Efficient Memory Crystallization for Graph Learning under Non-Stationary Distribution Shifts

Yue Hou[1,2], Ruomei Liu[1], Yingke Su[2], Junran Wu[1*], Ke Xu[1]

[1]State Key Laboratory of Complex & Critical Software Environment, Beihang University
[2]Shen Yuan Honors College, Beihang University *Corresponding author

NEURAL INFORMATION PROCESSING SYSTEMS

北京航空航天大學 BEIHANG UNIVERSITY

Paper | Code | WeChat

1. Introduction

Graph models face **continual distribution shifts**, intra-domain redundancy, inter-domain dependencies, and costly generative replay at test time.

► **EMC**: crystallize compact memories in closed form, then evolve them across domains to balance generalizability, adaptability, and forgetting.

2. Preliminaries

Adapt a source-pretrained GNN $f(\cdot;\Theta_0)$ to unlabeled graphs $\mathcal{T} = \{G^t\}_{t=1}^T$, without source data, under covariate shift:

$$\mathbb{P}_{\mathcal{S}}(G) \neq \mathbb{P}_{\mathcal{T}}(G), \qquad \mathbb{P}_{\mathcal{S}}(Y \mid G) = \mathbb{P}_{\mathcal{T}}(Y \mid G). \tag{1}$$

3. Methodology

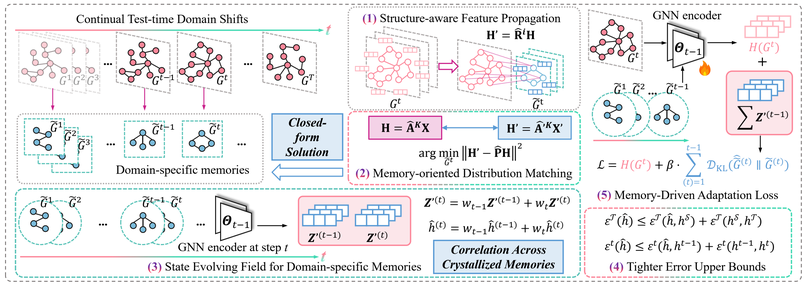


Figure 1. EMC: memory crystallization and memory-driven adaptation.

3.1 Efficient Memory Crystallization

Structure-aware Feature Propagation. Propagate features over K hops and form confidence-weighted semantic groups using predicted classes as anchors:

$$\mathbf{H} = \hat{\mathbf{A}}^K\mathbf{X}, \qquad \mathbf{H}' = \hat{\mathbf{P}}\mathbf{H}, \qquad \hat{\mathbf{P}} = \mathrm{diag}(\hat{\mathbf{R}}^1, \ldots, \hat{\mathbf{R}}^C). \tag{2}$$

Here $\hat{\mathbf{R}}^i$ is the row-normalized class-wise aggregation matrix, down-weighting uncertain nodes.

Memory-oriented Distribution Matching.

Proposition 3.1 (Distribution Matching). *For encoder weight $\Psi \sim \Phi$ and ordered memory nodes ($\mathbf{P}' = \mathbf{I}$), the embedding discrepancy satisfies*

$$\begin{aligned}&\mathbb{E}_{\Psi\sim\Phi}\|(\mathbf{H}' - \hat{\mathbf{P}}\mathbf{H})\Psi\|^2 \\ &\quad\le \mathbb{E}_{\Psi\sim\Phi}\|\mathbf{H}' - \hat{\mathbf{P}}\mathbf{H}\|^2\|\Psi\|^2.\end{aligned} \tag{3}$$

Thus memory construction minimizes $\|\mathbf{H}' - \hat{\mathbf{P}}\mathbf{H}\|^2$, aligning essential semantic information without optimizing a generator.

Redundancy-eliminated Memory Construction. Set $A'_{ij} = \mathbf{1}[\cos(\mathbf{H}'_i, \mathbf{H}'_j) > \tau]$, and recover smooth features with $\mathbf{Q} = \hat{\mathbf{A}}'^K$ and memory Laplacian $\mathbf{L}'$:

$$\min_{\mathbf{X}'} \|\mathbf{Q}\mathbf{X}' - \mathbf{H}'\|^2 + \alpha\,\mathrm{tr}(\mathbf{X}'^\top\mathbf{L}'\mathbf{X}'). \tag{4}$$

Proposition 3.2 (Closed-Form Memory Features). *The feature-reconstruction objective admits the closed-form solution*

$$\mathbf{X}' = (\mathbf{Q}^\top\mathbf{Q} + \alpha\mathbf{L}')^{-1}\mathbf{Q}^\top\mathbf{H}'. \tag{5}$$

The memory system has size $N'_t \times N'_t$, with $N'_t = 1\%N_t$.

3.2 Memory-Driven Adaptation with Tighter Bound

State Evolving Field over Crystallized Memories. Continually drifting domains are correlated; matching each marginal alone does not explicitly model their dependencies. EMC mixes consecutive memory states:

$$\mathbf{Z}'^{(t)} \leftarrow w_{t-1}\mathbf{Z}'^{(t-1)} + w_t\mathbf{Z}'^{(t)}, \tag{6}$$

where $w_{t-1} + w_t = 1$ and $w_{t-1}, w_t \ge 0$. The corresponding mixed hypothesis is $\hat{h} = w_{t-1}\hat{h}^{t-1} + w_t\hat{h}^t$.

Theorem 3.5 (Error Bound with Crystallized Memories). *For $\hat{h} = w_0\hat{h}^{\mathcal{S}} + w_1\hat{h}^{\mathcal{T}}$, with nonnegative weights summing to one, the target risk is bounded by*

$$\epsilon^{\mathcal{T}}(\hat{h}) \le \epsilon^{\mathcal{T}}(\hat{h}, h^{\mathcal{S}}) + \epsilon^{\mathcal{T}}(h^{\mathcal{S}}, h^{\mathcal{T}}). \tag{7}$$

Theorem 3.6 (Continual Error Bound). *For consecutive target domains and the state-evolving mixed hypothesis, the per-step target risk satisfies*

$$\epsilon^t(\hat{h}) \le \epsilon^t(\hat{h}, h^{t-1}) + \epsilon^t(h^{t-1}, h^t). \tag{8}$$

Derivation of the Adaptation Loss. Minimize current-domain predictive entropy and replay-based KL regularization on previous memories:

$$\mathcal{L} = H(G^t) + \beta\sum_{j=1}^{t-1}\mathcal{D}_{\mathrm{KL}}(\hat{\tilde{G}}^{(j)}\|\tilde{G}^{(j)}). \tag{9}$$

Here $\tilde{G}^{(j)}$ provides a frozen memory anchor, and $\hat{\tilde{G}}^{(j)}$ denotes current predictions on the same memory. Update $\Theta_{t-1} \to \Theta_t$, keeping the classifier fixed.

Proposition 3.7 (Stability of the Adaptation Loss). *Assume temperature-τ softmax, class probabilities at least $\gamma > 0$, and encoder/classifier Lipschitz constants L_h, L_g in the model parameters. Then*

$$\mathcal{D}_{\mathrm{KL}}(\hat{\tilde{G}}^{(j)}\|\tilde{G}^{(j)}) \le \frac{L_g^2L_h^2}{4\gamma\tau^2}\|\Theta_t - \Theta_{t-1}\|_2^2. \tag{10}$$

Visualization of Crystallized Memories

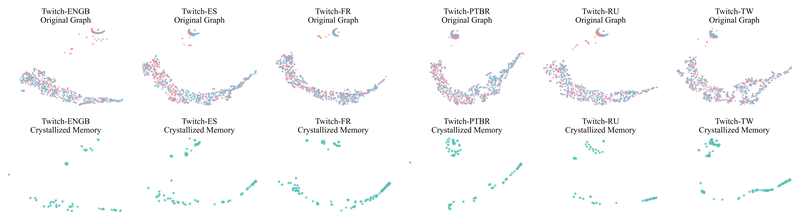


Figure 2. t-SNE of original graphs and crystallized memories on Twitch / GCN.

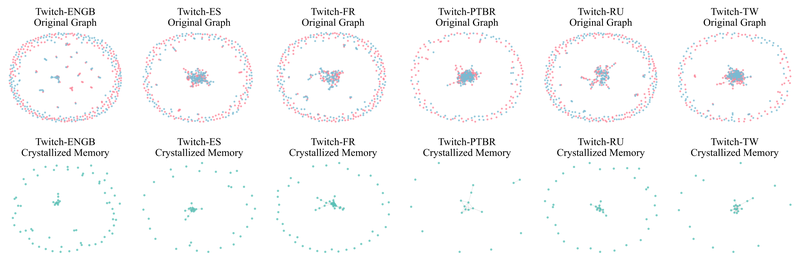


Figure 3. NetworkX views on Twitch / GCN: sampled original graphs (top), full crystallized memories (bottom).

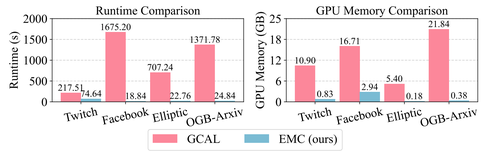
4. Experiments

4.1 Experimental Setup. Four datasets under regional and temporal shifts; five runs.

4.2 Overall Performance Comparison

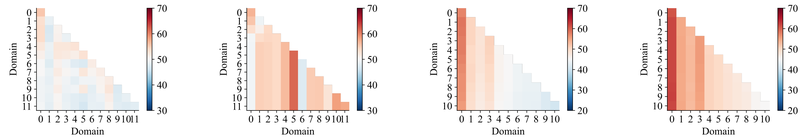
Table 1. GCN performance (%, mean ± std., five runs), runtime (s), and GPU memory (MiB).

Dataset	Methods	Backbone: GCN					
		AP (↑)	AF (↑)	TDG (↑)	TDA (↑)	Time (↓)	GPU (↓)
Twitch	LEBED	51.06±0.05	0.24±0.05	51.87±0.06	48.90±0.04	114.97	1440.57
	GCAL	55.65±0.09	0.42±0.13	54.78±0.60	53.20±0.62	217.51	11159.8
	EMC	**55.95±0.15**	**0.47±0.08**	**55.56±0.13**	**55.83±0.14**	**74.64**	**854.97**
Facebook	LEBED	46.90±0.00	-0.25±0.00	47.72±0.00	46.38±0.00	93.59	3481.53
	GCAL	52.72±0.36	0.72±0.19	49.37±0.28	48.96±0.29	1675.2	17107.84
	EMC	**53.10±0.00**	**2.15±0.27**	**53.49±0.27**	**52.30±0.04**	**18.84**	**3006.08**
Elliptic	LEBED	49.44±0.19	-4.86±0.20	52.10±0.06	53.82±0.02	114.01	745.36
	GCAL	56.57±0.14	0.88±0.13	53.08±2.54	50.31±2.98	707.24	5533.99
	EMC	**56.76±0.46**	**0.96±0.30**	**58.73±0.40**	**55.92±0.31**	**22.76**	**189.36**
Arxiv	LEBED	19.09±0.04	-0.04±0.05	14.18±0.02	17.44±0.02	227.80	839.39
	GCAL	14.15±6.94	-1.80±2.06	17.90±8.58	13.77±5.57	1362.76	22342.47
	EMC	**19.32±1.57**	**3.38±1.99**	**19.39±1.47**	**20.39±1.71**	**24.95**	**299.18**



87.4% less runtime
92.4% less GPU memory
Average reductions vs. GCAL.

Figure 4. Runtime and GPU memory cost: GCAL vs. EMC.



(a) GCAL / Facebook (b) EMC / Facebook (c) GCAL / Arxiv (d) EMC / Arxiv

Figure 5. Performance matrices of EMC and GCAL.

4.3 Scaling Efficiency Analysis

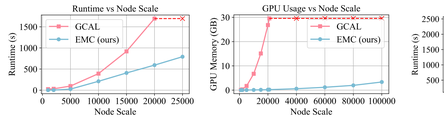


Figure 6. Node scaling efficiency.

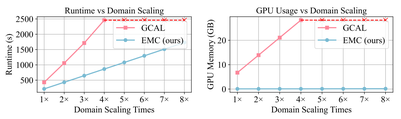


Figure 7. Domain scaling efficiency.

4.4 Ablation Studies

Table 2. Ablation study (%, s, MiB); GCN, SAGE for Arxiv.

Dataset	Method	AP(↑)	TDG(↑)	TDA(↑)	Time(↓)	GPU(↓)
Twitch	*w/o DM*	52.12	53.77	51.35	99.43	1486.81
	w/o EA	51.50	52.31	51.21	60.13	783.14
	EMC	**55.95**	**55.56**	**55.83**	74.64	854.97
Facebook	*w/o DM*	48.78	48.96	48.92	85.45	13481.55
	w/o EA	51.30	50.85	50.34	17.65	2965.72
	EMC	**53.10**	**53.49**	**52.30**	18.84	3006.08
Elliptic	*w/o DM*	47.44	49.20	47.57	97.60	747.93
	w/o EA	47.35	50.55	50.00	20.52	183.57
	EMC	**56.76**	**58.73**	**55.92**	22.76	189.36
Arxiv	*w/o DM*	12.61	16.45	17.58	30.52	1083.73
	w/o EA	20.65	20.50	20.51	13.26	390.12
	EMC	**46.07**	**47.11**	**43.52**	24.84	391.02

hou_yue@buaa.edu.cn

NeurIPS 2026 | Sydney | Hall 1–4 | Tue 17:00 AEDT | Poster Session 2