SELECTIVE COARSE-LAYER DISCRETE DISTILLATION FOR CONTINUAL EEG PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained EEG models may be adapted continually as data arrive from heterogeneous domains. For models that use discrete tokenizers, this adaptation risks both catastrophic forgetting and drift in the token assignments consumed by downstream models. We propose CAP-CL (Coarse Assignment Preservation for Continual Learning), which exploits the hierarchy of residual vector quantization to regularize coarse assignments while allowing finer representations to adapt. CAP-CL distills coarse-layer (L0) assignments from the previous-stage tokenizer on current and replay data, with replay capped separately for each domain and finer layers exempt from the distillation loss. On a stream of 21 EEG datasets, using an autoregressive readout trained with replay on hard token indices, CAP-CL achieves a mean final balanced accuracy of 0.421. It outperforms the tested replay and distillation controls in all five paired seeds, across three fixed stage orders, and under matched coarse-layer loss weights. Preserving assignments between successive stages has different effects from anchoring them to the initial tokenizer: a fixed initial teacher reduces cumulative coarse-layer flips from 0.876 to 0.568 but lowers final accuracy to 0.402. These results support coarse-only discrete distillation as an effective regularizer for continual EEG pretraining with an adaptive readout and show that stronger long-term index preservation alone does not explain downstream performance.

1 INTRODUCTION

Large-scale EEG models are increasingly built in two stages: a vector-quantized tokenizer maps raw multichannel signals to discrete codes, and a larger backbone is trained over those codes (Jiang et al., 2024; Cui et al., 2026; Bao et al., 2022; Peng et al., 2022). In this setup, freezing the tokenizer means that the backbone always sees a fixed code interface.

EEG corpora accumulate continually, and successive releases differ in subjects, montage, acquisition hardware, and task paradigm (Pfurtscheller & Neuper, 2001; Lotte et al., 2018; Jayaram & Barachant, 2018). These shifts require a model to absorb new signal statistics without discarding representations that remain useful for earlier domains. As the number of subjects and domains grows, repeatedly collecting all historical recordings into a pooled retraining job increases storage, data-movement, preprocessing, and optimization costs. We therefore study a staged setting in which new domains are incorporated as they become available and a controlled replay sample from earlier domains provides continuity (Gururangan et al., 2020; Jin et al., 2022; Cossu et al., 2024). In this setting, both the tokenizer and its downstream readout are updated as new domains arrive. The central question is how to preserve the parts of a discrete interface that should remain stable while retaining enough plasticity to learn from each new domain. Our protocol uses a fixed cap per prior domain rather than a fixed global memory budget. The stored replay set therefore grows with the number of completed stages, although no stage requires full-corpus joint retraining.

Updating a tokenizer introduces an additional failure mode beyond the usual stability–plasticity problem for continuous representations: *discrete assignment drift*. After an update, the same waveform may receive a different integer even when its underlying feature has moved only modestly, so continual learning must manage both representation change and the discrete interface presented downstream. Recent tokenizer-adaptation work seeks to limit disruptive code changes or revise discrete vocabularies as data evolve (Feng et al., 2026; Purason et al., 2025). A *residual* tokenizer

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.

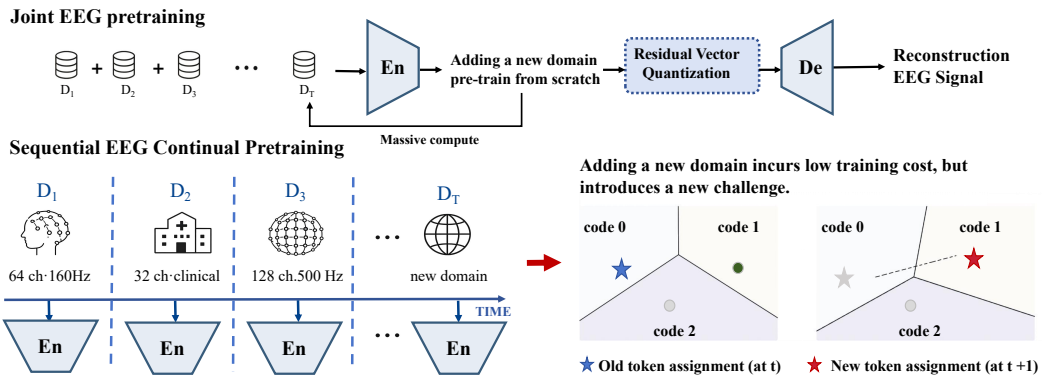


Figure 1: **Continual EEG pretraining exposes two pressures on a discrete tokenizer.** (a) Datasets arrive as a stream and replay is maintained with a per-domain cap of 1024 trials, so the stored replay set grows with the number of completed stages. (b) A tokenizer update can change the discrete assignment of the same waveform (assignment flip), creating drift in the indices consumed downstream. These pressures motivate the coarse-only, previous-stage index distillation introduced in Figure 2.

encodes a waveform as a coarse index followed by successive refinements (van den Oord et al., 2017; Lee et al., 2022; Zeghidour et al., 2022); the coarse and residual assignments need not play the same role. We examine two design axes. The first is **scope**: the coarse layer defines the initial quantization partition, whereas finer layers encode residual detail. The second is **temporal reference**: a constraint can preserve continuity with the immediately preceding tokenizer or enforce identity with the original tokenizer throughout the stream. We test both axes.

We examine both axes on Main-21, a stream of 21 EEG datasets, using a TUEG-pretrained residual tokenizer (Obeid & Picone, 2016; Cui et al., 2026) and an autoregressive readout retrained on hard indices after each stage. **CAP-CL** (Coarse-Anchored Plasticity for Continual Learning) applies a hard-index cross-entropy term to the coarse layer (L0) only, using the previous-stage tokenizer as teacher on current and replay batches. The finer layers continue to receive the base objective and are not directly constrained by the teacher (Figure 2). Our comparisons show that extending the constraint across residual layers or anchoring it to the initial tokenizer lowers downstream utility relative to CAP-CL, even though initial-tokenizer anchoring preserves more long-term index identity. CAP-CL preserves continuity of the coarse partition across adjacent stages while allowing the representation to reorganize over longer horizons.

Contributions.

- We identify **scope** and **temporal reference** as two separable design axes for discrete constraints on hierarchical tokenizers under continual pretraining, and give a total-weight-matched formulation (Eq. 4) that makes layer subsets directly comparable at equal regularization budget.
- We instantiate the axes as **CAP-CL**. Under three fixed stage orders, it has the highest mean Final accuracy among the matched continual-pretraining recipes (Table 2 and Appendix Tables 12–13).
- We show that lower cumulative coarse-layer flip does not necessarily improve downstream accuracy: the initial-teacher configuration preserves more long-term identity but performs worse than CAP-CL (Table 3).
- We will provide the stream protocol, including the dataset registry, preprocessing contract, frozen stage orders, and baseline configurations, to support reproduction of the layer, teacher, batch-scope, and lookup controls reported here.

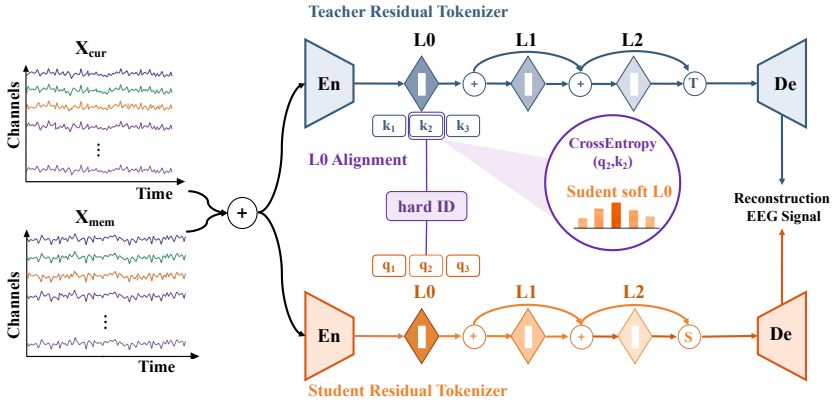


Figure 2: **The added index loss acts only on L0.** A frozen previous-stage teacher provides hard L0 indices for waveform x; L1/L2 continue to train through the base tokenizer objective.

2 Related Work

Discrete tokenizers as learned interfaces. Vector-quantized tokenizers are commonly trained once and then treated as fixed interfaces for a backbone trained on their integer codes (Bao et al., 2022; Dong et al., 2023; Peng et al., 2022; Esser et al., 2021; Yu et al., 2024; Jiang et al., 2024). This convention avoids interface drift, but it leaves open how a discrete interface should be updated when new data cannot be pooled with the original corpus. Our work uses an implementation-specific tokenizer that we call LexVQ, based on the dual-domain residual-quantization design of BrainRVQ (Cui et al., 2026), as a testbed for that update problem.

Continual adaptation of EEG and pretrained models. EEG continual-learning studies typically adapt a downstream backbone while leaving the tokenizer or codebook untouched (Zhou et al., 2025; 2026; Li et al., 2026). Continual-pretraining methods for other modalities instead regulate parameters, continuous features, logits, or replayed examples as data arrive (Wu et al., 2024; Jin et al., 2022; Cossu et al., 2024; Gururangan et al., 2020; Kirkpatrick et al., 2017; Li & Hoiem, 2016; Rolnick et al., 2019; Buzzega et al., 2020). These approaches do not directly address the discrete failure mode in which the same waveform receives a different integer after a tokenizer update. We study unlabeled updates to a hierarchical tokenizer on a heterogeneous EEG stream, followed by readout adaptation under a shared budget.

Discrete distillation and hierarchical constraint design. Related work updates or adapts discrete representations through codebook expansion, iterative cluster refresh, or drift-aware code reassignment (Zhang et al., 2025; Hsu et al., 2021; Feng et al., 2026). The unresolved design question is where and how far such supervision should act in a residual tokenizer: coarse and residual layers have different roles, and a previous-stage teacher imposes a different temporal constraint from a fixed initial teacher. CAP-CL addresses this question with coarse-layer-only distillation from the previous stage, and we compare it with a total-weight-matched all-layer control and continuous feature distillation under the same EEG pipeline. This comparison tests whether continual discrete adaptation benefits from selective coarse-layer anchoring instead of uniform preservation across the residual hierarchy.

3 Method

CAP-CL changes the tokenizer training objective within an otherwise shared pipeline. We first describe LexVQ and the evaluation protocol, then introduce the coarse-layer distillation loss and its controls.

3.1 LexVQ: a residual EEG tokenizer

LexVQ is our implementation-specific name for the BrainRVQ-style residual EEG tokenizer used throughout our experiments (Cui et al., 2026). Its architecture is held fixed across methods; our contribution is CAP-CL's continual-adaptation objective. A shared encoder feeds parallel temporal/frequency residual quantizers (van den Oord et al., 2017; Lee et al., 2022; Zeghidour et al., 2022) with layers $\ell \in \{0, 1, 2\}$. Writing $r^{(0)}$ for the encoder embedding and $v_k^{(\ell)}$ for code vectors,

$$z^{(\ell)} = \arg\min_k \left\| r^{(\ell)} - v_k^{(\ell)} \right\|_2^2, \qquad r^{(\ell+1)} = r^{(\ell)} - v_{z^{(\ell)}}^{(\ell)}, \tag{1}$$

so L0 selects a coarse atom and L1/L2 refine residuals. We write $i^{(\ell)}(x; \Theta)$ for hard indices and $q^{(\ell)}(x; \Theta)$ for soft logits against the current codebook. The base tokenizer objective combines reconstruction and quantization losses; continual-learning regularizers are added during adaptation, without class-label targets. Main-21 initializes LexVQ at Θ_0 and the AR model at Φ_0 from TUEG-pretrained checkpoints, with no intermediate motor-imagery (MI) adaptation (Appendix C).

3.2 Setup and readout protocol

At stage t, replay-using methods update LexVQ on $\mathcal{D}_t$ with a per-domain buffer $\mathcal{M}_{<t}$ (1024 trials per prior domain). The stored replay set grows with the number of completed domains, but each update avoids pooling the full historical corpus. Each stage then freezes LexVQ, produces hard indices $i^{(\ell)}(x; \Theta)$, and fine-tunes the autoregressive readout and task-incremental (Task-IL) heads under a shared budget (Appendix B.2); continual-pretraining methods act only in the preceding unlabeled tokenizer update. Flips and LexVQ probes evaluate the tokenizer directly, whereas Final and backward transfer (BWT) evaluate the complete adapted pipeline. CAP-CL, Feat-KD@Θ_{t-1}, and DER++ weights are selected with a one-factor-at-a-time (OFAT) screen and transferred to Main-21; the OFAT probe trials are isolated from the Main-21 evaluation pools (Appendices F, J.7).

3.3 Token assignment drift

For fixed waveforms and reference Θ_{ref},

$$\mathrm{flip}_{L\ell}(\Theta; \Theta_{\mathrm{ref}}) = \mathbb{E}\Big[\mathbf{1}\big\{i^{(\ell)}(x; \Theta) \neq i^{(\ell)}(x; \Theta_{\mathrm{ref}})\big\}\Big]. \tag{2}$$

flip_{L0} relative to the previous-stage tokenizer Θ_{t-1} measures local assignment drift; relative to the initial tokenizer Θ_0, it measures cumulative assignment drift. **Freeze-CB** freezes codebook tables with a trainable encoder. If flips were purely table-driven, Freeze-CB would eliminate them; Section 4.2 shows they remain large.

3.4 Coarse-only discrete distillation

CAP-CL regularizes the L0 assignments toward the previous-stage tokenizer while leaving L1/L2 free of index cross-entropy. The teacher is a frozen snapshot refreshed once per stage; it provides targets and adds no trainable module. These two choices define the constraint design that we test: L0 is the coarse partition on which the residual sequence is organized, while the delayed teacher limits the constraint to the current stage transition. CAP-CL therefore anchors coarse assignments locally, without imposing a fixed initial code identity on every later update. The residual layers still receive the base reconstruction and quantization objectives, but are not directly required to reproduce the teacher's indices. Figure 2 shows this loss within LexVQ. With stop-grad teacher Θ_{t-1}, current batch x_{cur}, and replay batch x_{mem},

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{recon}}(x_{\mathrm{cur}}; \Theta) + \lambda_{\mathrm{rep}}\, \mathcal{L}_{\mathrm{recon}}(x_{\mathrm{mem}}; \Theta) \\ + \frac{\lambda_{\mathrm{cap}}}{2}\, \mathrm{CE}\big(q^{(0)}(x_{\mathrm{mem}}; \Theta),\, i^{(0)}(x_{\mathrm{mem}}; \Theta_{t-1})\big) \\ + \frac{\lambda_{\mathrm{cap}}}{2}\, \mathrm{CE}\big(q^{(0)}(x_{\mathrm{cur}}; \Theta),\, i^{(0)}(x_{\mathrm{cur}}; \Theta_{t-1})\big), \end{aligned} \tag{3}$$

where $\mathcal{L}_{\mathrm{recon}}$ denotes the base LexVQ objective, including reconstruction and quantization terms, with λ_{rep}=1 and λ_{cap}=0.75 unless noted (Appendix F). Setting λ_{cap}=0 recovers ER, giving a direct removal ablation of the proposed regularizer without changing replay or the downstream training schedule. Student logits use temperature τ=0.1; the cross-entropy targets the teacher's hard L0

index. Splitting λ_{cap} equally over memory and current batches keeps the *total* index-loss weight identical to a memory-only variant. For a layer set $\mathcal{S} \subseteq \{0, 1, 2\}$, the total-weight-matched generalization replaces the two L0 terms by

$$\frac{\lambda_{\text{cap}}}{2|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \Big[\text{CE}\big(q^{(\ell)}(x_{\text{mem}}; \Theta),\, i^{(\ell)}(x_{\text{mem}}; \Theta_{t-1})\big) + \text{CE}\big(q^{(\ell)}(x_{\text{cur}}; \Theta),\, i^{(\ell)}(x_{\text{cur}}; \Theta_{t-1})\big) \Big], \quad (4)$$

so CAP-CL is $\mathcal{S}=\{0\}$ and the all-layer discrete control is $\mathcal{S}=\{0, 1, 2\}$ under the same total index-loss weight λ_{cap} ("total-weight-matched" in tables). Figure 2 summarizes the L0-only delayed-teacher path. Table 4 compares layer choices under matched-total and unnormalized coefficient rules. Encoder and all codebooks remain trainable.

3.5 Baselines (continual-pretraining recipes on LexVQ)

We fix LexVQ and the autoregressive stack, and vary only the unlabeled continual-pretraining (CPT) recipe (Appendix Table 8, Appendix B.3). Naive-CPT updates LexVQ on the current stage only. Freeze-Lex skips the LexVQ update. Freeze-CB freezes the codebooks while keeping the encoder trainable; we use it only as a diagnostic and do not include it in the Main-21 comparison. Matched replay methods share the 1024-trial buffer and head rehearsal: ER, LwF (adapted), Feat-KD@Θ_0, Feat-KD@Θ_{t-1}, EWC (adapted), DER++ (adapted), CAP-CL, and the total-weight-matched all-layer discrete control (Eq. 4 with $\mathcal{S}=\{0, 1, 2\}$). Joint-Offline pools all stages under matched caps; foundation-model bodies (BIOT / LaBraM / CBraMod) appear in Table 2 as architecture-unmatched references. Let $z(x; \Theta)$ be LexVQ pre-quantization features. Feat-KD@Θ_0 is memory-only MSE to an initial teacher (no memory reconstruction; its coefficient is fixed a priori),

$$\mathcal{L}_{\text{feat}} = \mathcal{L}_{\text{recon}}(x_{\text{cur}}; \Theta) + \lambda\, \mathbb{E}_{x \sim \mathcal{M}_{<t}} \big\| z(x; \Theta) - z(x; \Theta_0) \big\|_2^2, \quad (5)$$

with λ=20. Feat-KD@Θ_{t-1} is a delayed-feature control with the same teacher, replay reconstruction, and current/replay scope, but pre-quantization MSE instead of L0 index cross-entropy,

$$\begin{aligned} \mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{recon}}(x_{\text{cur}}; \Theta) &+ \lambda_{\text{rep}}\, \mathcal{L}_{\text{recon}}(x_{\text{mem}}; \Theta) \\ &+ \frac{\lambda_{\text{kd}}}{2} \big\| z(x_{\text{mem}}; \Theta) - z(x_{\text{mem}}; \Theta_{t-1}) \big\|_2^2 \\ &+ \frac{\lambda_{\text{kd}}}{2} \big\| z(x_{\text{cur}}; \Theta) - z(x_{\text{cur}}; \Theta_{t-1}) \big\|_2^2, \end{aligned} \quad (6)$$

with λ_{rep}=1, λ_{kd}=1. Its coefficient is selected with the shared tuning protocol; MSE and cross-entropy are not assumed to have equivalent scales or gradient magnitudes. Adapted LwF / DER++ / EWC objectives are in Appendix B.3. Feat-KD answers "discrete L0 vs. continuous features under matched scope"; Eq. 4 with $\mathcal{S}=\{0, 1, 2\}$ answers "L0-only vs. full residual index loss under matched total weight".

4 Experiments

4.1 Setup

Main-21 has 21 stages (14 imagery, 6 motor-execution, 1 arithmetic), plus Yang2025 as the primary corpus-unseen held-out set and Jeong2020MI as a Jeong-family paradigm-transfer probe (Appendix J.7). Neither dataset was used in TUEG pretraining, as an update stage, or for hyperparameter selection. Within-file 80/20 splits are trial-disjoint but may share subject and recording file. A separate five-stage imagery stream is used to select the loss weights for CAP-CL, Feat-KD@Θ_{t-1}, and DER++ before the Main-21 runs (Appendix F); these three recipes use the same prespecified OFAT grid, probe criterion, seed set, and compute budget. LwF, EWC, and Feat-KD@Θ_0 retain their predeclared fixed coefficients rather than receiving a separate Main-21 search. Matched LexVQ methods share the tokenizer, autoregressive stack, and 1024-trial buffer; they differ only in the unlabeled LexVQ objective. Stream utility uses the index-consuming readout on those checkpoints (Appendix B.2). We report mean $\pm$ sample std over shared seeds 0–4.

Table 1: **Five-stage tokenizer screen and design alternatives** (5 seeds; mean ± sample std). $\text{flip}_{L\ell}$ vs. frozen TUEG Θ_0 equation 12; ranking column mean-probe equation 13. Within each Setting block, best mean-probe in **bold**. Each block is a separate matched run; probe scores are compared only within a block. Not a Main-21 Task-IL ranking table; OFAT uses old-probe in Table 10.

Setting	Method	flip_{L0}	flip_{L1}	flip_{L2}	mean-probe↑
Primary screen	Naive-CPT (primary/current-only)	0.901 ±0.003	0.953 ±0.001	0.990 ±0.004	0.348 ±0.022
	Freeze-CB (frozen codebook; primary)	0.854 ±0.005	0.920 ±0.008	0.981 ±0.010	0.361 ±0.016
	CAP-CL (ours)	0.880 ±0.004	0.979 ±0.012	0.995 ±0.005	**0.393** ±0.015
Assignment variants	No-anchor CPT (assignment block)	0.924 ±0.003	0.966 ±0.012	0.993 ±0.003	0.356 ±0.009
	Feat-KD@Θ_0	0.765 ±0.004	0.904 ±0.004	0.976 ±0.001	**0.365** ±0.010
	Feat-KD@Θ_{t-1}	0.873 ±0.003	0.909 ±0.001	0.989 ±0.009	0.352 ±0.019
	Commit@Θ_{t-1}	0.916 ±0.004	0.976 ±0.001	0.993 ±0.004	0.348 ±0.013
	Feat+Commit@Θ_{t-1}	0.884 ±0.003	0.954 ±0.001	0.991 ±0.006	0.359 ±0.014
Codebook variants	Trainable-CB (codebook block)	0.925 ±0.004	0.958 ±0.007	0.994 ±0.006	**0.371** ±0.026
	Proximal (usage-gated)	0.912 ±0.002	0.965 ±0.009	0.992 ±0.008	0.350 ±0.022
	Proximal (all codes)	0.885 ±0.002	0.962 ±0.001	0.992 ±0.007	0.357 ±0.016

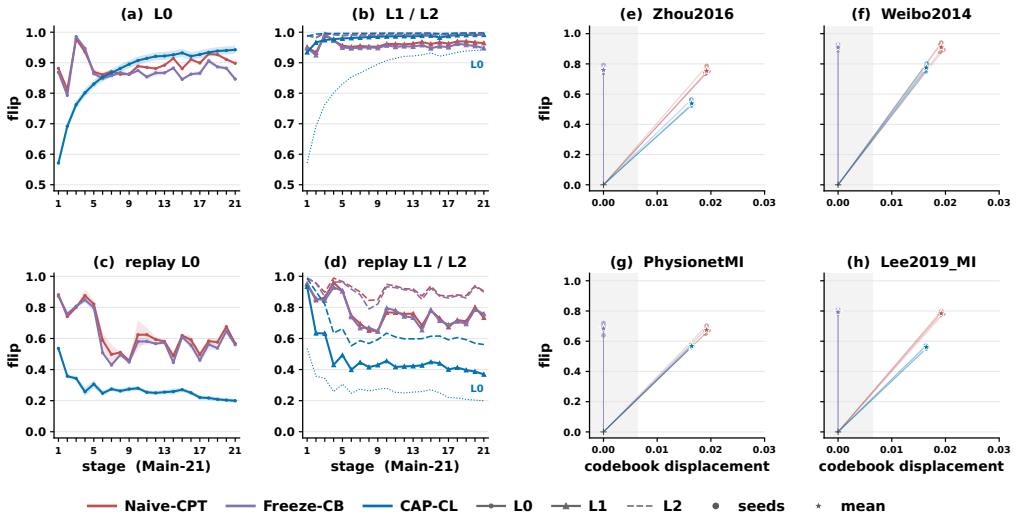


Figure 3: **Local consistency can coexist with cumulative assignment drift.** (a)–(d) Main-21: CAP-CL constrains local L0 flip relative to Θ_{t-1} on replay while assignments continue to diverge from Θ_0. (e)–(h) A short Zhou update shows that frozen codebook tables do not ensure fixed assignments.

4.2 Tokenizer-level evidence: freezing codebooks does not freeze indices

On the five-stage tokenizer screen (Table 1), Naive-CPT reaches flip_{L0}=0.901 relative to Θ_0, while Freeze-CB still yields 0.854 with little pre-quantization probe gain (0.361 vs. 0.348). This control establishes that fixed codebook tables are insufficient for assignment stability, because encoder updates alone can move examples across codebook boundaries. CAP-CL improves the pre-quantization probe (0.393 ± 0.015) while cumulative L0 flip remains high (0.880), showing that feature quality and agreement with the initial tokenizer need not improve together. On Main-21, Figure 3(a)–(d) shows lower local L0 flip for CAP-CL than for the plotted updating controls, alongside continued fine-layer reassignment. The short-update panels (e)–(h) further show that freezing codebook tables alone does not stabilize assignments.

Table 2: **CAP-CL improves final accuracy beyond replay alone and the tested distillation controls.** After each LexVQ update stage, LexVQ is frozen and produces hard indices for the autoregressive fine-tuning stage (Appendix B.2). EEG foundation-model rows are architecture-unmatched references: same Main-21 order and budgets. Held-out probes: Yang2025 (2 classes, chance 0.5) and Jeong2020MI (11 classes, chance ≈0.091). Seeds 0–4, mean ± sample std. Best matched *LexVQ* streaming entry in **bold** on Final / BWT / Yang2025 / Jeong2020MI (foundation models and Joint-Offline excluded).

Setting	Method	Final ↑	BWT ↑	Yang2025	Jeong2020MI
EEG foundation	BIOT	0.390 ±0.016	−0.102 ±0.013	0.605 ±0.018	0.193 ±0.012
	LaBraM	0.368 ±0.006	−0.063 ±0.011	0.599 ±0.016	0.165 ±0.012
	CBraMod	0.367 ±0.008	−0.071 ±0.010	0.600 ±0.016	0.160 ±0.009
LexVQ extremes	Naive-CPT	0.372 ±0.015	−0.041 ±0.011	0.582 ±0.017	0.188 ±0.016
	Freeze-Lex	0.394 ±0.010	0.005 ±0.008	0.599 ±0.014	0.191 ±0.018
Continual pretraining	ER	0.401 ±0.009	0.004 ±0.009	0.606 ±0.012	0.199 ±0.015
	LwF	0.407 ±0.007	0.009 ±0.008	0.614 ±0.012	0.203 ±0.021
	Feat-KD@Θ_0	0.410 ±0.007	0.008 ±0.009	0.612 ±0.011	0.190 ±0.022
	EWC	0.399 ±0.007	0.006 ±0.007	0.607 ±0.013	0.178 ±0.027
	Feat-KD@Θ_{t-1}	0.398 ±0.008	0.003 ±0.008	0.611 ±0.011	0.200 ±0.011
	DER++	0.403 ±0.009	0.005 ±0.011	0.608 ±0.013	0.187 ±0.013
	All-layer	0.405 ±0.009	0.003 ±0.012	0.616 ±0.012	0.195 ±0.014
	CAP-CL (ours)	**0.421** ±0.008	**0.011** ±0.008	**0.633** ±0.011	**0.224** ±0.014
Reference	Joint-Offline	0.434 ±0.012	–	0.619 ±0.014	0.193 ±0.014

4.3 Downstream accuracy

Table 2 tests whether coarse-only distillation improves the shared EEG pipeline. Removing the proposed loss gives ER; adding it raises mean Final from 0.401 to 0.421 with the same tokenizer replay and downstream training budget. CAP-CL also exceeds the strongest continuous-distillation baseline by 1.1 percentage points and the all-layer discrete control by 1.6 percentage points. It also has the highest BWT and held-out probe scores among the matched streaming methods. Joint-Offline remains higher on Final by 1.3 percentage points; the foundation-model rows provide an architecture reference, and the held-out probes quantify transfer. Table 3 isolates the teacher and batch-scope choices. Replacing the previous-stage teacher with the initial teacher lowers Final from 0.421 to 0.412, despite reducing cumulative L0 flip from 0.876 to 0.568. Thus, lower cumulative flip relative to Θ_0 is not sufficient for higher downstream utility under this readout.

CAP-CL exceeds all seven continual-pretraining baselines in all five paired seeds. The mean Final differences range from 0.011 against Feat-KD@Θ_0 to 0.023 against Feat-KD@Θ_{t-1}; every paired 95% t-interval excludes zero (Table 14).

The advantage is also stable to the tested stage orderings; the full method-by-method tables are in Appendix H.

4.4 Depth and residual-layer ablations

Table 3 reports the teacher and batch-scope controls before Table 4 combines the depth configurations under two explicit coefficient rules. With a fixed total index-loss weight, CAP-CL exceeds the all-layer control (0.421 versus 0.405). When the L0 coefficient remains 0.75, adding fine-layer cross-entropy also lowers Final (0.413, 0.412, and 0.409). The teacher and current/replay scope are shared across the depth comparisons.

With the total index-loss weight held fixed, L0-only distillation gives the best Final and BWT values; adding L1 or L2 reduces performance. The preference for L0-only remains when the L0 coefficient is fixed at 0.75, which argues against a simple loss-mass explanation. Forcing residual assignments to follow the same teacher therefore reduces downstream utility in these comparisons.

Table 3: **Teacher and batch-scope controls for L0-only distillation.** Main-21, index-consuming readout, five seeds. flip_{L0} vs. Θ_0 is diagnostic only (not bolded).

Teacher	CE scope	CE layers	Final ↑	BWT ↑	flip_{L0}
Θ_{t-1} (delayed)	mem+cur	L0	**0.421** ±0.008	**0.011** ±0.008	0.876 ±0.021
Θ_{t-1} (delayed)	mem only	L0	0.414 ±0.007	0.007 ±0.009	0.925 ±0.004
Θ_0 (init)	mem+cur	L0	0.412 ±0.009	0.006 ±0.008	0.568 ±0.010
Θ_0 (init)	mem only	L0	0.409 ±0.005	0.005 ±0.011	0.639 ±0.005

Table 4: **Layer-design grid under two coefficient rules.** Teacher Θ_{t-1}, memory+current, Main-21, seeds 0–4; ER is the loss-removal reference from Table 2. The total-weight-matched block fixes λ_{cap}=0.75 (Eq. 4); the unnormalized block applies the listed layer weights directly. flip_{L0} vs. Θ_0 is diagnostic only. Best matched streaming entry is in **bold** on Final/BWT.

Setting	Config	Name	Final ↑	BWT ↑	flip_{L0}
Loss removal	λ_{cap}=0	ER	0.401 ±0.009	0.004 ±0.009	–
Fixed total index-loss weight	$\{0\}$	**CAP-CL**	**0.421** ±0.008	**0.011** ±0.008	0.876 ±0.021
	$\{0,1\}$	L0+L1 (matched)	0.408 ±0.011	0.006 ±0.010	0.868 ±0.020
	$\{0,1,2\}$	All-layer	0.405 ±0.009	0.003 ±0.012	0.852 ±0.022
	$\{1,2\}$	fine-only	0.394 ±0.010	−0.001 ±0.009	0.918 ±0.015
Unnormalized	$(0.75, 0.75, 0)$	L0+L1 (direct)	0.413 ±0.012	0.005 ±0.009	0.871 ±0.019
	$(0.75, 0.25, 0.25)$	+fine@0.25	0.412 ±0.010	0.005 ±0.011	0.869 ±0.020
	$(0.75, 0.75, 0.75)$	+fine@0.75	0.409 ±0.008	0.002 ±0.008	0.855 ±0.021
	$(0, 0.75, 0.75)$	fine-only	0.404 ±0.014	0.002 ±0.011	0.920 ±0.015

4.5 Shared codebook-lookup control

Table 5 tests whether the gap to the all-layer control requires method-specific lookup vectors. With both readouts using the same initial codebook table, CAP-CL retains a higher Final (0.419 versus 0.397). The AR model and task heads still train in this control, so it does not test compatibility with a frozen downstream readout.

4.6 Frozen-consumer compatibility diagnostic

The main evaluation retrains the AR readout after each tokenizer update. To separate this adaptive-readout result from compatibility with an existing consumer, we additionally freeze the AR backbone and all previously learned Task-IL heads immediately before each tokenizer update. We then evaluate the same old tasks before and after replacing Θ_{t-1} with Θ_t, without readout training or head rehearsal. Table 6 reports the seed-level aggregate change and the corresponding layer-wise assignment flips.

With the AR consumer frozen, ER loses 0.036 in BAcc after tokenizer updates, compared with 0.022 for All-layer and 0.009 for CAP-CL. All-layer has the smallest flips in every layer, yet does not have the smallest consumer drop. CAP-CL instead lowers L0 flip relative to ER (0.256 vs. 0.402) while retaining substantial L1/L2 plasticity (0.552/0.788).

5 Conclusion

On Main-21, CAP-CL reaches higher Final accuracy than replay alone, the continuous-feature controls, and all-layer discrete distillation. Its loss constrains only the coarse assignment from one stage to the next; the layer ablations show that extending the index loss to residual layers lowers utility. A fixed initial teacher preserves more long-term index identity but obtains lower Final accuracy, indicating that cumulative code preservation is not sufficient for this task. The results support coarse, stage-local anchoring together with adaptation in the residual layers.

Table 5: **CAP-CL's advantage persists with shared initial lookup tables.** Index sequences are emitted by the named update recipe. Lookup uses (i) the stage codebook of that recipe (default index-consuming protocol); (ii) that recipe's own lagged codebook $\Theta_{t-1}^{(\text{recipe})}$ (one-stage lag; geometry not cross-aligned); or (iii) a *shared* frozen Θ_0 table for every recipe (geometry-aligned fixed-lookup). The AR model and task heads are still fine-tuned under Algorithm 3; only the lookup tables are fixed within each readout phase. Best matched streaming entry in **bold** on Final/BWT.

Index source	Lookup codebook	Final ↑	BWT ↑	Yang2025	Jeong2020MI
CAP-CL	stage Θ (default)	**0.421** ±0.008	**0.011** ±0.008	**0.633** ±0.011	**0.224** ±0.014
CAP-CL	own Θ_{t-1}	0.414 ±0.012	0.007 ±0.010	0.603 ±0.007	0.208 ±0.016
CAP-CL	shared Θ_0	0.419 ±0.007	0.009 ±0.012	0.625 ±0.019	0.212 ±0.012
All-layer	stage Θ (default)	0.405 ±0.009	0.003 ±0.012	0.616 ±0.012	0.195 ±0.014
All-layer	own Θ_{t-1}	0.399 ±0.010	−0.002 ±0.008	0.596 ±0.012	0.161 ±0.009
All-layer	shared Θ_0	0.397 ±0.011	0.001 ±0.009	0.598 ±0.018	0.193 ±0.011

Table 6: **Frozen-consumer compatibility after tokenizer updates.** The AR backbone and old-task heads remain fixed while the tokenizer changes from Θ_{t-1} to Θ_t. We report held-out BAcc before/after the update, Δ, and layer-wise flips; values are first averaged over stages $t \geq 1$ and old tasks within each seed, then over seeds.

Method	Pre	Post	Δ BAcc ↑	L0 flip	L1 flip	L2 flip
ER	0.383 ±0.015	0.347 ±0.008	−0.036 ±0.013	0.402 ±0.016	0.574 ±0.017	0.768 ±0.012
All-layer	0.373 ±0.011	0.351 ±0.012	−0.022 ±0.011	0.209 ±0.006	0.238 ±0.006	0.279 ±0.005
CAP-CL	0.386 ±0.013	0.377 ±0.009	−0.009 ±0.008	0.256 ±0.008	0.552 ±0.019	0.788 ±0.014

Limitations. The within-file trial-level split prevents the held-out trials from entering training, replay, or readout fitting, but train and evaluation can still share subjects and recording files; these results therefore do not establish subject-disjoint or session-disjoint generalization. We study one residual-tokenizer architecture and one 64-channel preprocessing contract, so broader architectural and montage robustness remains open.

AI USE STATEMENT

Generative AI tools assisted with language editing and code development. The authors verified all scientific claims, experimental procedures, numerical results, citations, and final text.

ETHICS STATEMENT

This work performs secondary analysis of previously collected human EEG under each dataset's license and citation requirements. We do not redistribute restricted raw signals; personally identifying fields are unused. Original ethics approval and consent remain with the source studies. We acknowledge limitations of neural-data reuse, including potential re-identification risks from residual biometric structure in EEG or discrete codes.

REPRODUCIBILITY STATEMENT

We will release a dataset registry, preprocessing manifests, frozen task-order configurations, split definitions, and evaluation scripts for the Main-21 CPT protocol. Appendix G documents streams, inventory, and governance; Appendix J documents the common tensor format and leakage controls. Appendix B summarizes training budgets, the LexVQ→AR computation path (Appendix B.1), index-consuming readout and total-weight-matched discrete controls (Appendix B.2), and hardware; Appendix C records the TUEG foundation initialization and corpus-level overlap audit; Appendix J.7 records OFAT isolation and Jeong ME/MI held-out scope; Appendix F documents the shared five-stage OFAT for CAP-CL / Feat-KD@Θ_{t-1} / DER++ continuous weights, CE weight accounting, and the locked CAP-CL corner $(\lambda_{\text{rep}}, \lambda_{\text{cap}}, \tau) = (1, 0.75, 0.1)$; Appendix A.5 states seed-level uncertainty units. Anonymous code, configuration files, and evaluation scripts will be released with the supplementary material.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

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 (NeurIPS)*, 2020.

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492, 2024. doi: 10.1016/j.neunet.2024.106492.

Mingzhe Cui, Tao Chen, Yang Jiao, Yiqin Wang, Lei Xie, Yi Pan, and Luca Mainardi. BrainRVQ: A high-fidelity EEG foundation model via dual-domain residual quantization and hierarchical autoregression. *arXiv preprint arXiv:2602.16951*, 2026.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. PeCo: Perceptual codebook for BERT pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 552–560, 2023. doi: 10.1609/aaai.v37i1.25130.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yuebo Feng, Jiahao Liu, Mingzhe Han, Dongsheng Li, Hansu Gu, Peng Zhang, Tun Lu, and Ning Gu. Drift-aware continual tokenization for generative recommendation. *arXiv preprint arXiv:2603.29705*, 2026.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Vinay Jayaram and Alexandre Barachant. MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 1–16, 2022. doi: 10.18653/v1/2022.bigscience-1.1.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Weibin Li, Wendu Li, Yushan You, Chen Wei, and Quanying Liu. NeuroOnline: Bridging pretraining and online adaptation for EEG foundation models. *arXiv preprint arXiv:2607.03925*, 2026.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, 2016. Extended version in IEEE TPAMI 2018.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Gert Pfurtscheller and Christa Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001.

Taido Purason, Pavel Chizhov, Ivan P. Yamshchikov, and Mark Fishel. Teaching old tokenizers new words: Efficient tokenizer adaptation for pre-trained models. *arXiv preprint arXiv:2512.03989*, 2025.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994. Also available as arXiv:2107.03312.

Yizhou Zhang, Yuan Gao, Wangjin Zhou, Zicheng Yuan, Keisuke Imoto, and Tatsuya Kawahara. SONAR: Self-distilled continual pre-training for domain adaptive audio representation. *arXiv preprint arXiv:2509.15703*, 2025.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. BrainUICL: An unsupervised individual continual learning framework for EEG applications. In *International Conference on Learning Representations (ICLR)*, 2025.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Shijian Li, and Gang Pan. EvoBrain: Continual learning of EEG foundation models across heterogeneous BCI tasks. *arXiv preprint arXiv:2606.01767*, 2026.

A Evaluation Metrics and Formulas

This appendix states closed-form definitions for continual-learning and tokenizer-drift metrics used in the main text. Unless noted, dataset indices $i \in \{0, \ldots, T-1\}$ and training stages $t \in \{0, \ldots, T-1\}$ are 0-indexed; T is the stream length. All task-level scores use *balanced accuracy* (BAcc) unless a table caption states otherwise. Tokenizer-axis and backbone/Task-IL-axis numbers are never mixed in one comparison table.

A.1 BALANCED ACCURACY

For true labels y and predictions $\hat{y}$ with label set $\mathcal{C}$,

$$\mathrm{BAcc}(y, \hat{y}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\{j : y_j = c\}|} \sum_{j:\, y_j = c} \mathbf{1}[\hat{y}_j = c], \tag{7}$$

i.e., the unweighted mean of per-class recalls. Empty classes are skipped. Frozen linear probes and Task-IL heads both report BAcc on held-out trials of the evaluated dataset.

A.2 CONTINUAL EVALUATION MATRIX

Let $A \in \mathbb{R}^{T \times T}$ store

$$A_{t,i} = \text{score on dataset } i \text{ after finishing stage } t, \tag{8}$$

where the score is frozen-probe BAcc (tokenizer axis) or Task-IL head BAcc. Entries for domains not yet observed are stored as NaN and excluded from averages. Final and related means are *macro* averages over datasets (equal weight per dataset), not trial-count-weighted averages.

A.3 TASK / BACKBONE CONTINUAL METRICS

Final accuracy.

$$\mathrm{Final} = \frac{1}{T} \sum_{i=0}^{T-1} A_{T-1,\, i}. \tag{9}$$

Backward transfer (BWT). Mean change between the score on dataset i immediately after it is learned and its final score (Lopez-Paz & Ranzato, 2017) (higher is better; typically negative under forgetting):

$$\mathrm{BWT} = \frac{1}{T-1} \sum_{i=0}^{T-2} \big(A_{T-1,\, i} - A_{i,\, i}\big). \tag{10}$$

Held-out transfer. Let $\mathcal{H}$ collect post-stream probes never used in TUEG pretraining, never used as CPT stages, and never used for hyperparameter choice. Yang2025 is the primary corpus-unseen release (2 classes / chance 0.5). Jeong2020MI (11 classes / chance ≈ 0.091) is a same-study-family paradigm-transfer probe: its H5 file is held out while Jeong2020ME appears as a Main-21 companion stage (Appendix J.7). After the final Main-21 stage, we report per-dataset BAcc on each $h \in \mathcal{H}$ under a frozen linear probe on AR features (Table 2 and the index-consuming order-sensitivity panels). These columns are not LexVQ pre-quantization $\mathrm{mean\text{-}probe}$ equation 13 and are not the stream Task-IL head score.

A.4 TOKENIZER-AXIS DRIFT METRICS

Assignment flip. Fix an evaluation set of waveforms under the common preprocessing contract. Let $c(\cdot; \Theta)$ denote discrete LexVQ indices under tokenizer parameters Θ, with the same packing layout as in training (residual layers $L{=}0, 1, 2$; temporal/frequency branches when present). Relative to a reference tokenizer Θ_{ref}, either the previous-stage snapshot Θ_{t-1} or the TUEG initialization Θ_0 (Appendix C),

$$\mathrm{flip}(\Theta; \Theta_{\mathrm{ref}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\big[c(x_n; \Theta) \neq c(x_n; \Theta_{\mathrm{ref}})\big], \tag{11}$$

where the indicator is evaluated position-wise over reported index slots and then averaged. Layer-wise flips restrict the average to residual layer ℓ:

$$\mathrm{flip}_{L\ell}(\Theta; \Theta_{\mathrm{ref}}) = \frac{1}{N_\ell} \sum_{n=1}^{N} \sum_{p \in \mathcal{P}_\ell} \mathbf{1}\big[c_p(x_n; \Theta) \neq c_p(x_n; \Theta_{\mathrm{ref}})\big], \tag{12}$$

with $\mathcal{P}_\ell$ the index positions belonging to layer ℓ and $N_\ell = N|\mathcal{P}_\ell|$ the number of layer-ℓ index positions across the N waveforms. We treat flip_{L0} vs. a stated reference as a layer-wise stability diagnostic, not a ranking metric. Under Freeze-Lex, Θ is unchanged across stages, so $\mathrm{flip}_{L0}(\Theta_t; \Theta_{t-1}) = 0$ identically and is reported as such. When the reference is frozen Θ_0, the scalar measures cumulative drift from the initial tokenizer; it coincides with the CAP-CL objective only if the CE teacher is also Θ_0. Fine-layer flips (flip_{L1}, flip_{L2}) measure fine-layer reassignment; high flip alone does not demonstrate useful adaptation.

Mean probe (tokenizer axis). After stage t (or at stream end), we freeze LexVQ, extract mean-pooled *pre-quantization* features, and train a linear classifier with BAcc equation 7. Let p_i be the probe score on stream dataset i. At the end of a T-stage stream,

$$\text{mean-probe} = \frac{1}{T} \sum_{i=0}^{T-1} p_i. \tag{13}$$

Higher is better; this probe diagnoses whether pre-quantization features remain linearly usable, and is *not* interchangeable with backbone BWT. For hyperparameter screens that separate retention from plasticity on the same $\{p_i\}$, we also report

$$\text{old-probe} = \frac{1}{T-1} \sum_{i=0}^{T-2} p_i, \qquad \text{new-probe} = p_{T-1}, \tag{14}$$

so that $\text{mean-probe} = \big((T-1)\,\text{old-probe} + \text{new-probe}\big)/T$. OFAT locking uses old-probe only; new-probe is reported for transparency.

Codebook displacement (secondary). For a single codebook table $C \in \mathbb{R}^{K \times d}$ before/after an update, let $\pi^\star$ be the optimal permutation aligning rows by ℓ_2 cost (Hungarian matching). Mean displacement is

$$\Delta_{\mathrm{cb}} = \frac{1}{K} \sum_{k=1}^{K} \big\| C^{\mathrm{new}}_{\pi^\star(k)} - C^{\mathrm{old}}_{k} \big\|_2. \tag{15}$$

We do not use Δ_{cb} as a Main-21 or five-stage ranking metric; it is the horizontal axis of the short-probe panels in Figure 3 (e)–(h). The primary evidence that table freeze does not stop flips is five-stage Freeze-CB in Table 1; the short probe only illustrates high flip_{L0} at $\Delta_{\mathrm{cb}} \approx 0$.

A.5 Multi-seed reporting and comparison units

For each *fixed* stage schedule we run $S=5$ independent training seeds and report mean $\pm$ sample standard deviation over seeds for each scalar metric. Matched LexVQ CPT/CL rows share the same seed set $\{0, 1, 2, 3, 4\}$. The independent unit for training-randomness uncertainty is the seed under that schedule: Main-text tables use this multi-seed summary; single-seed trajectories are for qualitative plots only.

We do *not* treat the T stages of one stream as T i.i.d. replicates, and we do not resample within a single stream to manufacture extra degrees of freedom. Order sensitivity (Appendix H) uses two additional frozen permutations of the same 21 datasets under matched budgets and the same five seeds; those tables are separate conditions, not pooled with the locked-order seeds to enlarge N. Held-out columns are corpus-level transfer scores from a frozen AR probe after the stream and are not additional stage replicates. Final is the primary ranking metric for matched streaming LexVQ rows.

B Implementation Details

This appendix gives the training and evaluation budgets for the LexVQ comparisons. The implementation combines the BrainRVQ-based LexVQ tokenizer (Cui et al., 2026) with a hierarchical AR readout; the component details are given in Section 3.1.

Table 7: Parameter schedule under the locked LexVQ→AR stack. **U** denotes updated parameters, F denotes fixed parameters carried forward from the preceding checkpoint, and – denotes a module unused in that phase.

Phase	LexVQ encoder	Codebooks	AR backbone	Task-IL heads
Foundation / LexVQ pretraining	**U**	**U**	–	–
Foundation / AR index CE	F^{ID}	F	**U**	–
Main-21 / tokenizer update	**U**	**U**	F	F
Freeze-CB control	**U**	F	F	F
Freeze-Lex control	$\mathrm{F}^{\mathrm{no\ CPT}}$	F	$\mathrm{F}^{\dagger}$	$\mathrm{F}^{\dagger}$
Main-21 / AR readout	$\mathrm{F}^{\mathrm{emit}}$	$\mathrm{F}^{\mathrm{lookup}}$	$\mathbf{U}^{\mathrm{slow}}$	$\mathbf{U}^{\mathrm{fast}}$
Evaluation / held-out probe	F	F	F	*fresh linear*

$^{\mathrm{ID}}$ LexVQ is fixed while providing index targets; $^{\mathrm{emit}}$ and $^{\mathrm{lookup}}$ indicate the frozen tokenizer operations during AR training. $^{\mathrm{no\ CPT}}$ Freeze-Lex skips the LexVQ update; † it then follows the same AR readout and head protocol as the other rows. $^{\mathrm{slow}}$/$^{\mathrm{fast}}$ denote the backbone/head learning-rate split. Naive-CPT uses no LexVQ replay or head rehearsal, but shares the AR optimizer, learning-rate, and step schedule with replay-using rows. Index-consuming step, learning-rate, and rehearsal budgets are fixed in Appendix B.2.

Main-21 Task-IL (utility ranking). Locked 21-stage order (Appendix G.2); within-file 80/20 train/hold-out splits (disjoint trials; the same recording file and subject may occur in both); for the five stages that also appear in OFAT, reserved OFAT probe trial indices are removed from the Main-21 evaluation pool before the 80/20 draw (Appendix J.7); per stage at most 64 recording files, a load pool of 10,000 trials, 4096 training trials, and 512 evaluation trials; batch size 8; LexVQ CPT for 400 steps per stage except Freeze-Lex (0 CPT steps); AdamW learning rate 1×10^{-4}, weight decay 0; replay buffer of 1024 trials per prior domain for replay-using methods (Naive-CPT has no LexVQ buffer); amplitude scale $1/100$; TUEG-pretrained LexVQ/AR checkpoints (Appendix C); seeds 0–4. Continuous regularizer weights for CAP-CL, Feat-KD@Θ_{t-1}, and DER++ are selected under one shared five-stage OFAT protocol and compute budget (Appendix F): CAP-CL transfers $(\lambda_{\mathrm{rep}}, \lambda_{\mathrm{cap}}, \tau) = (1, 0.75, 0.1)$; Feat-KD@$\Theta_{t-1}$ transfers λ_{kd}=1; DER++ transfers λ_{der}=1. DER++ (adapted) stores the same raw waveforms as ER and adds delayed pre-quantization MSE on memory; it does not cache logits or code indices (Eq. 17). LwF (adapted) applies delayed pre-quantization MSE on the current batch plus memory reconstruction (Eq. 16); EWC (adapted) uses a reconstruction-Fisher penalty on LexVQ parameters (Eq. 18). Feat-KD@Θ_{t-1} uses the same delayed teacher, replay reconstruction, current/replay coefficient split, CPT/AR step budget, and OFAT weight-selection protocol as CAP-CL, with pre-quantization MSE in place of L0 index-CE.

Joint-Offline (pooled reference). Joint-Offline uses the same initialization, splits, per-domain caps (4096 train / 512 eval trials), and AR fine-tuning readout as the streaming table. All 21 training subsets are mixed in one pool with domain-uniform sampling (not trial-count weighted). LexVQ CPT uses reconstruction only (no CL regularizer, no replay) for 21×400=8400 steps; AR fine-tuning uses 21×800=16800 steps with LexVQ frozen; each Task-IL head is then trained for 800 steps with LexVQ and AR frozen. BWT is undefined.

AR fine-tuning readout. After each CPT stage, LexVQ is frozen. **Primary (index-consuming).** The AR input is the frozen LexVQ hard-index sequence (Appendix B.2): AdamW with learning rates 1×10^{-5} (AR backbone) and 1×10^{-4} (head), weight decay 0.05, 800 steps per stage, batch size 8; replay-using methods use prior-domain head rehearsal with probability 0.5, whereas Naive-CPT trains only on the current-task head batch. Order-sensitivity and Table 3 use this same index-consuming readout. We do not insert an extra AR-SSL phase between CPT and readout. Algorithm 3 records the primary path. EEG FM references in Table 2 replace LexVQ+AR with each model's native Task-IL stack under the same stage order and step/head budgets (Appendix B.3).

B.1 Computation path and training pseudocode

Weight schedule. Table 7 lists which blocks are updated, held fixed, or unused at each stage. The TUEG foundation first trains LexVQ and then trains AR to predict its hierarchical indices with

Algorithm 1 Main-21 stage loop (locked protocol)

Require: TUEG checkpoints Θ_0 (LexVQ), Φ_0 (AR); stages $\mathcal{D}_{1:T}$; method $\mathcal{R}$
$\Theta \leftarrow \Theta_0$, $\Phi \leftarrow \Phi_0$, $\mathcal{M} \leftarrow \emptyset$, heads $\leftarrow \emptyset$
$\Theta_{\text{delay}} \leftarrow \text{StopGrad}(\Theta)$ ▷ delayed teacher for CAP-CL / LwF / Feat-KD / DER++
for $t = 1$ to T **do**
  Build train/eval splits of $\mathcal{D}_t$; refresh per-domain waveform buffer $\mathcal{M}$
  if $\mathcal{R}$ is not Freeze-Lex **then**
    $\Theta \leftarrow \text{LEXVQ-CPT}(\Theta, \Theta_{\text{delay}}, \Theta_0, \mathcal{D}_t^{\text{tr}}, \mathcal{M}; \mathcal{R})$ ▷ Alg. 2 for CAP-CL
  end if
  Freeze all of Θ
  if $\mathcal{R}$ uses delayed teacher **then**
    $\Theta_{\text{delay}} \leftarrow \text{StopGrad}(\text{Copy}(\Theta))$
  end if
  $(\Phi, \text{heads}) \leftarrow \text{AR-ID-FINETUNE}(\Phi, \Theta, \text{heads}, \mathcal{D}_t^{\text{tr}}, \mathcal{M})$ ▷ Alg. 3
  Evaluate Task-IL matrix row t with frozen (Φ, heads) on each $\mathcal{D}_i^{\text{eval}}$
end for
Held-out probes: freeze Φ; fit linear classifiers on AR mean features
return Final/BWT from Task-IL matrix; held-out BAcc; optional flip / mean-probe on Θ

Algorithm 2 CAP-CL LexVQ CPT step (one minibatch)

Require: Live student Θ; stop-grad teacher Θ_{t-1}; current batch x_{cur}; memory batch x_{mem}
$\mathcal{L} \leftarrow \mathcal{L}_{\text{recon}}(x_{\text{cur}}; \Theta) + \lambda_{\text{rep}} \mathcal{L}_{\text{recon}}(x_{\text{mem}}; \Theta)$
$i_{\text{mem}}^{(0)} \leftarrow \text{StopGrad}\big(i^{(0)}(x_{\text{mem}}; \Theta_{t-1})\big)$ ▷ hard L0 indices; both branches
$i_{\text{cur}}^{(0)} \leftarrow \text{StopGrad}\big(i^{(0)}(x_{\text{cur}}; \Theta_{t-1})\big)$
$\mathcal{L} \leftarrow \mathcal{L} + \frac{\lambda_{\text{cap}}}{2} \text{CE}\big(q^{(0)}(x_{\text{mem}}; \Theta), i_{\text{mem}}^{(0)}\big)$
$\mathcal{L} \leftarrow \mathcal{L} + \frac{\lambda_{\text{cap}}}{2} \text{CE}\big(q^{(0)}(x_{\text{cur}}; \Theta), i_{\text{cur}}^{(0)}\big)$
Update encoder and all codebook tables of Θ by $\nabla_\Theta \mathcal{L}$ ▷ no index-CE on L1/L2
return Θ

code-ID cross-entropy. The Main-21 tokenizer-update phase updates LexVQ only; the readout phase freezes LexVQ and fine-tunes the AR model and heads under the index-consuming graph.

B.2 INDEX-CONSUMING READOUT AND TOTAL-WEIGHT-MATCHED DISCRETE CONTROLS

This subsection describes the readout and the layer and lookup controls reported above.

Index-consuming AR input. Let $i(x; \Theta) = \{i^{(\ell)}(x; \Theta)\}_{\ell \in \{0,1,2\}}$ be the hard residual indices (temporal and frequency branches packed as in TUEG AR code-ID pretraining). The primary readout replaces raw-patch $\Phi.\text{ForwardFeatures}(x)$ with an embedding lookup of $i(x; \Theta)$ under the frozen LexVQ codebook tables, then mean-pools AR features for the Task-IL head. LexVQ encoder/codebook weights do not receive Task-IL gradients. Optimizer, step count (800), batch size (8), LR split (1×10^{-5} / 1×10^{-4}), and weight decay (0.05) are fixed for all LexVQ index-consuming rows; head-rehearsal probability is fixed at 0.5 for replay-using rows and 0 for Naive-CPT. Default lookup uses the stage codebook of the emitting LexVQ Θ; this couples index sequences with codebook geometry and AR re-adaptation.

Total-weight-matched all-layer discrete and depth rows. The total-weight-matched depth block in Table 4 uses teacher Θ_{t-1}, mem+cur scope, λ_{rep}=1, τ=0.1, and total CE coefficient λ_{cap}=0.75 distributed as $\lambda_{\text{cap}}/(2|\mathcal{S}|)$ per (layer, batch) term in Eq. 4.

- $\mathcal{S}$=$\{0\}$: CAP-CL (same as Eq. 3).
- $\mathcal{S}$=$\{0, 1\}$: L0+L1 under fixed total CE weight.
- $\mathcal{S}$=$\{0, 1, 2\}$: all-layer discrete control.
- $\mathcal{S}$=$\{1, 2\}$: fine-only control (no L0 CE).

Algorithm 3 Index-consuming AR readout training (one stage)

Require: Frozen LexVQ Θ; AR Φ; Task-IL heads; labeled set $\mathcal{D}_t^{\mathrm{tr}}$; optional head-rehearsal pool

1: **for** step $= 1$ to N_{AR} **do**
2: Sample batch (x, y, j) from $\mathcal{D}_t^{\mathrm{tr}}$ with $j = t$, or from a prior-task rehearsal pool with probability 0.5 and task label j
3: $i \leftarrow \mathrm{StopGrad}\big(i(x;\Theta)\big)$ ▷ hard residual indices; both branches
4: $h \leftarrow \mathrm{MeanPool}\big(\Phi.\mathrm{ForwardFromCodes}(i)\big)$
5: $\mathcal{L}_{\mathrm{ce}} \leftarrow \mathrm{CE}(\mathrm{head}_j(h), y)$
6: Update Φ and head j at LR $1{\times}10^{-5}$ and $1{\times}10^{-4}$, respectively
7: **end for**
8: **return** (Φ, heads)

Each readout is trained from its corresponding CPT checkpoint under Algorithm 3. The coefficient λ_{cap} is not retuned for different layer sets. Seeds 0–4; report mean $\pm$ sample std as in Appendix A.5.

Fixed-λ_{L0} residual-split screen. The bottom block of Table 4 keeps λ_{L0}=0.75 fixed and sets λ_{L1}=$\lambda_{L2} \in \{0, 0.25, 0.75\}$ without weight renormalization. This checks whether the preference for L0-only persists when its coefficient is unchanged; adding fine-layer terms still increases the total regularization weight.

Codebook-lookup control. Table 5 reuses emitted hard indices from a CPT recipe during the readout phase (the AR model is still fine-tuned) under three lookup regimes: the stage codebook of that recipe; that recipe's own lagged $\Theta_{t-1}^{(\mathrm{recipe})}$; and a shared frozen Θ_0 table for every recipe. The shared-Θ_0 rows are the geometry-aligned fixed-lookup control; cross-tokenizer variants are left for follow-up.

Five-stage tokenizer screens. Zhou2016, Weibo2014, PhysionetMI, Lee2019_MI, then GrosseWentrup2009; 400 LexVQ steps per stage; batch size 8; at most 16 files and 256 training trials per stage (128 eval / 512 probe trials); the OFAT probe subset is fixed at split construction and excluded from Main-21 eval pools for these five names (Appendix J.7); length 800 samples at 200 Hz on the 64-channel contract; codebook size matched to TUEG (6000); amplitude scale $1/100$; 5 seeds. Full design matrix: Appendix D; continuous OFAT: Appendix F.

Hardware. All reported LexVQ CPT/CL and AR fine-tuning runs were executed on a single node with eight NVIDIA A800-SXM4-80GB GPUs (CUDA 13.2). Individual jobs used one GPU; multi-seed and multi-method sweeps were parallelized across the eight devices.

B.3 Adapted baseline objectives and implementation details

This subsection describes the adapted baseline objectives within the tokenizer-update and AR-readout protocol. Delayed teachers Θ_{t-1} are stop-grad deep copies refreshed *once per stage* after CPT (not every optimizer step). The buffer stores raw waveforms only (cap 1024 trials per prior domain). Locked recipe identifiers for the code release will be listed in the supplementary configs (not repeated here).

Adapted continuous / parameter baselines. Classic LwF distills task logits; classic DER++ caches logits; classic EWC estimates Fisher from supervised task loss. Under LexVQ CPT there are no class logits in the tokenizer-update phase, so we use the adapted objectives below (z as in Section 3.5; λ_{rep}=1 unless noted):

$$\begin{aligned}\mathcal{L}_{\mathrm{LwF}} = \mathcal{L}_{\mathrm{recon}}(x_{\mathrm{cur}};\Theta) + \lambda_{\mathrm{rep}}\,\mathcal{L}_{\mathrm{recon}}(x_{\mathrm{mem}};\Theta) \\ + \lambda_{\mathrm{lwf}}\left\|z(x_{\mathrm{cur}};\Theta) - z(x_{\mathrm{cur}};\Theta_{t-1})\right\|_2^2,\end{aligned} \tag{16}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{DER++}} = \mathcal{L}_{\mathrm{recon}}(x_{\mathrm{cur}};\Theta) + \lambda_{\mathrm{rep}}\,\mathcal{L}_{\mathrm{recon}}(x_{\mathrm{mem}};\Theta) \\ + \lambda_{\mathrm{der}}\left\|z(x_{\mathrm{mem}};\Theta) - z(x_{\mathrm{mem}};\Theta_{t-1})\right\|_2^2,\end{aligned} \tag{17}$$

Table 8: LexVQ CPT recipes and implementation details (architecture and AR readout fixed). Replay-using CPT/CL rows share a capped *waveform* buffer and Task-IL head rehearsal. LwF / DER++ / EWC are LexVQ-adapted: DER++ does not cache logits or code indices (Buzzega et al., 2020). Freeze-CB is a diagnostic screen (Table 1, Figure 3), not a Main-21 Task-IL row. Joint-Offline is a pooled reference, not a streaming competitor. Here, Mem denotes replay reconstruction in the tokenizer objective, and Replay denotes prior-task head rehearsal. ✓ indicates use; – indicates none; † is screen-only; ‡ is pooled.

Method	Constraint / target	Scope	Mem	Replay
Naive-CPT	current-stage recon. only	cur	–	–
Freeze-Lex	LexVQ frozen; AR/heads only	—	–	–
Freeze-CB	codebook tables frozen	screen	–	†
ER (Rolnick et al., 2019)	memory reconstruction	mem+cur	✓	✓
LwF (ad.) (Li & Hoiem, 2016)	delayed z MSE to Θ_{t-1}	cur KD	✓	✓
Feat-KD@Θ_0	initial-teacher z MSE to Θ_0 equation 5	mem KD	–	✓
Feat-KD@Θ_{t-1}	delayed z MSE equation 6	mem+cur KD	✓	✓
EWC (ad.) (Kirkpatrick et al., 2017)	recon-Fisher on LexVQ params	params	✓	✓
DER++ (ad.) (Buzzega et al., 2020)	delayed z MSE to Θ_{t-1}	mem KD	✓	✓
All-layer	index-CE, $\mathcal{S}=\{0,1,2\}$ equation 4	mem+cur CE	✓	✓
CAP-CL	L0 index-CE to Θ_{t-1} equation 3	mem+cur CE	✓	✓
Joint-Offline	pooled fit on all stages	pooled	–	‡

$$\mathcal{L}_{\mathrm{EWC}} = \mathcal{L}_{\mathrm{recon}}(x_{\mathrm{cur}};\Theta) + \lambda_{\mathrm{rep}}\,\mathcal{L}_{\mathrm{recon}}(x_{\mathrm{mem}};\Theta) + \lambda_{\mathrm{ewc}}\,\tfrac{1}{2}\sum_i F_i\,(\theta_i-\theta_i^{\star})^2, \tag{18}$$

with λ_{lwf}=1 and λ_{ewc}=10 fixed a priori, while DER++ uses the OFAT-transferred λ_{der}=1; LwF and EWC are not retuned on Main-21. Diagonal Fisher F is estimated from LexVQ reconstruction on the just-finished stage, accumulated across stages, and globally mean-normalized before the penalty; $\theta^{\star}$ is the post-stage parameter snapshot. ER is reconstruction only: $\mathcal{L}_{\mathrm{recon}}(x_{\mathrm{cur}}) + \lambda_{\mathrm{rep}}\mathcal{L}_{\mathrm{recon}}(x_{\mathrm{mem}})$.

Eq. 6 and memory reconstruction. Feat-KD@Θ_{t-1} includes the λ_{rep} memory-reconstruction term in both the formula and the code path. Feat-KD@Θ_0 equation 5 uses the memory buffer only for the initial-teacher MSE term and does *not* add memory reconstruction (intentionally asymmetric to the delayed mem+cur control).

Teacher refresh. All delayed teachers are refreshed at stage boundaries; the reported runs do not use per-step exponential moving averages.

C Foundation Initialization and Overlap Audit

Foundation corpus and stack. The implementation-specific LexVQ tokenizer and hierarchical AR readout are initialized from checkpoints pretrained on TUEG; BrainRVQ provides the architectural precedent (Cui et al., 2026), while the TUEG corpus is documented by Obeid and Picone (Obeid & Picone, 2016). Pretraining uses reconstruction and quantization objectives only: clinical event labels from TUEG are never LexVQ classification targets. The fixed Main-21, order-sensitivity, and five-stage CPT runs load the resulting TUEG foundation checkpoints as Θ_0 (LexVQ) and as the AR fine-tuning initialization. There is *no* pooled MI adaptation stage between TUEG foundation pretraining and the first CPT stage: MI waveforms enter only as CPT stream stages, Task-IL head training, or post-stream held-out probes.

Corpus-level overlap audit. The TUEG source releases and files used for foundation pretraining are distinct from the Main-21 and held-out releases in the corpus audit; Yang2025 and Jeong2020MI

Table 9: Initialization source for LexVQ CPT and the hierarchical AR readout on TUEG.

Role	Corpus / stack	What is trained	Labels / contact
LexVQ Θ_0	TUEG; LexVQ	encoder + residual codebooks	recon. + quant. only; no class CE
AR init	TUEG; hierarchical AR	hierarchical AR on TUEG tokens	predicts LexVQ indices; no MI labels
Main-21 CPT	MI/ME stream stages	LexVQ from Θ_0 under each recipe	Task-IL heads after CPT; within-file eval
Held-out probes	Yang2025, Jeong2020MI	frozen AR linear probe only	never in TUEG, stream, or tuning

are unused for stream updates and tuning. Main-21 uses the within-file trial-disjoint split, while OFAT probe trials are excluded from the corresponding ranking pools (Appendix J.7); Jeong2020MI is a same-family paradigm-transfer probe rather than an independent cohort.

D Tokenizer Design Alternatives

We use a five-stage tokenizer screen for assignment diagnostics and OFAT selection (Table 1). The fixed stream is Zhou2016, Weibo2014, PhysionetMI, Lee2019_MI, then GrosseWentrup2009 (Appendix G.3): 5 seeds, matched step and subset budgets, tokenizer-axis readouts only. Table 1 reports the primary Naive-CPT / Freeze-CB / CAP-CL comparison together with additional assignment and codebook diagnostics. The method labels state the intervention explicitly (current-only, no assignment term, frozen codebook, or trainable codebook). Main-text utility ranking remains the index-consuming Main-21 Task-IL table (Table 2), not this screen. Ranking within every Setting block is by $\mathrm{mean\text{-}probe}$ equation 13; lower flip_{L0} is not treated as success. Each Setting block is a separate matched run, so absolute $\mathrm{mean\text{-}probe}$ scores are compared only within a block; the single reported Freeze-CB row is the primary frozen-codebook run, and no duplicate rerun is included. OFAT selection uses $\mathrm{old\text{-}probe}$ on this stream (Table 10).

For the diagnostic rows in Table 1, No-anchor CPT denotes current-stage CPT without an assignment-preservation term; Trainable-CB leaves the codebook tables trainable; Commit@Θ_{t-1} applies the delayed teacher to the commitment term; Feat+Commit combines that term with feature distillation; and Proximal regularizes codebook displacement on either used codes or all codes, as indicated.

E Teacher and Batch-Scope Controls

Table 3 varies teacher and batch scope with CE restricted to L0; Table 4 contains the layer-design grid under both coefficient rules. Restricting the previous-stage loss to replay also lowers Final (0.414). These comparisons describe useful teacher and batch choices without identifying why the regularizer helps.

Figure 4 plots ten Main-21 domains for delayed×mem+cur to illustrate trough-to-end recovery under capped replay. Evaluations use the within-file $80/20$ hold-out split, not the replay buffer.

F CAP-CL Loss Hyperparameters

The continuous scalars in Eq. 3 are replay weight λ_{rep}, total index-CE weight λ_{cap} (split equally between current and memory batches when both are used), and logit temperature τ. Teacher indices are stop-gradient targets and gradients flow only through student logits.

CE weight accounting. Per-branch CE is mean-reduced over slots; temporal and frequency branches are summed (two branch terms, not a branch mean). Scope-matched mem+cur keeps the same intended total CE coefficient λ_{cap} as mem-only; only the batch support changes. The total-weight-matched layer rule divides λ_{cap} by $|\mathcal{S}|$; unnormalized layer coefficients and the L0-only teacher/scope variants are reported in Tables 4 and 3.

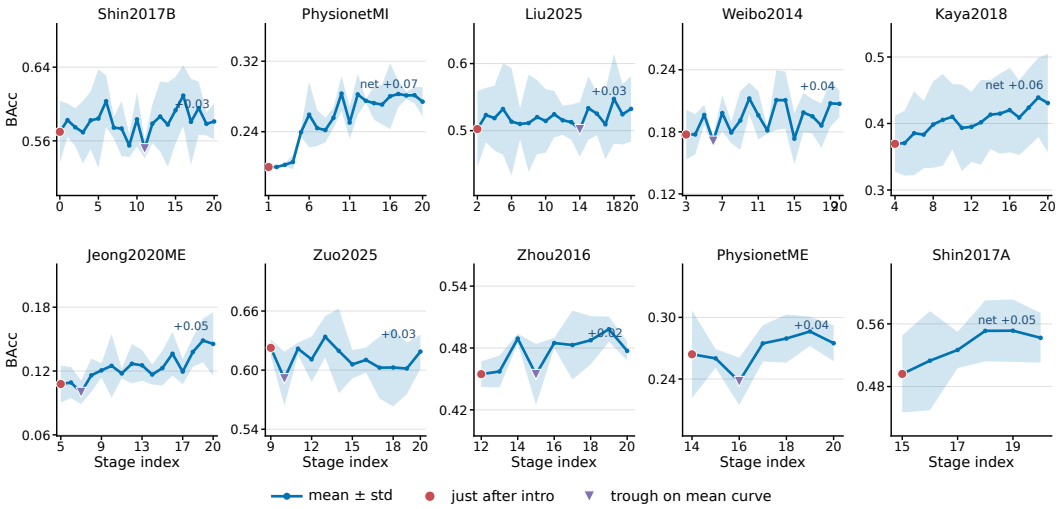


Figure 4: Main-21 hold-out BAcc trajectories for CAP-CL (delayed teacher, matched mem+cur L0 index-CE; mean±std over seeds). Ten selected domains illustrate recovery under the full pipeline, including AR training and replay; these trajectories do not isolate the effect of the assignment loss. Red marker: just after introduction; purple triangle: trough on the mean curve; annotated deltas are trough-to-end recovery (or net gain when the trough is the intro point).

Table 10: **CAP-CL continuous hyperparameter screens (OFAT).** Same five-stage stream, seeds, and pre-quantization probe protocol as Table 1; mean ± sample std. Feat-KD@Θ_{t-1} (λ_{kd}) and DER++ (λ_{der}) use this same OFAT stream budget and old-probe lock rule (transferred corners λ_{kd}=1, λ_{der}=1). Columns are old-probe / new-probe equation 14 (not mean-probe: OFAT locks on retention of prior stages). Within each block the other two scalars are held at $(\lambda_{\text{rep}}, \lambda_{\text{cap}}, \tau) = (1, 0.75, 0.1)$, so the three (def.) cells are the same corner. Selection: maximize old-probe; new-probe reported only. All three recipes use the same OFAT budget and selection criterion; Main-21 performance is evaluated only after these settings are transferred. Default values marked (def.).

Factor	Value	old-probe↑	new-probe↑	Factor	Value	old-probe↑	new-probe↑
λ_{cap}	0	0.329±0.018	0.498±0.024	λ_{rep}	0	0.342±0.015	0.505±0.023
	0.3	0.338±0.016	0.502±0.022		**1** (def.)	**0.362±0.010**	0.517±0.025
	0.5	0.348±0.014	0.510±0.020		2	0.346±0.014	0.509±0.022
	0.75 (def.)	**0.362±0.010**	0.517±0.025	τ	0.05	0.350±0.014	0.508±0.024
	1.0	0.349±0.013	0.508±0.021		**0.1** (def.)	**0.362±0.010**	0.517±0.025
	2.0	0.321±0.019	0.475±0.028		0.5	0.332±0.017	0.491±0.026

Shared one-factor-at-a-time selection protocol (CAP-CL / Feat-KD@Θ_{t-1} / DER++). The three recipes use the same prespecified OFAT protocol on the five-stage screen: the same stream, seeds, probe budget, and old-probe selection criterion, with only the recipe-specific scalar varied within its declared grid. Main-21 and held-out metrics are not used, and shared-stage probe trials are isolated from Main-21 evaluation pools (Appendix J.7). Table 10 documents the CAP-CL screen and the transferred settings $(\lambda_{\text{rep}}, \lambda_{\text{cap}}, \tau) = (1, 0.75, 0.1)$, $\lambda_{\text{kd}} = 1$, and $\lambda_{\text{der}} = 1$; the comparator rows are summarized by their selected settings.

All three transferred corners are used for the corresponding Main-21 cells (including Table 3 for CAP-CL).

Table 11: Experiment corpus used in this paper (Main-21 stream + held-out). Hours are estimated from retained trials × analysis-window length; held-out sets are Yang2025 and Jeong2020MI. Training uses the locked per-stage budgets, not every trial of every recording.

Split	# datasets	Subjects	Trials	Hours	Source channels	Sampling rate (Hz)
Main-21 imagery	14	446	226,812‡	209.1‡	14–128	160–1000
Main-21 motor execution	6	206	43,632‡	46.7‡	14–128	160–2500
Main-21 mental arithmetic	1	29	1,740	4.8	30	1000
Main-21 total	**21**	**681**	**272,184‡**	**260.6‡**	**14–128**	**160–2500**
Held-out imagery (Yang2025)	1	51	46,291	51.4	32	250
Held-out mixed MI/ME (Jeong2020MI)	1	25	39,051	43.4	60	2500
Held-out total	**2**	**76**	**85,342**	**94.8**	**32–60**	**250–2500**
Experiment total	**23**	**757**	**357,526‡**	**355.4‡**	**14–128**	**160–2500**

‡Trial/hour aggregates use curated inventory counts after event-level QC; Jeong2020ME, Zenodo-5882500, and Zuo2025 contribute to the dataset count but not to these aggregates. Stieger2021 alone accounts for $\approx 47\%$ of the inventory-counted Main-21 trials; per-stage sampling caps prevent single-set dominance.

G Corpus, Streams, and Data Governance

This appendix documents inclusion rules, the experiment inventory, fixed streams, and governance. For training budgets, see Appendix B.

G.1 Registry scope and inclusion

We construct a multi-dataset MI-EEG registry from public or authorized sources. Releases are included when they provide usable event timing, auditable subject/session and channel metadata, a defined paradigm and sampling rate, and compatible research-use terms; releases with irrecoverable timing or metadata, duplicates, incompatible tasks, or restrictive terms are excluded. Derived releases and same-cohort MI/ME companions receive a single canonical identity; selected motor-execution companions remain in the locked Main-21 stream (Table 11).

Experiment corpus (Main-21 + held-out). Table 11 summarizes the releases used in the reported experiments:

21 Main-21 stream stages (272,184 inventory-counted trials, ≈260.6 hours for releases with complete event counts) plus 2 held-out sets (85,342 trials, ≈94.8 hours). Per-stage sampling caps limit exposure so that large releases (notably Stieger2021) cannot dominate training.

G.2 Named inventory and QC notes

Main-21 stages (locked order). Shin2017B, PhysionetMI, Liu2025, Weibo2014, Kaya2018, Jeong2020ME, Ofner2017ME, Liu2024, Zenodo-5882500, Zuo2025, TrianaGuzman2024, SHU-2022, Zhou2016, GrosseWentrup2009, PhysionetME, Shin2017A, GuttmannFlury2025_ME, Schirrmeister2017, Lee2019_MI, Wairagkar2018, Stieger2021.

Held-out. Yang2025, Jeong2020MI.

Availability. All Main-21 and held-out releases used above can be converted to our common preprocessing format. Largest single stream set: Stieger2021 (127,313 trials).

Alternative Main-21 orders. Two metadata-derived permutations of the same 21 datasets (progressive nearest-neighbor and adversarial farthest-neighbor) were frozen before evaluation; construction and index-consuming tables are in Appendix H.

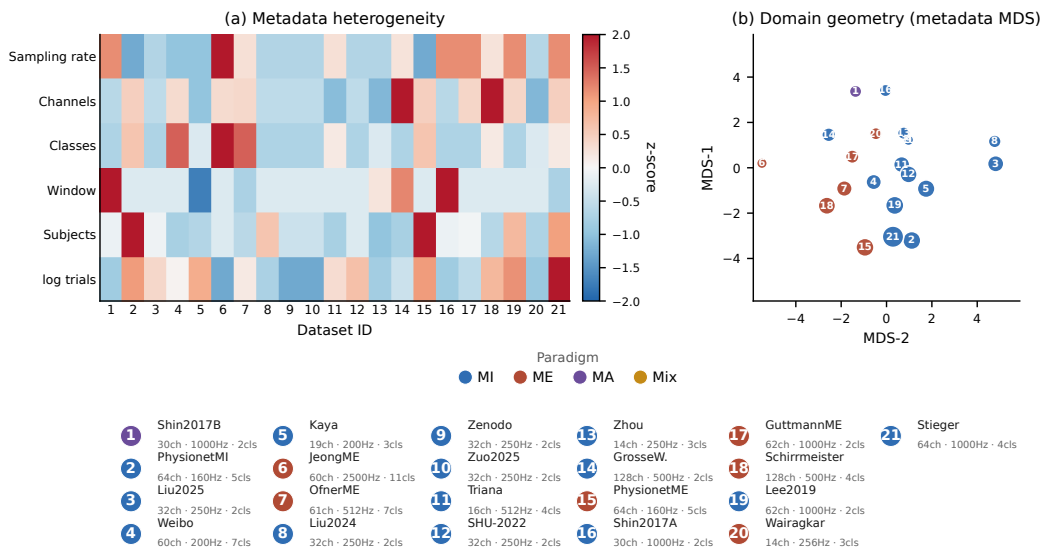


Figure 5: Main-21 stream heterogeneity (acquisition metadata and fingerprint distances).

G.3 Evaluation streams

The fixed Main-21 order and the five-stage diagnostic stream are listed in Appendix G.2 and used for the utility, tokenizer-screen, and OFAT experiments described in Appendices B and F. The held-out sets Yang2025 and Jeong2020MI remain outside TUEG pretraining, stream updates, and tuning, and appear only as frozen AR-probe columns.

G.4 Heterogeneity summary

For each dataset we record acquisition metadata (device/system, sampling rate, reference, channel names/locations), participant fields, paradigm fields, signal status, and governance fields (source URL, version, license, citation).

The Main-21 stream exhibits substantial heterogeneity along axes that motivate continual learning rather than pooled i.i.d. training (Figure 5):

- **Channels:** original montages range from 14 (Zhou2016, Wairagkar2018) to 128 (GrosseWentrup2009, Schirrmeister2017) electrodes before mapping to a shared 64-channel contract;
- **Sampling rate:** original rates range from 160 Hz (PhysionetMI/ME) to 2500 Hz (Jeong2020ME; held-out Jeong2020MI);
- **Label cardinality:** 2–11 classes across stream stages; held-out Yang2025 uses 2 classes (chance 0.5), and Jeong2020MI uses 11 mixed MI/ME classes (chance $\approx$0.091);
- **Population:** predominantly healthy adults, including stroke patients (Liu2024, Liu2025);
- **Protocol:** offline MI, online SMR (Stieger2021), motor execution companions, mental arithmetic (Shin2017B), and lower-limb MI;
- **Epoch length:** after resampling to 200 Hz, trial lengths $L_{\text{trial}} \in \{600, 800, 1000, 1400, 2000\}$ samples.

G.5 Licensing, ethics, and data governance

All datasets used in this work were obtained through public or authorized channels. Original participant consent and ethics approval remain the responsibility of the source studies; we perform

secondary analysis under the corresponding licenses and citation requirements. Personally identifying metadata are not used. This paper does *not* redistribute raw EEG by default: we will release registry metadata, download/conversion scripts, split files, and processed-feature contracts where redistribution is permitted, and provide index-only references for restricted releases.

H Order Sensitivity under Main-21 Permutations

We evaluate two alternative permutations of the 21 datasets under the *index-consuming* AR readout of Table 2. Schedules were frozen before we inspected the corresponding metrics. Primary locked-order utility ranking remains Table 2; Tables 12–13 repeat the same Setting/Method layout (including EEG FM references) under progressive and adversarial schedules.

Construction. The *progressive* order starts at the most central release (SHU-2022) and greedily appends the unused neighbor with smallest metadata distance, so successive stages tend to change gradually. The *adversarial* order starts at a peripheral release (Jeong2020ME) and greedily appends the farthest unused neighbor, so successive stages tend to jump farther in metadata space. These constructions specify smooth versus abrupt schedules; they are not a full combinatorial sweep over all 21! orders.

Matched protocol. Aside from stage order, the tokenizer-update methods and index-consuming AR fine-tuning follow Appendix B.2 and Table 2, including LwF, EWC, Feat-KD@Θ_0, Feat-KD@Θ_{t-1}, and Joint-Offline. Held-out Yang2025 and Jeong2020MI remain outside both streams and outside tuning.

Progressive order. SHU-2022 → TrianaGuzman2024 → Zhou2016 → Zenodo-5882500 → Zuo2025 → Wairagkar2018 → GuttmannFlury2025_ME → Ofner2017ME → Weibo2014 → Kaya2018 → Lee2019_MI → Stieger2021 → PhysionetMI → PhysionetME → Schirrmeister2017 → GrosseWentrup2009 → Shin2017A → Shin2017B → Jeong2020ME → Liu2024 → Liu2025.

Adversarial order. Jeong2020ME → Liu2025 → Shin2017B → Liu2024 → Schirrmeister2017 → Shin2017A → PhysionetME → GrosseWentrup2009 → Kaya2018 → PhysionetMI → Wairagkar2018 → Stieger2021 → Zhou2016 → Ofner2017ME → Lee2019_MI → Weibo2014 → GuttmannFlury2025_ME → TrianaGuzman2024 → Zuo2025 → SHU-2022 → Zenodo-5882500.

The smoother schedule preserves CAP-CL's Final ranking, suggesting that its advantage is not confined to the primary order.

The abrupt schedule also preserves CAP-CL's Final ranking, but Feat-KD@Θ_0 has higher mean BWT. Thus the observed accuracy advantage is consistent across the three tested orders, while the method with the best BWT depends on the order.

I Seed-Level Robustness of the Main Comparison

J Unified Preprocessing Pipeline

We convert heterogeneous releases into a common tensor format so that continual stages differ in neural and domain statistics rather than file format. Common preprocessing steps are: resample to 200 Hz, map to a fixed 64-channel montage, estimate any normalization ϕ_t only from stage-t training data (no future datasets), and store float32 tensors with provenance manifests. Dataset-specific importers, windows, and exceptions are recorded in per-dataset manifests that will be released with the code.

J.1 Signal import and event reconstruction

Raw files (EDF, MAT, EEGLAB SET, etc.) are loaded with dataset-specific importers. Event markers are validated against source documentation / MOABB conventions when available; invalid or incomplete trials (e.g., out-of-bound windows at recording ends) are rejected and counted in retention

Table 12: **Main-21 progressive order under index-consuming readout.** Same Setting/Method layout and matched method set as Table 2; only the stage schedule differs (progressive order above). Joint-Offline is order-independent and is repeated from Table 2. EEG FM rows remain architecture-unmatched native-feature references. Seeds 0–4, mean ± sample std. Best matched *LexVQ* streaming entry in **bold** on Final / BWT / Yang2025 / Jeong2020MI (FM bodies and Joint-Offline excluded).

Setting	Method	Final ↑	BWT ↑	Yang2025	Jeong2020MI
EEG FM	BIOT	0.369 ±0.013	−0.125 ±0.010	0.608 ±0.021	0.201 ±0.027
	LaBraM	0.362 ±0.008	−0.074 ±0.007	0.584 ±0.035	0.169 ±0.021
	CBraMod	0.360 ±0.005	−0.066 ±0.004	0.611 ±0.007	0.162 ±0.021
LexVQ extremes	Naive-CPT	0.368 ±0.012	−0.052 ±0.014	0.578 ±0.018	0.171 ±0.014
	Freeze-Lex	0.389 ±0.011	−0.008 ±0.012	0.596 ±0.015	0.178 ±0.021
CPT / CL	ER	0.399 ±0.009	0.008 ±0.008	0.603 ±0.014	0.191 ±0.024
	LwF	0.404 ±0.010	0.010 ±0.009	0.609 ±0.013	0.180 ±0.011
	Feat-KD@Θ_0	0.407 ±0.009	0.011 ±0.010	0.608 ±0.012	0.193 ±0.014
	EWC	0.397 ±0.011	0.007 ±0.008	0.604 ±0.014	0.192 ±0.013
	Feat-KD@Θ_{t-1}	0.396 ±0.008	0.005 ±0.007	0.607 ±0.012	0.186 ±0.010
	DER++	0.402 ±0.010	0.012 ±0.009	0.605 ±0.013	0.183 ±0.014
	All-layer	0.403 ±0.009	0.006 ±0.010	0.612 ±0.013	0.185 ±0.012
	CAP-CL (ours)	**0.418** ±0.008	**0.015** ±0.007	**0.622** ±0.015	**0.198** ±0.011
Reference	Joint-Offline	0.434 ±0.012	–	0.619 ±0.014	0.193 ±0.014

Table 13: **Main-21 adversarial order under index-consuming readout.** Same Setting/Method layout and matched method set as Table 2; only the stage schedule differs (adversarial order above). Joint-Offline is order-independent and is repeated from Table 2. EEG FM rows remain architecture-unmatched native-feature references. Seeds 0–4, mean ± sample std. Best matched *LexVQ* streaming entry in **bold** on Final / BWT / Yang2025 / Jeong2020MI (FM bodies and Joint-Offline excluded).

Setting	Method	Final ↑	BWT ↑	Yang2025	Jeong2020MI
EEG FM	BIOT	0.370 ±0.006	−0.121 ±0.004	0.604 ±0.020	0.195 ±0.004
	LaBraM	0.375 ±0.006	−0.073 ±0.007	0.596 ±0.010	0.173 ±0.005
	CBraMod	0.360 ±0.002	−0.088 ±0.007	0.600 ±0.004	0.164 ±0.008
LexVQ extremes	Naive-CPT	0.375 ±0.011	−0.038 ±0.013	0.585 ±0.017	0.181 ±0.014
	Freeze-Lex	0.392 ±0.010	−0.004 ±0.010	0.601 ±0.014	0.197 ±0.020
CPT / CL	ER	0.403 ±0.010	0.006 ±0.008	0.610 ±0.015	0.206 ±0.028
	LwF	0.406 ±0.009	0.008 ±0.011	0.611 ±0.012	0.193 ±0.008
	Feat-KD@Θ_0	0.409 ±0.007	**0.012** ±0.008	0.609 ±0.011	0.206 ±0.028
	EWC	0.400 ±0.008	0.004 ±0.007	0.606 ±0.012	0.196 ±0.013
	Feat-KD@Θ_{t-1}	0.399 ±0.008	0.004 ±0.008	0.610 ±0.012	0.202 ±0.012
	DER++	0.404 ±0.009	0.007 ±0.010	0.607 ±0.014	0.190 ±0.022
	All-layer	0.406 ±0.008	0.005 ±0.009	0.614 ±0.012	0.198 ±0.015
	CAP-CL (ours)	**0.420** ±0.007	0.010 ±0.011	**0.624** ±0.012	**0.215** ±0.023
Reference	Joint-Offline	0.434 ±0.012	–	0.619 ±0.014	0.193 ±0.014

logs. Cue-onset definitions follow the source protocol (annotation onset, cue latency, or documented MI onset inside longer epochs). Variable-length recordings are cut to the dataset's analysis window rather than forcibly padded to a global duration; the resulting sample length L_{trial} therefore remains dataset-specific after temporal standardization of the sampling rate. Signal units are converted to μV when needed; polarity and channel-name aliases (e.g., T3→T7) are normalized before montage mapping.

Table 14: **Seed-level Final scores and paired differences for the main continual-pretraining comparison.** Same LexVQ checkpoints and autoregressive protocol as Table 2. Δ is CAP-CL minus the named baseline on the same seed. MeanΔ [95% CI] uses the t interval over seeds 0–4.

Method	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Mean±std
CAP-CL	0.410	0.426	0.419	0.430	0.420	0.421 ±0.008
ER	0.390	0.402	0.395	0.412	0.406	0.401 ±0.009
EWC	0.391	0.399	0.394	0.410	0.401	0.399 ±0.007
Feat-KD@Θ_0	0.403	0.418	0.407	0.417	0.405	0.410 ±0.007
LwF	0.398	0.415	0.403	0.412	0.407	0.407 ±0.007
All-layer	0.395	0.417	0.401	0.411	0.401	0.405 ±0.009
Feat-KD@Θ_{t-1}	0.388	0.408	0.395	0.402	0.397	0.398 ±0.008
DER++	0.392	0.415	0.401	0.409	0.398	0.403 ±0.009
Δ vs. ER	0.020	0.024	0.024	0.018	0.014	0.020 [0.015, 0.025]
Δ vs. EWC	0.019	0.027	0.025	0.020	0.019	0.022 [0.017, 0.027]
Δ vs. Feat-KD@Θ_0	0.007	0.008	0.012	0.013	0.015	0.011 [0.007, 0.015]
Δ vs. LwF	0.012	0.011	0.016	0.018	0.013	0.014 [0.010, 0.018]
Δ vs. all-layer discrete	0.015	0.009	0.018	0.019	0.019	0.016 [0.011, 0.021]
Δ vs. Feat-KD@Θ_{t-1}	0.022	0.018	0.024	0.028	0.023	0.023 [0.019, 0.027]
Δ vs. DER++	0.018	0.011	0.018	0.021	0.022	0.018 [0.013, 0.023]

J.2 Temporal standardization

All datasets are resampled to a common target rate $f_{\text{target}} = 200\,\text{Hz}$, chosen as a trade-off between retaining MI-relevant μ/β content and controlling compute/storage for long continual streams. Anti-alias filtering is applied via polyphase resampling (resample_poly) with dataset-specific up/down factors. Primary analysis windows follow source conventions (commonly $[0, 3]$–$[0, 4]$ s from cue/annotation onset; longer windows such as $[0, 7]$ s or $[0, 10]$ s are preserved when they are part of the original MI protocol). Window heterogeneity is reported with the curated summary (Table 11) and treated as a protocol factor.

J.3 Channel and montage harmonization

The primary design uses a **canonical 64-channel montage** (extended 10–20 / 10–10 layout used by the LexVQ stack). Source channels are name-normalized and mapped to standard_1020/standard_1005 coordinates; missing electrodes are reconstructed by spherical-spline interpolation before an explicit reorder to the fixed 64-channel list. Coverage relative to the 64-channel bank varies across datasets (e.g., near-complete overlap for PhysionetMI vs. lower overlap for sparse 14–32 channel caps); interpolation QC (Pearson / NRMSE on held-out leave-one-out probes) is logged per dataset.

We acknowledge that interpolated missing channels can induce montage bias. The main benchmark uses the 64-channel contract for tokenizer compatibility; a native-montage + channel-mask track is a natural sensitivity analysis outside the main comparison.

J.4 Filtering, referencing, and artifact policy

We avoid redundant aggressive cleaning when source releases are already filtered or referenced; documented dataset-specific corrections are retained in the manifests. ICA and automated trial rejection are not part of the default track, and corrupted or flat channels are excluded through logged QC checks.

J.5 Normalization and tensor construction

The standardized sample at stage t is

$$\widetilde{\boldsymbol{X}}_n^{(t)} = \mathcal{P}_t\Big(\boldsymbol{X}_n^{(t)};\, \phi_t\Big) \in \mathbb{R}^{C_{\text{std}} \times L_{\text{std}}}, \tag{19}$$

with $C_{\text{std}} = 64$ and $L_{\text{std}} = T_t$ determined by the dataset window at 200 Hz. Any dataset-level normalization statistics ϕ_t are estimated only when $\mathcal{D}_t$ becomes available and are stored with that dataset's metadata; global normalization using future datasets is forbidden. Final tensors are stored as float32 HDF5 arrays with keys for EEG, labels, channel names, sampling rate, and provenance attributes.

J.6 Label harmonization

Continual LexVQ updates use reconstruction, quantization, and the specified continual-learning regularizer, without class-label targets. Labels are retained for evaluation: frozen linear probes (tokenizer-axis diagnostics) and Task-IL supervised heads under the fixed readout protocol. We maintain an ontology mapping from original integer codes to canonical MI categories (e.g., left_hand, right_hand, feet), noting that integer codes are *not* commensurate across datasets. Main-21 Task-IL and held-out probes use *dataset-specific* heads/probes under this ontology; we do not force a single global class space across stages. Ambiguous or rest-only labels are documented and filtered per evaluation protocol rather than forcibly merged.

J.7 Split construction, OFAT isolation, and held-out scope

Within-stage train/eval (Main-21 and five-stage screens). Both protocols use within-file $80/20$ train/hold-out splits: for each loaded H5 file, a seeded shuffle holds out $\approx$20% of trials; those evaluation indices never enter CPT, AR fine-tuning, head training, or replay for that run. For the five stages shared with OFAT, reserved OFAT probe trial indices are removed from the Main-21 candidate pool before this draw (paragraph below). Train and eval may share recording file and subject. This is a leak-controlled *trial* split for comparing CPT recipes under a shared readout, not a subject-disjoint or session-disjoint generalization claim.

OFAT versus Main-21 ranking. Continuous weights for CAP-CL / Feat-KD@Θ_{t-1} / DER++ are locked once by five-stage $\mathrm{old\text{-}probe}$ (Appendix F) and transferred unchanged to Main-21. Selection never uses Main-21 $\mathrm{Final/BWT}$ and never uses held-out sets in $\mathcal{H}$. The five OFAT stages share dataset identities with Main-21 (Zhou2016, Weibo2014, PhysionetMI, Lee2019_MI, GrosseWentrup2009). At split construction we reserve a fixed OFAT probe subset for those names and exclude those trial indices from every Main-21 evaluation pool (Appendix B), so OFAT selection probes and Main-21 ranking evals are trial-disjoint even where dataset names coincide. Preprocessing statistics and early stopping likewise never see Main-21 ranking metrics or $\mathcal{H}$.

Held-out $\mathcal{H}$. Yang2025 and Jeong2020MI are excluded from TUEG pretraining, CPT stages, and tuning. Yang2025 is the corpus-unseen transfer probe; Jeong2020MI is a same-family paradigm-transfer probe paired with the Jeong2020ME stream companion (both have n=25, 60 channels, and 2500 Hz). Duplicate or derived releases are assigned a single registry identity.

Replay capacity. Replay-using methods use a per-domain trial cap (1024 waveforms per prior domain).

J.8 Quality control and reproducibility

Automated QC records retention, channel and shape checks, label/event consistency, and interpolation summaries; deterministic seeds, software versions, and manifests support reproducible audit trails.