Rethinking Graph Backdoor Defense: A Topological, Coarse-to-Fine Perspective

Jiecheng Zhai Xuzeng Li Jian Wang* Jiqiang Liu
Beijing Jiaotong University, Beijing, China
{jc_zhai,22110130,wangjian,jqliu}@bjtu.edu.cn

Abstract

Graph Neural Networks (GNNs) power applications from social and financial networks to biology, yet they are vulnerable to backdoor attacks where tiny trigger subgraphs force targeted misclassification while preserving clean accuracy. We present **TCF**, a *T*opological *C*oarse-to-*F*ine defense that relies only on structure. First, *Coarse Structural Pruning* (CSP) screens nodes via three near-linear tests—local spectral moments, one-step 1-WL color rarity, and ego-density Z-scores—merged by a unified p-value rule with finite-sample FPR control. Second, a *structure-based* detector is trained on clean d-hop subgraphs versus compact synthetic triggers from small-world and preferential-attachment priors. Finally, *label-flip verification pruning* removes a subgraph only if its deletion flips the node's prediction. On Cora, PubMed, Flickr, and OGB-Arxiv under three state-of-the-art attacks, TCF typically reduces ASR to $< 5\%$ while maintaining clean accuracy, indicating topology alone can deliver accurate, scalable graph backdoor defense.

1 Introduction

Graphs underpin key applications in social media[Fan et al., 2019, Guo et al., 2022], finance[Cheng et al., 2023, Innan et al., 2024], biology[Lee et al., 2020, Li et al., 2022], recommendation, and knowledge graphs. Graph Neural Networks (GNNs)[Kipf and Welling, 2016, Veličković et al., 2017] have become the standard tool for learning on such data via message passing over topology[Yang et al., 2021, Zhang et al., 2021a, Yu et al., 2021, Yasunaga et al., 2022, Jia et al., 2023, Li et al., 2023], delivering strong results on node/graph classification[Yang et al., 2022, Yao et al., 2022, Liu et al., 2021, Wang et al., 2024] and link prediction[Li et al., 2024, Xiong et al., 2024]. However, recent work shows GNNs are vulnerable to *backdoor attacks*[Zhang et al., 2021b, Xi et al., 2021]: an adversary implants tiny trigger subgraphs during training so that any node carrying the trigger is mapped to an attacker-chosen label while clean accuracy remains high.

Prior attacks have evolved from early trigger designs to adaptive and in-distribution variants that improve stealth and reduce budgetDai et al. [2023], Zhang et al. [2024a]. In response, most defenses learn heavy detectors[Yu et al., 2025] that *depend on node attributes* (or joint feature–structure models), often using explainers[Jiang and Li, 2022] or perturbation procedures. This creates three practical issues: (i) *computational cost*—per-node analysis scales poorly on large graphs; (ii) *limited transferability*—feature distributions vary across domains; and (iii) *underuse of topology*—triggers are discrete structural insertions and must disturb local graph structure, even when globally subtle. These observations motivate a defense that is *topology-only*, *lightweight*, and *coarse-to-fine*.

We propose **TCF**, a *T*opological *C*oarse-to-*F*ine framework for graph backdoor defense. First, *Coarse Structural Pruning (CSP)* performs near-linear screening with three complementary signals—local spectral moments, one-step 1-WL color rarity, and ego-density Z-scores—merged via a unified p-value test with finite-sample false-positive control. CSP retains a small candidate set. Second,

*Corresponding author.

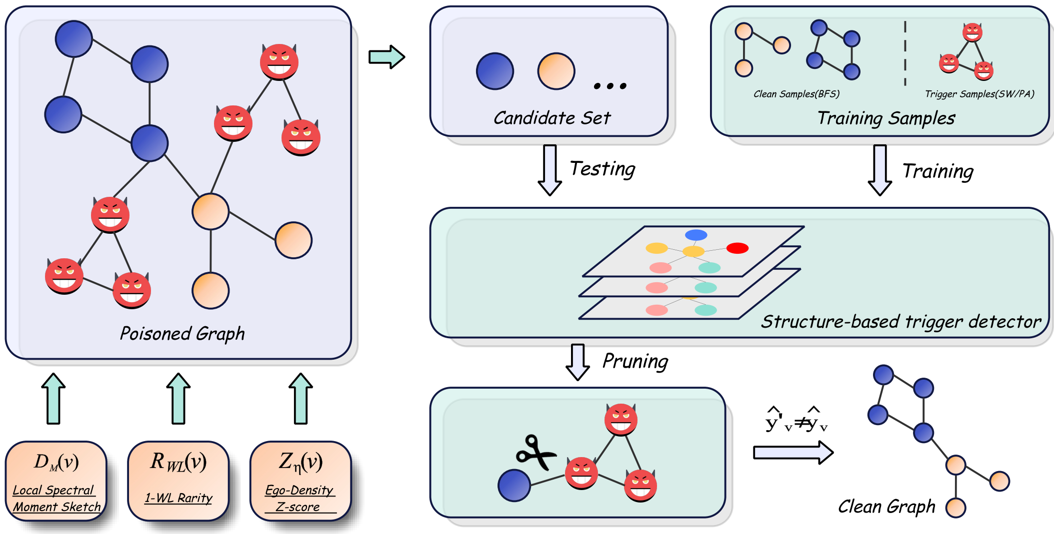

Figure 1: Overview of the TCF pipeline: coarse structural screening, structure-only detection, and label-flip verification.

a *structure-only* detector (GCN + classifier) is trained on clean d-hop subgraphs versus compact synthetic triggers generated from small-world and preferential-attachment priors, with features neutralized to enforce structure bias. Finally, *label-flip verification pruning* removes a subgraph only if its deletion flips the node's prediction, ensuring causal relevance.

2 Method

2.1 Preliminaries

We consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ with node set $\mathcal{V}$, edge set $\mathcal{E}$, and node-feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d_x}$, where $N = |\mathcal{V}|$ and the i-th row of $\mathbf{X}$ is the feature of node $v_i \in \mathcal{V}$. The adjacency matrix is $\mathbf{A} \in \mathbb{R}^{N \times N}$. $\mathbf{D}$ is the degree matrix, and $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ is the normalized Laplacian. The task is node classification with a GNN f_θ producing class probabilities $f_\theta(\mathcal{G}, v)$ for node v. A *trigger* is a small connected subgraph $g_t = (\mathcal{V}_t, \mathcal{E}_t, \mathbf{X}_t)$ with size $n = |\mathcal{V}_t| \le \mathcal{B}$ (budget) and edge count $e = |\mathcal{E}_t| \in \left[n-1, \frac{n(n-1)}{2}\right]$. Attaching g_t to a target node via a few new edges yields a poisoned graph, denoted $\mathcal{G} \oplus g_t$. All symbols introduced here will be used consistently in subsequent sections.

2.2 Coarse Structural Pruning (CSP)

2.2.1 Local Spectral Moment Sketch

Setup. Let $\tilde{\mathbf{L}}$ be the normalized Laplacian of $\mathcal{G}$ affinely rescaled to $[-1, 1]$. For node v and order r, define the Chebyshev moment sketch

$$\phi_{\text{SPEC}}(v) = \left[e_v^\top T_1(\tilde{\mathbf{L}})e_v, \ldots, e_v^\top T_r(\tilde{\mathbf{L}})e_v\right] \in \mathbb{R}^r, \tag{1}$$

and let $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the mean and covariance of ϕ_{SPEC} estimated on clean data. The spectral deviation is the Mahalanobis distance

$$D_{\text{M}}(v) = \left(\phi_{\text{SPEC}}(v) - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\phi_{\text{SPEC}}(v) - \boldsymbol{\mu}\right). \tag{2}$$

Computing $\phi_{\text{SPEC}}(\cdot)$ for all nodes requires r sparse matrix–vector multiplies, i.e., $O(rE)$.

Lemma 1 (Clean FPR control and detection power)**.** *Let $r \ge 1$. (a)* Clean FPR. *If clean-node sketches satisfy $\phi_{\text{SPEC}}(v) \overset{clean}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D_{\text{M}}(v) \overset{clean}{\sim} \chi_r^2$. Thus choosing $\tau_{\text{M}} = F_{\chi_r^2}^{-1}(1-\delta)$ ensures* $\Pr\left(D_{\text{M}}(v) > \tau_{\text{M}} \mid v \text{ clean}\right) \le \delta$*. If $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are estimated from n_0 clean sketches, then*

$$\frac{n_0 - r}{r(n_0 - 1)} (\phi - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\phi - \hat{\boldsymbol{\mu}}) \overset{clean}{\sim} F_{r, n_0 - r}, \tag{3}$$

yielding an exact finite-sample threshold. (b) Power. *If a trigger induces a mean shift* $\Delta = \mathbf{\Sigma}^{-1/2}\big(\phi_{\text{SPEC}}(v^\star) - \boldsymbol{\mu}\big)$ *with noncentrality* $\lambda = \|\Delta\|_2^2 > 0$*, then* $D_{\mathrm{M}}(v^\star) \sim \chi_r^2(\lambda)$ *and*

$$\Pr\big(D_{\mathrm{M}}(v^\star) > \tau_{\mathrm{M}}\big) = 1 - F_{\chi_r^2(\lambda)}(\tau_{\mathrm{M}}), \tag{4}$$

which increases monotonically with λ.

Proof. See Appendix B for a detailed derivation based on (i) asymptotic normality of bounded Chebyshev filters, (ii) the χ^2 law of Mahalanobis distances, and (iii) the Hotelling–T^2 correction.

Each coordinate $e_v^\top T_k(\tilde{\mathbf{L}})e_v$ summarizes short closed walks around v (triangles, short cycles). Triggers inject extra local connectivity, shifting ϕ_{SPEC} and enlarging D_{M}. Lemma 1 provides a calibrated threshold for clean false alarms and an explicit power characterization via the noncentrality λ, while retaining near-linear runtime $O(rE)$.

2.2.2 1-WL Rarity

Setup. Run one iteration of Weisfeiler–Lehman (1-WL) color refinement on G in $O(E)$ time. Let $c(v)$ be the WL color of node v. From a clean calibration set $\mathcal{D}^+_{\text{cal}}$, compute the empirical color frequency

$$\hat{\pi}(c) = \frac{\#\{u \in \mathcal{D}^+_{\text{cal}} : \ c(u) = c\}}{|\mathcal{D}^+_{\text{cal}}|}. \tag{5}$$

Define a conformal p-value

$$p(v) = \frac{1 + \#\{u \in \mathcal{D}^+_{\text{cal}} : \ \hat{\pi}(c(u)) \le \hat{\pi}(c(v))\}}{1 + |\mathcal{D}^+_{\text{cal}}|}, \tag{6}$$

and the logarithmic rarity score $R_{\text{WL}}(v) = -\log p(v)$ (larger is rarer).

Lemma 2 (Finite-sample FPR control and detection power)**.** *(a) Clean FPR. If the clean calibration set and a clean test node are exchangeable w.r.t. 1-WL colors, then*

$$\Pr\big(p(v) \le \delta \mid v \text{ clean}\big) \ \le \ \delta, \qquad \forall\, \delta \in (0,1). \tag{7}$$

Equivalently, thresholding $R_{\text{WL}}(v)$ *at* $-\log\delta$ *controls the clean false positive rate at* δ*. (b) Power. If a trigger alters the degree–multiset pattern so that* $c(v^\star)$ *shifts to lower clean-frequency bins, then* $p(v^\star)$ *stochastically decreases (and* $R_{\text{WL}}(v^\star)$ *increases); in particular,*

$$\Pr\big(R_{\text{WL}}(v^\star) > -\log\delta\big) \ \ge \ \Pr\big(\hat{\pi}(c(v^\star)) \le q_\delta\big), \tag{8}$$

where q_δ *is the* δ*-quantile of clean frequencies.*

Proof. See Appendix C. Part (a) follows from exchangeability of ranks in conformal prediction; part (b) follows from monotonicity of the rank statistic with respect to $\hat{\pi}(c(v))$.

A single 1-WL step encodes each node's one-hop degree–multiset pattern. Trigger insertions (e.g., small cliques or stars) perturb this pattern, making the resulting color unusually infrequent under the clean reference. The conformal construction yields distribution-free, finite-sample FPR guarantees while keeping runtime linear in E.

2.2.3 Ego-Density Z-score

Setup. For node v, let $\mathcal{N}_1(v)$ be its 1-hop neighbors and let $G[\{v\}\cup\mathcal{N}_1(v)]$ be the ego network. Denote by $m_{\text{EGO}}(v)$ the number of edges inside this ego network (including v–neighbor and neighbor–neighbor edges). Define the ego-density

$$\eta(v) \ = \ \frac{2\, m_{\text{EGO}}(v)}{|\mathcal{N}_1(v)|\big(|\mathcal{N}_1(v)| - 1\big) + 2|\mathcal{N}_1(v)|} \ \in [0,1], \tag{9}$$

and its standardized score $Z_\eta(v) = (\eta(v) - \bar{\eta})/\sigma_\eta$, where $(\bar{\eta}, \sigma_\eta)$ are estimated from clean data (optionally conditioned on degree/community). We compute m_{EGO} in linear time via neighbor-pair sampling, preserving $O(E)$ runtime.

Lemma 3 (Two-sided FPR control and detection power)**.** *(a) Clean FPR. Let $\mathcal{D}^{+}_{\text{cal}}$ be a clean calibration set and define scores $s(u) = |Z_\eta(u)|$ for $u \in \mathcal{D}^{+}_{\text{cal}}$. For a test node v, set the conformal p-value*

$$p(v) \;=\; \frac{1 + \#\{u \in \mathcal{D}^{+}_{\text{cal}} : \; s(u) \geq s(v)\}}{1 + |\mathcal{D}^{+}_{\text{cal}}|}. \tag{10}$$

Under exchangeability (i.i.d. clean sampling), $\Pr\big(p(v) \leq \delta \mid v \text{ clean}\big) \leq \delta$ for any $\delta \in (0, 1)$. Equivalently, thresholding $|Z_\eta(v)|$ by the $(1-\delta)$-quantile of $\{s(u)\}$ controls the clean FPR at δ. (b) Power. If a trigger induces an ego-density shift $\eta(v^\star) = \eta_0(v^\star) + \Delta_\eta$ with $|\Delta_\eta|/\sigma_\eta \geq \gamma > 0$, then

$$\Pr\big(p(v^\star) \leq \delta\big) \;\geq\; 1 - \Phi\big(z_{1-\delta} - \gamma\big), \tag{11}$$

where $z_{1-\delta}$ is the $(1-\delta)$-quantile of $\mathcal{N}(0, 1)$ and Φ its CDF; thus power increases with the standardized shift γ.

Proof. See Appendix D. Part (a) follows from rank-exchangeability of conformal scores; part (b) uses a CLT/Bernstein approximation for ego-density under local edge additions/removals.

$\eta(v)$ quantifies how crowded (triangle/clique-like) or sparse (star/chain-like) the 1-hop neighborhood is. Triggers that add neighbor–neighbor edges increase η; star/chain-like additions decrease it. The two-sided standardized score $|Z_\eta|$ captures both effects, and the rank-based calibration provides distribution-free FPR control with linear-time computation.

2.2.4 Coarse Score and Candidate Set

We aggregate the three CSP signals into a single score

$$S(v) = \lambda_1 \cdot \mathbb{1}\{D_{\text{M}}(v) > \tau_{\text{M}}\} + \lambda_2 \cdot \mathbb{1}\{R_{\text{WL}}(v) > \tau_{\text{WL}}\} + \lambda_3 \cdot \mathbb{1}\{|Z_\eta(v)| > \tau_\eta\}. \tag{12}$$

We keep the top-ρ% nodes by $S(v)$,

$$C = \big\{\, v \in \mathcal{V} \; : \; S(v) \text{ in top-}\rho\% \,\big\}, \tag{13}$$

and run the detector and verification pruning only for $v \in C$. This reduces downstream cost by a factor ρ while CSP itself remains near-linear: $O(rE)$ for spectral moments plus $O(E)$ for 1-WL and ego-density.

2.3 Refined Pruning

2.3.1 Sample collection.

After CSP's coarse screening, we assemble supervised data for fine-grained inspection. Positives are d-hop subgraphs randomly sampled from a clean graph (BFS with a node cap). Negatives are small connected subgraphs synthesized under Small-World (SW) and Preferential Attachment (PA) priors with size $n \leq B$; trigger node features are set to constants to enforce structure-only cues.

2.3.2 Structure-based trigger detector.

We train a lightweight structure-only GCN encoder with mean readout and a linear classifier to separate clean vs. trigger subgraphs using class-weighted cross-entropy. At test time, only CSP-flagged nodes are evaluated: for each candidate v, extract $G[v; d]$, embed, and score; high-score subgraphs proceed to verification.

2.3.3 Label-flip verification pruning.

Let f_θ be the downstream node classifier. For candidate node v and suspicious subgraph $g \subseteq G[v; d]$, compute the predicted label before/after temporary removal:

$$\hat{y}_v = \arg\max_c \; f_\theta(G, v)_c, \qquad \hat{y}'_v = \arg\max_c \; f_\theta(G \setminus g, \; v)_c. \tag{14}$$

We permanently prune g if a label flip occurs,

$$\hat{y}'_v \neq \hat{y}_v, \tag{15}$$

This causality check preserves precision, while CSP limits how many nodes reach this stage, keeping the pipeline scalable.

Table 1: Defense performance of TCF compared with baseline methods.

Dataset	Attack	No-Defense		Prune		Prune-LD		RIGBD		DShield		TCF (Ours)	
		ASR	**ACC**	**ASR**	**ACC**	**ASR**	**ACC**	**ASR**	**ACC**	**ASR**	**ACC**	**ASR**	**ACC**
Cora	GTA	0.900	0.823	0.243	0.788	0.199	0.789	0.056	0.802	**0.007**	0.819	0.047±0.003	0.800±0.003
	UGBA	0.941	0.834	0.894	0.790	0.876	0.776	0.076	0.815	0.037	0.821	**0.035±0.004**	0.814±0.002
	DPGBA	0.946	0.826	0.899	0.801	0.882	0.782	0.153	0.809	0.027	0.848	**0.025±0.003**	0.848±0.003
PubMed	GTA	0.843	0.855	0.312	0.775	0.255	0.750	0.070	0.813	0.042	0.858	**0.013±0.002**	0.850±0.002
	UGBA	0.891	0.861	0.877	0.810	0.830	0.801	0.062	0.781	0.034	0.847	**0.033±0.004**	0.851±0.003
	DPGBA	0.897	0.864	0.890	0.800	0.881	0.810	0.120	0.755	0.048	0.820	**0.035±0.003**	0.786±0.004
Flickr	GTA	0.877	0.452	0.082	0.430	0.068	0.402	0.081	0.409	0.071	0.520	**0.062±0.005**	0.433±0.004
	UGBA	0.912	0.461	0.800	0.431	0.850	0.421	0.091	0.401	0.071	0.501	**0.044±0.004**	0.452±0.004
	DPGBA	0.921	0.455	0.876	0.407	0.853	0.410	0.138	0.408	0.056	0.505	**0.037±0.004**	0.432±0.003
OGB-Arxiv	GTA	0.753	0.634	0.124	0.630	0.119	0.633	0.108	0.612	0.092	0.620	**0.058±0.005**	0.639±0.003
	UGBA	0.964	0.665	0.936	0.661	0.901	0.660	0.099	0.604	**0.001**	0.619	0.024±0.004	0.642±0.003
	DPGBA	0.971	0.651	0.945	0.627	0.928	0.657	0.117	0.648	**0.003**	0.655	0.017±0.003	0.632±0.004

3 Experiments

3.1 Experimental Setup.

We evaluate on Cora, PubMed, Flickr, and OGB-Arxiv against three attacks (GTA[Xi et al., 2021], UGBA[Dai et al., 2023], DPGBA[Zhang et al., 2024a]) and four defenses (PruneDai et al. [2023], Prune-LD[Dai et al., 2023], RIGBD[Zhang et al., 2024b], DShield[Yu et al., 2025]). Our CSP uses one 1-WL iteration and Chebyshev order r=4 for spectral moments; we target a global clean-FPR δ=0.03 via conformal calibration of the component thresholds, aggregate with the weighted indicator score $S(v)$, and—for the budgeted variant—select candidates by the top-ρ=5% nodes ranked by $S(v)$. Conformal calibration for 1-WL and ego-density uses a label-stratified 10% slice of training nodes as the clean calibration set. Ego-density Z is computed with neighbor–neighbor pair sampling capped at M=2000 pairs per node and degree-binned standardization (5 bins). For the refined detector in **TCF** (Topological Coarse-to-Fine Defense), positives are clean d-hop subgraphs (BFS, seed rate α=0.2, depth d=5); negatives are SW/PA triggers with node budget B=10 and constant features. The detector is a 2-layer GCN (16 hidden), trained for 300 epochs with learning rate 0.01; we use an 80/20 train/test split and report the mean over 5 runs. Backdoor injection strictly follows each attack's original settings. *Full dataset statistics, baseline details, and hyperparameters appear in the Appendix.*

3.2 Results.

Across all datasets and attacks, **TCF** attains consistently low attack success rate (ASR; typically $< 5\%$) while preserving high clean accuracy. Compared to DShield, TCF shows similar or within 1–2% higher ASR in a few cases but matches or improves clean accuracy and exhibits notably stronger cross-dataset transfer: models trained on one dataset maintain low ASR and stable accuracy when deployed to another, whereas DShield degrades. Overall, the topological coarse-to-fine pipeline yields robust detection and causal pruning with near-linear runtime.

4 Discussion

We present a topology-only, coarse-to-fine defense that combines broad structural screening, structure-based detection, and label-flip verification to reduce attack success while maintaining clean accuracy and showing solid transfer across datasets. While results are encouraging, several limits remain. The study focuses on small, localized trigger patterns; other attack styles or adaptive strategies that intentionally resemble normal structure could be more challenging. The screening stage relies on a modest amount of clean calibration and a handful of hyperparameters, whose defaults worked well in our tests but might benefit from automatic tuning in unusual graphs. The detector deliberately ignores node attributes to promote transfer, which can trade some recall in settings where features are stable and informative. Finally, the verification step uses predictions of the downstream model, and additional safeguards may help when decisions are uncertain.Overall, our work offers a clear, topology-based way to defend GNNs and provides a practical base for future extensions to richer settings such as heterogeneous graphs and multi-relation networks.

Acknowledgments and Disclosure of Funding

This work was supported by the Science and Technology Research and Development Program of China State Railway Group Co., Ltd. (Grant No. K2024W004), and the Fundamental Research Funds for the Central Universities (Grant No. 2025JBZY025).

References

D. Cheng, Z. Niu, J. Zhang, Y. Zhang, and C. Jiang. Critical firms prediction for stemming contagion risk in networked-loans through graph-based deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14205–14213, 2023.

E. Dai, M. Lin, X. Zhang, and S. Wang. Unnoticeable backdoor attacks on graph neural networks. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, pages 2263–2273, 2023.

W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin. Graph neural networks for social recommendation. In *Proceedings of the World Wide Web Conference (WWW)*, pages 417–426, 2019.

Z. Guo, K. Yu, A. Jolfaei, G. Li, F. Ding, and A. Beheshti. Mixed graph neural network-based fake news detection for sustainable vehicular social networks. *IEEE Transactions on Intelligent Transportation Systems*, 24(12): 15486–15498, 2022.

N. Innan, A. Sawaika, A. Dhor, S. Dutta, S. Thota, H. Gokal, N. Patel, M. A.-Z. Khan, I. Theodonis, and M. Bennai. Financial fraud detection using quantum graph neural networks. *Quantum Machine Intelligence*, 6 (1), 2024.

M. Jia, J. Hu, Y. Liu, Z. Gao, and Y. Yao. Topology-guided graph learning for process fault diagnosis. *Industrial & Engineering Chemistry Research*, 62(7):3238–3248, 2023.

B. Jiang and Z. Li. Defending against backdoor attack on graph neural network by explainability. *arXiv preprint*, 2022.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint*, 2016.

J.-Y. Lee, N. C. Sadler, R. G. Egbert, C. R. Anderton, K. S. Hofmockel, J. K. Jansson, and H.-S. Song. Deep learning predicts microbial interactions from self-organized spatiotemporal patterns. *Computational and Structural Biotechnology Journal*, 18:1259–1269, 2020.

F. Li, J. Feng, H. Yan, G. Jin, F. Yang, F. Sun, D. Jin, and Y. Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1): 1–21, 2023.

J. Li, H. Shomer, H. Mao, S. Zeng, Y. Ma, N. Shah, J. Tang, and D. Yin. Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

M. M. Li, K. Huang, and M. Zitnik. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12):1353–1369, 2022.

N. Liu, S. Jian, D. Li, Y. Zhang, Z. Lai, and H. Xu. Hierarchical adaptive pooling by capturing high-order dependency for graph representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 35 (4):3952–3965, 2021.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint*, 2017.

Y. Wang, X. Luo, C. Chen, X.-S. Hua, M. Zhang, and W. Ju. Disensemi: Semi-supervised graph classification via disentangled representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Z. Xi, R. Pang, S. Ji, and T. Wang. Graph backdoor. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1523–1540, 2021.

S. Xiong, Y. Yang, A. Payani, J. C. Kerce, and F. Fekri. Teilp: Time prediction over knowledge graphs via logical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16112–16119, 2024.

H. Yang, H. He, W. Zhang, and Y. Bai. Mtgk: Multi-source cross-network node classification via transferable graph knowledge. *Information Sciences*, 589:395–415, 2022.

X. Yang, Y. Lyu, T. Tian, Y. Liu, Y. Liu, and X. Zhang. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1417–1423, 2021.

W. Yao, K. Guo, Y. Hou, and X. Li. Hierarchical structure-feature aware graph neural network for node classification. *IEEE Access*, 10:36846–36855, 2022.

M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, and J. Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pages 37309–37323, 2022.

H. Yu, Y. Liu, X. Zhang, L. Sun, and J. Li. Dshield: Defending against backdoor attacks on graph neural networks via discrepancy learning. In *Network and Distributed System Security Symposium (NDSS)*, 2025.

X. Yu, S.-H. Wang, and Y.-D. Zhang. Cgnet: A graph-knowledge embedded convolutional neural network for detection of pneumonia. *Information Processing & Management*, 58(1):102411, 2021.

X. Zhang, C. Huang, Y. Xu, L. Xia, P. Dai, L. Bo, J. Zhang, and Y. Zheng. Traffic flow forecasting with spatial-temporal graph diffusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15008–15015, 2021a.

Z. Zhang, J. Jia, B. Wang, and N. Z. Gong. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 15–26, 2021b.

Z. Zhang, M. Lin, E. Dai, and S. Wang. Rethinking graph backdoor attacks: A distribution-preserving perspective. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 4386–4397, 2024a.

Z. Zhang, M. Lin, J. Xu, Z. Wu, E. Dai, and S. Wang. Robustness-inspired defense against backdoor attacks on graph neural networks. *arXiv preprint*, 2024b.

A Related Works

Backdoor attacks on Graph Neural Networks (GNNs) inject small trigger subgraphs into training data and assign target labels, causing models to misclassify any test input that contains the trigger while keeping clean accuracy largely unchanged. Early studies introduced universal subgraph triggers and motif-based triggers to increase effectiveness. Subsequent lines of work developed adaptive trigger generators that tailor triggers to the data, selected target nodes using centrality measures to maximize impact, and explored clean-label settings that enhance stealth by avoiding explicit label changes on poisoned nodes.

Defenses generally aim to discover and mitigate the influence of triggers without harming overall task performance. Representative approaches learn robust node embeddings in a self-supervised manner, cluster nodes to reveal distributional irregularities, and prune suspicious nodes or edges based on inter-cluster discrepancies. Other lines use explanation- or perturbation-based analyses to identify abnormal structural patterns. These methods provide useful baselines but often rely on node attributes or heavy training procedures, motivating topology-focused defenses that emphasize structural signals and causal verification.

B Derivation for Lemma 1

B.1 Asymptotic normality of moment sketches. Let $\tilde{\mathbf{L}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{\Lambda} \in [-1, 1]^{N\times N}$. Chebyshev filters satisfy $|T_k(x)| \le 1$ on $[-1, 1]$ and admit the expansion $T_k(\tilde{\mathbf{L}}) = \sum_{j=0}^{k} \alpha_{k,j} \tilde{\mathbf{L}}^j$. Hence

$$e_v^\top T_k(\tilde{\mathbf{L}}) e_v = \sum_{j=0}^{k} \alpha_{k,j}\, e_v^\top \tilde{\mathbf{L}}^j e_v,$$

where $e_v^\top \tilde{\mathbf{L}}^j e_v$ counts weighted closed walks of length j rooted at v. For sparse random graphs (e.g., configuration models with $\Delta = O(\log N)$) these counts are sums of weakly dependent bounded terms and obey a multivariate CLT:

$$\sqrt{N}\big(\phi_{\text{SPEC}}(v) - \boldsymbol{\mu}\big) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

B.2 Mahalanobis distance laws. If $\phi \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D_{\mathrm{M}} = (\phi - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\phi - \boldsymbol{\mu}) \sim \chi_r^2$. If $\phi \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\Delta, \boldsymbol{\Sigma})$, then $D_{\mathrm{M}} \sim \chi_r^2(\lambda)$ with $\lambda = \|\Delta\|_2^2$. When $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are replaced by empirical estimates from n_0 i.i.d. clean samples, Hotelling's theorem yields

$$\frac{n_0 - r}{r(n_0 - 1)}\,(\phi - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\phi - \hat{\boldsymbol{\mu}}) \;\sim\; F_{r, n_0 - r}.$$

B.3 Thresholds and power. Set the clean FPR to δ via $\tau_{\mathrm{M}} = F_{\chi_r^2}^{-1}(1 - \delta)$ (or the finite-sample F-quantile above). Under a trigger, the mean shift Δ generated by extra short cycles increases the noncentrality λ, and the detection probability $1 - F_{\chi_r^2(\lambda)}(\tau_{\mathrm{M}})$ grows monotonically with λ.

C Derivation for Lemma 2

C.1 Conformal p-value validity. Under exchangeability of $\{c(u) : u \in \mathcal{D}_{\text{cal}}^+\} \cup \{c(v)\}$, the multiset $\{\hat{\pi}(c(u))\} \cup \{\hat{\pi}(c(v))\}$ is also exchangeable. Hence the rank of $\hat{\pi}(c(v))$ among $1 + |\mathcal{D}_{\text{cal}}^+|$ values is uniform; the smoothed rank $p(v)$ is super-uniform: $\Pr(p(v) \le \delta) \le \delta$, giving FPR control for any δ.

C.2 Detection power. If a trigger moves $c(v^\star)$ toward colors with smaller clean frequency, then $\hat{\pi}(c(v^\star))$ tends to be lower than calibration values, increasing its extremal rank and decreasing $p(v^\star)$. Since $R_{\text{WL}} = -\log p$ is monotone in $1/p$, the probability of exceeding a fixed threshold grows with the shift toward rarer colors, yielding the stated inequality.

D Derivation for Lemma 3

D.1 Moments under a local independence model. For fixed $d_v = |\mathcal{N}_1(v)|$, write $m_{\text{EGO}}(v) = d_v + X_v$, where X_v counts neighbor–neighbor edges. Under an ER/locally independent approximation, $X_v \sim \mathrm{Binomial}\big(\binom{d_v}{2},\, p_v\big)$. Hence

$$\mathbb{E}[\eta(v) \mid d_v] = \frac{2\big(d_v + \binom{d_v}{2} p_v\big)}{d_v(d_v - 1) + 2d_v}, \qquad \mathrm{Var}[\eta(v) \mid d_v] = \frac{4\binom{d_v}{2} p_v(1 - p_v)}{\big(d_v(d_v - 1) + 2d_v\big)^2}.$$

D.2 Normal/Bernstein approximation. When $\binom{d_v}{2}$ is moderate, by CLT

$$\frac{\eta(v) - \mathbb{E}[\eta(v) \mid d_v]}{\sqrt{\mathrm{Var}[\eta(v) \mid d_v]}} \;\approx\; \mathcal{N}(0, 1).$$

For finite d_v, Bernstein's inequality yields for any $t > 0$:

$$\Pr\Big(\big|\eta(v) - \mathbb{E}[\eta(v) \mid d_v]\big| > t\Big) \;\le\; 2\exp\left(-\frac{\binom{d_v}{2} t^2}{2\,p_v(1 - p_v)/(d_v(d_v - 1) + 2d_v)^2 + \frac{2}{3}t}\right).$$

D.3 Finite-sample FPR via conformal ranks. Define $s(u) = |Z_\eta(u)|$ on $\mathcal{D}_{\text{cal}}^+$. Under exchangeability of $\mathcal{D}_{\text{cal}}^+ \cup \{v\}$, the (smoothed) rank of $s(v)$ among $1 + |\mathcal{D}_{\text{cal}}^+|$ values is uniform; thus $\Pr(p(v) \le \delta) \le \delta$.

D.4 Power under standardized shift. If a trigger changes neighbor–neighbor connectivity so that $\eta(v^\star) = \eta_0(v^\star) + \Delta_\eta$, then under the normal approximation $|Z_\eta(v^\star)| \approx |Z_0(v^\star) + \Delta_\eta/\sigma_\eta|$. For a one-sided exceedance, $\Pr(|Z_\eta| > z_{1-\delta}) \ge 1 - \Phi(z_{1-\delta} - |\Delta_\eta|/\sigma_\eta)$, yielding the stated bound.

Table 2: Dataset statistics

Dataset	Nodes	Edges	Features	Classes
Cora	2,708	5,429	1,443	7
PubMed	19,717	44,338	500	3
Flickr	89,250	899,756	500	7
OGB-Arxiv	169,343	1,116,243	128	40

E Experimental Details

Datasets

We use four public benchmarks covering small/medium/large graphs: **Cora** and **PubMed** (citation networks for semi-supervised node classification), **Flickr** (an image-related social graph with higher sparsity/heterophily), and **OGB-Arxiv** (a large-scale citation network); dataset statistics are in Table 2. Unless otherwise noted, we adopt an 80/20 train/test split; within training, 10% of nodes are reserved as a clean calibration set for CSP's conformal procedures. For refined detection, positives are random d-hop BFS subgraphs with seed rate α=0.2 and depth d=5, and negatives are small connected triggers synthesized from SW/PA priors with node budget B=10. All node features inside synthesized triggers are overwritten to constants to enforce structure-only learning.

Baselines

Attack baselines. We evaluate three representative backdoor attacks: (i) GTA (adaptive triggers using features and topology), (ii) UGBA (imperceptible triggers for stealth), and (iii) DPGBA (distribution-preserving triggers to reduce OOD effects). Backdoor injection strictly follows each original setting (target label, poison rate, and trigger size).

Defense baselines. We compare against: Prune (cosine-similarity edge pruning), Prune+LD (edge pruning plus dropping labels connected to low-similarity edges), RIGBD (random edge dropping with robust training), and DShield (self-supervised contrastive pretraining with discrepancy-based purification). Our method **TCF** uses CSP for screening (Chebyshev order r=4, Bonferroni aggregation with clean-FPR target δ=0.03, candidate budget ρ=5% in the budgeted variant), followed by a structure-only GCN detector (2 layers, 16 hidden units, 300 epochs, learning rate 0.01) and label-flip verification. All results are averaged over five runs with different seeds.

Compute Resources

Experiments are conducted on a workstation running **Ubuntu 20.04** with **2× NVIDIA RTX 3090** GPUs and **64 GB** RAM. We use PyTorch/pyG implementations with mixed precision enabled where applicable.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our contributions and conclusions in the Abstract and Introduction, and restate the main findings in the Discussion section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the method's scope and limitations (e.g., applicable threat models, tuning/assumption sensitivities, and failure modes) in the Limitations/Discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical statements list their assumptions and are accompanied by proofs or proof sketches; details and auxiliary lemmas are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We report datasets, baselines, hyperparameters, train/validation/test splits, and evaluation protocols in Experiments and provide further configuration details in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: To preserve double-blind review, we do not include a public link in the anonymized submission; we will release code (and scripts to obtain datasets) upon acceptance, consistent with conference policy.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Training settings (optimizer, learning rate, epochs), model/backbone choices, search budgets, and key thresholds are described in Experiments with complete values listed in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean ± 95% confidence intervals over n=5 independent seeds (varying initialization, split sampling, and attack sampling), computed as Student-t CIs (df=4) and used consistently for all ASR/ACC tables.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support th e main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We cover this topic in the Compute Resources section of the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We use publicly available academic datasets and standard research baselines and do not involve human subjects or sensitive data; we follow NeurIPS ethics guidelines and discuss potential misuse in Discussion.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: We do not include a dedicated broader-impacts section in the anonymized draft; we will add a concise statement on potential societal implications and safeguards in the camera-ready.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release a high-risk foundation model or a newly collected dataset; the work proposes a research method and evaluation procedure without assets requiring gated access.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: While we rely on standard public datasets/code, we do not enumerate licenses and versions in the paper; we will document sources and licenses in the repository README and supplemental material.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce a new dataset or pretrained model in this submission; if code is released, it will include a license and environment specification.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or human-subject studies are conducted in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No IRB approval is required because the work does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not used as integral or non-standard components of the proposed method or evaluation pipeline; any incidental writing assistance does not affect technical content.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.