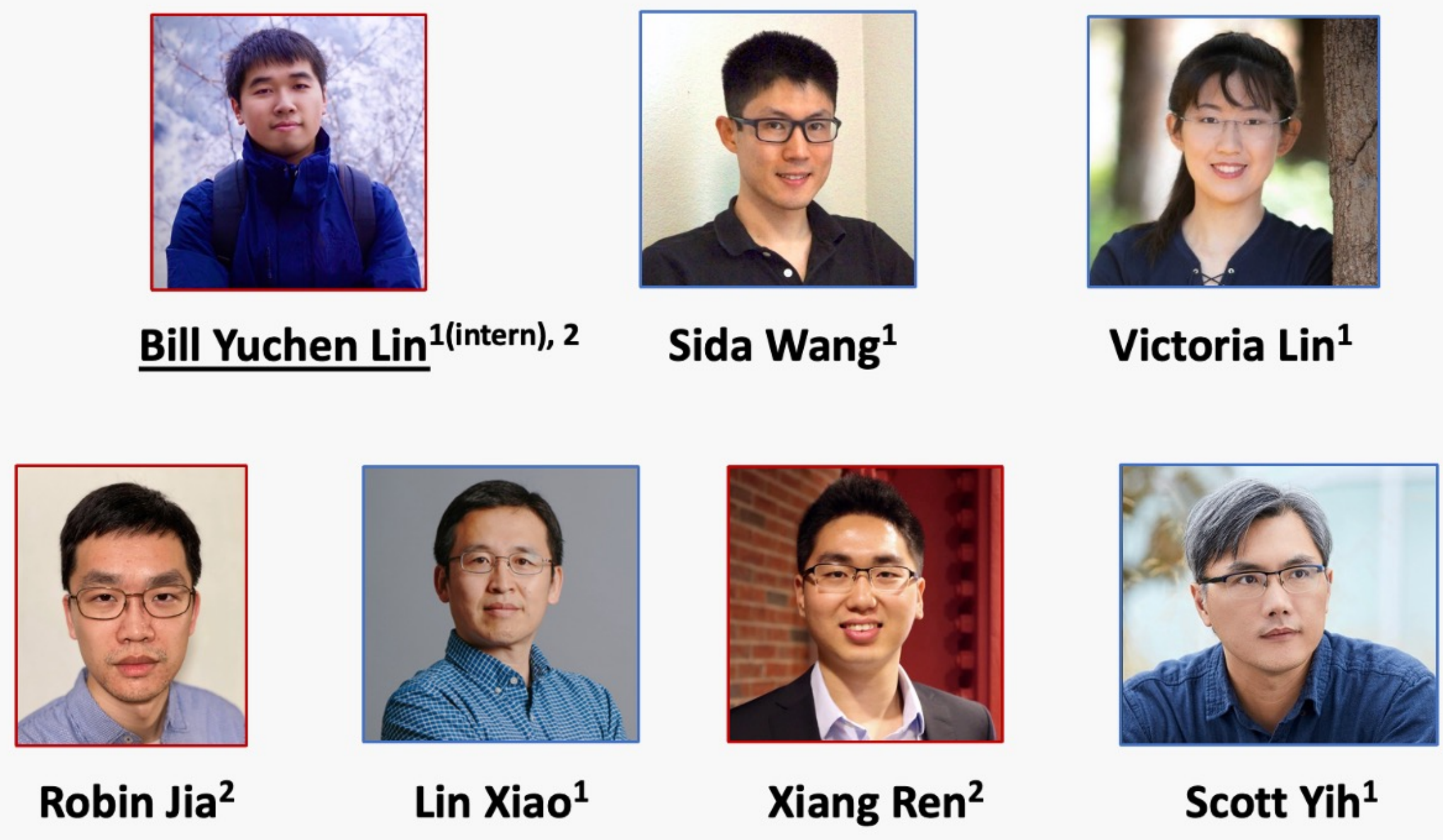


On *Continual Model Refinement* in Out-of-Distribution Data Streams

Continual Model Refinement (CMR)



Bill Yuchen Lin^{1(intern), 2}

Sida Wang¹

Victoria Lin¹

Robin Jia²

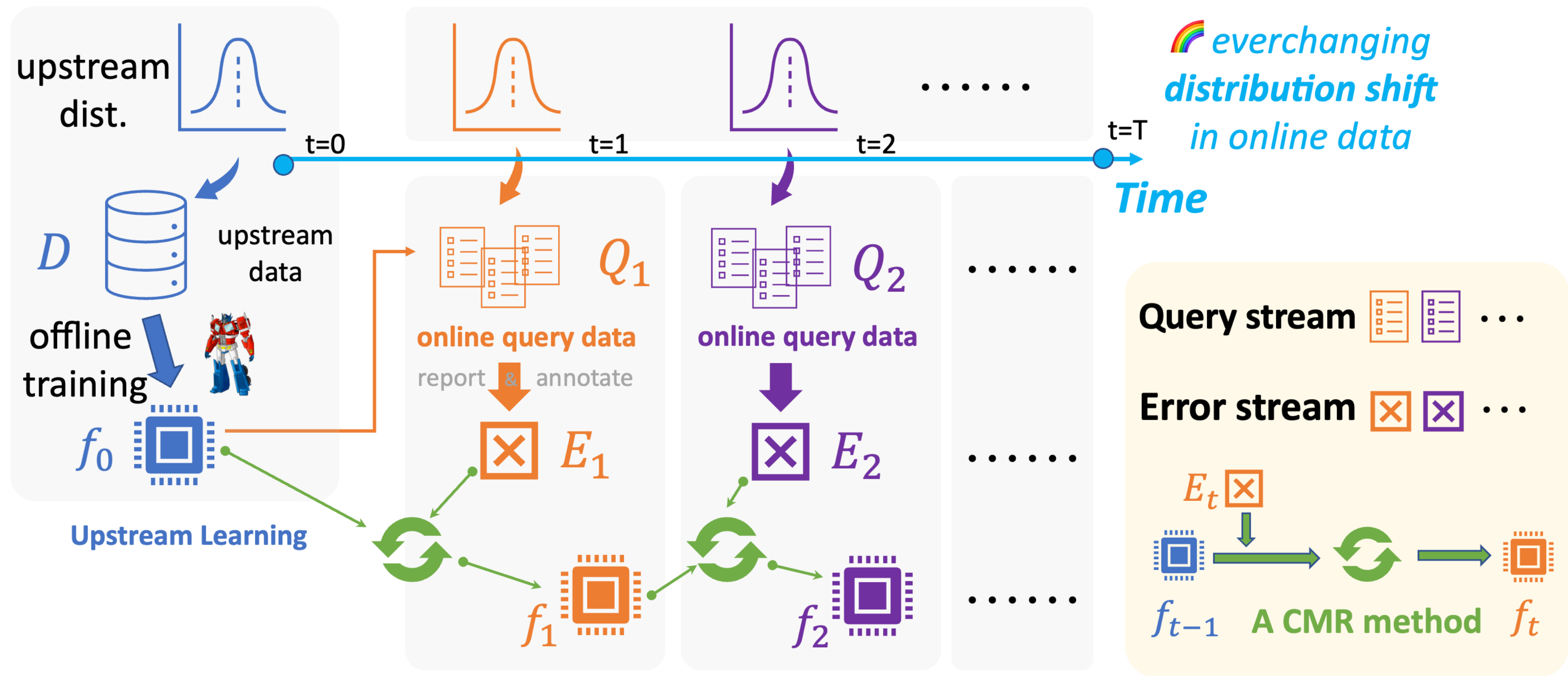
Lin Xiao¹

Xiang Ren²

Scott Yih¹

¹facebook AI Research

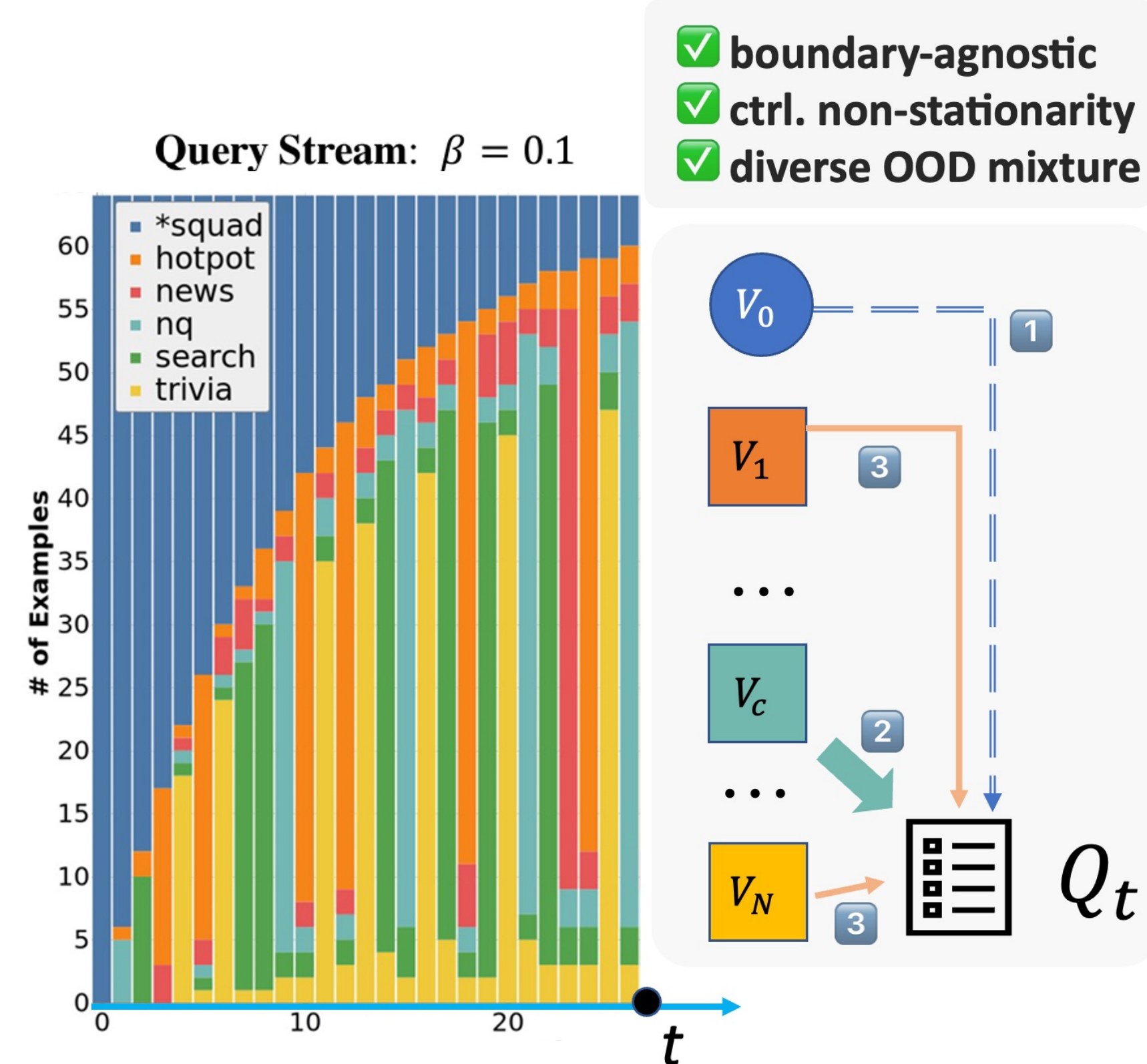
²USC



Evaluating CMR: OOD Data Stream

Algorithm 1: Sampling query streams with *controllable non-stationarity* from multiple data clusters.

Input Data Clusters: V_0, V_1, \dots, V_N
Configuration Arguments: $T, b, (\alpha, \beta, \gamma)$.
Output: A query stream $\{Q_1, Q_2, \dots, Q_T\}$
foreach t **in** $\text{range}(1, T)$ **do**
 $b_u = \lfloor b * \alpha^{t-1} \rfloor$; $b_o = b - b_u$; $b'_o = \lfloor b_o * \gamma \rfloor$
 $c_t \sim P(c|c_{t-1}; \beta)$
 /* The prob. of switching the major OOD data cluster is $1 - \beta$, i.e., $P(c_t \neq c_{t-1}) = 1 - \beta$ */
 $V_{c_t} = \bigcup_{k \in [1, N] | k \neq c_t} V_k$
 1 $Q_t \leftarrow \text{sample}(V_0, b_u)$
 /* $V_0 \sim \mathcal{U}$; from upstream distribution */
 2 $Q_t \leftarrow \text{sample}(V_{c_t}, b'_o)$
 /* from the current major OOD data cluster */
 3 $Q_t \leftarrow \text{sample}(V_{\neq c_t}, b_o - b'_o)$
 /* from non-major data clusters */



Methods ↓ Metrics →	EFR	UKR ^(T)	OKR ^(T)
Frozen Upstream ($f_t \equiv f_0$)	0.00	80.27	36.13
● Continual Fine-Tuning	97.36	66.21	77.73
■ Online L2Reg.	97.18	71.09	83.59
▲ Online EWC	97.49	68.55	85.74
⊕ Exp. Replay	97.07	72.46	87.30
◆ MaxLoss	97.43	75.00	84.77
▶ MIR	97.08	75.78	87.50
Offline Refining ($f_0 \rightarrow f_T$)	95.62	83.78	93.75

all can **fix** the errors pretty well (on average)

forget **upstream** knowledge

forget **online** knowledge

Error Fixing Rate (EFR)

$$\text{EFR}(t) = \frac{|\{(x, y) \in E_t \mid f_t(x) = y\}|}{|E_t|}$$

Cumulative Success Rates (CSR)

$$\text{CSR}(t) = 1 - \frac{|E_{<t}|}{|Q_{<t}|}$$

Upstream Knowledge Retention (UKR)

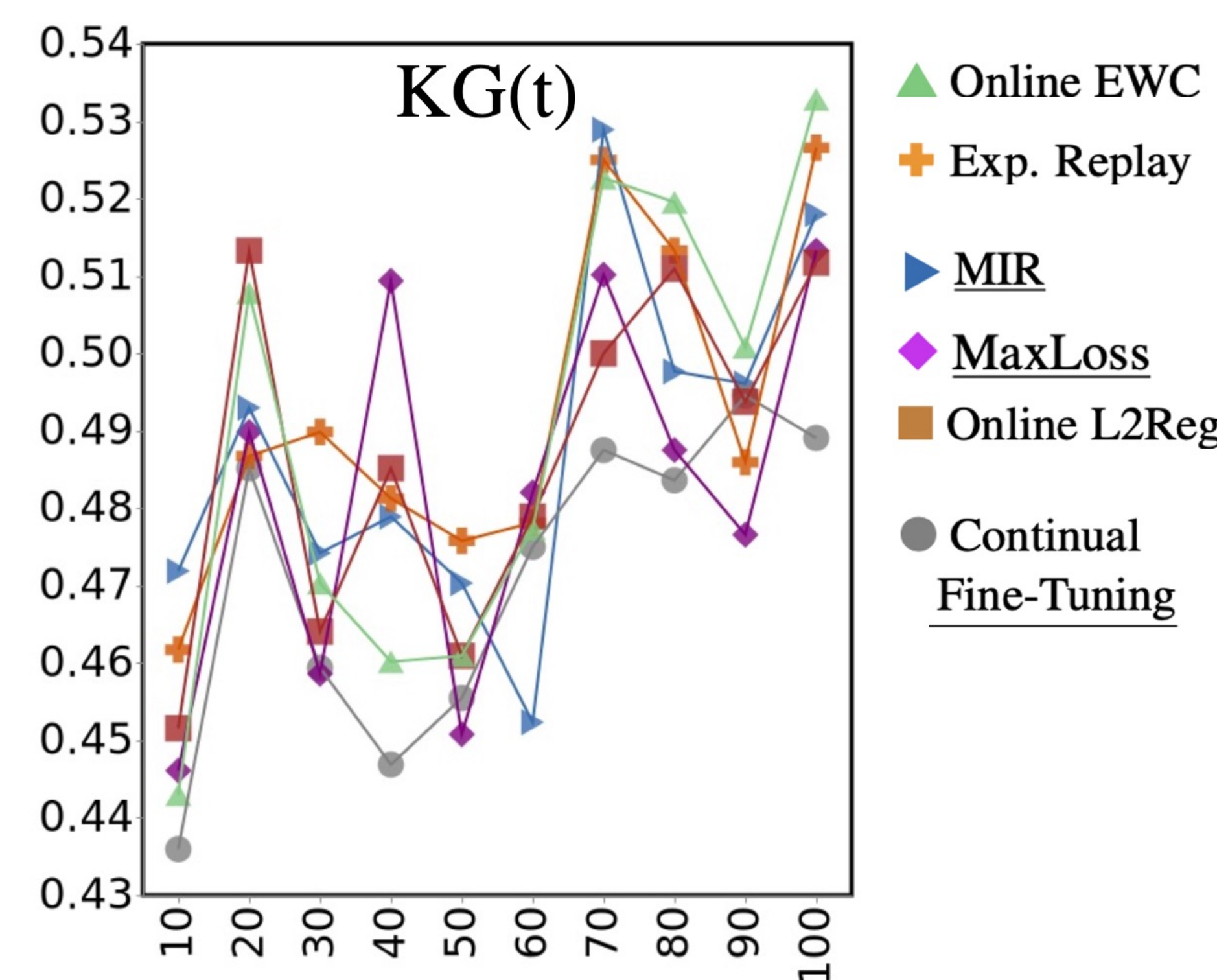
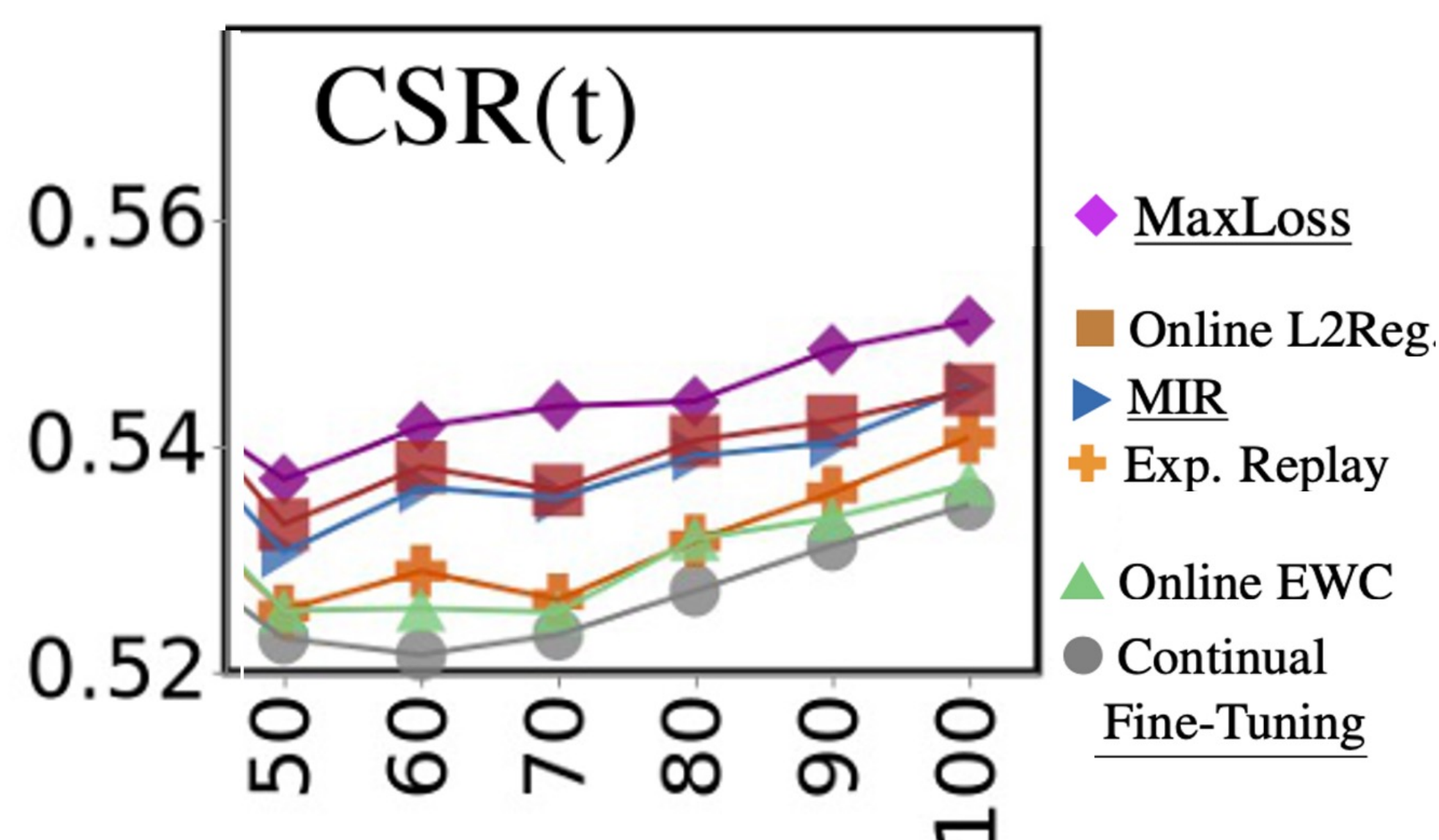
$$\text{UKR}(t) = \text{Acc}(f_t, D)$$

Online Knowledge Retention (OKR)

$$\text{OKR}(t) = \text{Acc}(f_t, Q_{<t})$$

Knowledge Generalization (KG)

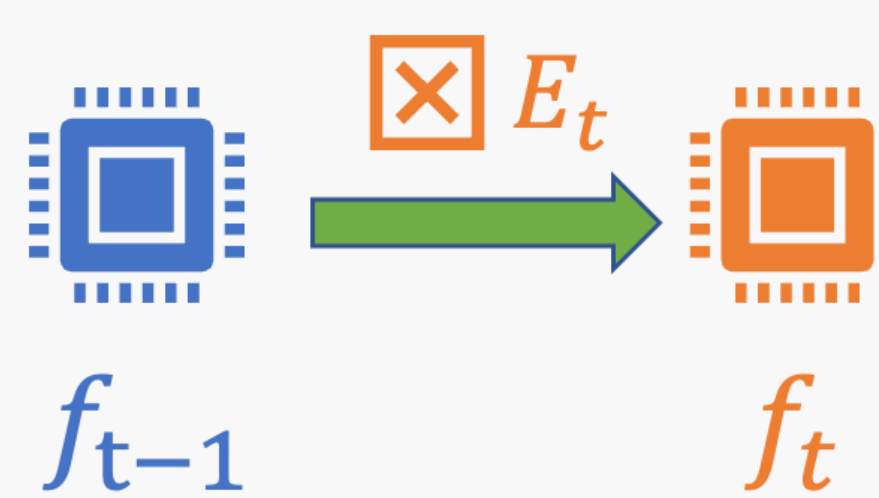
$$\text{KG}(t) = \text{Acc}(f_t, H)$$



CMR Methods: Regularization | Replay

$$\mathcal{L}_{\text{total}}(t) = \mathcal{L}_{\text{Error}}(t) + \lambda \mathcal{L}_{\text{Reg}}(t)$$

● Continual Fine-Tuning



■ Online L2Reg.

$$\mathcal{L}_{\text{L2Reg}}(t) = \sum_i (\theta_t^i - \theta_{t-1}^i)^2$$

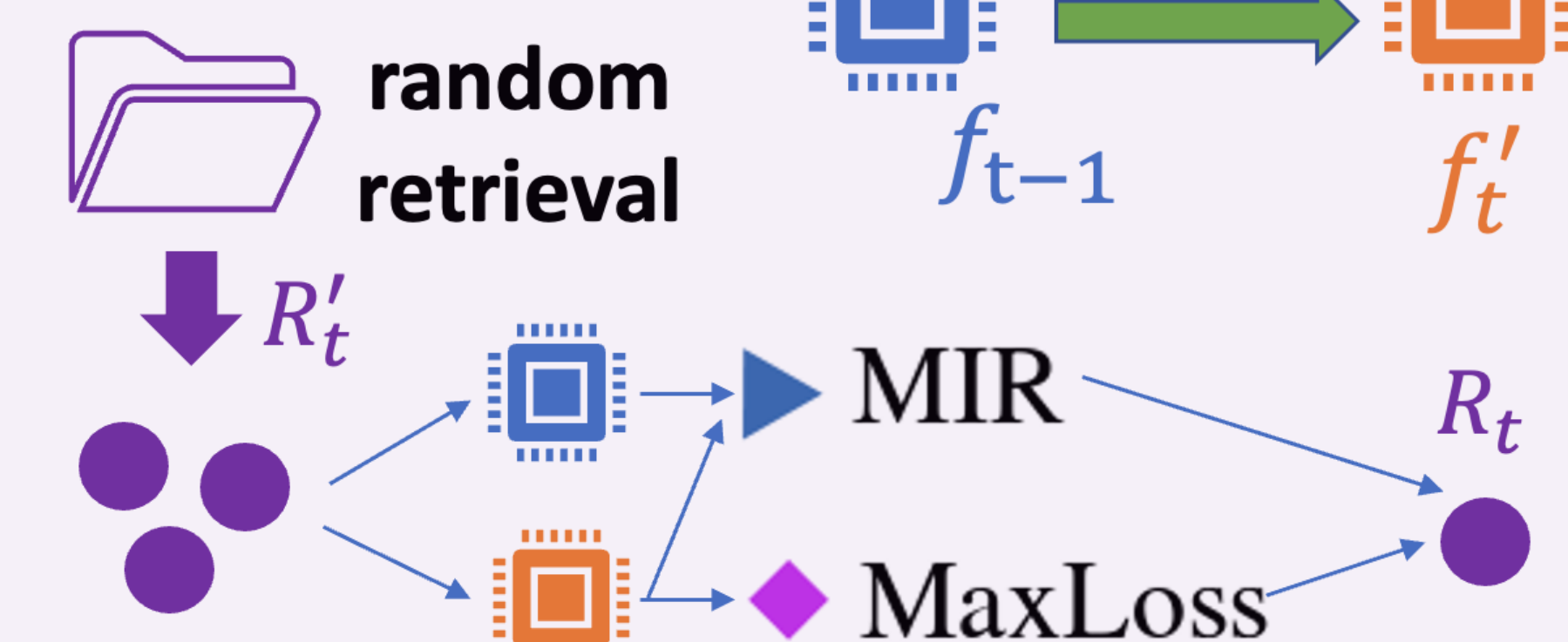
▲ Online EWC

$$\mathcal{L}_{\text{EWC}}(t) = \sum_{j=1}^{t-1} \left(\frac{1}{2} \sum_i F_{ii}^{(j)} (\theta_t^i - \theta_{t-1}^i)^2 \right)$$

⊕ Exp. Replay



Conditional Replay



The typical CL setup

- boundary-aware
- pre-defined data streams
- disallowing revisiting the past dist.
- single dist. to learn in a period

CMR: a more realistic CL setup

- boundary-agnostic
- dynamic streams of pred. errors
- allowing revisiting the past dist.
- diverse mixtures of OOD data

<https://cmr-nlp.github.io/>