GRACE: GAUSSIAN PROTOTYPICAL CONSENSUS WITH PERTURBATION-AWARE ADAPTATION FOR TEST-TIME PROMPT TUNING IN GRAPH FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the problem of test-time prompt tuning for graph foundation models (GFMs), which aims to improve GFMs at test time by optimizing prompts. Existing approaches typically learn prompts from a small number of support data to facilitate GFM adaptation. Despite recent progress, the performance of these approaches is far from satisfactory due to overconfident pseudo-labels of target nodes and thus unstable prompt optimization. Towards this end, we propose a novel approach named Gaussian Prototypical Consensus with Perturbation-aware Adaptation (**GRACE**) for test-time prompt tuning in GFMs. The core of our **GRACE** is to improve the stability of test-time prompt tuning from two complementary perspectives, i.e., *probabilistic prototype measurement* and *selective perturbation-aware adaptation*. In particular, our **GRACE** first estimates Gaussian prototypes from support nodes to characterize the representation spaces across different layers in GFMs. More importantly, we aggregate covariance-aware Mahalanobis distances between each target node and its corresponding prototypes via majority voting, thereby filtering out unreliable pseudo-labels. Furthermore, we select target nodes with strong prototypical distance stability under graph perturbations and estimate prompt update directions through state-consistent symmetric perturbations, coupling reliability probing and prompt adaptation within a unified perturbation-aware procedure. Experiments on benchmark datasets validate the effectiveness of **GRACE** against extensive baselines. Our code is available at https://anonymous.4open.science/r/GRACE-03D3.

1 INTRODUCTION

Graph foundation models (GFMs) aim to transfer reusable graph representations across tasks and domains, reducing the need to train a separate encoder for each downstream graph (Mao et al., 2024; Liu et al., 2025; Zhao et al., 2024; Wang et al., 2025; Yuan et al., 2025). We study *few-shot test-time adaptation for node classification* on an unseen target domain, where a pre-trained GFM remains frozen and adaptation uses only a small labeled support set together with the unlabeled target graph (Jiang et al., 2026a; Zhang et al., 2025). The central challenge is to exploit target-domain evidence without retraining the encoder or accessing target-test labels at any stage.

Existing work provides lightweight interfaces for adapting frozen graph representations. Graph prompting introduces trainable target-side modules (Sun et al., 2022; Liu et al., 2023; Fang et al., 2023), while multi-domain GFMs improve transferability through cross-domain pre-training and transferable prompts (Yu et al., 2025; Wang et al., 2025). GFMate more directly adapts centroid and layer prompts using labeled and unlabeled target nodes (Jiang et al., 2026a). In parallel, test-time adaptation exploits uncertainty, augmentation consistency, and target-side reliability cues under distribution shift (Wang et al., 2021; Zhang et al., 2022; Shu et al., 2022).

Despite these advances, two coupled limitations remain in few-shot GFM adaptation: ❶ ***Failure to explicitly characterize class-conditional distributions.*** Prototype-based adaptation typically represents each class by a point estimate, which captures a center but not class-specific dispersion or anisotropic geometry (Snell et al., 2017; Lee et al., 2018; Turazza et al., 2026). On an unseen target graph, nodes at similar centroid distances can therefore have different compatibility with the

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.

target class distribution. The problem is compounded across GFM layers, which encode different neighborhood ranges and structural semantics (Xu et al., 2018; Chen et al., 2025): a confident fused pseudo-label need not be supported by every layer. Updating all layers indiscriminately can then contaminate the class statistics used by later predictions. Reliable adaptation thus requires both explicit class-conditional geometry and selective sample–layer updates. ❷ ***Perturbation instability under evolving target evidence.*** Graph perturbations can reveal target-node reliability, but prompt updates become difficult to interpret when the supervision state changes across perturbation queries. If pseudo-labels, statistics, sample weights, or graph views drift between paired prompt perturbations, the objective difference conflates prompt sensitivity with a changing supervision state. Reliable perturbation-aware adaptation therefore requires support-calibrated reliability probing together with state-consistent prompt updates. Figure 1 summarizes both of these gaps in current approaches.

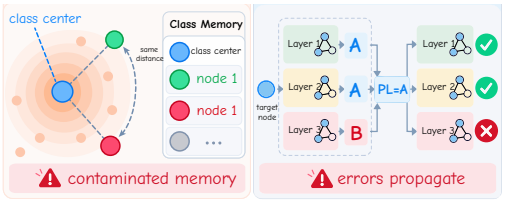


Figure 1: Layer-wise Gaussian class geometry and selective inter-layer commitment expose information hidden by point prototypes and fused confidence before perturbation-aware target adaptation.

To address these issues, we introduce Gaussian Prototypical Consensus with Perturbation-aware Adaptation for Test-time Prompt Tuning in Graph Foundation Models (`GRACE`). `GRACE` maintains layer-wise Gaussian prototype memory through class-specific first and second raw moments, explicitly modeling target-class location and dispersion in each frozen representation space. Layer-wise predictions are probabilistically fused, while perturbed labeled supports provide virtual references for estimating pseudo-label reliability. Consensus-guided selective updating allows an unlabeled node to refine only layers that agree with the fused pseudo-label, limiting cross-layer contamination. For optimization, `GRACE` separates target-evidence construction from prompt updates: after refreshing Gaussian memory and sample–layer supervision, it freezes the adaptation state within each paired query and updates only lightweight post-encoder prompts through symmetric perturbation-based function evaluations (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017). The resulting difference reflects prompt variation under fixed supervision, coupling graph-view reliability probing with state-consistent prompt adaptation under probabilistic inter-layer reliability controls. Our contributions are:

❶ ***New Perspective***: We formulate few-shot GFM adaptation around two coupled requirements: explicitly characterizing class-conditional target distributions across frozen representation layers and coordinating reliability-aware graph perturbations with state-consistent prompt perturbations.

❷ ***Novel Methodology***: We introduce `GRACE`, which combines layer-wise Gaussian first/second-moment prototype memory, probabilistic layer fusion, virtual-reference confidence calibration, and exact sample–layer consensus with perturbation-aware optimization of post-encoder prompts under a target adaptation state frozen within each paired query throughout test-time prompt optimization.

❸ ***Extensive Experiments.*** We conduct extensive experiments on a diverse set of benchmark datasets, showing that `GRACE` is: ***(i) robust***, consistently outperforming strong baselines across diverse settings, and ***(ii) well-grounded***, with comprehensive analyses validating its adaptation components.

2 Related Work

Graph Foundation Models. Graph foundation models learn reusable graph representations for transfer across tasks and domains (Shi et al., 2024; Mao et al., 2024; Liu et al., 2025). Graph prompting provides lightweight adaptation interfaces, including task reformulation and learnable feature or representation prompts (Sun et al., 2022; Liu et al., 2023; Fang et al., 2023; Sun et al., 2023; Chen et al., 2025). Multi-domain GFMs further transfer knowledge through cross-domain coordination, structural alignment, and transferable prompting, as exemplified by GCOPE, MDGPT, SAMGPT, MDGFM, BRIDGE, and RiemannGFM (Zhao et al., 2024; Yu et al., 2024; 2025; Wang et al., 2025; Yuan et al., 2025; Sun et al., 2025). Closest to our setting, GFMate tunes post-encoder centroid and layer prompts using few-shot labels and unlabeled target data (Jiang et al., 2026a). Its centroid-based class representation and learned layer weighting provide a target-side adaptation interface. Our distinction is to maintain class-conditioned first and second moments in each layer and to control which unlabeled sample–layer pairs can refine them. Unlike global layer weighting alone,

this selective feedback regulates target-evidence assimilation at the level where class statistics are maintained during every target-memory refinement step at test time consistently.

Test-time Adaptation Methods. Test-time adaptation uses target inputs to adapt to distribution shift without retraining on source data. Representative approaches use entropy minimization, augmentation consistency, or uncertainty-aware objectives (Wang et al., 2021; Zhang et al., 2022; 2025). Graph-specific methods address structural shifts through test-time neighborhood augmentation or structure transformation, including GraphPatcher and GrapHoST (Ju et al., 2023; Jiang et al., 2026b). Perturbation-based consistency can further expose prediction instability under distribution shift (Zhang et al., 2022). In our setting, graph perturbations and prompt perturbations play complementary roles: the former calibrate target-node reliability against labeled supports, whereas the latter update post-encoder prompts under a target-evidence state fixed across each symmetric pair. This coupling connects evolving target statistics to state-consistent perturbation-aware adaptation during few-shot GFM transfer on previously unseen target graphs at test time.

3 The Proposed GRACE

Preliminaries. We consider transductive few-shot node classification on an unseen target graph using a frozen GFM. Let $\mathcal{G} = (V, E, \boldsymbol{X})$ contain C target classes, $\mathcal{S} \subset V$ be the labeled support set, and $\mathcal{U} = V \setminus \mathcal{S}$ contain the remaining nodes. Each class has $K \geq 1$ support nodes, $K = 1$ is the default one-shot setting. Only support labels enter adaptation. Labels of nodes in $\mathcal{U}$ are hidden during adaptation. Target node features and graph structure remain visible throughout transductive adaptation. For selected encoder layers $l = 1, \ldots, L$, let $\boldsymbol{h}_i^{(l)} \in \mathbb{R}^{d_l}$ be the bounded, normalized representation of node i. An input perturbation $T_\xi^{(\sigma_p)}$ applies edge dropout and feature masking at rate σ_p while preserving node identities (Rong et al., 2020; You et al., 2020; Hou et al., 2022).

3.1 Framework Overview

Adapting a frozen GFM to an unseen graph from only a handful of labeled nodes is challenging for two coupled reasons. First, class geometry can vary across graphs and encoder layers, making predictions from any single representation space unreliable on an unseen target graph. Second, once erroneous pseudo-labels are fed back into adaptation, their noise can be amplified and steer the learned prompts away from the target semantics. These observations motivate a target-adaptation framework that jointly refines class-level geometric memory and lightweight prompts, while keeping the pretrained encoder fixed. As Figure 2 shows, GRACE alternates between refining a probabilistic target-class memory and optimizing lightweight prompts against the resulting supervision. The first component (Section 3.2) maintains layer-wise Gaussian prototypes, fuses predictions, and updates only layers agreeing with the fused pseudo-label. The second (Section 3.4) uses perturbed labeled references to filter candidates and paired prompt-perturbation queries to optimize prompts. Each round computes predictions and weights, updates Gaussian moments with clean admitted features, and freezes evidence for prompt queries.

3.2 Layer-wise Covariance-aware Distances from Gaussian Prototypes

Reliable pseudo-labels are crucial for effective model adaptation, yet predictions derived from a single representation layer can be unreliable due to layer-specific noise and semantic bias. To obtain more trustworthy supervision, we exploit the complementary semantic evidence encoded across different layers and seek cross-layer consensus through class-wise prototypes. To capture layer-specific class geometry under scarce target labels, we model each class at each selected layer with an identity-regularized Gaussian prototype rather than a single centroid, providing covariance-aware evidence for subsequent cross-layer consensus. For class c and layer l, the Gaussian Prototype Memory (GPM) maintains a first raw moment $\boldsymbol{\mu}_c^{(l)}$ and a second raw moment $\boldsymbol{Q}_c^{(l)}$. With $\mathcal{S}_c = \{s \in \mathcal{S} : y_s = c\}$, we initialize them from clean support representations as follows:

$$\boldsymbol{\mu}_c^{(l)} = \frac{1}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} \boldsymbol{h}_s^{(l)}, \qquad \boldsymbol{Q}_c^{(l)} = \frac{1}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} \boldsymbol{h}_s^{(l)} \boldsymbol{h}_s^{(l)\top}. \tag{1}$$

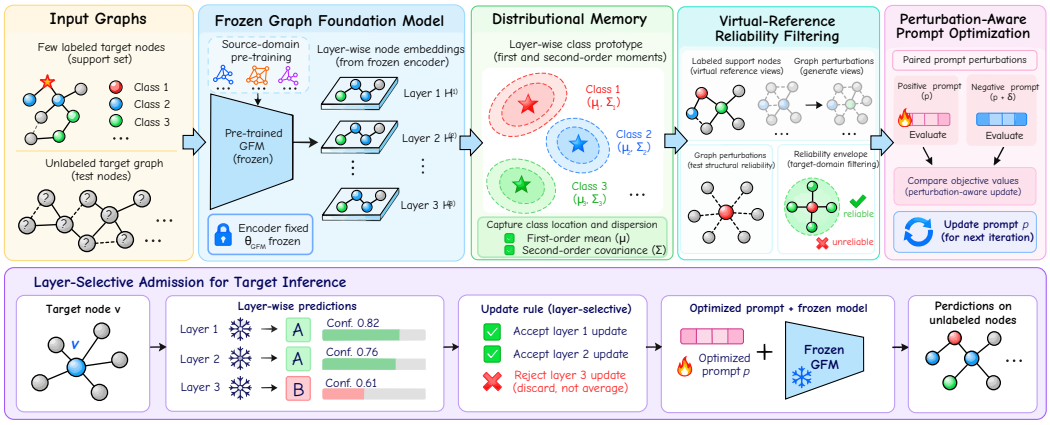


Figure 2: Frozen GFM features form layer-wise Gaussian class states, refined by perturbation-based calibration and inter-layer agreement for reliable pseudo-labeling and perturbation-aware prompt adaptation on an unseen target graph without updating the frozen encoder.

The corresponding identity-regularized class covariance is

$$\boldsymbol{\Sigma}_c^{(l)} = \boldsymbol{Q}_c^{(l)} - \boldsymbol{\mu}_c^{(l)} \boldsymbol{\mu}_c^{(l)\top} + \boldsymbol{I}, \tag{2}$$

where the identity term serves as an isotropic shrinkage prior, regularizing unreliable few-shot second-order estimates while allowing admitted target evidence to progressively refine the class geometry. The pair $(\boldsymbol{\mu}_c^{(l)}, \boldsymbol{\Sigma}_c^{(l)})$ defines the Gaussian class memory $\mathcal{N}(\boldsymbol{\mu}_c^{(l)}, \boldsymbol{\Sigma}_c^{(l)})$. In the one-shot setting, covariance starts from the identity geometry, selected target evidence enriches it through the moment updates in Section 3.3 as reliable target evidence accumulates over rounds. A learnable centroid prompt $\boldsymbol{b}_c^{(l)}$ adjusts the prototype location to $\boldsymbol{\phi}_c^{(l)} = \boldsymbol{\mu}_c^{(l)} + \boldsymbol{b}_c^{(l)}$. The squared Mahalanobis distance from node i to this prompted prototype is defined as follows:

$$d_{ic}^{(l)} = \left(\boldsymbol{h}_i^{(l)} - \boldsymbol{\phi}_c^{(l)}\right)^\top \left(\boldsymbol{\Sigma}_c^{(l)}\right)^{-1} \left(\boldsymbol{h}_i^{(l)} - \boldsymbol{\phi}_c^{(l)}\right). \tag{3}$$

This distance scales deviations by the identity-regularized class covariance, assigning greater penalties to deviations along low-variance directions. We combine it with directional similarity to obtain the class score

$$s_{ic}^{(l)} = \cos(\boldsymbol{h}_i^{(l)}, \boldsymbol{\phi}_c^{(l)}) - \frac{\lambda_{\text{fuse}}}{2d_l} \left(d_{ic}^{(l)} + \log\det \boldsymbol{\Sigma}_c^{(l)}\right). \tag{4}$$

The cosine term captures angular alignment, while the Gaussian term incorporates covariance-scaled proximity and class-distribution volume (Lee et al., 2018). The factor d_l normalizes the Gaussian contribution by layer dimension, and $\lambda_{\text{fuse}} \geq 0$ controls its strength. We evaluate the quadratic form and log-determinant through Cholesky factorization and triangular solves. Normalizing the scores yields the layer-wise distribution $\boldsymbol{p}_i^{(l)} = \text{softmax}(\boldsymbol{s}_i^{(l)})$ used for cross-layer consensus.

3.3 Cross-Layer Prototypical Consensus for Reliable Pseudo-Labeling

To combine complementary layer-wise evidence without mixing inconsistent feature geometries, we fuse the layer-wise class distributions in their shared probability space through a learnable layer prompt $\boldsymbol{a} \in \mathbb{R}^L$, while retaining layer-specific Gaussian prototypes throughout adaptation:

$$\boldsymbol{p}_i = \sum_{l=1}^{L} \omega_l \boldsymbol{p}_i^{(l)}, \qquad \boldsymbol{\omega} = \text{softmax}(\boldsymbol{a}). \tag{5}$$

Fusion takes place in the shared class-probability space, while each Gaussian prototype retains its layer-specific feature geometry. For an unlabeled node, let $\widehat{y}_i = \arg\max_c p_{ic}$ be its clean-view fused pseudo-label and $\widehat{y}_i^{(l)} = \arg\max_c p_{ic}^{(l)}$ its layer-wise prediction. Given a candidate admission decision $g_i \in \{0, 1\}$ from Section 3.4, we define the following sample–layer gate

$$a_i^{(l)} = \mathbf{1}\{\widehat{y}_i^{(l)} = \widehat{y}_i\}, \qquad w_i^{(l)} = g_i \, p_{i\widehat{y}_i} \, a_i^{(l)}. \tag{6}$$

Here, g_i selects nodes according to perturbation-based reliability, $a_i^{(l)}$ assigns their evidence to agreeing layers, and $p_{i\widehat{y}_i}$ weights their contribution by fused confidence. The resulting weights jointly govern Gaussian-memory refinement and unlabeled prompt supervision. To refine the memory, group candidates by their fused pseudo-label as $\mathcal{U}_c = \{i \in \mathcal{U} : \widehat{y}_i = c\}$ and define the admitted mass $m_c^{(l)} = \sum_{i\in\mathcal{U}_c} w_i^{(l)}$. We maintain an effective mass $n_c^{(l)}$ together with the two raw moments, initialized by $n_c^{(l)} = |\mathcal{S}_c|$. Once candidate admission has been evaluated at the current adaptation round, the round-specific admitted evidence is accumulated into the persistent sufficient statistics through the following weighted update:

$$\boldsymbol{\mu}_c^{(l)+} = \frac{n_c^{(l)}\boldsymbol{\mu}_c^{(l)} + \sum_{i\in\mathcal{U}_c} w_i^{(l)}\boldsymbol{h}_i^{(l)}}{n_c^{(l)} + m_c^{(l)}}, \qquad \boldsymbol{Q}_c^{(l)+} = \frac{n_c^{(l)}\boldsymbol{Q}_c^{(l)} + \sum_{i\in\mathcal{U}_c} w_i^{(l)}\boldsymbol{h}_i^{(l)}\boldsymbol{h}_i^{(l)\top}}{n_c^{(l)} + m_c^{(l)}}. \tag{7}$$

The corresponding effective mass is updated as $n_c^{(l)+} = n_c^{(l)} + m_c^{(l)}$. The covariance is then reconstructed by Equation (2), pairs with zero admitted mass retain their existing statistics. This sufficient-statistic update lets reliable target evidence determine the memory update while tracking class location and dispersion. Consensus-selected evidence progressively shapes each layer's class geometry without an additional memory-mixing coefficient during adaptation.

3.4 Perturbation-aware Selective Prompt Tuning

Cross-layer consensus improves pseudo-label quality, but it does not by itself identify which target nodes are reliable enough to drive adaptation on an unseen graph. We therefore make adaptation explicitly perturbation-aware: labeled supports provide target-domain reference anchors for class-conditioned reliability filtering, which then supervises paired prompt-perturbation updates of the centroid and layer prompts while the GFM encoder remains frozen.

Perturbation-aware Data Selection. To identify trustworthy target evidence, we test whether each candidate preserves its assigned class geometry and prediction under mild graph perturbations, referencing these responses to labeled supports from the same target graph rather than relying on global hand-crafted thresholds. During graph probing, the current memory, prompts, and clean-view pseudo-label $\widehat{y}_i$ are held fixed. For the perturbed representation $\boldsymbol{h}_i^{(l)}(\xi)$, let $d_{ic}^{(l)}(\xi)$ denote Equation (3) evaluated on that view, and define $D_i^{(l)}(\xi) = d_{i\widehat{y}_i}^{(l)}(\xi)/d_l$. The candidate's distance response is

$$R_i = \mathbb{E}_\xi\left[\sum_{l=1}^{L} \omega_l \big(D_i^{(l)}(\xi) - D_i^{(l)}(0)\big)^2\right]. \tag{8}$$

We also measure its prediction flip frequency as $F_i = \Pr_\xi[\arg\max_c p_{ic}(\xi) \neq \widehat{y}_i]$. Both statistics are estimated from graph views generated by edge dropout and feature masking at rate σ_p. The response R_i measures changes in covariance-normalized class distance, and F_i measures changes in the assigned class. Applying the same graph perturbations to labeled support nodes produces *virtual references*. For each support node s, we compute the response and flip statistics relative to its known class y_s. The class-c reference bank defines a reliability envelope: τ_c^R is its largest labeled response, τ_c^F its largest labeled flip frequency, and τ_c^Q its smallest labeled true-class probability. A candidate is admitted when its response, flip frequency, and clean-view confidence satisfy

$$g_i = \mathbf{1}\{R_i \le \tau_{\widehat{y}_i}^R\}\,\mathbf{1}\{F_i \le \tau_{\widehat{y}_i}^F\}\,\mathbf{1}\{p_{i\widehat{y}_i} \ge \tau_{\widehat{y}_i}^Q\}. \tag{9}$$

This perturbation-aware confidence filtering (PaCF) derives class-specific selection criteria from labeled target behavior. Its three criteria couple geometric stability, predictive consistency, and class confidence. The admission decision g_i is combined with cross-layer agreement in Equation (6), after which the accepted clean features refine the Gaussian sufficient statistics through Equation (7).

Perturbation-aware Prompt Tuning. Given the selected target evidence, we adapt only the centroid prompts $\{\boldsymbol{b}_c^{(l)}\}$ and layer prompt $\boldsymbol{a}$, using symmetric prompt perturbations to estimate state-consistent normalized-coordinate update direction for class geometry and cross-layer aggregation. Let $\boldsymbol{\theta}$ collect these prompts. At round t, the refreshed moments, clean-graph representations, pseudo-labels, and complete weights $w_i^{(l)}$ are fixed during the paired evaluations. In scale-normalized prompt coordinates, we sample $\boldsymbol{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_\theta})$, where d_θ is the prompt dimension. Reliability probing uses

perturbed graph views, whereas both prompt queries reuse cached clean-graph representations and the same fixed supervision. The resulting normalized-coordinate Gaussian-smoothing zeroth-order update (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017) therefore isolates the objective response to prompt variation under the selected target evidence.

3.5 SUMMARIZATION

The complete optimization objective of **GRACE** at adaptation round t is

$$\mathcal{L}^{(t)}_{\mathrm{GRACE}}(\boldsymbol{\theta}) = -\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\log p_{s,y_s}(\boldsymbol{\theta}) - \frac{1}{W_t}\sum_{i\in\mathcal{U}}\sum_{l=1}^{L} w^{(l)}_{i,t}\log p^{(l)}_{i,\widehat{y}_{i,t}}(\boldsymbol{\theta}), \tag{10}$$

where $p_{s,y_s}(\boldsymbol{\theta}) = \sum_{l=1}^{L}\omega_l(\boldsymbol{\theta})p^{(l)}_{s,y_s}(\boldsymbol{\theta})$ is the fused support probability, $w^{(l)}_{i,t} = g_{i,t}\,p_{i\widehat{y}_{i,t}}\mathbf{1}\{\widehat{y}^{(l)}_{i,t} = \widehat{y}_{i,t}\}$ is the reliability- and consensus-aware pseudo-label weight, and $W_t = \sum_{i,l} w^{(l)}_{i,t}$ (with the second term set to zero when $W_t = 0$). Thus, the first term anchors adaptation to labeled target supports, whereas the second transfers only perturbation-stable, confidence-weighted, and cross-layer-consistent target evidence. The Gaussian prototype update and PaCF do not introduce additional additive losses; they determine the round-specific memory, pseudo-labels, and weights, all of which are fixed during each paired prompt query. Perturbation-aware prompt tuning then performs a normalized-coordinate symmetric perturbation update by evaluating the same objective,

$$F^{\pm}_t = \mathcal{L}^{(t)}_{\mathrm{GRACE}}(\boldsymbol{\theta}_t \pm \boldsymbol{u}_t), \qquad \widehat{\boldsymbol{g}}_t = \frac{F^{+}_t - F^{-}_t}{2}\boldsymbol{u}_t, \qquad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \widehat{\boldsymbol{g}}_t, \tag{11}$$

which gives the complete loss-driven adaptation rule while leaving the GFM encoder frozen.

3.6 THEORETICAL ANALYSIS

In this part, we analyze the update behavior of the Gaussian prototype memory (GPM). For a fixed class-layer pair (c, l), let $W_i = g_i p_{i\hat{y}_i} a^{(l)}_i$ for nodes with $\hat{y}_i = c$, let $m = \sum_i W_i$ be the admitted mass, and let n be the current effective mass from Equation (7). For $m > 0$, define $\widetilde{W}_i = W_i/m$, $\bar{\boldsymbol{h}} = \sum_i \widetilde{W}_i \boldsymbol{h}^{(l)}_i$, and $\overline{\boldsymbol{h}\boldsymbol{h}^\top} = \sum_i \widetilde{W}_i \boldsymbol{h}^{(l)}_i \boldsymbol{h}^{(l)\top}_i$.

$$\boldsymbol{\mu}^{+} = \frac{n\boldsymbol{\mu} + m\bar{\boldsymbol{h}}}{n+m}, \qquad \boldsymbol{Q}^{+} = \frac{n\boldsymbol{Q} + m\overline{\boldsymbol{h}\boldsymbol{h}^\top}}{n+m}, \tag{12}$$

with $\boldsymbol{\Sigma} = \boldsymbol{Q} - \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{I}$. Let $P^{-} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\widehat{P}_W = \sum_{i:\hat{y}_i=c}\widetilde{W}_i\mathcal{N}(\boldsymbol{h}^{(l)}_i, \boldsymbol{I})$.

Proposition 3.1. *For $n > 0$ and $m > 0$, the Gaussian produced by (12), $P^{+} = \mathcal{N}\big(\boldsymbol{\mu}^{+}, \boldsymbol{Q}^{+} - \boldsymbol{\mu}^{+}\boldsymbol{\mu}^{+\top} + \boldsymbol{I}\big)$, is the unique solution of*

$$P^{+} = \arg\min_{P\in\mathcal{G}_d}\Big[n\,\mathrm{KL}(P^{-}\|P) + m\,\mathrm{KL}(\widehat{P}_W\|P)\Big], \tag{13}$$

where $\mathcal{G}_d$ denotes the family of non-degenerate Gaussian distributions on $\mathbb{R}^d$. Furthermore, the covariance of P^{+} is positive definite because of the fixed identity stabilizer.

Proposition 3.1 interprets the joint moment update as a KL-anchored Gaussian projection driven by the commitment-weighted empirical distribution $\widehat{P}_W$, motivating the next result on incorrect pseudo-label mass under reliability-controlled inter-layer commitment during target adaptation.

For the candidate set $\mathcal{C}_c = \{i : \hat{y}_i = c\}$, define $Z_i = \mathbf{1}\{y_i = c\}$. Let N_1 and N_0 denote the numbers of correct and incorrect candidates, respectively, and $\epsilon^{\mathrm{raw}}_c = N_0/(N_0 + N_1)$. For $z \in \{0, 1\}$, let $m^{(l)}_z$ denote the average commitment weight W_i over candidates with $Z_i = z$, and define $\rho_{c,l} = m^{(l)}_0/m^{(l)}_1$.

Proposition 3.2. *Suppose $m^{(l)}_1 > 0$ and $\rho_{c,l} < 1$. The normalized incorrect weight entering the GPM update is*

$$\epsilon^{\mathrm{adm}}_{c,l} = \frac{\rho_{c,l}\epsilon^{\mathrm{raw}}_c}{1 - \epsilon^{\mathrm{raw}}_c + \rho_{c,l}\epsilon^{\mathrm{raw}}_c}, \tag{14}$$

and therefore $\epsilon^{\mathrm{adm}}_{c,l} < \epsilon^{\mathrm{raw}}_c$ whenever $0 < \epsilon^{\mathrm{raw}}_c < 1$.

Table 1: One-shot node-classification accuracy (%). Values are mean ± standard deviation over five runs. Best and runner-up results in the cross-domain block are highlighted.

Method	Texas	Cornell	Wisconsin	Chameleon	Squirrel	Arxiv-year	Cora	CiteSeer	Photo
Single-domain Training and Testing									
GCN	38.82±9.79	24.58±10.09	43.36±13.17	24.30±6.59	19.96±5.89	18.64±6.91	29.85±8.98	33.39±11.86	47.09±5.81
GAT	39.96±9.62	25.84±12.26	45.99±10.86	22.81±7.38	20.77±4.48	19.93±5.40	33.25±9.72	35.51±9.70	47.33±4.97
SAGE	40.38±8.21	32.55±13.36	47.41±11.36	31.29±8.87	22.34±4.32	20.75±4.99	35.76±8.89	38.80±9.73	49.72±3.81
GPR	37.60±9.54	30.77±13.42	49.56±15.58	28.44±6.65	21.06±6.14	10.81±9.94	38.99±15.77	29.77±13.22	43.39±4.66
H2GCN	47.75±12.89	35.58±15.02	45.55±10.61	28.01±9.50	21.10±3.06	15.54±7.33	30.90±9.98	30.91±12.79	45.81±6.09
Self-supervised Learning + Fine-tuning									
LP+FT	36.93±11.90	23.88±11.43	41.37±14.22	25.34±7.19	20.04±5.61	17.94±6.06	35.59±9.74	34.92±12.08	48.82±6.55
DGI+FT	34.46±12.92	22.89±11.32	38.89±15.77	25.78±7.34	20.85±6.57	18.06±6.22	32.38±8.86	33.96±11.57	45.17±7.34
GCL+FT	28.81±15.32	20.56±13.36	35.70±17.73	24.69±8.82	19.97±5.62	15.53±8.89	33.27±9.50	36.05±13.44	47.33±5.89
Self-supervised Learning + Prompt Tuning									
GPPT	44.47±10.88	29.74±8.32	35.16±9.07	29.91±6.48	21.16±5.95	19.96±7.63	40.62±8.69	39.79±10.67	50.19±7.74
GPF	47.34±12.72	52.15±14.53	40.19±14.49	30.95±9.18	22.71±4.87	20.89±8.20	45.75±9.61	40.51±12.79	49.38±6.56
ProNoG	55.31±12.92	48.49±11.54	46.29±17.74	31.19±8.09	24.25±4.79	OOM	56.54±12.33	37.79±13.35	47.72±6.60
GraphPrompt	53.27±9.95	55.13±13.39	45.03±11.83	33.29±9.19	23.02±4.89	22.61±4.66	49.77±8.82	38.69±13.98	46.65±6.53
DAGPrompt	68.27±10.02	59.11±10.04	50.49±11.59	37.79±6.62	25.67±6.34	23.08±8.14	54.88±9.24	47.24±9.59	52.96±6.07
All-In-One	63.79±15.91	57.24±12.70	55.35±18.36	27.94±6.31	21.18±7.06	15.29±7.55	49.92±11.75	40.69±15.88	52.25±7.33
Multi-domain Pre-training and Cross-domain Fine-tuning then Testing									
ProNoG*	48.25±9.60	37.18±10.04	39.70±12.32	26.48±8.69	22.10±5.52	OOM	40.07±13.35	35.56±10.63	45.28±7.02
GCOPE	64.76±14.84	60.98±14.66	54.66±12.86	30.58±7.44	22.16±5.77	17.98±5.51	39.06±12.52	42.26±14.19	55.69±4.68
MDGPT	59.76±12.44	54.19±13.08	50.40±15.07	28.04±4.28	24.41±7.01	OOM	44.52±11.39	41.98±12.24	54.96±10.25
SAMGPT	66.79±10.77	59.34±9.82	52.29±14.40	38.12±8.90	25.75±6.29	OOM	52.83±12.04	47.76±10.55	56.33±9.04
MDGFM	–	40.77±5.96	–	28.36±3.65	24.30±3.26	–	44.83±7.41	42.18±6.41	–
BRIDGE	53.35±11.90	49.28±12.56	48.85±13.35	32.75±6.62	19.89±9.02	24.47±6.68	44.08±9.54	38.89±8.72	58.79±11.58
RiemannGFM	58.60±15.27	46.35±11.92	47.32±16.58	29.68±9.95	20.13±8.58	OOM	37.91±16.13	38.02±9.58	49.69±13.32
GFMate	76.63±7.81	79.67±8.47	63.01±10.78	47.25±6.11	27.02±6.22	30.19±3.65	59.68±5.37	56.25±13.33	58.85±2.17
GRACE (Ours)	**79.21**±5.06	**81.46**±6.65	**74.39**±8.39	**53.48**±5.89	**27.47**±5.37	**32.47**±2.92	**60.39**±4.54	**58.93**±8.66	**62.93**±1.63

Furthermore, assume $\|\boldsymbol{h}_i^{(l)}\|_2 \le B$. *Let* $(\boldsymbol{\mu}_t, \boldsymbol{Q}_t, \boldsymbol{\Sigma}_t)$ *denote the resulting GPM sequence generated by* (12)*, and* $(\boldsymbol{\mu}_t^\star, \boldsymbol{Q}_t^\star, \boldsymbol{\Sigma}_t^\star)$ *the coupled oracle sequence obtained from the same initialization and update after removing the incorrectly pseudo-labeled component and renormalizing at each step. If* m_s *is the admitted mass at step* s *and* n_t *is the effective mass after* t *updates, define*

$$\mathcal{E}_t = \frac{1}{n_t} \sum_{s=0}^{t-1} m_s \epsilon_{c,l,s}^{\mathrm{adm}}. \tag{15}$$

Then

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^\star\|_2 \le 2B\mathcal{E}_t, \qquad \|\boldsymbol{Q}_t - \boldsymbol{Q}_t^\star\|_{\mathrm{op}} \le 2B^2\mathcal{E}_t, \tag{16}$$

and

$$\|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t^\star\|_{\mathrm{op}} \le 6B^2\mathcal{E}_t. \tag{17}$$

Proposition 3.2 complements Proposition 3.1 by quantifying how commitment controls the erroneous component of the admitted target evidence. When commitment preferentially suppresses incorrect candidates (i.e., $\rho_{c,l} < 1$), the contamination entering the projection is contracted, and its accumulated influence on both first- and second-order GPM statistics is bounded by the admitted-mass-weighted term $\mathcal{E}_t$. Appendix B provides detailed proofs of both propositions and their associated bounds.

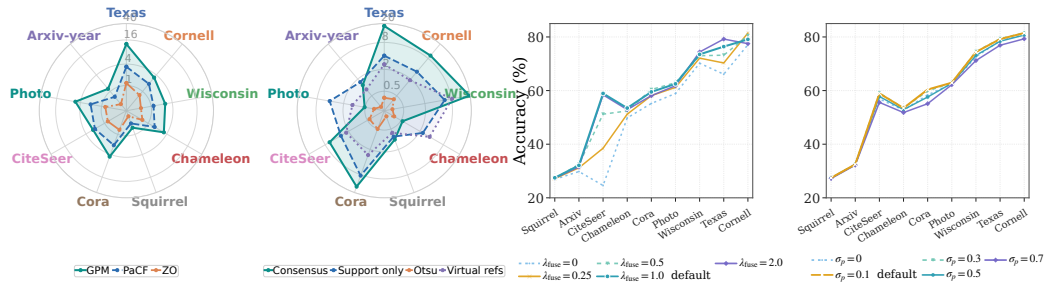
4 Experiments

4.1 Experiment Setup

We evaluate Texas, Cornell, Wisconsin (Pei et al., 2020), Chameleon and Squirrel (Rozemberczki et al., 2021), Cora and CiteSeer (Yang et al., 2016), Photo (Shchur et al., 2018), and arxiv-year derived from OGBN-Arxiv (Hu et al., 2020) under the transductive one-versus-all cross-domain protocol of GFMate (Jiang et al., 2026a). Each target is held out while the remaining source graphs are used for label-free GFM pre-training, the encoder is then frozen. We report mean ± standard deviation over five independent outer seeds. Each run uses an independently generated split with exactly K labeled nodes per class. Target features and structure are visible during adaptation, but test labels are never used for adaptation or model selection.

4.2 Empirical Performance

Performance Comparison. Table 1 reports the main one-shot node-classification results on nine held-out target graphs, from which we make three observations. ▷ **GRACE** *consistently improves the*



(a) Core mechanisms (b) Admission/supervision (c) Distribution fusion (d) Graph perturbation

Figure 3: Dataset-wise ablations and parameter analysis. The first two panels show core and admission/supervision ablations. (c)–(d) transpose the two parameter sweeps: the x-axis shows all nine target graphs ordered by their across-sweep mean accuracy, and each curve corresponds to one setting of distributional fusion weight λ_{fuse} or graph perturbation rate σ_p.

strongest cross-domain baseline across diverse target graphs. It raises the nine-dataset average from 55.39% with GFMate to 58.97%, an absolute gain of 3.58 points, and improves all nine targets. The largest gains occur on Wisconsin, Chameleon, and Photo, while improvements remain positive even on Squirrel, showing that explicit target-class distribution modeling and reliability-controlled adaptation transfer across different graph regimes. ▷ **GRACE** *achieves the strongest overall performance with improved run-to-run stability.* It obtains the highest mean accuracy on every target while reducing the average standard deviation from 7.10 to 5.46 relative to GFMate; the largest reduction is on CiteSeer, from 13.33 to 8.66. This indicates that the gains are not driven by a small subset of favorable runs or datasets. ▷ **GRACE** *remains effective across encoder architectures.* Table 2 shows consistently strong performance with different encoder. Full comparison are shown in Appendix C.1.

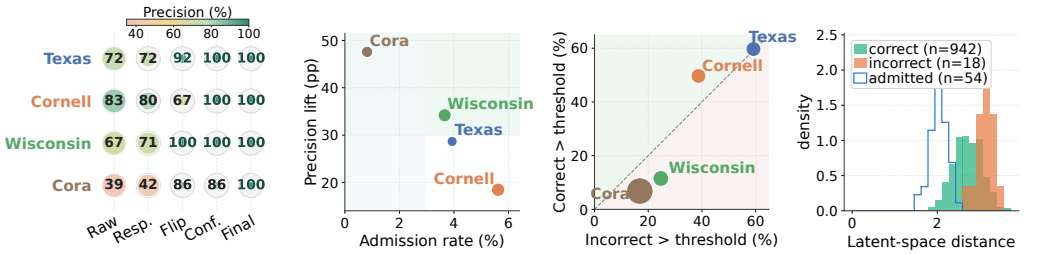
Table 2: Accuracy (%) of **GRACE** under the same target protocol across representative encoder backbones.

Backbone	Texas	Wisconsin	Cora	Photo
GCN	79.21±5.06	74.39±8.39	60.39±4.54	62.93±1.63
SAGE	72.40±7.13	72.10±6.74	59.20±4.96	61.50±1.76
GAT	74.10±6.11	72.80±6.40	60.80±4.54	62.10±1.63
H2GCN	80.30±4.07	76.20±5.05	60.10±4.96	61.80±1.76

Ablation Study. Starting from full **GRACE**, Figure 3 ablates one mechanism at a time. ▷ Gaussian prototype memory. Removing GPM reduces accuracy across targets, most notably on Texas, supporting the use of first- and second-order class statistics for few-shot target geometry. ▷ Perturbation-aware confidence filtering. Disabling PaCF consistently hurts performance, confirming the value of calibrating unlabeled evidence with labeled target anchors. ▷ Paired prompt perturbation. Replacing the symmetric perturbation update with Adam causes smaller but consistent drops, indicating additional gains from perturbation-aware prompt optimization. ▷ Layer consensus. Removing layer agreement is particularly harmful on challenging graphs, validating layer-specific commitment of fused pseudo-labels. ▷ Support-only learning. Using labeled supports alone underperforms the full model, demonstrating the benefit of selectively admitted unlabeled evidence. ▷ Global Otsu thresholding. Otsu does not improve mean accuracy over the full model, favoring support-referenced admission. ▷ Virtual references. Removing virtual support views degrades performance on challenging targets.

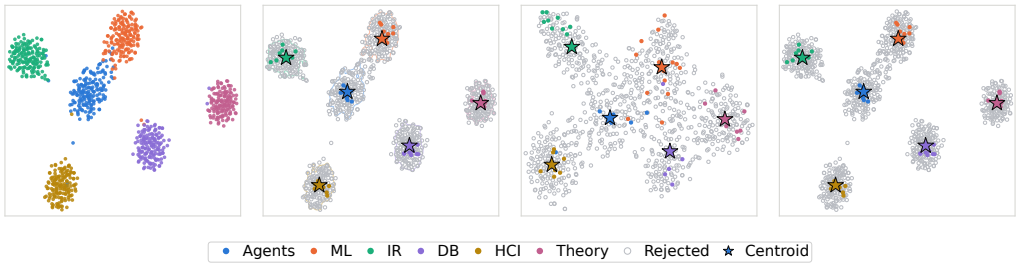
Parameter Analysis. Figure 3 studies the two explicit controls in **GRACE**. Panels (c)–(d) use a transposed sweep view: all nine target graphs, ordered by their across-sweep mean accuracy for readability, form the x-axis, while each curve is one parameter setting. The y-axis remains the unmodified absolute mean target accuracy; parameter sensitivity is therefore read from the vertical separation among settings at each target. ▷ Gaussian fusion weight λ_{fuse}. Removing the covariance-aware term ($\lambda_{\text{fuse}} = 0$) consistently degrades performance, while moderate nonzero weights generally remain effective, confirming the benefit of combining cosine similarity with class-distribution geometry. ▷ Graph perturbation rate σ_p. Without graph perturbation ($\sigma_p = 0$), the reliability probe cannot expose prediction or geometric instability, while moderate perturbations remain effective, overly strong perturbations move virtual-reference and candidate views too far from the clean graph and weaken local probes of target-domain reliability.

Selective Admission Diagnostics. We examine how reliability-aware filtering affects pseudo-label admission by progressively incorporating perturbation response, flip behavior, predictive confidence, and layer-wise consistency, and by analyzing the resulting purity–retention trade-off, selectivity gain,



(a) Purity retention (b) Selectivity gain (c) Response separation (d) Prototype distance

Figure 4: Admission diagnostics. *(a)* Purity–retention. *(b)* Selectivity–gain. *(c)* Perturbation-response discrimination. *(d)* Distance to the assigned prototype for admitted candidates.



(a) Adapted geometry (b) Adapted reliability (c) Frozen reliability (d) Admitted labels

Figure 5: Class-relative geometry and reliability diagnostics on CiteSeer. *(a)* Prompt-conditioned prototype-relative geometry over frozen post-encoder representations. *(b)* Reliability under the adapted prompt configuration. *(c)* The same admission rule under the unadapted prompt/prototype configuration. *(d)* Admitted pseudo-labels.

perturbation-response discrimination, and prototype distance. The results are shown in Figure 4. We observe that progressively combining these reliability signals yields a cleaner admitted subset with higher pseudo-label precision, while admitted candidates exhibit stronger perturbation stability and remain closer to their assigned class prototypes, indicating that the admission mechanism effectively filters out unreliable target samples.

Representation Geometry under Adaptation. We further study how prompt adaptation and selective admission reshape the class-relative structure within the frozen target representation space on CiteSeer by comparing the same frozen post-encoder representations under the unadapted and optimized prompt/prototype configurations, while keeping the encoder frozen, and by applying the same admission rule to both spaces. The results are shown in Figure 5, where ground-truth labels are used only for post-hoc visualization. We observe that under the optimized prompt/prototype configuration, admitted samples are more concentrated in class-coherent regions around their assigned prototypes, whereas the frozen geometry contains more erroneous candidates, suggesting that adaptation improves the representation geometry on which reliability is assessed and enables selective admission to focus updates on more trustworthy target evidence.

5 Conclusion

We presented **GRACE**, a reliability-controlled framework combining distributional prototype memory, support-referenced filtering, and perturbation-aware prompt tuning for few-shot graph adaptation. Graph perturbations calibrate the reliability of target evidence, while symmetric prompt perturbations update post-encoder prompts under a fixed supervision state. Across nine targets, **GRACE** raises average one-shot accuracy over GFMate from 55.39% to 58.97% while reducing the average standard deviation from 7.10 to 5.46. Gains remain positive across GCN, SAGE, GAT, and H2GCN on representative targets, and ablations consistently identify Gaussian memory and support-referenced admission as key contributors. The perturbation-aware design also lowers peak VRAM on Photo and arxiv-year while preserving the same four model evaluations per optimization step.

AI USE STATEMENT

AI assistance supported drafting, editing, and format checks.

REPRODUCIBILITY STATEMENT

Appendix A records the execution contract, and Appendix B gives the theoretical derivations. The experimental protocol uses five outer seeds, a persistent target split generated separately for each run, balanced K-shot-per-class support, and a 1:9 validation/test partition for the remaining target nodes in every run.

LIMITATIONS AND RESPONSIBLE USE

GRACE models target-class geometry from selected encoder-layer representations using class-conditioned first- and second-order statistics. In extremely small support regimes, the covariance floor regularizes estimation and keeps the Gaussian memory well conditioned. Deployment should include domain-specific evaluation for fairness, robustness, and safety.

REFERENCES

Qin Chen, Liang Wang, Bo Zheng, and Guojie Song. DAGPrompT: Pushing the limits of graph prompting with a distribution-aware graph prompt tuning approach. In *Proceedings of the ACM Web Conference 2025*, pp. 4346–4358, 2025. doi: 10.1145/3696410.3714917. URL https://doi.org/10.1145/3696410.3714917.

Taoran Fang, Yunchao Zhang, Yang Yang, Chunping Wang, and Lei Chen. Universal prompt tuning for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2285. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a4a1ee071ce0fe63b83bce507c9dc4d7-Abstract-Conference.html.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://doi.org/10.1137/120880811.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022. doi: 10.1145/3534678.3539321. URL https://doi.org/10.1145/3534678.3539321.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

Yan Jiang, Ruihong Qiu, and Zi Huang. GFMate: Empowering graph foundation models with test-time prompt tuning. *arXiv preprint arXiv:2605.14809*, 2026a. URL https://arxiv.org/abs/2605.14809. Accepted at ICML 2026.

Yan Jiang, Ruihong Qiu, and Zi Huang. Does homophily help in robust test-time node classification? In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 271–281, 2026b. doi: 10.1145/3773966.3777948. URL https://doi.org/10.1145/3773966.3777948.

Mingxuan Ju, Tong Zhao, Wenhao Yu, Neil Shah, and Yanfang Ye. GraphPatcher: Mitigating degree bias for graph neural networks via test-time augmentation. *arXiv preprint arXiv:2310.00800*, 2023. URL https://arxiv.org/abs/2310.00800.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. Graph foundation models: Concepts, opportunities and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):5023–5044, 2025. doi: 10.1109/TPAMI.2025.3548729. URL https://doi.org/10.1109/TPAMI.2025.3548729.

Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. GraphPrompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference 2023*, 2023. doi: 10.1145/3543507.3583386. URL https://doi.org/10.1145/3543507.3583386.

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34670–34692. PMLR, 2024. URL https://proceedings.mlr.press/v235/mao24a.html.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1e2agrFvS.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. DropEdge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkx1qkrKPr.

Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021. doi: 10.1093/comnet/cnab014. URL https://doi.org/10.1093/comnet/cnab014.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018. URL https://arxiv.org/abs/1811.05868.

Chuan Shi, Junze Chen, Jiawei Liu, and Cheng Yang. Graph foundation model. *Frontiers of Computer Science*, 18(6):186355, 2024. doi: 10.1007/s11704-024-40046-0. URL https://doi.org/10.1007/s11704-024-40046-0.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1038. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4077–4087, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Li Sun, Zhenhao Huang, Suyang Zhou, Qiqi Wan, Hao Peng, and Philip S. Yu. RiemannGFM: Learning a graph foundation model from riemannian geometry. In *Proceedings of the ACM Web Conference 2025*, pp. 1154–1165, 2025. doi: 10.1145/3696410.3714952. URL https://doi.org/10.1145/3696410.3714952.

Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. GPPT: Graph pre-training and prompt tuning to generalize graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1717–1727, 2022. doi: 10.1145/3534678.3539249. URL https://doi.org/10.1145/3534678.3539249.

Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, and Jihong Guan. All in one: Multi-task prompting for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2120–2131, 2023. doi: 10.1145/3580305.3599256. URL https://doi.org/10.1145/3580305.3599256.

Fabio Turazza, Marco Picone, and Marco Mamei. The gaussian-head ofl family: One-shot federated learning from client global statistics. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0da2c78fed88ad2d368994970a430c96-Abstract-Conference.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Shuo Wang, Bokui Wang, Zhixiang Shen, Boyan Deng, and Zhao Kang. Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64806–64821. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25dj.html.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5453–5462. PMLR, 2018. URL https://proceedings.mlr.press/v80/xu18c.html.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48. PMLR, 2016. URL https://proceedings.mlr.press/v48/yanga16.html.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-Abstract.html.

Xingtong Yu, Chang Zhou, Yuan Fang, and Xinming Zhang. Text-free multi-domain graph pre-training: Toward graph foundation models. *arXiv preprint arXiv:2405.13934*, 2024. URL https://arxiv.org/abs/2405.13934.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. SAMGPT: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM Web Conference 2025*, pp. 1142–1153, 2025. doi: 10.1145/3696410.3714828. URL https://doi.org/10.1145/3696410.3714828.

Haonan Yuan, Qingyun Sun, Junhua Shi, Xingcheng Fu, Bryan Hooi, Jianxin Li, and Philip S. Yu. How much can transfer? BRIDGE: Bounded multi-domain graph foundation model with generalization guarantees. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73604–73644. PMLR, 2025. URL https://proceedings.mlr.press/v267/yuan25h.html.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642, 2022. doi: 10.52202/068431-2799. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Haihong Zhao, Aochuan Chen, Xiangguo Sun, Hong Cheng, and Jia Li. All in one and one for all: A simple yet effective method towards cross-domain graph pretraining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4443–4454, 2024. doi: 10.1145/3637528.3671913. URL https://doi.org/10.1145/3637528.3671913.

A Implementation Contract

A.1 Target split and run contract

For each outer run r, an independent random split seed is generated and used to materialize a persistent `split_1`. The support index set $\mathcal{S}_r$ contains exactly K labeled nodes from every target class and is therefore class balanced. After those few-shot training nodes are removed, the remaining target nodes are randomly divided into validation and test partitions in a 1:9 ratio. Five runs yield the reported mean and standard deviation. Test-node features and graph structure remain visible during adaptation, test labels do not.

The GFM follows the one-versus-all cross-domain protocol: for each held-out target graph, the remaining source graphs are used for label-free pre-training. The default pre-training configuration is `pre_train_lr=1e-4`, `pre_train_wd=0`, `pre_train_epochs=10`, `gfm_dim=100`, `backbone=GCN`, and `pretrain_type=linkpred`. The target encoder stays frozen.

A.2 Execution order

Algorithm 1 `GRACE`: reliability-controlled perturbation-aware graph adaptation

Require: Target graph $\mathcal{G}$, balanced labeled support $\mathcal{S}$, frozen GFM encoder, distributional fusion control λ_{fuse} and graph perturbation control σ_p.

1: Initialize $\boldsymbol{\mu}_c^{(l)}$ and $\boldsymbol{Q}_c^{(l)}$ from labeled support features, initialize post-encoder prompts.
2: **for** $t = 1, \ldots, t_s$ **do**
3: Encode the clean graph and support/target views perturbed with strength σ_p.
4: Evaluate the distributional prototype scores in Equations (2)–(5).
5: Build class-wise virtual-reference stability envelopes from labeled support views.
6: Filter target candidates by Equation (9).
7: Apply Equation (6), discard each sample–layer pair with a conflicting pseudo-label.
8: Accumulate the current round's admitted, reliability-weighted evidence into the persistent first- and second-moment statistics using (7).
9: Freeze refreshed moments, pseudo-labels, complete weights $w_i^{(l)}$, and clean-graph features for the paired query.
10: Estimate $\widehat{\boldsymbol{g}}_t$ by Equation (11) and update the post-encoder prompts.
11: **end for**
12: **return** Clean-graph predictions from the final distributional memory and prompts.

A.3 Implementation details

Equation (9) computes the reliability envelope directly from labeled virtual references.

We place prompts after the frozen encoder so a fixed graph view is encoded once and shared by the positive/negative prompt queries. Graph perturbations and prompt-coordinate perturbations serve different roles: the former probe reliability, the latter estimate a state-consistent prompt update direction.

A.4 Perturbation-aware adaptation profile

With the distributional memory, pseudo-labels, complete sample–layer weights, and clean-graph features held fixed within each paired query, every adaptation step evaluates positive and negative prompt perturbations in normalized coordinates and updates only the post-encoder class/layer prompts. Graph perturbations are used to build support-calibrated reliability evidence, whereas prompt perturbations operate on cached clean-graph representations, fixing the evidence across

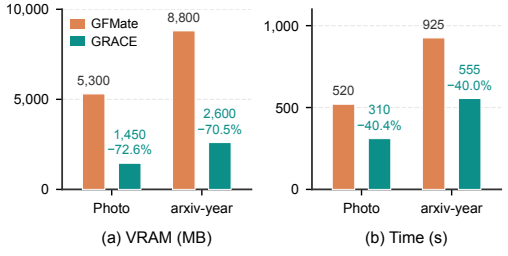


Figure 6: Peak VRAM and wall-clock time on Photo and arxiv-year, lower is better. Percentages show reductions from GFMate. Both methods use four model evaluations per step.

each pair isolates the objective response to the prompt perturbation.

Practical efficiency. Figure 6 shows that **GRACE** reduces peak VRAM from 5.30 to 1.45 GB on Photo and from 8.80 to 2.60 GB on arxiv-year, while shortening training from 520 to 310 s on Photo and from 925 to 555 s on arxiv-year. Both use four model evaluations per step, while **GRACE** incurs less overhead by operating on cached frozen-encoder representations during prompt perturbation updates.

B Proofs of Theoretical Results

Proof of Proposition 3.1. Let $M = \frac{n}{n+m}P^- + \frac{m}{n+m}\widehat{P}_W$. Multiplying an objective by the positive constant $n+m$ does not change its minimizer, so terms independent of P can be removed from (13), yielding $\arg\min_{P\in\mathcal{G}_d} -\mathbb{E}_{\boldsymbol{h}\sim M}[\log p(\boldsymbol{h})]$. Write $P = \mathcal{N}(\boldsymbol{r}, \boldsymbol{S})$, and denote by $\boldsymbol{\mu}_M$ and $\boldsymbol{C}_M$ the mean and covariance of M. Up to an additive constant, the objective is $\frac{1}{2}\{\log\det \boldsymbol{S} + \mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{C}_M) + (\boldsymbol{r}-\boldsymbol{\mu}_M)^\top \boldsymbol{S}^{-1}(\boldsymbol{r}-\boldsymbol{\mu}_M)\}$. For any $\boldsymbol{S} \succ 0$, the last term is uniquely minimized by $\boldsymbol{r} = \boldsymbol{\mu}_M$. Substituting $\boldsymbol{r} = \boldsymbol{\mu}_M$, the remaining objective is uniquely minimized at $\boldsymbol{S} = \boldsymbol{C}_M$.

Since $\mathbb{E}_{P^-}[\boldsymbol{h}] = \boldsymbol{\mu}$ and $\mathbb{E}_{P^-}[\boldsymbol{h}\boldsymbol{h}^\top] = \boldsymbol{Q} + \boldsymbol{I}$, while $\mathbb{E}_{\widehat{P}_W}[\boldsymbol{h}] = \bar{\boldsymbol{h}}$ and $\mathbb{E}_{\widehat{P}_W}[\boldsymbol{h}\boldsymbol{h}^\top] = \overline{\boldsymbol{h}\boldsymbol{h}^\top} + \boldsymbol{I}$, we obtain $\boldsymbol{\mu}_M = (n\boldsymbol{\mu} + m\bar{\boldsymbol{h}})/(n+m) = \boldsymbol{\mu}^+$ and $\mathbb{E}_M[\boldsymbol{h}\boldsymbol{h}^\top] = \boldsymbol{Q}^+ + \boldsymbol{I}$. Hence $\boldsymbol{C}_M = \boldsymbol{Q}^+ - \boldsymbol{\mu}^+\boldsymbol{\mu}^{+\top} + \boldsymbol{I}$, which proves the claimed form of P^+. The shared identity covariance also gives $\boldsymbol{C}_M \succeq \boldsymbol{I} \succ 0$. □

Proof of Proposition 3.2. By definition, the total incorrect and correct update masses are $N_0 m_0^{(l)}$ and $N_1 m_1^{(l)}$, respectively, thus the normalized incorrect weight is given by

$$\epsilon_{c,l}^{\mathrm{adm}} = \frac{N_0 m_0^{(l)}}{N_1 m_1^{(l)} + N_0 m_0^{(l)}}. \tag{18}$$

Using $N_0/(N_0+N_1) = \epsilon_c^{\mathrm{raw}}$ and $m_0^{(l)} = \rho_{c,l} m_1^{(l)}$ gives (14). Since $\rho_{c,l} < 1$, $\epsilon_{c,l}^{\mathrm{adm}} < \epsilon_c^{\mathrm{raw}}$.

Let $P_t^\star$ and P_t^e denote the normalized correct and incorrect components of the weighted admitted distribution. Then $P_t^{\mathrm{adm}} = (1-\epsilon_{c,l,t}^{\mathrm{adm}})P_t^\star + \epsilon_{c,l,t}^{\mathrm{adm}} P_t^e$. Because $\|\boldsymbol{h}\|_2 \le B$, both component means have norm at most B, and $\|\boldsymbol{h}\boldsymbol{h}^\top\|_{\mathrm{op}} \le B^2$, which yields the following moment bounds:

$$\|\bar{\boldsymbol{h}}_t - \bar{\boldsymbol{h}}_t^\star\|_2 \le 2B\epsilon_{c,l,t}^{\mathrm{adm}}, \qquad \|\overline{\boldsymbol{h}\boldsymbol{h}^\top}_t - \overline{\boldsymbol{h}\boldsymbol{h}^\top}_t^\star\|_{\mathrm{op}} \le 2B^2\epsilon_{c,l,t}^{\mathrm{adm}}. \tag{19}$$

Applying (12) to the actual and oracle sequences with their common support initialization and $n_{s+1} = n_s + m_s$ yields, after unrolling, the admitted-mass-weighted average in (15) and therefore (16). Finally, convex moment updates preserve $\|\boldsymbol{\mu}_t\|_2, \|\boldsymbol{\mu}_t^\star\|_2 \le B$, so

$$\|\boldsymbol{\mu}_t\boldsymbol{\mu}_t^\top - \boldsymbol{\mu}_t^\star\boldsymbol{\mu}_t^{\star\top}\|_{\mathrm{op}} \le (\|\boldsymbol{\mu}_t\|_2 + \|\boldsymbol{\mu}_t^\star\|_2)\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^\star\|_2 \le 4B^2\mathcal{E}_t. \tag{20}$$

The common covariance floor cancels when comparing $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_t^\star$, combining this inequality with (16) establishes the covariance bound in (17). □

C Additional Experimental Results

C.1 Complete backbone comparison

Table 3: Complete backbone comparison. Five-run mean ± standard deviation under the same target protocol. Each dataset block compares GFMate and **GRACE**, bold marks the better mean.

	Texas		Wisconsin		Cora		Photo	
Backbone	**GFMate**	**GRACE**	**GFMate**	**GRACE**	**GFMate**	**GRACE**	**GFMate**	**GRACE**
GCN	76.63±7.81	**79.21±5.06**	63.01±10.78	**74.39±8.39**	59.68±5.37	**60.39±4.54**	58.85±2.17	**62.93±1.63**
SAGE	65.80±9.51	**72.40±7.13**	65.40±11.79	**72.10±6.74**	58.10±5.78	**59.20±4.96**	57.20±2.31	**61.50±1.76**
GAT	68.90±8.49	**74.10±6.11**	66.70±11.12	**72.80±6.40**	59.90±5.37	**60.80±4.54**	57.90±2.17	**62.10±1.63**
H2GCN	78.60±5.43	**80.30±4.07**	70.80±8.76	**76.20±5.05**	59.30±5.37	**60.10±4.96**	57.60±2.31	**61.80±1.76**

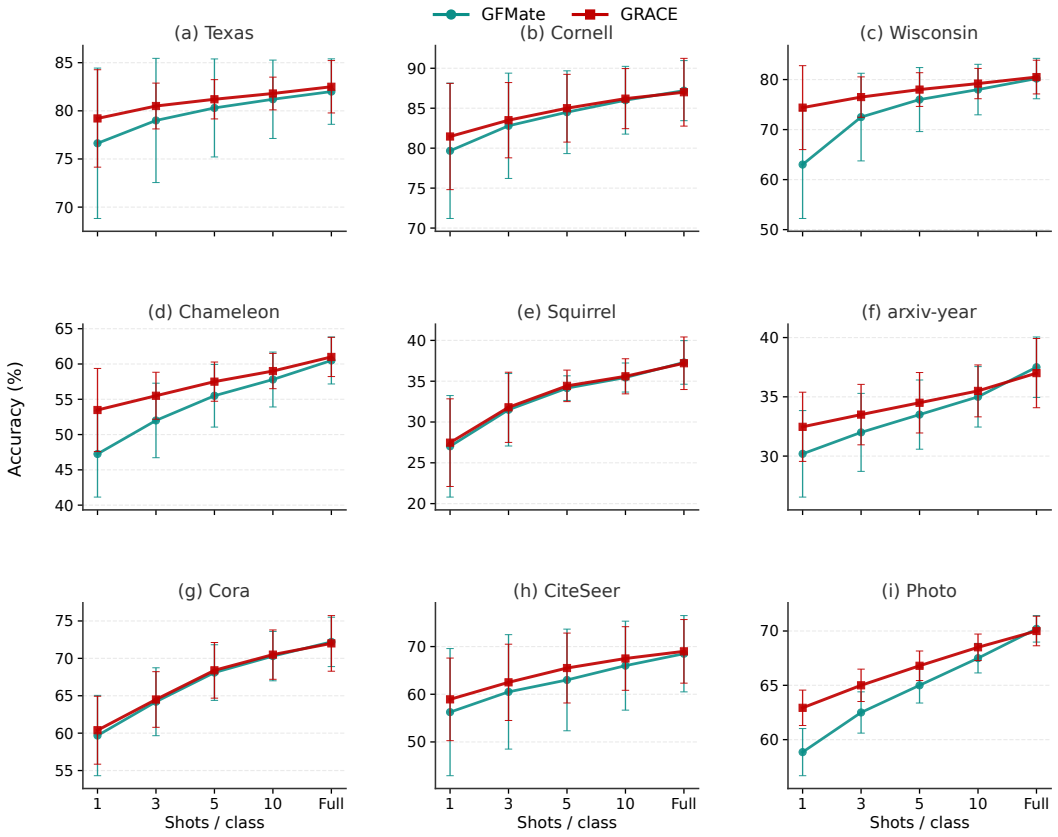


Figure 7: Label-budget comparison across nine target datasets. Panels (a)–(i) report node-classification accuracy as the number of labeled support nodes per class increases from 1-shot to full-shot, markers denote mean accuracy and error bars denote one standard deviation over five runs.

Across the 16 dataset–backbone pairs formed by Texas, Wisconsin, Cora, and Photo with GCN, SAGE, GAT, and H2GCN in Table 3, **GRACE** outperforms GFMate in every case by 4.12 points on average across four target domains and four backbones under the same evaluation protocol.

C.2 Label-budget comparisons

Every few-shot condition uses balanced K-shot-per-class supports and a 1:9 validation/test split. Figure 7 summarizes the 1-, 3-, 5-, 10-, and full-shot results across all nine targets.

C.3 Perturbation-aware paired-query consistency

Table 4: Perturbation-aware paired-query consistency. Comparison of the full and black-box **GRACE** configurations.

Dataset	GRACE full	GRACE black box
Texas	79.21 ± 5.06	79.21 ± 5.01
Wisconsin	74.39 ± 8.39	74.39 ± 8.42
Cora	60.39 ± 4.54	60.39 ± 4.55
Photo	62.93 ± 1.63	62.90 ± 1.51

Table 4 compares the full and black-box execution configurations. Their near-identical central values across the evaluated targets support the consistency of the perturbation-aware paired-query update.