EGOEXO-VIEWCL: VIEW-LABELED CONTINUAL ACTION LEARNING WITH STABLE ANALYTIC ANCHORS AND VIEW-FACTORIZED ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual video learning is usually evaluated as a sequence of new action classes, yet an embodied learner also moves between first- and third-person views. A model may preserve action semantics within an observed view yet lose cross-view transfer, or adapt by overwriting useful pretrained structure. We introduce EGOEXO-VIEWCL, a view-labeled continual action benchmark built from EgoMe, Assembly101, and Charades-Ego. Its three protocols isolate view shift under a fixed label space, joint class-and-view expansion, and sequential paired-view arrival. The benchmark preserves native single- and multi-label targets and evaluates both final recognition and continual dynamics, including view-conditioned interference, view-order robustness, and paired-view consistency. We further propose VFCT, which combines an order-stable analytic anchor, bounded semantic adaptation, and cumulative class–view prototype score mixing. Controlled evaluation keeps the observation set fixed when measuring view order, averages five class orders, and crosses paired-view arrival direction with true and pseudo pairing. Across the three datasets, VFCT consistently improves recognition under view switches, joint class–view expansion, and paired-view transitions. Continual and protocol-specific diagnostics show that retaining shared action evidence and adapting view-dependent cues are complementary throughout the stream. These results establish viewpoint order, pairing, and their interaction with class arrival as important axes for continual action learning.

1 INTRODUCTION

People routinely connect an action seen from outside with the same action performed from their own viewpoint. Video models do not receive this invariance for free. Egocentric footage emphasizes hands, manipulated objects, and camera motion, whereas exocentric footage exposes body pose and scene context. Paired datasets such as Charades-Ego (Sigurdsson et al., 2018), Assembly101 (Sener et al., 2022), and EgoMe (Qiu et al., 2025) make this distinction observable; Ego-Exo4D and EgoExoLearn broaden it to skilled and asynchronous procedural activity (Grauman et al., 2024; Huang et al., 2024). Their standard recognition protocols nevertheless assume a stationary training set. Conversely, most continual-learning (CL) benchmarks vary classes or datasets while leaving viewpoint implicit (De Lange et al., 2021).

An agent may first observe third-person demonstrations and later receive first-person experience, with earlier observations available only through bounded memory. Viewpoint is therefore a continual dimension rather than merely an offline domain shift: arrival order determines whether shared action knowledge survives changing visual evidence. Class-incremental evaluation cannot distinguish forgetting from a cross-view gap, and a final recognition score can hide retention restricted to the dominant view. Paired observations also make correspondence and arrival direction factors in sequential transfer. Offline view-invariant learning (Li et al., 2021; Luo et al., 2025) and test-time adaptation (Shi et al., 2026) do not address sequential training, while video CL emphasizes class arrival (Park et al., 2021; Villa et al., 2022; 2023). We study both within one recognition interface that controls view order, class order, and paired-view arrival direction.

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.

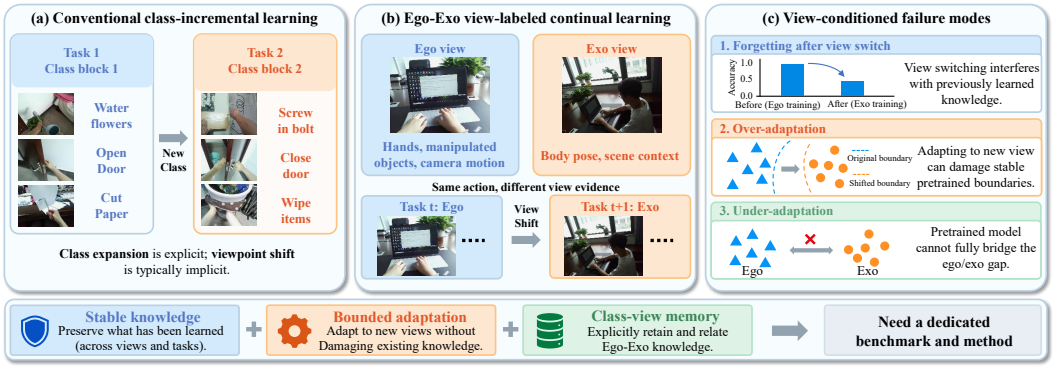

Figure 1: Motivation and challenges of ego–exo view-labeled continual learning.

Beyond defining the stream, this setting changes what a continual learner should preserve and adapt. Frozen pretrained representations can provide separable features for many downstream classes, so unrestricted sequential optimization may harm more than it helps. Methods such as RanPAC (McDonnell et al., 2023) avoid parameter drift through accumulated statistics, yet a fixed representation cannot repair a systematic ego/exo mismatch. This leads to a more specific question: *can a learner preserve shared action structure while adapting only the view-dependent evidence that it lacks?* We answer with VFCT, which factorizes continual knowledge into a shared analytic foundation, a bounded semantic correction, and view-conditioned prototype memory. Its Stable Anchor combines order-independent statistics with support-weighted view-specific scores, while the learned and memory paths adapt and retain complementary action evidence.

We make three contributions. **(1)** EGOEXO-VIEWCL defines three view-labeled CL protocols over substantially different paired video datasets, including a native multi-label formulation for Charades-Ego. **(2)** We propose VFCT, which organizes continual recognition into a Stable Anchor for order-independent analytic statistics and view-specific class evidence, Bounded Semantic Adaptation for controlled correction, and Class–View Prototype Memory for reusing historical evidence across views. **(3)** We expose three distinct continual failure modes—view interference, class–view entanglement, and paired-view transfer instability— and diagnose them through learning dynamics, view-order robustness, stability–plasticity, and paired-view consistency.

2 RELATED WORK

Egocentric and ego–exo video. Egocentric datasets support action understanding and episodic reasoning (Damen et al., 2022; Grauman et al., 2022). Paired datasets connect actor/observer, wearable/static-camera, and observing/imitating videos (Sigurdsson et al., 2018; Sener et al., 2022; Qiu et al., 2025); Ego-Exo4D and EgoExoLearn broaden this setting (Grauman et al., 2024; Huang et al., 2024). Prior work uses correspondence for transfer, language alignment, translation, and test-time adaptation (Li et al., 2021; Lin et al., 2022; Luo et al., 2025; Shi et al., 2026); these methods largely assume jointly available views rather than a memory-limited stream.

Video continual learning. Video CL studies sequential class arrival and prompting (Park et al., 2021; Villa et al., 2022; 2023). CE^{4}L evaluates continual ego/exo objectives with routed adapters and no test-time task identity (Yan et al., 2026). Neither this line nor offline ego–exo learning treats viewpoint itself as an axis of knowledge evolution. EGOEXO-VIEWCL closes this gap by controlling view-only, class–view, and paired arrival in one interface.

Continual-learning mechanisms. EWC and LwF constrain parameter change; replay and DER++ revisit examples or logits; and iCaRL combines exemplars with class means (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Rolnick et al., 2019; Buzzega et al., 2020; Rebuffi et al., 2017). GEM/A-GEM constrain gradients, while BiC/PODNet address classifier bias and distillation (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Wu et al., 2019; Douillard et al., 2020).

Table 1: Protocol definitions. All tasks are evaluated without task identity. "All 50" denotes the fixed 50-class label space.

Protocol	Tasks	Label support	View stream	Pairing and diagnostic focus
A	4	All 50/task	{(E, X, E, X), (X, E, X, E), (E, E, X, X), (X, X, E, E)}	One view/cohort; view-order interference
B	5	10 new classes/task	One fixed view/class block	Five class orders; class–view co-evolution
C	4	All 50/task	(X, E, X, E) or (E, X, E, X)	True/pseudo pairs; two arrival directions

Continual learning with pretrained representations. Prompt methods keep the backbone fixed and learn compact instructions (Wang et al., 2022c;b;a; Smith et al., 2023); SimpleCIL and APER use frozen or adapted features (Zhou et al., 2025); and EASE expands adapter subspaces (Zhou et al., 2024). ACIL and RanPAC accumulate analytic solutions (Zhuang et al., 2022; McDonnell et al., 2023). VFCT adds bounded correction and class–view prototypes to this stable path.

3 EGOEXO-VIEWCL

3.1 PROBLEM DEFINITION

Intuitively, each task reveals a subset of action–view combinations rather than only new semantic classes. Formally, at task t, the learner receives $\mathcal{D}_t = \{(v_i, \boldsymbol{y}_i, e_i, r_i)\}$, where v_i is a video observation, $C = |\mathcal{Y}|$, $\boldsymbol{y}_i \in \{0,1\}^C$ is one-hot for single-label data and multi-hot for multi-label data, $e_i \in \{\text{ego}, \text{exo}\}$ is the viewpoint, and r_i optionally identifies a cross-view counterpart. Let $\mathcal{S}_t \subseteq \mathcal{Y} \times \{\text{ego}, \text{exo}\}$ denote the action–view support exposed at task t. A continual ego–exo stream changes $\mathcal{S}_t$ by revisiting known actions from a different viewpoint, introducing new actions under a changing viewpoint, or presenting the two views of a paired cohort in sequence. These cases separate view-conditioned interference, class acquisition, and cross-view transfer rather than collapsing them into generic distribution shift.

The learner processes each task sequentially and can revisit previously supervised task instances only through a bounded memory. In multi-label Protocol B, a raw video may recur with a newly revealed label block; its previously supervised labels remain unavailable outside replay. The viewpoint label e_i is available during training and inference: view-aware methods may use it, whereas view-agnostic methods ignore it. Task identity is unavailable at test time. After each task, the learner is evaluated on every test stream seen so far. We report final recognition, retention, performance in each view, view gap, and paired-view consistency. For pairs with a common active label space, paired-view consistency is top-1 prediction agreement for single-label data and mean Jaccard agreement between predicted label sets for multi-label data, using one validation-selected threshold (two empty sets have agreement one).

3.2 DATASETS AND TARGETS

We instantiate the 50-class setting of EGOEXO-VIEWCL on three paired video datasets. Its classes are selected by training-pair frequency, ties are resolved lexicographically, and labels are remapped once before protocol construction. We retain the official train, validation, and test boundaries, and no cross-view pair spans two splits.

EgoMe pairs observing and imitating videos and defines single-label, video-level activity recognition. Assembly101 aligns wearable and static-camera segments, which we map to ego and exo views, and uses single-label coarse-action targets. Charades-Ego pairs actor and observer videos and retains its native video-level multi-label targets rather than forcing simultaneous actions into one class. Protocol B masks unintroduced Charades-Ego labels rather than treating them as negatives.

3.3 THREE PROTOCOLS

Let E denote an egocentric task and X an exocentric task. Table 1 separates the factors changed by each protocol: the class support available at a task, the order of observed views, and whether the two observations of a pair become available sequentially.

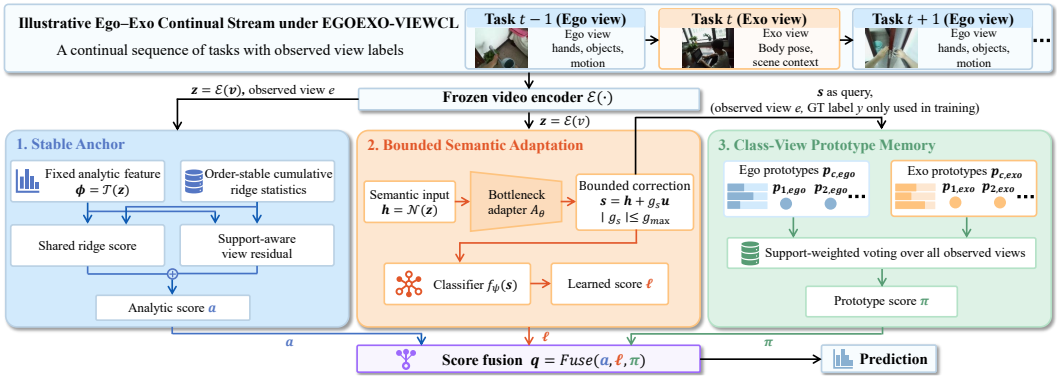

Figure 2: Overview of VFCT.

A: View-Order Continual Learning. Protocol A varies the order of four view-specific tasks while fixing the 50-class label space and the observation, label, and view multiset. Alternating and blocked schedules measure sensitivity to view order on the same evidence.

B: Joint Class-and-View Incremental Learning. Protocol B introduces five disjoint blocks of ten classes, each observed in one view, and evaluates prediction over all classes seen so far. We use frequency, positive-count-balanced, and three random class orders. Charades-Ego supervises only the current label block, treating other labels as unknown; a video may contribute to multiple blocks in the same view.

C: Sequential Paired-View Continual Learning. Protocol C divides paired observations into two cohorts and presents each cohort's ego and exo views in adjacent tasks. Both arrival directions are evaluated. A pseudo-pair control rearranges exo observations within class across cohorts, preserving the sample set, view counts, and task sizes while changing cross-view adjacency and task composition. Paired-view consistency is always evaluated using the original cross-view correspondences.

4 VFCT: Stable and View-Factorized Continual Learning

VFCT factorizes continual knowledge according to three challenges: view-order interference calls for shared, order-stable evidence; joint class–view expansion calls for controlled semantic correction; and paired-view arrival calls for persistent, view-conditioned class evidence.

Scalars are italic, vectors and matrices bold italic, fixed labels and named functions upright, and sets and maps calligraphic. For observation v_i, the frozen encoder gives $\boldsymbol{z}_i = \mathcal{E}(v_i) \in \mathbb{R}^{d_z}$. The Stable Anchor uses $\boldsymbol{\phi}_i = \mathcal{T}(\boldsymbol{z}_i) \in \mathbb{R}^{d_\phi}$ from the shared frozen random map; the learned and prototype paths use $\boldsymbol{h}_i = \mathcal{N}(\boldsymbol{z}_i) \in \mathbb{R}^{d_{\text{sem}}}$. The paths return $\boldsymbol{a}_i, \boldsymbol{\ell}_i, \boldsymbol{\pi}_i \in \mathbb{R}^C$, fused as $\boldsymbol{q}_i \in \mathbb{R}^C$. Analytic features remain fixed and $\mathcal{N}$ is identity in our experiments. Cross-view correspondences define Protocol C and optional paired replay; action labels supervise classification. We omit unambiguous task indices from cumulative states.

Three components. The corresponding factors are **(I) Stable Anchor**, which combines order-independent ridge statistics with view-specific score residuals; **(II) Bounded Semantic Adaptation**, which learns a limited shared correction; and **(III) Class–View Prototype Memory**, which retains view-conditioned class summaries (Figure 2).

Use of view labels. View labels have a restricted routing role. During training, they assign analytic observations and prototype updates to ego or exo cells; during inference, they select the residual head in Component I. The shared analytic head and the semantic adapter use action supervision but do not branch on the view. Component III votes over every supported view prototype for a class, so the current test view does not select a prototype. Thus view metadata structures the stable memories without creating task-specific prediction interfaces.

4.1 Component I: Stable Anchor

Let $\mathcal{I}_t$ index the samples in task t, and let $m_{i,c} \in \{0,1\}$ indicate whether the target for class c is observed for sample i. For single-label data every class target is observed; for multi-label Protocol B, labels not yet introduced have $m_{i,c} = 0$. The shared anchor solves the masked ridge problem

$$\min_{\boldsymbol{w}_c} \sum_{k \le t} \sum_{i \in \mathcal{I}_k} m_{i,c}(\boldsymbol{\phi}_i^\top \boldsymbol{w}_c - y_{i,c})^2 + \lambda \|\boldsymbol{w}_c\|_2^2, \tag{1}$$

a classical ridge model (Hoerl & Kennard, 1970). Rather than retaining past features, VFCT maintains label-wise sufficient statistics

$$\boldsymbol{G}_{t,c} = \boldsymbol{G}_{t-1,c} + \sum_{i \in \mathcal{I}_t} m_{i,c} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \quad \boldsymbol{b}_{t,c} = \boldsymbol{b}_{t-1,c} + \sum_{i \in \mathcal{I}_t} m_{i,c} \boldsymbol{\phi}_i y_{i,c}, \tag{2}$$

with $\boldsymbol{w}_{t,c}^{\mathrm{sh}} = (\boldsymbol{G}_{t,c} + \lambda \mathbf{I})^{-1} \boldsymbol{b}_{t,c}$. Here $\boldsymbol{G}_{t,c}, \mathbf{I} \in \mathbb{R}^{d_\phi \times d_\phi}$ and $\boldsymbol{b}_{t,c}, \boldsymbol{w}_{t,c} \in \mathbb{R}^{d_\phi}$. Classes with the same mask share one Gram accumulator.

Proposition 1 (Order-stable analytic memory)**.** *For fixed analytic features, fixed $\lambda > 0$, and fixed observations and label masks, accumulating Eq. 2 under any task partition or order gives the same classifier as masked batch ridge, in exact arithmetic.*

This follows from commutativity of the normal-equation sums, yielding a compressed, order-stable record (Zhuang et al., 2022).

Analytic view-score residual. The shared solution is stable but can underfit systematic ego/exo differences. We therefore accumulate the same statistics separately for each view as $\boldsymbol{G}_{t,c}^e$ and $\boldsymbol{b}_{t,c}^e$, with $\boldsymbol{w}_{t,c}^e = (\boldsymbol{G}_{t,c}^e + \lambda \mathbf{I})^{-1} \boldsymbol{b}_{t,c}^e$. These heads yield $a_{i,c}^{\mathrm{sh}} = \boldsymbol{\phi}_i^\top \boldsymbol{w}_{t,c}^{\mathrm{sh}}$ and $a_{i,c}^e = \boldsymbol{\phi}_i^\top \boldsymbol{w}_{t,c}^e$. With the number of positive labeled observations accumulated by the analytic branch, $N_{c,e}^{\mathrm{ana}}$, the Stable Anchor shrinks the view-specific score toward the shared score:

$$\begin{aligned} \kappa_{c,e} &= \frac{N_{c,e}^{\mathrm{ana}}}{N_{c,e}^{\mathrm{ana}} + \tau_{\mathrm{view}}}, \\ a_{i,c} &= a_{i,c}^{\mathrm{sh}} + \alpha_{\mathrm{view}} \mathbf{1}\{N_{c,e_i}^{\mathrm{ana}} > 0\} \kappa_{c,e_i} \big(a_{i,c}^{e_i} - a_{i,c}^{\mathrm{sh}}\big), \\ & 0 \le \alpha_{\mathrm{view}} \le 1, \qquad \tau_{\mathrm{view}} \ge 0. \end{aligned} \tag{3}$$

The residual is zero for unseen class–view cells. The support factor prevents a sparsely observed view head from immediately overriding the shared estimate. As support grows, $\kappa_{c,e}$ approaches one and the corresponding score difference is admitted up to the global multiplier α_{view}; larger τ_{view} makes this transition more conservative. Because the shared and view-specific heads operate on the same fixed feature $\boldsymbol{\phi}_i$, the residual adapts class evidence while leaving the pretrained representation unchanged.

4.2 Component II: Bounded Semantic Adaptation

Bounded semantic correction. A fixed anchor cannot recover class structure absent from the frozen descriptor, so the learned path supplies a deliberately limited correction. A bottleneck network produces $\boldsymbol{u}_i^{\mathrm{raw}}$, normalized as $\boldsymbol{u}_i = \|\boldsymbol{h}_i\|_2 \boldsymbol{u}_i^{\mathrm{raw}} / \max(\|\boldsymbol{u}_i^{\mathrm{raw}}\|_2, 10^{-12})$. A zero-initialized bounded gate then gives

$$\boldsymbol{s}_i = \boldsymbol{h}_i + g_{\mathrm{sem}} \boldsymbol{u}_i, \qquad g_{\mathrm{sem}} = g_{\max} \tanh(\gamma_{\mathrm{sem}}), \quad |g_{\mathrm{sem}}| \le g_{\max}. \tag{4}$$

γ_{sem} is a trainable scalar initialized to zero, and the fixed constant $g_{\max}$ sets the largest permitted correction. The class head reads only the semantic feature $\boldsymbol{s}_i$ and returns the logit vector $\boldsymbol{\ell}_i$. A cosine classifier with cross entropy is used for single-label targets; a linear classifier with masked binary cross entropy is used for multi-label targets.

Proposition 2 (Pointwise bounded adaptation)**.** *If $0 < \|\boldsymbol{h}_i\|_2$, $\|\boldsymbol{u}_i\|_2 \le \|\boldsymbol{h}_i\|_2$, and $0 \le g_{\max} < 1$, then $\|\boldsymbol{s}_i - \boldsymbol{h}_i\|_2 / \|\boldsymbol{h}_i\|_2 \le g_{\max}$ and the angle between $\boldsymbol{s}_i$ and $\boldsymbol{h}_i$ is at most $\arcsin(|g_{\mathrm{sem}}|) \le \arcsin(g_{\max})$.*

At initialization, $g_{\mathrm{sem}} = 0$ and hence $\boldsymbol{s}_i = \boldsymbol{h}_i$. The pointwise bound limits each current correction; it is a stability constraint, not by itself a guarantee against temporal forgetting.

4.3 COMPONENT III: CLASS–VIEW PROTOTYPE MEMORY

Support-weighted per-view score voting. To reuse action evidence across view changes, we maintain cumulative class statistics separately for ego and exo observations and combine their scores according to observed support. For class c and view e, VFCT keeps a cumulative semantic mean $\boldsymbol{p}_{c,e} \in \mathbb{R}^{d_{\text{sem}}}$, normalized when scored, and its support count $N_{c,e}^{\text{proto}}$. If an update contributes $m_{c,e}^{+}$ positive examples with mean $\boldsymbol{\mu}_{c,e}$, then

$$\boldsymbol{p}_{c,e} \leftarrow \frac{N_{c,e}^{\text{proto}} \boldsymbol{p}_{c,e} + m_{c,e}^{+} \boldsymbol{\mu}_{c,e}}{N_{c,e}^{\text{proto}} + m_{c,e}^{+}}, \qquad N_{c,e}^{\text{proto}} \leftarrow N_{c,e}^{\text{proto}} + m_{c,e}^{+}. \tag{5}$$

For multi-label data, an example contributes to cell (c, e) only when $m_{i,c} = 1$ and $y_{i,c} = 1$; unknown labels contribute neither features nor support counts. Counts refer to accumulated update contributions, including repeated training visits, rather than necessarily unique videos. Let $\mathcal{V}_c = \{e : N_{c,e}^{\text{proto}} > 0\}$ be the supported views of class c. For a supported class, the prototype branch assigns

$$\omega_{c,e} = \frac{N_{c,e}^{\text{proto}}}{\sum_{e' \in \mathcal{V}_c} N_{c,e'}^{\text{proto}}}, \qquad \pi_{i,c} = \eta_{\text{proto}} \sum_{e \in \mathcal{V}_c} \omega_{c,e} \cos(\boldsymbol{s}_i, \boldsymbol{p}_{c,e}), \tag{6}$$

with $\omega_{c,e} = 0$ outside $\mathcal{V}_c$; $\eta_{\text{proto}} > 0$ sets the scale and unseen classes are masked. With one observed view, this reduces to an ordinary prototype score. Here $\pi_{i,c}$ is a class score, not a probability. Support-weighted voting keeps both view summaries available while preventing a small class–view cell from contributing as strongly as a well-supported cell. It also avoids forcing ego and exo features into a single centroid: cross-view evidence is combined at the score level, after each view has retained its own semantic summary.

Same-cell stability and training. At the beginning of each task, the per-cell prototype snapshot is detached as $\boldsymbol{p}_{c,e}^{0}$. For the set Ω_t of class–view cells in the current training batch (including replay) that already exist in this snapshot, let $\boldsymbol{\mu}_{c,e}$ be the current semantic mean. We use

$$\mathcal{L}_{\text{st}} = \frac{1}{\max(1, |\Omega_t|)} \sum_{(c,e) \in \Omega_t} \big(1 - \cos(\boldsymbol{\mu}_{c,e}, \boldsymbol{p}_{c,e}^{0})\big), \tag{7}$$

which is zero when Ω_t is empty. This regularizes the current semantic mean toward the stored mean of the same class and view. The learned objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{st}} \mathcal{L}_{\text{st}}, \tag{8}$$

where $\mathcal{L}_{\text{cls}}$ is cross entropy or masked BCE. Analytic heads and prototype means are updated directly rather than by this gradient loss.

Task update. At the beginning of task t, we snapshot each existing prototype as $\boldsymbol{p}_{c,e}^{0}$. The semantic path is then optimized with Eq. 8, and current non-replay samples update Eq. 5 at each training visit using the detached semantic features computed for that batch. Replay affects the learned loss but is not re-counted in prototype updates. Consequently, stored means combine features from historical adapter states; the pointwise bound does not make them current-space historical centroids. At the end of the task, frozen descriptors update the analytic statistics in Eq. 2 and the ridge heads are solved.

Inference. Let $\text{Std}(\boldsymbol{x})$ standardize a sample's score vector across observed classes by subtracting its mean and dividing by $\max(\text{std}(\boldsymbol{x}), 10^{-6})$. For single-label prediction, fusion is performed in two stages:

$$\widetilde{\boldsymbol{q}}_i = (1 - \alpha_{\text{learn}}) \,\text{Std}(\boldsymbol{a}_i) + \alpha_{\text{learn}} \,\text{Std}(\boldsymbol{\ell}_i), \qquad \boldsymbol{q}_i = (1 - \alpha_{\text{proto}}) \,\text{Std}(\widetilde{\boldsymbol{q}}_i) + \alpha_{\text{proto}} \,\text{Std}(\boldsymbol{\pi}_i). \tag{9}$$

For multi-label prediction, the ridge branch retains its regression scores:

$$\begin{aligned} \boldsymbol{q}_i = {} & (1 - \alpha_{\text{proto}})(1 - \alpha_{\text{learn}}) \boldsymbol{a}_i \\ & + (1 - \alpha_{\text{proto}}) \alpha_{\text{learn}} \,\text{sigmoid}(\boldsymbol{\ell}_i) \\ & + \alpha_{\text{proto}} \,\text{sigmoid}(\boldsymbol{\pi}_i), \qquad \alpha_{\text{learn}}, \alpha_{\text{proto}} \in [0, 1]. \end{aligned} \tag{10}$$

where sigmoid acts elementwise. The fused values are ranking scores. Never-observed classes remain masked. Fusion and shrinkage parameters are selected on validation data and locked before test evaluation.

Table 2: EgoExo-ViewCL overall results (%). EgoMe and Assembly101 use accuracy; Charades-Ego uses macro-mAP. Each cell is mean ± std, with schedules averaged within each seed before the three-seed standard deviation. Bold is column-best.

Method	EgoMe			Assembly101			Charades-Ego		
	A	B	C	A	B	C	A	B	C
Sequential									
Fine-tuning	$40.32_{\pm.60}$	$11.18_{\pm1.81}$	$49.24_{\pm.58}$	$28.38_{\pm.19}$	$11.29_{\pm.08}$	$29.74_{\pm.14}$	$26.21_{\pm.11}$	$22.18_{\pm1.63}$	$25.45_{\pm.19}$
Regularization									
EWC (2017)	$42.50_{\pm1.44}$	$15.36_{\pm.22}$	$50.72_{\pm.49}$	$27.21_{\pm.20}$	$11.68_{\pm.33}$	$29.16_{\pm.20}$	$26.28_{\pm.05}$	$26.03_{\pm.83}$	$25.77_{\pm.13}$
LwF (2017)	$44.95_{\pm1.18}$	$23.69_{\pm.71}$	$52.98_{\pm.58}$	$28.55_{\pm.23}$	$18.54_{\pm.43}$	$31.13_{\pm.52}$	$26.08_{\pm.15}$	$28.67_{\pm1.11}$	$26.50_{\pm.11}$
Replay									
ER (2019)	$47.91_{\pm.41}$	$55.74_{\pm.76}$	$52.84_{\pm.29}$	$29.66_{\pm.36}$	$30.76_{\pm.98}$	$31.16_{\pm.18}$	$26.51_{\pm.47}$	$30.30_{\pm.16}$	$25.85_{\pm.15}$
DER++ (2020)	$48.65_{\pm.64}$	$56.80_{\pm.78}$	$56.24_{\pm1.25}$	$30.52_{\pm.30}$	$35.65_{\pm.11}$	$31.82_{\pm.31}$	$26.18_{\pm.24}$	$29.63_{\pm.09}$	$26.54_{\pm.12}$
Prototype									
iCaRL (2017)	$34.41_{\pm.51}$	$51.21_{\pm.15}$	$41.49_{\pm.80}$	$19.34_{\pm.94}$	$31.55_{\pm.38}$	$22.01_{\pm.56}$	$20.05_{\pm.35}$	$23.15_{\pm.20}$	$19.84_{\pm.16}$
RanPAC (2023)	$56.91_{\pm.31}$	$69.42_{\pm.36}$	$60.54_{\pm.59}$	$29.73_{\pm1.29}$	$38.44_{\pm.38}$	$31.19_{\pm.21}$	$26.67_{\pm.43}$	$34.03_{\pm.04}$	$26.07_{\pm.17}$
View-aware									
ViewLN-ER	$48.29_{\pm1.08}$	$57.22_{\pm.36}$	$54.15_{\pm.78}$	$27.75_{\pm.42}$	$30.31_{\pm.21}$	$29.79_{\pm.25}$	$26.27_{\pm.13}$	$28.02_{\pm.29}$	$25.25_{\pm.13}$
CE^4L (2026)	$49.00_{\pm.68}$	$47.64_{\pm.20}$	$56.48_{\pm.16}$	$29.49_{\pm.32}$	$26.49_{\pm.22}$	$32.18_{\pm.09}$	$27.14_{\pm.16}$	$32.67_{\pm.10}$	$27.20_{\pm.24}$
Ours									
VFCT	$\mathbf{58.49_{\pm.45}}$	$\mathbf{73.48_{\pm.76}}$	$\mathbf{65.98_{\pm.52}}$	$\mathbf{32.66_{\pm.46}}$	$\mathbf{41.26_{\pm.46}}$	$\mathbf{34.10_{\pm.27}}$	$\mathbf{30.61_{\pm.03}}$	$\mathbf{38.01_{\pm.03}}$	$\mathbf{30.19_{\pm.01}}$

5 Experiments

5.1 Setup

We compare fine-tuning, EWC, LwF, ER, DER++, iCaRL, RanPAC, ViewLN-ER, CE^4L, and VFCT on identical feature streams and validation protocols. Replay methods share a 500-example exemplar budget; analytic statistics and prototype banks constitute additional method state.

EgoMe and Assembly101 use top-1 accuracy as the task score; Charades-Ego uses macro-mAP over the labels active in the evaluated task.

For this experimental instantiation, EgoMe and Charades-Ego use frozen TSP descriptors (Alwassel et al., 2021), while Assembly101 uses frozen TSM descriptors (Lin et al., 2019). All model and fusion hyperparameters are selected on validation data and locked before test evaluation. We use three seeds and report their mean and standard deviation. Within each seed, Protocol A first averages four fixed-set schedules, B averages five class orders, and C averages true/pseudo pairing crossed with two arrival directions. The three resulting seed means determine the reported mean and standard deviation.

5.2 Results

Main comparison. VFCT ranks first among the ten compared methods in all nine dataset–protocol cells (Table 2). Gains in B concern adaptation under joint expansion, whereas C permits reuse of same-class cross-view history.

Continual-learning diagnostics. Figure 3 averages the three datasets equally. Protocol B averages five class orders within each seed; Protocol C averages both arrival directions and three seeds. In A, fine-tuning benefits past tasks after same-view updates but harms them after view switches, whereas VFCT preserves both responses. This contrast shows that view shift is not merely a generic distribution change: shared action evidence must remain stable while view-dependent cues adapt. The same complementarity sustains VFCT throughout B. In C, stable first-view performance shows that second-view learning preserves previously acquired evidence.

A: robustness to view order. Figure 4 (left) holds the sample, label, and view multiset fixed and varies only task assignment. VFCT's four-schedule range is 0.60 points, versus 1.66 for fine-tuning. RanPAC varies less but has lower accuracy: VFCT improves recognition while retaining limited order sensitivity.

B: stability–plasticity under joint change. The center panel separates retained performance on the first four Assembly101 blocks from learning of the fifth, averaging five class orders within each

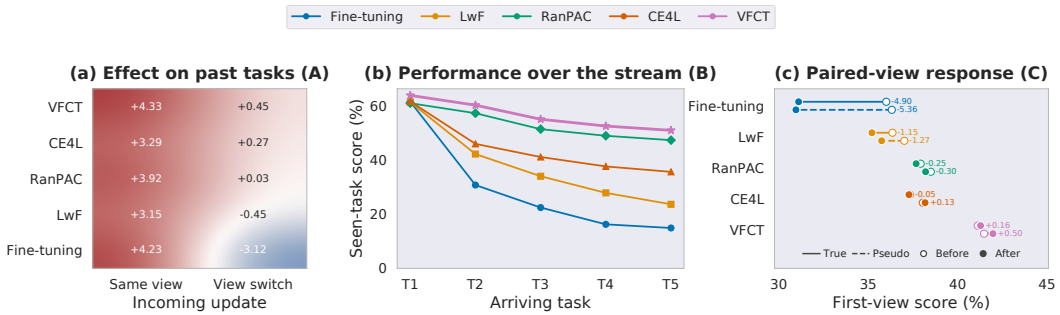

Figure 3: Continual-learning diagnostics. Left: Protocol A previous-task score changes. Center: Protocol B seen-task performance. Right: Protocol C first-view retention; open/filled denotes before/after second-view arrival, solid/dashed lines true/pseudo pairs.

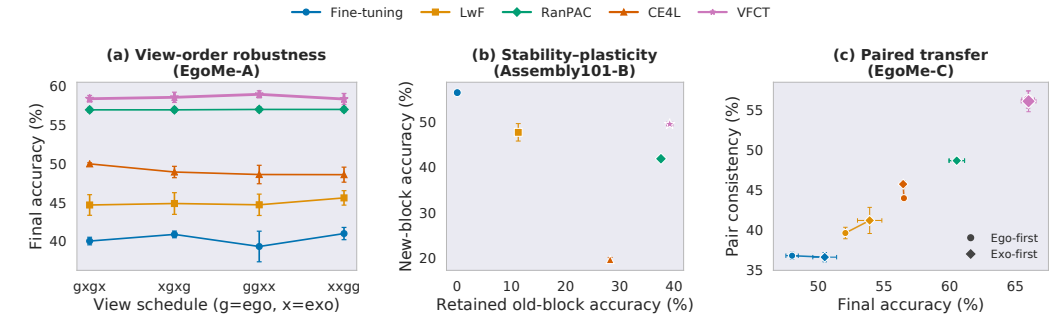

Figure 4: Protocol diagnostics. Left: EgoMe-A view-order robustness (mean ± seed SD). Center: Assembly101-B old/new-block accuracy. Right: EgoMe-C recognition and paired-view consistency; filled/open denotes true/pseudo cohorts, circles/diamonds ego-first/exo-first arrival.

seed. VFCT improves both old-block retention and new-block learning over RanPAC, with the larger gain in acquisition. This suggests useful plasticity beyond the analytic foundation: stability alone is insufficient when classes and views expand together. Fine-tuning instead learns the new block at the cost of old-block recognition.

C: paired-stream recognition and arrival direction. The right panel relates recognition to paired-view consistency across both arrival directions. VFCT occupies the high-accuracy, high-consistency region: its agreement accompanies correct recognition rather than merely similar predictions across views. True and pseudo conditions remain close within each direction, while exo-first arrival yields higher consistency. Here, arrival direction distinguishes cross-view behavior more clearly than cohort pairing.

Cumulative component analysis. We expand the Stable Anchor into its shared analytic prediction and its view-specific score residual, then add Bounded Semantic Adaptation and Class–View Prototype Memory. Each stage uses its corresponding training objective; learned-score-only provides a separate reference. Protocol scores average conditions within each seed and then the three seeds; Mean averages the protocols.

Before memory, semantic training uses classification loss alone; Full adds memory scores and stability training. The largest semantic increment occurs in B, while memory contributes most in C, linking adaptation to joint expansion and memory to cross-view reuse. Figure 5 shows six classes selected by training frequency. Frozen and semantic projections are fitted separately; prototypes use the semantic projection.

Path	A	B	C	Mean
Learned only (ref.)	26.97	27.43	28.54	27.65
Shared anchor	31.35	39.27	32.63	34.42
+ View residual	31.92	39.71	32.86	34.83
+ Semantic	32.60	40.84	33.11	35.52
+ Memory (Full)	32.66	41.26	34.10	36.01

Table 3: Assembly101 cumulative training analysis (accuracy, %). Rows successively add components.

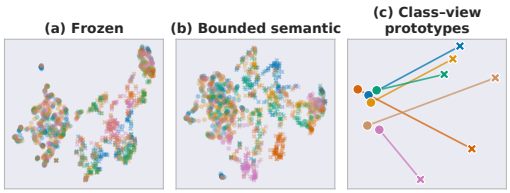

Figure 5: Assembly101-C representations. Colors denote actions; circles/crosses denote ego/exo views.

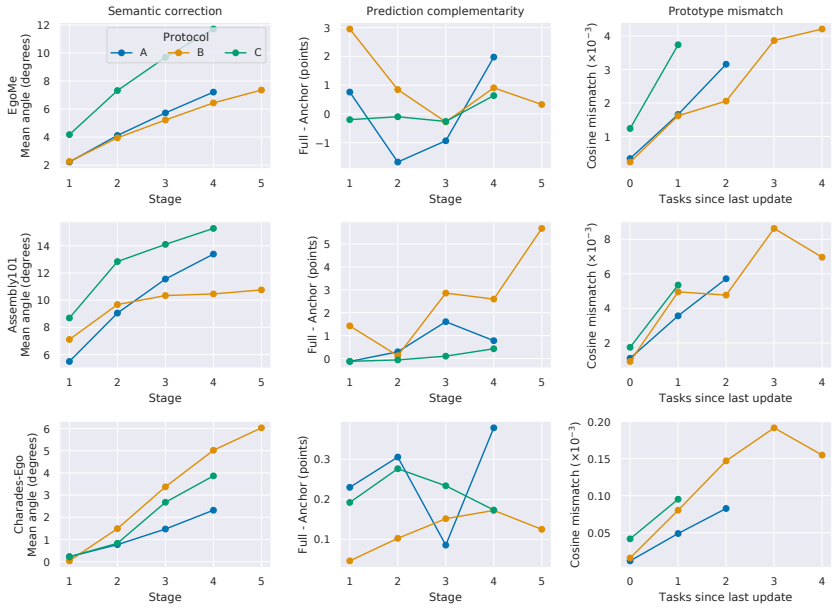

Figure 6: Mechanism diagnostics by dataset (rows) and protocol (lines). Left: semantic correction angle. Center: Full–Anchor score change. Right: prototype cosine mismatch as a function of the number of tasks since its last update.

Mechanism diagnostics. Figure 6 traces how the three components evolve at each task boundary along one prespecified validation stream per protocol and dataset, averaged over three seeds. We use blocked ego-first for A, frequency order for B, and true-pair ego-first for C. As shown in Figure 6, Assembly101 requires the largest semantic correction and obtains the strongest final complement beyond the Stable Anchor. EgoMe occupies an intermediate regime, while Charades-Ego stays closer to the frozen representation and exhibits a smaller, yet positive, final Full–Anchor difference. Prototype mismatch generally grows with the time since a cell was last updated, but remains small in absolute cosine distance across all three datasets.

6 Conclusion

We introduced EgoExo-ViewCL with fixed-observation view-order, joint class–view, and paired-stream protocols, and proposed VFCT with a Stable Anchor, Bounded Semantic Adaptation, and Class–View Prototype Memory. Across the evaluated schedules, class orders, and arrival directions, the diagnostics show that shared action evidence and view-dependent adaptation are complementary. Our findings suggest that viewpoint should be treated as an explicit axis of continual knowledge evolution rather than merely an appearance variation: an effective learner must preserve shared structure, adapt view-specific evidence, and reuse cross-view history.

Ethics Statement

We reuse publicly released human-activity datasets under their respective licenses and introduce no participants or annotations. Ego/exo video can contain identifiable people and private environments; deployment should respect consent, dataset purpose, and privacy constraints. The benchmark is intended for research evaluation, not surveillance or individual assessment.

AI Assistance

Generative AI tools were used to improve the clarity, grammar, concision, and organization of the manuscript. The authors reviewed and revised all assisted text, verified that it accurately reflects the methods and experimental results, and take full responsibility for the final content of the paper.

References

Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3166–3176. IEEE, 2021.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European conference on computer vision*, pp. 86–102. Springer, 2020.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22072–22086. IEEE, 2024.

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 (13):3521–3526, 2017.

Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6939–6949. IEEE, 2021.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15802–15812. IEEE, 2025.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13678–13687. IEEE, 2021.

Heqian Qiu, Zhaofeng Shi, Lanxiao Wang, Huiyu Xiong, Xiang Li, and Hongliang Li. Egome: A new dataset and challenge for following me via egocentric view in real world. *arXiv preprint arXiv:2501.19061*, 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542. IEEE, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21064–21074. IEEE, 2022.

Zhaofeng Shi, Heqian Qiu, Lanxiao Wang, Qingbo Wu, Fanman Meng, Lili Pan, and Hongliang Li. Test-time ego-exo-centric adaptation for action anticipation via multi-label prototype growing and dual-clue consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16988–16999, June 2026.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7396–7404. IEEE, 2018.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Andrés Villa, Kumail Alhamoud, Victor Escorcia, Fabian Caba, Juan León Alcázar, and Bernard Ghanem. vclimb: A novel video class incremental learning benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19035–19044, 2022.

Andrés Villa, Juan León Alcázar, Motasem Alfarra, Kumail Alhamoud, Julio Hurtado, Fabian Caba Heilbron, Alvaro Soto, and Bernard Ghanem. Pivot: Prompting for video continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24214–24223. IEEE, 2023.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022a.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022c.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 374–382. IEEE, 2019.

Hongwei Yan, Kanglei Zhou, Yuchen Liu, Qingyu Shi, Yi Zhong, and Liyuan Wang. Ce^{4}l: Continual ego, exo, and ego-exo learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23554–23564. IEEE, 2024.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. Acil: Analytic class-incremental learning with absolute memorization and privacy protection. *Advances in Neural Information Processing Systems*, 35:11602–11614, 2022.

A BENCHMARK DETAILS

A.1 SPLIT CONSTRUCTION AND AUDITING

EgoMe uses the official train/validation/test files before selecting the 50-class setting of EGOEXO-VIEWCL. The source release contains 7,902 pairs; after official-split and feature validation, our full track retains 7,418/1,600/3,286 view samples, 184 classes, and 6,152 complete pairs. The 50-class setting keeps 2,497 selected pairs (4,994 view samples), split as 3,066/638/1,290 samples. Assembly101 groups all ego and static cameras associated with the same coarse action interval, then samples a deterministic ego/exo pair per group for the 50-class protocol. It contains 6,565 selected pairs, or 7,446/2,190/3,494 view samples across train/validation/test. Charades-Ego uses paired video identity, preserves all active labels among the selected 50 classes, and contains 5,178/908/1,668 videos across the three splits.

Selection is based only on training counts. For every track, the build verifies that sample identifiers are unique, features have the declared shape, labels are in range, pair membership is symmetric, and no pair identifier appears in more than one data split. The task builder is deterministic given the split version, class order, protocol order, and seed. Protocol A additionally verifies equality of the sample–label–view multiset across its four schedules. Protocol B audits source-video reuse, current-block label masking, per-task positive counts, and zero class–view support violations. Protocol C verifies equality of the true and pseudo sample sets and records the fraction of true pairs broken by the deterministic within-class reassignment. Strict subject/scene/group-heldout splits are diagnostic variants, not substitutes for the official main splits.

A.2 METRIC DEFINITIONS

Let $A_{t,j}$ be the task-j score after learning task t. We report

$$\mathrm{Avg}_T = \frac{1}{T}\sum_{j=1}^{T} A_{T,j}, \tag{11}$$

$$\mathrm{Forget}_T = \frac{1}{T-1}\sum_{j<T}\left(\max_{j\le k<T} A_{k,j} - A_{T,j}\right), \tag{12}$$

$$\mathrm{BWT}_T = \frac{1}{T-1}\sum_{j<T}(A_{T,j} - A_{j,j}). \tag{13}$$

For single-label tracks, A is top-1 accuracy; for Charades-Ego it is macro-mAP over labels valid for that task. ViewGap is $|A_T^{\mathrm{ego}} - A_T^{\mathrm{exo}}|$. Single-label pair consistency is the fraction of valid pairs receiving the same prediction. Multi-label pair consistency is Jaccard agreement after selecting one global threshold on validation data. For an eligible pair (u, v), let $\widehat{\mathcal{Y}}_u$ and $\widehat{\mathcal{Y}}_v$ be the labels above the threshold within their common active label set. Agreement is $|\widehat{\mathcal{Y}}_u \cap \widehat{\mathcal{Y}}_v|/|\widehat{\mathcal{Y}}_u \cup \widehat{\mathcal{Y}}_v|$; a pair of empty predicted sets is assigned one. We average over eligible cross-view pairs, so high consistency alone need not indicate correctness. Pairs are valid only when identifiers match, views differ, and active label spaces agree.

For a set $\mathcal{P}$ of eligible pairs, the two consistency metrics are

$$\mathrm{PC}_{\mathrm{single}} = \frac{1}{|\mathcal{P}|}\sum_{(u,v)\in\mathcal{P}} \mathbf{1}\{\arg\max \boldsymbol{q}_u = \arg\max \boldsymbol{q}_v\}, \qquad \mathrm{PC}_{\mathrm{multi}} = \frac{1}{|\mathcal{P}|}\sum_{(u,v)\in\mathcal{P}} \frac{|\widehat{\mathcal{Y}}_u \cap \widehat{\mathcal{Y}}_v|}{|\widehat{\mathcal{Y}}_u \cup \widehat{\mathcal{Y}}_v|}, \tag{14}$$

with the empty-union convention given above. Retention on an old-task set $\mathcal{J}$ is $|\mathcal{J}|^{-1}\sum_{j\in\mathcal{J}} A_{t,j}$. The immediate change for a previously seen task $j < t$ is $A_{t,j} - A_{t-1,j}$. Figure 3(a) averages these changes over all eligible (t, j) pairs within each stream, grouping by whether tasks t and j share a view, then averages streams within each seed. Figure 3 averages within dataset before giving each dataset equal weight. These averages combine track-specific percentage scores rather than pooling individual predictions.

B Implementation Details

Fixed analytic map. All datasets and protocols use the same form of fixed random feature map. For a frozen encoder feature $\boldsymbol{z}_i \in \mathbb{R}^{d_z}$, we first compute $\widehat{\boldsymbol{z}}_i = \boldsymbol{z}_i / \max(\|\boldsymbol{z}_i\|_2, 10^{-12})$ and then set

$$\boldsymbol{\phi}_i = \mathcal{T}(\boldsymbol{z}_i) = \mathrm{ReLU}(\boldsymbol{R}^\top \widehat{\boldsymbol{z}}_i) \in \mathbb{R}^{d_\phi}, \qquad \boldsymbol{R} \in \mathbb{R}^{d_z \times d_\phi}. \tag{15}$$

The entries of $\boldsymbol{R}$ are drawn once from a standard normal distribution using the run seed. Both $\boldsymbol{R}$ and the output dimension d_ϕ are fixed before the continual stream and receive no gradient updates. The shared and per-view analytic heads use this same $\boldsymbol{\phi}_i$; only their cumulative ridge statistics differ. Thus $\mathcal{T}$ introduces no trainable temporal module or dataset-specific analytic branch.

Counts and score fusion. $N_{c,e}^{\mathrm{ana}}$ counts positive, observed class targets in the analytic accumulation pass over each arriving task. It is independent of semantic epochs and excludes replay. $N_{c,e}^{\mathrm{proto}}$ instead counts non-replay contributions at semantic training visits, including repeated visits across epochs. Both follow label masks; neither definition assumes global deduplication of raw videos across tasks.

For observed classes $\mathcal{C}_t$ and a score vector $\boldsymbol{x}$, define

$$\mathrm{Std}(\boldsymbol{x})_c = \frac{x_c - \mu(\boldsymbol{x})}{\sqrt{\max(\nu(\boldsymbol{x}), 10^{-12})}}, \quad \mu(\boldsymbol{x}) = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} x_c, \quad \nu(\boldsymbol{x}) = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} (x_c - \mu(\boldsymbol{x}))^2. \tag{16}$$

Single-label fusion standardizes the analytic and learned endpoints, blends them, and standardizes that blend before mixing with the standardized prototype endpoint. At a fusion weight of zero or one, the implementation selects the corresponding endpoint directly; unseen classes retain their exclusion mask. All reported VFCT fusion weights are strictly between zero and one. For multi-label fusion, the transforms are $f_{\mathrm{ana}}(\boldsymbol{a}) = \boldsymbol{a}$, $f_{\mathrm{learn}}(\boldsymbol{\ell}) = \mathrm{sigmoid}(\boldsymbol{\ell})$, and $f_{\mathrm{proto}}(\boldsymbol{\pi}) = \mathrm{sigmoid}(\boldsymbol{\pi})$, where sigmoid is applied elementwise; the ridge branch is neither clipped nor sigmoid-transformed in the reported configuration. Fusion weights are chosen on validation data; the resulting scores are used for ranking, without assuming calibrated probabilities.

Baseline summary. All methods use the same frozen video descriptors and no test-time task identity. Fine-tuning updates the shared MLP; EWC and LwF add a Fisher penalty and previous-model distillation, respectively. ER replays features and labels, DER++ also stores logits, and iCaRL combines class-balanced exemplars, distillation, and class means. RanPAC fits an analytic classifier on fixed projected features. ViewLN-ER sends ego/exo inputs through separate LayerNorms and a shared classifier, while CE^{4}L uses routed task adapters under this recognition interface. Replay methods share a 500-example budget, and Charades-Ego losses mask labels not yet revealed.

B.1 Metric Reference

Let $A_{t,j}$ be the score on task j after task t, and A_t^e the score evaluated on view e. All scores are percentages; multi-label scores use the evaluated task's observed label set.

Table 4: Retention and view diagnostics.

Quantity	Definition
Old-task retention	$\lvert\mathcal{J}\rvert^{-1} \sum_{j \in \mathcal{J}} A_{t,j}$ for the stated old-task set
Immediate old-task change	$A_{t,j} - A_{t-1,j}$ for $j < t$, grouped by the two tasks' views
First-view retention	First-view task score before and after its second-view arrival
Old/new-block accuracy	Mean of the first four block scores / fifth block score at the final stage of B
View gap	$\lvert A_t^{\mathrm{ego}} - A_t^{\mathrm{exo}} \rvert$

C Extended Main Results

Tables 5–7 extend the main comparison with ego- and exo-view performance. "All" is the task-macro score used in the main table. Conditions are averaged within each seed before reporting the mean and standard deviation across three seeds.

Table 5: View-resolved results on EgoMe (accuracy, %). Bold denotes the best method in each column.

Method	A			B			C		
	All	Ego	Exo	All	Ego	Exo	All	Ego	Exo
Sequential									
Fine-tuning	$40.32_{\pm.60}$	$42.63_{\pm.32}$	$38.08_{\pm.93}$	$11.18_{\pm1.81}$	$0.00_{\pm.00}$	$18.29_{\pm2.69}$	$49.24_{\pm.58}$	$54.56_{\pm.72}$	$43.85_{\pm.44}$
Regularization									
EWC	$42.50_{\pm1.44}$	$45.31_{\pm2.72}$	$39.81_{\pm.45}$	$15.36_{\pm.22}$	$1.54_{\pm.53}$	$23.73_{\pm.17}$	$50.72_{\pm.49}$	$55.67_{\pm.90}$	$45.75_{\pm.97}$
LwF	$44.95_{\pm1.18}$	$49.70_{\pm2.13}$	$40.35_{\pm.25}$	$23.69_{\pm.71}$	$17.11_{\pm1.33}$	$27.21_{\pm.61}$	$52.98_{\pm.58}$	$59.39_{\pm.64}$	$46.47_{\pm.54}$
Replay									
ER	$47.91_{\pm.41}$	$53.40_{\pm.67}$	$42.48_{\pm.49}$	$55.74_{\pm.76}$	$67.04_{\pm1.50}$	$47.64_{\pm.24}$	$52.84_{\pm.29}$	$58.54_{\pm.30}$	$47.11_{\pm.27}$
DER++	$48.65_{\pm.64}$	$53.79_{\pm1.61}$	$43.61_{\pm.29}$	$56.80_{\pm.78}$	$69.43_{\pm1.84}$	$47.66_{\pm.88}$	$56.24_{\pm1.25}$	$61.99_{\pm1.89}$	$50.43_{\pm.64}$
Prototype									
iCaRL	$34.41_{\pm.51}$	$37.57_{\pm1.29}$	$31.28_{\pm1.66}$	$51.21_{\pm.15}$	$65.15_{\pm.32}$	$41.64_{\pm.35}$	$41.49_{\pm.80}$	$46.11_{\pm.92}$	$36.86_{\pm.71}$
RanPAC	$56.91_{\pm.31}$	$63.94_{\pm1.06}$	$49.73_{\pm1.63}$	$69.42_{\pm.36}$	$83.86_{\pm.43}$	$59.43_{\pm.35}$	$60.54_{\pm.59}$	$67.20_{\pm.34}$	$53.85_{\pm1.39}$
View-aware									
ViewLN-ER	$48.29_{\pm1.08}$	$53.82_{\pm1.96}$	$42.75_{\pm.33}$	$57.22_{\pm.36}$	$72.52_{\pm.51}$	$46.27_{\pm.75}$	$54.15_{\pm.78}$	$60.28_{\pm1.07}$	$47.96_{\pm.82}$
CE^{4}L	$49.00_{\pm.68}$	$53.22_{\pm1.50}$	$44.77_{\pm.54}$	$47.64_{\pm.20}$	$60.12_{\pm.98}$	$38.94_{\pm.35}$	$56.48_{\pm.16}$	$62.78_{\pm.35}$	$50.14_{\pm.20}$
Ours									
VFCT	$\mathbf{58.49}_{\pm.45}$	$\mathbf{65.55}_{\pm.45}$	$\mathbf{51.29}_{\pm.71}$	$\mathbf{73.48}_{\pm.76}$	$\mathbf{86.33}_{\pm.65}$	$\mathbf{64.43}_{\pm.82}$	$\mathbf{65.98}_{\pm.52}$	$\mathbf{72.67}_{\pm.50}$	$\mathbf{59.29}_{\pm.53}$

Table 6: View-resolved results on Assembly101 (accuracy, %). Bold denotes the best method in each column.

Method	A			B			C		
	All	Ego	Exo	All	Ego	Exo	All	Ego	Exo
Sequential									
Fine-tuning	$28.38_{\pm.19}$	$18.84_{\pm.15}$	$37.92_{\pm.27}$	$11.29_{\pm.08}$	$0.00_{\pm.00}$	$14.99_{\pm.24}$	$29.74_{\pm.14}$	$18.84_{\pm.26}$	$40.64_{\pm.53}$
Regularization									
EWC	$27.21_{\pm.20}$	$17.92_{\pm.66}$	$36.49_{\pm.26}$	$11.68_{\pm.33}$	$0.00_{\pm.00}$	$15.80_{\pm.01}$	$29.16_{\pm.20}$	$18.55_{\pm.25}$	$39.78_{\pm.25}$
LwF	$28.55_{\pm.23}$	$19.97_{\pm.31}$	$37.14_{\pm.43}$	$18.54_{\pm.43}$	$9.45_{\pm.16}$	$22.42_{\pm.59}$	$31.13_{\pm.52}$	$21.07_{\pm.36}$	$41.19_{\pm.75}$
Replay									
ER	$29.66_{\pm.36}$	$19.75_{\pm.26}$	$39.57_{\pm.47}$	$30.76_{\pm.98}$	$24.86_{\pm1.58}$	$38.72_{\pm.58}$	$31.16_{\pm.18}$	$20.87_{\pm.58}$	$41.45_{\pm.23}$
DER++	$30.52_{\pm.30}$	$20.48_{\pm.26}$	$40.57_{\pm.34}$	$35.65_{\pm.11}$	$28.96_{\pm.49}$	$43.81_{\pm.40}$	$31.82_{\pm.31}$	$21.72_{\pm.22}$	$41.91_{\pm.40}$
Prototype									
iCaRL	$19.34_{\pm.94}$	$12.42_{\pm1.01}$	$26.25_{\pm1.07}$	$31.55_{\pm.38}$	$21.43_{\pm.97}$	$39.81_{\pm.28}$	$22.01_{\pm.56}$	$13.88_{\pm.68}$	$30.15_{\pm.54}$
RanPAC	$29.73_{\pm1.29}$	$19.95_{\pm.71}$	$39.50_{\pm1.94}$	$38.44_{\pm.38}$	$30.31_{\pm.55}$	$50.98_{\pm.39}$	$31.19_{\pm.21}$	$20.91_{\pm.70}$	$41.47_{\pm.88}$
View-aware									
ViewLN-ER	$27.75_{\pm.42}$	$17.84_{\pm.08}$	$37.65_{\pm.82}$	$30.31_{\pm.21}$	$20.41_{\pm.54}$	$36.63_{\pm.59}$	$29.79_{\pm.25}$	$18.90_{\pm.22}$	$40.68_{\pm.40}$
CE^{4}L	$29.49_{\pm.32}$	$20.45_{\pm.16}$	$38.53_{\pm.50}$	$26.49_{\pm.22}$	$24.24_{\pm.41}$	$35.84_{\pm.67}$	$32.18_{\pm.09}$	$21.87_{\pm.18}$	$42.49_{\pm.08}$
Ours									
VFCT	$\mathbf{32.66}_{\pm.46}$	$\mathbf{22.06}_{\pm.68}$	$\mathbf{43.27}_{\pm.24}$	$\mathbf{41.26}_{\pm.46}$	$\mathbf{32.16}_{\pm.48}$	$\mathbf{52.77}_{\pm.50}$	$\mathbf{34.10}_{\pm.27}$	$\mathbf{23.36}_{\pm.08}$	$\mathbf{44.85}_{\pm.60}$

Table 7: View-resolved results on Charades-Ego (macro-mAP, %). Bold denotes the best method in each column.

Method	A			B			C		
	All	Ego	Exo	All	Ego	Exo	All	Ego	Exo
Sequential									
Fine-tuning	$26.21_{\pm.11}$	$24.05_{\pm.28}$	$25.95_{\pm.12}$	$22.18_{\pm1.63}$	$20.06_{\pm1.41}$	$23.60_{\pm1.78}$	$25.45_{\pm.19}$	$24.25_{\pm.18}$	$25.04_{\pm.23}$
Regularization									
EWC	$26.28_{\pm.05}$	$24.00_{\pm.01}$	$26.10_{\pm.05}$	$26.03_{\pm.83}$	$24.18_{\pm1.47}$	$27.26_{\pm.58}$	$25.77_{\pm.13}$	$24.57_{\pm.11}$	$25.29_{\pm.18}$
LwF	$26.08_{\pm.15}$	$23.73_{\pm.23}$	$25.91_{\pm.24}$	$28.67_{\pm1.11}$	$27.14_{\pm.76}$	$29.69_{\pm1.36}$	$26.50_{\pm.11}$	$25.23_{\pm.14}$	$26.08_{\pm.06}$
Replay									
ER	$26.51_{\pm.47}$	$24.20_{\pm.22}$	$26.21_{\pm.67}$	$30.30_{\pm.16}$	$28.57_{\pm.16}$	$31.46_{\pm.28}$	$25.85_{\pm.15}$	$24.63_{\pm.12}$	$25.38_{\pm.22}$
DER++	$26.18_{\pm.24}$	$23.63_{\pm.24}$	$26.21_{\pm.38}$	$29.63_{\pm.09}$	$25.78_{\pm.23}$	$32.19_{\pm.14}$	$26.54_{\pm.12}$	$25.06_{\pm.16}$	$26.30_{\pm.12}$
Prototype									
iCaRL	$20.05_{\pm.35}$	$18.05_{\pm.51}$	$19.61_{\pm.20}$	$23.15_{\pm.20}$	$20.91_{\pm.17}$	$24.64_{\pm.24}$	$19.84_{\pm.16}$	$18.77_{\pm.19}$	$19.47_{\pm.37}$
RanPAC	$26.67_{\pm.43}$	$24.48_{\pm.35}$	$26.00_{\pm.26}$	$34.03_{\pm.04}$	$32.72_{\pm.27}$	$34.90_{\pm.25}$	$26.07_{\pm.17}$	$24.78_{\pm.30}$	$25.84_{\pm.23}$
View-aware									
ViewLN-ER	$26.27_{\pm.13}$	$24.25_{\pm.19}$	$25.70_{\pm.14}$	$28.02_{\pm.29}$	$26.27_{\pm.25}$	$29.19_{\pm.34}$	$25.25_{\pm.13}$	$24.12_{\pm.26}$	$24.77_{\pm.16}$
CE^{4}L	$27.14_{\pm.16}$	$24.84_{\pm.25}$	$26.71_{\pm.23}$	$32.67_{\pm.10}$	$32.85_{\pm.34}$	$32.56_{\pm.20}$	$27.20_{\pm.24}$	$26.20_{\pm.28}$	$26.45_{\pm.19}$
Ours									
VFCT	$\mathbf{30.61}_{\pm.03}$	$\mathbf{28.46}_{\pm.07}$	$\mathbf{30.11}_{\pm.03}$	$\mathbf{38.01}_{\pm.03}$	$\mathbf{36.96}_{\pm.01}$	$\mathbf{38.72}_{\pm.04}$	$\mathbf{30.19}_{\pm.01}$	$\mathbf{29.02}_{\pm.01}$	$\mathbf{29.71}_{\pm.02}$

C.1 STAGE-WISE PERFORMANCE

Tables 8–16 report the seen-task score after each update in one prespecified stream per protocol: blocked ego-first for A, frequency order for B, and true-pair ego-first for C. These are the same streams used for the mechanism diagnostics. A and C are not averaged position-wise across opposing view orders; their columns instead name the arriving ego (E) or exo (X) cohort. B names the arriving class blocks. Each entry is the mean and standard deviation over three seeds.

Table 8: EgoMe seen-task accuracy (%) along the Protocol-A stream $E_1 \rightarrow E_2 \rightarrow X_1 \rightarrow X_2$. Bold denotes the best method after each arrival.

Method	E_1	E_2	X_1	X_2
Sequential				
Fine-tuning	$47.41_{\pm 2.00}$	$48.25_{\pm .67}$	$36.87_{\pm 1.38}$	$39.34_{\pm 1.96}$
Regularization				
EWC	$44.51_{\pm 1.56}$	$47.15_{\pm .96}$	$39.32_{\pm 2.13}$	$42.77_{\pm 1.95}$
LwF	$43.89_{\pm .95}$	$49.66_{\pm .93}$	$43.87_{\pm 1.06}$	$44.70_{\pm 1.37}$
Replay				
ER	$50.31_{\pm 4.07}$	$54.13_{\pm 1.76}$	$48.23_{\pm 1.33}$	$47.97_{\pm .91}$
DER++	$46.17_{\pm 1.56}$	$53.90_{\pm 1.33}$	$47.35_{\pm 1.29}$	$48.48_{\pm 1.54}$
Prototype				
iCaRL	$31.26_{\pm 1.79}$	$38.43_{\pm 1.55}$	$30.57_{\pm .33}$	$32.93_{\pm .91}$
RanPAC	$\mathbf{54.24}_{\pm \mathbf{.72}}$	$65.04_{\pm .47}$	$57.23_{\pm .63}$	$56.94_{\pm .31}$
View-aware				
ViewLN-ER	$50.31_{\pm 1.86}$	$55.37_{\pm .60}$	$46.86_{\pm 2.09}$	$47.53_{\pm 1.11}$
CE^4L	$49.48_{\pm .72}$	$52.65_{\pm .56}$	$47.17_{\pm .80}$	$48.58_{\pm 1.18}$
Ours				
VFCT	$54.04_{\pm .00}$	$\mathbf{65.11}_{\pm \mathbf{.17}}$	$\mathbf{58.03}_{\pm \mathbf{.53}}$	$\mathbf{58.88}_{\pm \mathbf{.45}}$

Table 9: EgoMe seen-task accuracy (%) after each Protocol-B class-block arrival. Bold denotes the best method after each arrival.

Method	Block 1	Block 2	Block 3	Block 4	Block 5
Sequential					
Fine-tuning	$60.35_{\pm 2.64}$	$40.56_{\pm 2.22}$	$19.59_{\pm 3.10}$	$12.14_{\pm 8.53}$	$9.07_{\pm 7.08}$
Regularization					
EWC	$61.22_{\pm 3.60}$	$39.44_{\pm 1.71}$	$24.11_{\pm 1.57}$	$23.65_{\pm 1.17}$	$16.94_{\pm 1.74}$
LwF	$60.35_{\pm .75}$	$45.95_{\pm 2.90}$	$35.43_{\pm 1.68}$	$34.64_{\pm 3.40}$	$25.64_{\pm 2.99}$
Replay					
ER	$57.73_{\pm 2.95}$	$67.23_{\pm 2.13}$	$59.90_{\pm 1.47}$	$59.59_{\pm 2.18}$	$58.55_{\pm 1.77}$
DER++	$59.91_{\pm .38}$	$66.03_{\pm .52}$	$62.39_{\pm 1.59}$	$59.12_{\pm 1.48}$	$57.35_{\pm 1.86}$
Prototype					
iCaRL	$58.17_{\pm 2.26}$	$63.72_{\pm 2.10}$	$54.98_{\pm 1.31}$	$54.58_{\pm .71}$	$51.53_{\pm .79}$
RanPAC	$56.86_{\pm 1.31}$	$69.13_{\pm .66}$	$65.29_{\pm 1.12}$	$69.01_{\pm .81}$	$69.03_{\pm 1.10}$
View-aware					
ViewLN-ER	$59.26_{\pm 1.36}$	$68.47_{\pm 2.30}$	$61.09_{\pm .62}$	$58.19_{\pm .99}$	$58.86_{\pm 1.00}$
CE^4L	$59.91_{\pm 2.30}$	$47.33_{\pm 1.79}$	$49.07_{\pm 1.79}$	$50.19_{\pm .67}$	$51.45_{\pm .76}$
Ours					
VFCT	$\mathbf{61.44}_{\pm \mathbf{.65}}$	$\mathbf{72.51}_{\pm \mathbf{.40}}$	$\mathbf{70.74}_{\pm \mathbf{.92}}$	$\mathbf{73.96}_{\pm \mathbf{1.18}}$	$\mathbf{74.02}_{\pm \mathbf{.17}}$

Table 10: EgoMe seen-task accuracy (%) along the Protocol-C stream $E_1 \rightarrow X_1 \rightarrow E_2 \rightarrow X_2$. Bold denotes the best method after each arrival.

Method	E_1	X_1	E_2	X_2
Sequential				
Fine-tuning	$54.27_{\pm .96}$	$43.06_{\pm 1.72}$	$52.01_{\pm .44}$	$48.84_{\pm .59}$
Regularization				
EWC	$55.06_{\pm 1.03}$	$41.62_{\pm 2.46}$	$53.46_{\pm .43}$	$50.12_{\pm .91}$
LwF	$54.76_{\pm .52}$	$46.78_{\pm 1.28}$	$55.03_{\pm .81}$	$52.44_{\pm .19}$
Replay				
ER	$55.95_{\pm 1.19}$	$45.04_{\pm .37}$	$54.39_{\pm .70}$	$53.41_{\pm 1.46}$
DER++	$57.24_{\pm 1.24}$	$49.45_{\pm 1.74}$	$57.59_{\pm .66}$	$56.89_{\pm 2.44}$
Prototype				
iCaRL	$47.22_{\pm 1.53}$	$37.25_{\pm 1.34}$	$48.18_{\pm 1.59}$	$40.22_{\pm 1.93}$
RanPAC	$64.58_{\pm 1.57}$	$54.96_{\pm .37}$	$61.71_{\pm .59}$	$60.51_{\pm .58}$
View-aware				
ViewLN-ER	$53.27_{\pm .79}$	$45.68_{\pm 1.86}$	$53.73_{\pm 1.08}$	$54.52_{\pm 1.80}$
CE^4L	$55.56_{\pm 1.79}$	$47.77_{\pm .74}$	$56.54_{\pm 1.36}$	$56.64_{\pm .55}$
Ours				
VFCT	$\mathbf{66.87}_{\pm \mathbf{.91}}$	$\mathbf{58.28}_{\pm \mathbf{.37}}$	$\mathbf{66.14}_{\pm \mathbf{.49}}$	$\mathbf{65.89}_{\pm \mathbf{.63}}$

Table 11: Assembly101 seen-task accuracy (%) along the Protocol-A stream $E_1 \rightarrow E_2 \rightarrow X_1 \rightarrow X_2$. Bold denotes the best method after each arrival.

Method	E_1	E_2	X_1	X_2
Sequential				
Fine-tuning	$15.91_{\pm.48}$	$19.00_{\pm.37}$	$22.57_{\pm.85}$	$28.03_{\pm.94}$
Regularization				
EWC	$14.61_{\pm1.65}$	$18.59_{\pm.23}$	$21.94_{\pm.85}$	$26.54_{\pm.25}$
LwF	$14.61_{\pm1.65}$	$18.13_{\pm.40}$	$24.57_{\pm1.15}$	$28.39_{\pm.97}$
Replay				
ER	$16.59_{\pm.70}$	$18.78_{\pm.65}$	$24.22_{\pm.58}$	$28.72_{\pm.43}$
DER++	$16.29_{\pm.53}$	$19.46_{\pm.70}$	$25.16_{\pm.13}$	$29.92_{\pm.15}$
Prototype				
iCaRL	$8.30_{\pm1.05}$	$12.50_{\pm.73}$	$15.74_{\pm.98}$	$19.48_{\pm1.71}$
RanPAC	$16.06_{\pm.70}$	$19.80_{\pm.74}$	$25.14_{\pm1.17}$	$29.73_{\pm1.29}$
View-aware				
ViewLN-ER	$15.75_{\pm1.05}$	$17.79_{\pm1.12}$	$23.13_{\pm.24}$	$27.71_{\pm.91}$
CE^4L	$16.67_{\pm1.37}$	$19.04_{\pm.52}$	$24.22_{\pm.29}$	$28.01_{\pm.22}$
Ours				
VFCT	$\mathbf{16.74}_{\pm.92}$	$\mathbf{21.17}_{\pm.35}$	$\mathbf{27.79}_{\pm.51}$	$\mathbf{32.34}_{\pm.07}$

Table 12: Assembly101 seen-task accuracy (%) after each Protocol-B class-block arrival. Bold denotes the best method after each arrival.

Method	Block 1	Block 2	Block 3	Block 4	Block 5
Sequential					
Fine-tuning	$69.20_{\pm.68}$	$16.18_{\pm.64}$	$19.07_{\pm1.20}$	$6.37_{\pm.81}$	$6.19_{\pm1.35}$
Regularization					
EWC	$69.20_{\pm.68}$	$16.18_{\pm.72}$	$19.18_{\pm1.27}$	$6.33_{\pm1.54}$	$6.06_{\pm2.10}$
LwF	$69.39_{\pm.58}$	$36.16_{\pm.53}$	$28.78_{\pm1.30}$	$16.67_{\pm2.22}$	$12.91_{\pm3.07}$
Replay					
ER	$69.78_{\pm.54}$	$48.48_{\pm.68}$	$41.62_{\pm3.47}$	$24.15_{\pm1.82}$	$15.50_{\pm2.83}$
DER++	$\mathbf{71.24}_{\pm.48}$	$51.78_{\pm1.63}$	$\mathbf{43.52}_{\pm1.31}$	$27.91_{\pm1.51}$	$25.73_{\pm.36}$
Prototype					
iCaRL	$69.82_{\pm.40}$	$47.97_{\pm1.56}$	$40.47_{\pm1.76}$	$31.60_{\pm2.11}$	$28.46_{\pm.51}$
RanPAC	$68.63_{\pm1.84}$	$50.55_{\pm.13}$	$38.43_{\pm.97}$	$28.76_{\pm.43}$	$24.65_{\pm.53}$
View-aware					
ViewLN-ER	$69.78_{\pm.47}$	$47.47_{\pm1.22}$	$43.29_{\pm1.37}$	$32.03_{\pm1.25}$	$28.38_{\pm2.06}$
CE^4L	$70.81_{\pm.98}$	$47.81_{\pm.46}$	$34.28_{\pm1.10}$	$25.66_{\pm1.00}$	$21.48_{\pm1.11}$
Ours					
VFCT	$70.89_{\pm.77}$	$\mathbf{52.22}_{\pm.51}$	$43.50_{\pm.93}$	$\mathbf{33.92}_{\pm.17}$	$\mathbf{30.26}_{\pm.30}$

Table 13: Assembly101 seen-task accuracy (%) along the Protocol-C stream $E_1 \rightarrow X_1 \rightarrow E_2 \rightarrow X_2$. Bold denotes the best method after each arrival.

Method	E_1	X_1	E_2	X_2
Sequential				
Fine-tuning	$19.61_{\pm.39}$	$27.54_{\pm.34}$	$26.74_{\pm.86}$	$29.92_{\pm.24}$
Regularization				
EWC	$19.42_{\pm1.26}$	$27.35_{\pm.20}$	$26.64_{\pm.25}$	$28.55_{\pm.53}$
LwF	$19.73_{\pm.80}$	$28.70_{\pm.31}$	$27.12_{\pm.07}$	$30.80_{\pm.75}$
Replay				
ER	$19.04_{\pm.80}$	$29.17_{\pm.39}$	$27.74_{\pm.76}$	$31.25_{\pm.53}$
DER++	$19.27_{\pm1.46}$	$30.06_{\pm.32}$	$27.85_{\pm.48}$	$31.94_{\pm.73}$
Prototype				
iCaRL	$11.97_{\pm1.61}$	$22.39_{\pm.63}$	$16.64_{\pm.66}$	$23.55_{\pm.70}$
RanPAC	$19.54_{\pm.63}$	$29.51_{\pm.40}$	$27.51_{\pm.38}$	$31.19_{\pm.22}$
View-aware				
ViewLN-ER	$18.32_{\pm.17}$	$29.15_{\pm.98}$	$26.07_{\pm.45}$	$29.16_{\pm.30}$
CE^4L	$\mathbf{21.09}_{\pm.34}$	$30.05_{\pm.10}$	$28.05_{\pm.25}$	$31.42_{\pm.07}$
Ours				
VFCT	$20.83_{\pm.47}$	$\mathbf{31.85}_{\pm.54}$	$\mathbf{29.53}_{\pm.19}$	$\mathbf{34.27}_{\pm.44}$

Table 14: Charades-Ego seen-task macro-mAP (%) along the Protocol-A stream $E_1 \rightarrow E_2 \rightarrow X_1 \rightarrow X_2$. Bold denotes the best method after each arrival.

Method	E_1	E_2	X_1	X_2
Sequential				
Fine-tuning	$21.01_{\pm.41}$	$25.57_{\pm.21}$	$25.41_{\pm.40}$	$26.37_{\pm.18}$
Regularization				
EWC	$21.01_{\pm.41}$	$25.65_{\pm.40}$	$25.63_{\pm.44}$	$26.20_{\pm.14}$
LwF	$21.01_{\pm.41}$	$23.98_{\pm.27}$	$25.25_{\pm.20}$	$26.18_{\pm.13}$
Replay				
ER	$20.09_{\pm.29}$	$25.01_{\pm.37}$	$26.20_{\pm.51}$	$26.71_{\pm.54}$
DER++	$20.24_{\pm.89}$	$24.07_{\pm.47}$	$25.75_{\pm.39}$	$26.55_{\pm.43}$
Prototype				
iCaRL	$18.53_{\pm.42}$	$17.92_{\pm.23}$	$19.55_{\pm.18}$	$20.75_{\pm.47}$
RanPAC	$23.43_{\pm.21}$	$25.14_{\pm.27}$	$26.07_{\pm.33}$	$26.67_{\pm.43}$
View-aware				
ViewLN-ER	$18.10_{\pm.36}$	$25.72_{\pm.24}$	$26.40_{\pm.09}$	$26.37_{\pm.16}$
CE^4L	$24.29_{\pm.08}$	$25.99_{\pm.14}$	$26.35_{\pm.11}$	$26.82_{\pm.18}$
Ours				
VFCT	$\mathbf{25.50}_{\pm.04}$	$\mathbf{28.43}_{\pm.08}$	$\mathbf{30.01}_{\pm.02}$	$\mathbf{30.66}_{\pm.10}$

Table 15: Charades-Ego seen-task macro-mAP (%) after each Protocol-B class-block arrival. Bold denotes the best method after each arrival.

Method	Block 1	Block 2	Block 3	Block 4	Block 5
Sequential					
Fine-tuning	$47.44_{\pm.12}$	$35.55_{\pm.52}$	$26.79_{\pm1.34}$	$25.55_{\pm3.66}$	$23.39_{\pm1.81}$
Regularization					
EWC	$47.44_{\pm.12}$	$36.36_{\pm1.00}$	$29.11_{\pm2.62}$	$30.49_{\pm2.02}$	$26.38_{\pm1.34}$
LwF	$47.44_{\pm.12}$	$41.35_{\pm.31}$	$35.75_{\pm1.07}$	$34.20_{\pm1.24}$	$29.61_{\pm1.22}$
Replay					
ER	$46.15_{\pm.34}$	$39.08_{\pm.38}$	$34.42_{\pm.02}$	$34.17_{\pm.19}$	$31.10_{\pm.59}$
DER++	$47.26_{\pm.36}$	$39.01_{\pm.24}$	$34.45_{\pm.41}$	$33.07_{\pm.78}$	$30.32_{\pm.77}$
Prototype					
iCaRL	$32.95_{\pm.99}$	$28.77_{\pm.25}$	$26.11_{\pm.62}$	$24.74_{\pm.78}$	$23.46_{\pm.62}$
RanPAC	$44.90_{\pm1.31}$	$40.59_{\pm.62}$	$37.33_{\pm.30}$	$37.75_{\pm.02}$	$35.45_{\pm.06}$
View-aware					
ViewLN-ER	$45.96_{\pm.26}$	$36.66_{\pm1.00}$	$31.86_{\pm.77}$	$30.89_{\pm.18}$	$28.18_{\pm.22}$
CE^4L	$46.03_{\pm.31}$	$41.10_{\pm.70}$	$36.54_{\pm.42}$	$37.12_{\pm.15}$	$34.11_{\pm.18}$
Ours					
VFCT	$\mathbf{48.43}_{\pm.01}$	$\mathbf{44.11}_{\pm.01}$	$\mathbf{40.79}_{\pm.03}$	$\mathbf{41.77}_{\pm.04}$	$\mathbf{39.09}_{\pm.06}$

Table 16: Charades-Ego seen-task macro-mAP (%) along the Protocol-C stream $E_1 \rightarrow X_1 \rightarrow E_2 \rightarrow X_2$. Bold denotes the best method after each arrival.

Method	E_1	X_1	E_2	X_2
Sequential				
Fine-tuning	$24.35_{\pm.48}$	$24.02_{\pm.08}$	$25.41_{\pm.26}$	$25.26_{\pm.17}$
Regularization				
EWC	$24.35_{\pm.48}$	$24.27_{\pm.13}$	$25.54_{\pm.33}$	$25.63_{\pm.41}$
LwF	$24.35_{\pm.48}$	$24.74_{\pm.19}$	$25.94_{\pm.24}$	$26.58_{\pm.30}$
Replay				
ER	$22.92_{\pm.49}$	$24.71_{\pm.55}$	$25.41_{\pm.39}$	$25.97_{\pm.14}$
DER++	$22.98_{\pm.36}$	$24.71_{\pm.35}$	$26.03_{\pm.43}$	$26.71_{\pm.21}$
Prototype				
iCaRL	$16.88_{\pm.38}$	$18.51_{\pm.52}$	$19.12_{\pm.15}$	$20.17_{\pm.09}$
RanPAC	$23.67_{\pm.16}$	$24.25_{\pm.03}$	$25.22_{\pm.21}$	$26.08_{\pm.17}$
View-aware				
ViewLN-ER	$20.96_{\pm1.61}$	$24.18_{\pm.44}$	$25.05_{\pm.25}$	$25.12_{\pm.13}$
CE^4L	$26.24_{\pm.08}$	$25.42_{\pm.26}$	$26.32_{\pm.28}$	$27.06_{\pm.27}$
Ours				
VFCT	$\mathbf{27.03}_{\pm.02}$	$\mathbf{27.86}_{\pm.05}$	$\mathbf{29.22}_{\pm.01}$	$\mathbf{30.18}_{\pm.02}$

D PERFORMANCE UNDER MATCHED MEMORY

We compare VFCT, RanPAC, and ViewLN-ER under identical learner-owned persistent-state budgets on Assembly101. The budget includes model parameters and buffers, analytic state, prototypes, replay, frozen teachers, and optimizer state; the common frozen encoder and disk feature cache are excluded. We test 16, 64, and 128 MiB on Protocol A with alternating ego-first arrival and Protocol C with true exo-first arrival. All 54 runs fall within 1% of the target budget or one indivisible tensor row.

Table 17: Assembly101 accuracy (%) under matched persistent-state budgets. Each entry is mean ± standard deviation over three seeds; bold denotes the best method at the same budget and evaluation context.

	Protocol A: alternating ego-first			Protocol C: true exo-first		
Budget (MiB)	RanPAC	ViewLN-ER	VFCT	RanPAC	ViewLN-ER	VFCT
16	$29.84 \pm .56$	$27.65 \pm .54$	$26.00 \pm .48$	$30.20 \pm .07$	$29.74 \pm .88$	$26.29 \pm .23$
64	$29.49 \pm .41$	$29.29 \pm .80$	$32.40 \pm .18$	$31.74 \pm .13$	$30.40 \pm .30$	$33.14 \pm .41$
128	$29.29 \pm .28$	$29.11 \pm .41$	$32.24 \pm .32$	$32.04 \pm .49$	$31.47 \pm .20$	$33.96 \pm .49$

At 16 MiB, RanPAC performs best in both contexts. Once the budget reaches 64 MiB, VFCT gives the strongest result in both A and C, and it retains the lead at 128 MiB. Thus, the advantage of the complete method depends on sufficient persistent state rather than holding uniformly at every budget.

E PROTOTYPE HISTORY AND REPRESENTATION DIAGNOSTICS

Cumulative semantic means store training-visit contributions evaluated at historical adapter states. They are not generally the mean obtained by re-encoding all historical observations with the current adapter. Same-cell stability constrains only cells present in a training batch, and the pointwise adaptation constraint alone does not establish temporal consistency. Offline centroid audits use historical training data exclusively for measurement; the learner receives no extra replay or refreshed prototypes.

Table 18: Independent historical-centroid mismatch. Each audit averages supported cells; the table then averages all task boundaries and three seeds for the preselected representative stream of each dataset/protocol.

Dataset	Protocol	Stages	Cosine mismatch	Relative L2 mismatch
assembly101	A	12	0.00195	0.06141
assembly101	B	15	0.00362	0.08330
assembly101	C	12	0.00300	0.07998
charadesego	A	12	0.00002	0.00618
charadesego	B	15	0.00007	0.00902
charadesego	C	12	0.00006	0.00899
egome	A	12	0.00084	0.03511
egome	B	15	0.00130	0.04447
egome	C	12	0.00206	0.05794