	Coupled shift	52.87	54.99	17.79	−23.72
43	Clean	73.94	28.50	44.37	−37.68
43	Audio shift	53.94	54.52	18.35	−20.92
43	Coupled shift	51.03	57.28	15.59	−20.42
44	Clean	76.77	31.82	39.13	−38.16
44	Audio shift	51.13	53.83	17.12	−20.63
44	Coupled shift	49.34	53.19	17.76	−22.80

C.2 POOLED-SCORE DECOMPOSITION

For pooled evaluation, old and new groups each contain 15 classes. Their contributions to the 30-class Macro-F1 difference are therefore one half of the respective groupwise class-averaged F1 changes. Table 6 reports these contributions. They must not be equated with the task-wise Ret_{old} and P_{new}: both the evaluation aggregation and, for P_{new}, the evaluation time differ.

C.3 SEQUENTIAL-TRAINING FLOOR

Table 7 records the no-replay control. All old-task retention scores are zero. Since retention is at its lower bound, a small clean/shift difference cannot establish robustness to replay reliability or to corruption generally. These results are retained as a floor-effect control, rather than used to support the retrieval method.

Table 6: Contributions to pooled final Macro-F1 change versus the clean stream (pp). Small summation discrepancies reflect rounding.

Seed	Shift	Old-class contribution	New-class contribution	Total change
42	Audio	2.0105	−5.1058	−3.0953
43	Audio	3.4930	−4.1751	−0.6821
44	Audio	1.6079	−3.6810	−2.0731
Mean	Audio	2.3705	−4.3206	−1.9502
42	Coupled	2.0152	−4.5891	−2.5739
43	Coupled	3.6866	−6.0983	−2.4116
44	Coupled	2.2471	−6.1083	−3.8612
Mean	Coupled	2.6496	−5.5985	−2.9489

Table 7: Sequential training without replay. Old-task retention is zero in every run.

Seed	Stream	P_{new}	Ret_{old}	F_{old}	BWT
42	Clean	71.57	0.00	73.74	−73.50
42	Audio shift	71.58	0.00	73.74	−73.38
42	Coupled shift	72.06	0.00	73.74	−75.90
43	Clean	76.88	0.00	71.66	−74.47
43	Audio shift	75.60	0.00	71.66	−74.25
43	Coupled shift	74.47	0.00	71.66	−72.24
44	Clean	70.78	0.00	70.10	−70.36
44	Audio shift	75.51	0.00	70.10	−73.29
44	Coupled shift	76.50	0.00	70.10	−72.84

D Factorial Intervention and Task Trajectories

D.1 Contrasts

Table 8 expands the controlled comparison in the main text. The current-shift effect with clean replay (B–A) has a different direction from the replay-shift effect with clean current inputs (C–A). D–A changes both streams and cannot isolate either pathway. The interaction contrasts the replay effect under shifted and clean current inputs.

Table 8: Source-reported factorial contrasts (pp), computed before display rounding. BWT covers all five historical tasks.

Contrast	Accuracy	Final F1	P_{new}	Ret_{old}	F_{old}	BWT
C–A	−8.31	−10.91	−1.55	−15.90	15.90	−15.93
B–A	−0.69	2.13	−25.67	14.74	−14.74	21.59
D–A	−1.13	−0.68	−3.92	3.23	−3.23	2.39
D–B–C+A	7.87	8.10	23.29	4.39	−4.39	−3.28

Final accuracies are 35.56, 34.87, 27.25, and 34.43 for A–D, respectively. Corresponding standard BWT values are −37.09, −15.50, −53.02, and −34.70. The old-task BWT values equal the negative forgetting scores in Table 1.

D.2 Recorded task matrices

Table 9 gives the complete shared prefix and all continuation rows. The dash denotes a task that has not yet been introduced, rather than an unreported experiment. The accompanying `data/task_matrices.csv` also includes all three-seed ER stream matrices.

Table 9: Task-wise Macro-F1 matrix for the factorial intervention. A common clean prefix is used by all four conditions. Columns are evaluation tasks; rows are the last training task.

Condition	After task	T1	T2	T3	T4	T5	T6
Shared	1	72.87	—	—	—	—	—
Shared	2	50.28	76.84	—	—	—	—
Shared	3	49.29	44.41	68.65	—	—	—
A	4	34.63	34.80	46.88	78.51	—	—
A	5	30.29	44.18	46.06	57.12	80.14	—
A	6	25.77	32.42	39.08	37.23	57.05	64.17
B	4	38.25	50.53	53.78	63.54	—	—
B	5	43.63	52.64	59.92	58.63	48.18	—
B	6	40.11	51.43	49.96	46.94	64.14	34.10
C	4	24.64	27.25	29.87	78.76	—	—
C	5	24.82	18.12	29.03	20.99	67.36	—
C	6	14.28	12.66	22.64	11.94	37.88	72.07
D	4	41.70	42.39	46.54	74.67	—	—
D	5	36.47	47.79	53.31	55.53	76.01	—
D	6	31.20	39.46	36.32	36.92	51.64	60.39

Table 10: Final task-wise ER scores in the three-seed stream protocol. These final rows do not define P_{new}, which uses diagonal scores.

Seed	Stream	T1	T2	T3	T4	T5	T6
42	Clean	25.77	32.42	39.08	37.23	57.05	64.17
42	Audio	46.92	52.27	64.55	17.15	32.91	41.75
42	Coupled	45.34	57.24	62.39	22.93	32.64	37.81
43	Clean	20.72	33.21	31.58	43.69	57.05	65.79
43	Audio	48.96	55.54	59.05	21.07	41.49	49.71
43	Coupled	48.69	61.87	61.29	12.21	40.38	45.20
44	Clean	31.45	27.40	36.62	34.90	54.23	67.78
44	Audio	48.91	50.78	61.78	18.69	37.36	45.56
44	Coupled	45.77	52.27	61.52	7.97	37.12	42.19

E Feature-Weighting Negative Controls

These earlier controls act on replay representations and are distinct from the independent-head selector in Section 4. Let z_a, z_s denote the projected features and C the fusion classifier. Their support estimates are obtained by masking one modality:

$$p_a = \mathrm{softmax}(C(z_a, 0))_y, \qquad p_s = \mathrm{softmax}(C(0, z_s))_y, \qquad \alpha_m = \frac{p_m}{p_a + p_s}. \tag{15}$$

Direct weighting uses the concatenation $w = [\alpha_a z_a; \alpha_s z_s]$. The norm-preserving version forms $v = [2\alpha_a z_a; 2\alpha_s z_s]$ and rescales it to the original concatenation norm:

$$z^{\mathrm{NP}} = v \frac{\|[z_a; z_s]\|_2}{\|v\|_2} \tag{16}$$

for nonzero v, with a zero-vector guard. The common factor two cancels after exact renormalization. Weights are detached, and only replay features are transformed. Current, development, and test inputs retain the unweighted representation.

Table 11: Replay-only feature weighting from the shared clean task-three checkpoint under condition C (seed 42). These controls use the ER protocol, not the fixed-eight MIR objective.

Replay representation	Final F1	P_{new}	$\mathrm{Ret}_{\mathrm{old}}$	F_{old}	$\mathrm{BWT}_{\mathrm{old}}$
Unweighted	20.76	72.73	16.53	56.26	−56.26
Direct weighting	13.69	76.11	10.52	62.26	−62.26
Norm-preserving weighting	14.67	77.67	7.27	65.52	−65.52

Across 12,000 replay uses, the norm-preserving implementation records a maximum relative norm error of 1.65×10^{-7} and norm ratios in $[0.999999835, 1.000000165]$. Nevertheless, retention falls further than with direct weighting. A reduction in the overall concatenated feature norm is therefore insufficient to explain the unfavorable retention result. Renormalization does not control feature direction, relative modality contributions, or the gradient of the classifier, so this control does not exclude all optimization effects.

F Audit of Label-Conditioned Predictive Support

The audit reuses the saved heads and temperatures; it introduces no head retraining or test-based tuning. For each paired observation, define $\Delta r^m = r^{m,\mathrm{clean}} - r^{m,\mathrm{shift}}$ and $\Delta \widehat{R} = \widehat{R}^{\mathrm{clean}} - \widehat{R}^{\mathrm{shift}}$. Positive differences favor clean inputs. Historical stage–sample pairs may repeat an identity across stages and should not be interpreted as 1,200 independent subjects.

Table 12: Paired clean-minus-shift support differences. The final column is the percentage of pairs with a strictly positive difference. Training pairs are evaluated at task six.

Set	Modality	Mean	Std. dev.	Median	Positive (%)
Training (6,165)	Audio	0.00136	0.01068	−0.00237	36.40
	Text	0.00371	0.03759	−0.00018	48.11
	Combined	0.00253	0.02082	−0.00085	43.78
Historical (1,200)	Audio	−0.00242	0.00673	−0.00225	26.50
	Text	−0.00334	0.03246	−0.00219	38.58
	Combined	−0.00288	0.01696	−0.00208	35.92
Held-out (30)	Audio	−0.00147	0.01450	−0.00440	30.00
	Text	0.00869	0.04571	0.00144	56.67
	Combined	0.00361	0.02670	−0.00145	43.33

Predictive-correctness discrimination on the task-six training pairs is strong: audio AUROC is 0.984 for clean and 0.978 for shifted inputs; text AUROC is 0.997 and 0.994. These are in-sample diagnostics of the relationship between true-class support and correctness, not held-out proof of input-quality estimation. The 30-pair held-out check is too small for a strong generalization claim; its

audio results contain only three correct clean predictions and two correct shifted predictions. A high AUROC on such a set is not robust evidence. The algebraic relation between support and cross-entropy is likewise not a separate empirical validation.

G Retrieval Ablations and Selection Behavior

G.1 Development selection and continuation

At the end of continuation, development Macro-F1 is 15.93 for MIR, 19.41 for additive support with $\gamma = 0.5$, and 17.26 with $\gamma = 1.0$. The value 0.5 is selected and reused in full-sequence evaluation without another coefficient search. This selection procedure uses final development performance, not a criterion specifically optimized for retention.

The continuation comparison records 80,000 candidate evaluations and 8,000 replay selections per selector, with 80 candidates and eight selected examples per update. Mean candidate support is 0.07005. Selected support averages 0.06390 under MIR and 0.10367 under additive correction, and the reported selection-overlap rate is 18.86%. The selector therefore changes which examples are replayed in the expected score direction. This descriptive selection change does not by itself identify why retention improves in continuation. Mean historical support is 0.06668 on clean versions and 0.06956 on shifted versions, consistent with the negative paired audit mean.

G.2 Full-sequence results

Table 13: Additional full-sequence results for MIR and additive predictive support. Deltas retain the precision reported in the experiment summary. Average stage F1 is the recorded average of stage-level pooled scores.

Metric	MIR	Additive support	Reported change
Final accuracy	25.97	30.10	4.13
Final pooled Macro-F1	18.66	22.18	3.51
Average stage F1	33.67	35.43	1.77
P_{new}	68.89	70.64	1.75
Ret_{old}	16.56	13.09	-3.47
F_{old}	50.40	59.98	9.59
Standard BWT	-50.56	-48.58	1.98
BWT_{old}	-50.40	-59.98	-9.59

Standard BWT improves slightly, whereas old-group BWT deteriorates. These summaries weight different task groups and cannot be substituted for one another. Table 14 exposes the final task-wise distribution: later tasks 4 and 5 improve substantially under additive support, while tasks 1 and 2 deteriorate. This distribution explains why a pooled score alone is an inadequate retention assessment.

Table 14: Final task-wise Macro-F1 after complete six-task training.

Selector	T1	T2	T3	T4	T5	T6
MIR	14.69	17.36	17.64	15.38	25.11	64.56
Additive support	8.31	7.82	23.14	39.66	52.85	56.43

The reported overall overlap is 17.85%. Selecting higher-support examples is therefore an observable property of the method; uniform retention improvement is not. Full-sequence runs cannot be interpreted as continuations of the same task-three state, since the selector already changes training in tasks two and three.

G.3 Privileged paired-reference diagnostics

The paired-degradation score compares linked clean and shifted observations and is used only as an externally supplied diagnostic ranking. Its recorded range is 0.482–1.000; identical undegraded

Table 15: Full-sequence selection diagnostics. Candidate support is shared because the independently calibrated scoring heads and candidate identities are matched; learner interference scores can diverge as training proceeds. Overlap is the selection-overlap percentage reported by the audit.

Task	Replay selections	Candidate support	MIR selected	RAM-CIL selected	Overlap (%)
2	2,480	0.10287	0.09634	0.14096	13.31
3	1,680	0.08310	0.08982	0.11533	27.02
4	2,400	0.06670	0.06452	0.09651	16.50
5	2,400	0.07429	0.06906	0.11229	16.21
6	3,200	0.07835	0.06777	0.11709	18.81

pairs receive 1.000. Mean scores in the least- and most-degraded quintiles are 1.000 and 0.868, respectively. These checks establish the intended direction of that diagnostic, not calibrated ground-truth replay utility. The oracle experiments do not modify modality heads or temperatures, and clean reference features do not enter learner updates.

Over the full six-task run, additive paired-oracle retrieval changes final F1 by -0.07 pp, acquisition by -1.81 pp, retention by $+4.08$ pp, and forgetting by $+3.62$ pp relative to MIR. Its $\mathrm{BWT}_{\mathrm{old}}$ is -54.01. This result does not support a general acquisition or forgetting benefit from substituting a degradation-oriented score.

The MIR-first two-stage diagnostic is a separate task-three continuation, using the same checkpoint, memory identities, candidate sequences, batch order, saved heads, and temperatures as its MIR control. It narrows 80 candidates to 16 using MIR and then selects eight using the paired score. Its reported overlap with MIR top-eight is 21.91%. The final scores are 19.11 pooled F1, 75.79 acquisition, 17.56 retention, 55.23 forgetting, and -55.23 old-task BWT. These observations justify the specific acquisition contrast in the main text. They do not establish that replacing the privileged paired score with $\widehat{R}$ in Eq. (7) would reproduce it.

H Evidence Boundaries

The experiments support an effect of paired replay-observation replacement and a separation between stability and acquisition endpoints. They do not directly measure sample-level causal replay utility, gradient-conflict reduction, convergence acceleration, or robustness across datasets and corruption strengths. No learning curves, additional seeds for the selectors, or unperformed probe-utility studies are presented as measured results. The independent-head method in Eqs. (3)–(5), the masked-head feature controls, and the privileged paired-reference diagnostics answer different questions and are deliberately identified separately.