Synthetic Sparsity Pre-Training for Tail/Cold-Start Robust Graph Prompting under Topology Shift

Liz Lemma Future Detective

January 20, 2026

Abstract

Few-shot learning on graphs faces not only label scarcity but also structure scarcity: tail nodes with small neighborhoods and cold-start nodes with missing edges. Prior graph few-shot methods (meta-learning, pre-training, prompting) largely assume that training and testing tasks are i.i.d. or that topology is sufficiently informative; the survey source material highlights that this assumption breaks for structure-scarce settings and flags robustness and scalability as key 2026-era challenges. This paper proposes a structure-scarcity simulator for self-supervised graph pre-training: a degree-aware neighborhood deletion and attribute masking operator that explicitly trains encoders to remain predictive under topology degradation. We then perform prompt-only few-shot adaptation, freezing the backbone and updating <1

Table of Contents

5. 5. Pre-Training Objective: combined contrastive + generative loss using simulator views; discussion of why it targets topology-shift stability; implementation notes (negative sampling, masking policies) to be validated experimentally.

6. 6. Prompt-Only Adaptation: unified similarity/matching template; prompt parameterization options (vector prompts vs prompt-graphs); constrained ERM for K-shot learning; connection to parameter-efficient fine-tuning.

7. 7. Theory I Ž014 Stability-Based Transfer Bound: define representation stability under Shift; prove excess-risk bound decomposing prompt complexity and shift penalty; interpretability of terms; conditions for tightness.

8. 8. Theory II Ž014 Controlling Stability via Synthetic Sparsity: analysis in a linearized message passing/NTK-like regime; show simulator-augmented objective upper-bounds the stability term; provide explicit dependence on simulator strength.

9. 9. Lower Bounds and Hardness: matching $K = \Omega(P/\varepsilon^2)$ labeled-sample lower bound for P-parameter prompts; impossibility for fully isolated nodes without informative attributes; NP-hardness sketch for worst-case robust prediction under adversarial edge deletions.

10. 10. Experimental Plan (Flagged as Strengthening): controlled sparsity-shift benchmarks; real-world long-tail graphs; ablations of simulator schedules; compare prompting vs adapters/LoRA vs fine-tuning vs meta-learning; correlate measured stability proxy with accuracy under shift.

11. 11. Discussion and Limitations: when synthetic sparsity helps/hurts; relation to cross-domain graphs and complex graphs (hetero/dynamic); open directions (local-access deployment, interpretable prompt-graphs).

1 Introduction

We study node-level prediction in regimes where *structure scarcity* is the dominant obstacle rather than label scarcity alone. In many deployed graph systems, the nodes of interest at inference time are not typical members of the training population: they are tail entities with few interactions, newly arrived users or items with incomplete neighborhoods, or nodes whose incident edges are partially missing due to privacy, logging, or sampling. In such settings, a learned predictor is queried on an ego-neighborhood that is systematically *sparser* than the neighborhoods that carried supervision during training. We formalize this mismatch as a *topology shift*: if training labels are observed on structure-rich head nodes, then test instances correspond to tail/cold-start nodes whose r-hop induced neighborhoods have been perturbed by a sparsification operator Shift (edge deletion and feature masking being canonical examples). The salient point is that the covariates available to the learner—the local subgraph $\mathsf{Sub}_r(G, v)$ together with its attributes—change in a manner that is neither negligible nor symmetric across the node population.

This perspective exposes a limitation of the usual i.i.d. episodic assumptions in few-shot learning on graphs. Standard N-way K-shot evaluations implicitly posit that support and query examples are drawn from the same conditional distribution over observations given labels. On graphs, however, the observation associated with a node is itself a random object (a local neighborhood) whose distribution depends strongly on node degree, community position, and data-collection artifacts. In particular, the training pipeline often privileges high-degree, information-rich nodes (head nodes) because these are easier to label and yield lower-variance gradient estimates; meanwhile, the operational objective is frequently to perform well on precisely those nodes where the topology provides less context (tail nodes). Consequently, even when labels are scarce in the classical sense, the more severe shift is structural: the learner adapts on support samples whose neighborhoods exhibit one connectivity regime and is evaluated on query samples whose neighborhoods exhibit another.

The year 2026 adds a pragmatic constraint that makes this shift unavoidable in large-scale applications. Modern graph encoders are commonly pre-trained on massive unlabeled corpora and are expensive to update per downstream task. At the same time, downstream tasks are numerous and heterogeneous, and their labeled sets are small (often hundreds or fewer labeled nodes per task). This combination encourages *parameter-efficient* adaptation mechanisms, where one freezes a strong encoder f_θ and adapts a small number of parameters. We adopt this constraint explicitly: during downstream adaptation, the learner may only tune a prompt $p \in \mathbb{R}^P$ (or an equivalent prompt-graph), subject to a norm budget $\|p\| \le B$, while the backbone $f_\theta : \mathsf{Sub}_r(G, v) \to \mathbb{R}^d$ remains fixed. In this regime, the statistical

question is not whether full fine-tuning can recover performance, but rather what guarantees (and limitations) are attainable by optimizing over a low-dimensional prompt class under distribution shift in the local neighborhood.

A second constraint is *local access*. In many deployment settings, full-graph inference is infeasible: the graph is too large, continuously updated, or partitioned across services. We therefore restrict attention to algorithms that may query only $\mathsf{Sub}_r(G, v)$ for requested nodes v at adaptation and test time. This restriction is not merely computational; it shapes what information is even observable about a tail node. A cold-start node may have $\deg(v)$ close to zero, in which case its neighborhood contains little beyond its own attributes. Hence, any general theory must distinguish between performance limits stemming from lack of labels and performance limits stemming from lack of structure.

Within this setting, the key technical quantity is the *stability* of the frozen representation to topology shift. If the encoder output $f_\theta(\mathsf{Sub}_r(G, v))$ changes substantially when edges are deleted or features are masked, then no amount of prompt tuning on head nodes can guarantee good performance on tail nodes, because the prompt only sees head-style embeddings at adaptation time. Conversely, if f_θ is stable under a family of plausible sparsifications, then the shift from head to tail becomes a controlled perturbation in embedding space, and prompt-based ERM can transfer with a bounded penalty. This reframes robustness to structure scarcity as a representation-learning problem: we should pre-train f_θ so that it maps multiple corrupted versions of the same underlying neighborhood to nearby embeddings.

To achieve this, we incorporate a degree-aware simulator Sim_η during self-supervised pre-training. The simulator produces sparsified-and-masked views of neighborhoods, with corruption probabilities depending on $\deg(v)$ so that low-degree nodes are more aggressively perturbed. By training f_θ to be consistent across two independently simulated views, and simultaneously requiring sufficient information retention via a generative reconstruction objective, we encourage invariances that are aligned with tail/cold-start conditions. Importantly, this is not an adversarial robustness objective; it is a distributional one. We do not aim to be correct under arbitrary deletions, which can encode intractable combinatorial properties, but rather under shifts whose intensity and degree-dependence resemble the simulator family.

Finally, our objective is to connect these modeling choices to explicit tradeoffs between (i) the prompt parameter budget P, (ii) the number of labeled support nodes K, and (iii) the severity of topology shift. Prompt-only adaptation imposes an unavoidable statistical price: with too many prompt degrees of freedom relative to K, generalization on the support distribution degrades, and no representation stability can remove this dependence. Conversely, even with favorable P and K, a large shift in representation caused by neighborhood sparsification induces an irreducible transfer penalty. Our

framework is designed to make both effects transparent under the local-access constraint, thereby matching the realities of tail-heavy graph prediction and the operational preference for frozen encoders with lightweight adaptation.

2 Related Work

Our setting sits at the intersection of (i) few-shot adaptation on graphs, (ii) self-supervised graph representation learning, (iii) parameter-efficient tuning via prompts, and (iv) long-tail and cold-start graph prediction. We focus on the combination of a frozen encoder, local-access constraints, and an explicit head-to-tail topology shift, which is only partially addressed when these literatures are considered in isolation.

Meta-learning and few-shot learning on graphs. Graph meta-learning adapts models to new tasks with limited labeled nodes by exploiting shared structure across tasks. Representative approaches include gradient-based methods (e.g., MAML-style adaptations) and metric/prototype-based methods built atop GNN encoders, often with episodic training on subgraphs or node neighborhoods **?????**. Graph-specific extensions incorporate neighborhood sampling, hierarchical pooling, or task-conditioned message passing to better handle varying graph sizes and label semantics **??**. A recurring implicit assumption is that support and query nodes are drawn from the same distribution over ego-neighborhoods given the label, so that adaptation performance reflects label scarcity rather than a systematic change in neighborhood topology. Several works do address structural variability via augmentations, subgraph transformations, or uncertainty-aware aggregation, yet the mismatch between structure-rich training nodes and structure-poor test nodes is typically not isolated as a first-class distribution shift **??**. Our formulation makes this mismatch explicit through a shift operator acting on local neighborhoods and evaluates transfer to tail/cold-start regimes under frozen-backbone constraints.

Self-supervised pre-training: contrastive and generative objectives. Large-scale unlabeled pre-training has become the standard route to strong graph encoders. Contrastive methods learn representations invariant to augmentations such as edge dropping, feature masking, subgraph sampling, and diffusion, including DGI, MVGRL, GRACE, GraphCL, and GCC **?????**. Generative and reconstruction-based methods instead predict masked attributes, recover adjacency, or reconstruct local motifs, often improving feature utilization and preventing collapse **???**. Hybrid objectives combining contrastive consistency with reconstruction have also been explored to balance invariance and information preservation **??**. Our use of synthetic sparsification aligns with this tradition, but we emphasize a *degree-aware* cor-

ruption process designed to simulate tail-like neighborhoods and to control a stability quantity under topology shift. In particular, while many augmentations are chosen for empirical performance, our simulator is motivated by the desideratum that the induced invariances translate to bounded representation perturbations for the tail distribution.

Prompting and parameter-efficient tuning for graphs. Parameter-efficient fine-tuning (PEFT) methods, including adapters, prefixes, and low-rank updates, are widely used in language and vision to adapt large frozen backbones with limited task-specific parameters **???**. Analogous ideas have recently appeared in graph learning under the name of graph prompting: one freezes a pre-trained GNN and learns small prompt modules such as virtual prompt nodes/edges, feature prompts, or prototype vectors, with the goal of fast task adaptation and better transfer across datasets **???**. These approaches empirically support the premise that a strong self-supervised encoder can be reused across tasks with lightweight tuning. However, existing work largely studies prompting under matched train/test conditions and reports average-case improvements, whereas our focus is on *prompt-only* adaptation under a *structural* distribution shift. Moreover, we connect prompt dimension and support-set size through explicit sample-complexity terms, and we separate the statistical penalty of prompt expressivity from the transfer penalty induced by representation instability under sparsification.

Long-tail, cold-start, and robustness to structural scarcity. Learning on graphs with imbalanced degree or label distributions has been studied in long-tail node classification, cold-start recommendation, and semi-supervised settings where low-degree nodes receive weak propagation from neighbors **???**. Proposed remedies include degree-aware reweighting, topology or feature augmentation, neighbor imputation, self-training on pseudo-labels, and incorporating side information beyond the observed edges (e.g., text, profiles, or knowledge graphs) **???**. A related but distinct line considers adversarial or worst-case robustness to edge perturbations and poisoning attacks **??**. Our emphasis differs in two ways. First, we treat structure scarcity as a *distributional* phenomenon (tail/cold-start neighborhoods are systematically sparser) rather than an adversarial one, consistent with operational shifts caused by logging, privacy filtering, or sampling. Second, we impose local-access constraints and a frozen encoder, which rule out methods requiring repeated full-graph propagation at adaptation time. The resulting theory highlights an unavoidable limitation: when neighborhoods contain little or no information (e.g., extreme cold-start), performance is bounded by what attributes remain observable, motivating stability-aware pre-training rather than purely label-efficient adaptation mechanisms.

3 Problem Setup and Computational Model

We study node-level prediction under a combination of *label scarcity*, *structural scarcity*, and *parameter-efficiency* constraints. The ambient object is an attributed graph $G = (V, E, X)$ with node features $X \in \mathbb{R}^{|V| \times d_x}$. For a node $v \in V$ and an integer radius $r \geq 1$, we write $\mathsf{Sub}_r(G, v)$ for the induced r-hop ego-subgraph around v, i.e., the subgraph on nodes within graph distance at most r from v (together with their features). Throughout, we work in a local-access model: an algorithm may request $\mathsf{Sub}_r(G, v)$ for particular nodes v, but it is not permitted to execute full-graph message passing or to repeatedly traverse the entire edge set at adaptation or test time. This captures large-scale and privacy-filtered settings where only bounded neighborhoods can be materialized per query.

Head-to-tail evaluation under topology shift. We formalize the mismatch between structure-rich training nodes and structure-poor test nodes via two distributions. The *head* distribution $\mathcal{D}_{\text{head}}$ governs labeled examples with comparatively informative neighborhoods (e.g., moderate-to-high degree, dense local connectivity, unmasked attributes). The *tail* distribution $\mathcal{D}_{\text{tail}}$ governs evaluation examples in which the observable neighborhood is structurally degraded, as in low-degree nodes, partially logged edges, or cold-start entities. To model this degradation explicitly, we introduce a (possibly stochastic) *topology shift* operator Shift acting on graphs (or directly on ego-neighborhoods). Given an original downstream graph G^*, we may evaluate on neighborhoods extracted from $\mathsf{Shift}(G^*)$, or equivalently on pairs $\big(\mathsf{Sub}_r(G^*, v), \mathsf{Sub}_r(\mathsf{Shift}(G^*), v)\big)$ that share the same center node v and label but differ in observable edges and/or features. The essential assumption is that labels are preserved by Shift while the structural statistics of neighborhoods change in a way that emphasizes tail or cold-start regimes.

Frozen encoder and prompt-only adaptation. We assume access to a graph encoder

$$f_\theta : \mathsf{Sub}_r(G, v) \to \mathbb{R}^d$$

whose parameters θ are learned offline by self-supervised pre-training on unlabeled data from a pre-training distribution $\mathcal{D}_{\text{pre}}$. At downstream time, θ is frozen: the only trainable component is a *prompt* $p \in \mathbb{R}^P$ (or an equivalent prompt-graph parameterization) constrained by $\|p\| \leq B$ and a parameter budget P. A prompted predictor takes the form

$$v \;\longmapsto\; h_p(f_\theta(\mathsf{Sub}_r(G^*, v)))\,,$$

where h_p is a fixed template (e.g., a similarity-to-prototypes classifier, a small MLP head with prompt-controlled biases, or a prompt-conditioned linear

separator) whose dependence on task data enters only through p. This abstraction isolates the statistical role of prompt expressivity from representation learning in the encoder, and it enforces a parameter-efficient adaptation regime.

Downstream protocol and support/query access. A downstream task instance provides a labeled support set

$$S = \{(v_i, y_i)\}_{i=1}^{K},$$

drawn from head nodes of G^* (or more generally from $\mathcal{D}_{\text{head}}$), together with a query set Q drawn from tail or cold-start nodes (modeled by $\mathcal{D}_{\text{tail}}$). During adaptation, we may query only the r-hop neighborhoods $\mathsf{Sub}_r(G^*, v_i)$ for $v_i \in S$; during evaluation, each query node $v \in Q$ is presented through its shifted neighborhood $\mathsf{Sub}_r(\mathsf{Shift}(G^*), v)$ (or through an equivalent tail sampling rule). In particular, we do not assume availability of labels for tail nodes, and we do not assume that head and tail neighborhoods are identically distributed conditional on the label.

Risk functionals under shift. Let $\ell(\hat{y}, y)$ denote a bounded prediction loss (e.g., cross-entropy), and write $\mathcal{R}_{\text{head}}(p)$ for the expected risk when examples are drawn from the head distribution:

$$\mathcal{R}_{\text{head}}(p) \;=\; \mathbb{E}_{(v,y)\sim\mathcal{D}_{\text{head}}}\Big[\ell\big(h_p(f_\theta(\mathsf{Sub}_r(G^*, v))), y\big)\Big].$$

The target of interest is the test risk on the tail distribution, which we view as induced by a shift applied to neighborhoods:

$$\mathcal{R}_{\text{test}}(p) \;=\; \mathbb{E}_{(v,y)\sim\mathcal{D}_{\text{tail}}}\Big[\ell\big(h_p(f_\theta(\mathsf{Sub}_r(\mathsf{Shift}(G^*), v))), y\big)\Big].$$

Given a finite support set S, we define the empirical risk

$$\hat{\mathcal{R}}_S(p) \;=\; \frac{1}{K}\sum_{i=1}^{K} \ell\big(h_p(f_\theta(\mathsf{Sub}_r(G^*, v_i))), y_i\big),$$

and we perform prompt-only ERM (or any prompt-only optimizer) over $\{p : \|p\| \le B\}$, producing a task-adapted prompt p_{ERM}. This yields a precise separation between (i) a *statistical estimation* term controlled by P and K, and (ii) a *transfer* term controlled by how stable f_θ is to the shift Shift.

Cold-start and observability. The local-access model makes explicit the information available at test time. If $\deg(v)$ is small, then $\mathsf{Sub}_r(\mathsf{Shift}(G^*), v)$ may contain very few nodes or edges; in the extreme cold-start case $\deg(v) = 0$, the model reduces to attribute-only prediction from X_v. Consequently,

the feasibility of tail performance depends on whether (a) the remaining attributes are label-informative and (b) the frozen encoder produces embeddings that vary smoothly with respect to edge/feature deletions typical of the tail regime. The latter consideration motivates equipping pre-training with explicit synthetic structure scarcity so that the learned representation is not brittle to sparsification; we formalize this next by introducing a degree-aware simulator Sim_η and its controllable design parameters.

4 Structure-Scarcity Simulator Sim_η

We now specify the stochastic operator Sim_η used to generate tail-like views during self-supervised pre-training. Formally, for an input graph (or extracted neighborhood) $G = (V, E, X)$, the simulator outputs a random corrupted view $\tilde{G} = (V, \tilde{E}, \tilde{X})$ distributed as $\tilde{G} \sim \mathsf{Sim}_\eta(\cdot \mid G)$. The hyperparameters η control (i) degree-aware edge deletion, (ii) attribute masking/corruption, and (iii) optional conditioning on the ego-center and hop distance. Our design goal is to create perturbations that are *stronger on structurally weak regions* while remaining implementable by local operations on r-hop neighborhoods.

Degree-aware edge deletion. Let $e = (u, w) \in E$ be an edge. The simulator deletes e with probability

$$q_E(e; G, \eta) \in [0, 1], \qquad \tilde{E} = \{e \in E : \xi_e = 1\}, \quad \xi_e \sim \text{Bernoulli}(1 - q_E(e; G, \eta)).$$

The defining feature is that $q_E(e; G, \eta)$ increases as the incident degrees decrease, thereby producing views in which low-degree nodes become even sparser. A convenient parameterization is

$$q_E\big((u, w); G, \eta\big) = \text{clip}_{[0,1]}\left(\rho_E \cdot \Big(\frac{\deg_G(u) + 1}{\bar{d} + 1}\Big)^{-\gamma} \cdot \Big(\frac{\deg_G(w) + 1}{\bar{d} + 1}\Big)^{-\gamma} \cdot \omega\big(\text{dist}(u, v_0), \text{dist}(w, v_0)\big)\right),$$

where $\rho_E \in [0, 1]$ is a global sparsification strength, $\gamma \geq 0$ controls the degree sensitivity, $\bar{d}$ is a normalizing constant (e.g. the mean degree in the current sampled graph or batch), v_0 is the ego-center when simulating on $\mathsf{Sub}_r(G, v_0)$, and ω is an optional hop-dependent weight. Taking $\omega \equiv 1$ yields degree-only deletion; choosing ω increasing in hop index emphasizes deleting edges farther from the center (a common logging artifact), whereas decreasing ω stresses immediate-neighbor loss (a cold-start-like regime). The additive $+1$ prevents degeneracy at deg $= 0$ and ensures continuity in the schedule.

Conditional deletion and connectivity constraints. Purely independent Bernoulli deletion is analytically and computationally convenient, but

one may also incorporate mild conditioning to better match deployment shifts. Two examples are: (i) *per-node budget deletion*, where for each node u we retain at most $k(u)$ incident edges sampled proportionally to $(1 - q_E)$, and (ii) *minimum-degree preservation*, where we resample deletions until each node keeps at least one incident edge when $\deg_G(u) > 0$. Such conditioning breaks edge-wise independence but remains local and retains the essential monotonicity: nodes with smaller $\deg_G(u)$ experience larger expected relative loss of incident edges.

Attribute masking and corruption. Edge sparsification alone does not model attribute incompleteness, so Sim_η also corrupts X. We distinguish node-level masking from feature-dimension masking. For each node $u \in V$, we draw a node mask indicator

$$m_u \sim \text{Bernoulli}\big(1-q_V(u;G,\eta)\big), \qquad q_V(u;G,\eta) \;=\; \text{clip}_{[0,1]}\Bigg(\rho_V \cdot \Big(\frac{\deg_G(u)+1}{\bar{d}+1}\Big)^{-\gamma_V} \cdot \omega_V(\text{dist}(u,v_0))$$

and we set $\tilde{X}_u = X_u$ if $m_u = 1$ and otherwise apply a corruption operator Corr, e.g. $\tilde{X}_u = 0$, $\tilde{X}_u = \text{MASK}$, or $\tilde{X}_u = X_u + \epsilon$ with ϵ subgaussian. In addition, for feature vectors that remain present ($m_u = 1$), we may mask individual coordinates: for each dimension $j \in [d_x]$, draw $m_{u,j} \sim \text{Bernoulli}(1 - \rho_X)$ and set $\tilde{X}_{u,j} = X_{u,j}$ if $m_{u,j} = 1$ and otherwise replace with a learned mask token or zero. This two-level scheme allows the simulator to represent both missing-at-random attributes (coordinate masking) and missing-at-entity attributes (node masking), with degree-aware intensification when desired.

Locality and ego-subgraph compatibility. Although Sim_η is defined on graphs, in practice we apply it to sampled neighborhoods. Given $\mathsf{Sub}_r(G, v_0)$, we compute degrees either with respect to the ambient G (if available in pre-training) or with respect to the induced neighborhood (if only local access is permitted). Both choices preserve the intended bias: in either case, nodes near the center with few observed incident edges receive higher deletion/masking probability, producing synthetic tails inside the same pre-training distribution.

Controllable parameters and regimes. The simulator hyperparameters can be grouped as

$$\eta = (\rho_E, \gamma, \omega;\ \rho_V, \gamma_V, \omega_V;\ \rho_X;\ \mathsf{Corr};\ \text{conditioning flags}),$$

where (ρ_E, ρ_V, ρ_X) set the overall corruption strengths, (γ, γ_V) set the degree dependence, and (ω, ω_V) shape the hop profile. Varying ρ_E interpolates between fully observed neighborhoods ($\rho_E = 0$) and severe sparsification; increasing γ concentrates corruption on low-degree regions and is the principal mechanism by which we emulate tail/cold-start structure. Importantly, these

parameters admit calibration to an anticipated test-time shift: if we estimate empirical edge-retention curves as a function of degree in the downstream system, we can choose η so that the simulator matches (or intentionally upper-bounds) those curves.

A domination property. For our subsequent stability arguments, it is useful to ensure that the test-time operator Shift is *dominated* by Sim_η: informally, any edge or feature deletion that can occur under Shift occurs with at most the same probability under Sim_η. One sufficient condition (stated at the level of single edges/features) is

$$\forall e \in E:\ q_{\mathsf{Shift}}(e; G) \le q_E(e; G, \eta), \qquad \forall (u, j):\ q_{\mathsf{Shift}}(u, j; G) \le q_X(u, j; G, \eta),$$

with analogous inequalities for node-level masking. This condition can be enforced conservatively by choosing ρ_E, ρ_V, ρ_X slightly larger than the expected deployment corruption. The practical interpretation is that pre-training observes neighborhoods at least as sparse as those encountered at evaluation, thereby providing the encoder with repeated exposure to tail-like perturbations.

5 Pre-Training Objective: Contrastive–Generative Learning Under Synthetic Scarcity

Given unlabeled samples from $\mathcal{D}_{pre}$, we pre-train the encoder f_θ by repeatedly drawing two independent simulator views of the same underlying neighborhood and enforcing (i) *representation consistency* across these views and (ii) *reconstructability* of masked structure/attributes. Concretely, for a sampled pair (G, v) we draw $\tilde{G}^{(1)}, \tilde{G}^{(2)} \sim \mathsf{Sim}_\eta(\cdot \mid G)$ independently and compute

$$z^{(a)} \;=\; f_\theta\big(\mathsf{Sub}_r(\tilde{G}^{(a)}, v)\big) \in \mathbb{R}^d, \qquad a \in \{1, 2\}.$$

The pre-training loss is

$$\mathcal{L}_{pre}(\theta, \phi) \;=\; \mathcal{L}_{\text{contrast}}(\theta) \;+\; \lambda\, \mathcal{L}_{\text{gen}}(\theta, \phi),$$

where ϕ denotes parameters of the decoder(s) used for reconstruction, and $\lambda \ge 0$ controls the discriminative–generative tradeoff.

Contrastive invariance to simulator corruption. For the contrastive term, we adopt an InfoNCE-style objective over mini-batches of size m. Writing $z_i^{(a)}$ for the embedding of the i-th sampled neighborhood under view a, a standard choice is

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{2m} \sum_{i=1}^{m} \left[\log \frac{\exp\big(\langle z_i^{(1)}, z_i^{(2)} \rangle / \tau\big)}{\sum_{j=1}^{m} \exp\big(\langle z_i^{(1)}, z_j^{(2)} \rangle / \tau\big)} + \log \frac{\exp\big(\langle z_i^{(2)}, z_i^{(1)} \rangle / \tau\big)}{\sum_{j=1}^{m} \exp\big(\langle z_i^{(2)}, z_j^{(1)} \rangle / \tau\big)} \right],$$

with temperature $\tau > 0$. The only role of this term in our analysis is that it explicitly penalizes variation of f_θ across simulator-induced perturbations of the same underlying neighborhood, thereby biasing f_θ toward being insensitive to degree-dependent edge loss and attribute masking. This is precisely the invariance required to make $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ small when Shift is dominated by Sim_η.

Generative reconstruction under missingness. Contrastive learning alone can encourage invariances that discard task-relevant information. We therefore include a masked reconstruction term that forces embeddings to retain information sufficient to predict the uncorrupted neighborhood. A simple instantiation uses a decoder d_ϕ that takes as input the embedding and (optionally) local positional identifiers (e.g. hop index), and outputs reconstructions of masked node features and/or removed edges inside the r-hop induced subgraph. Denoting by $\mathcal{M}$ the set of corrupted feature coordinates (or nodes) and by $\mathcal{E}^-$ the set of deleted edges under the simulator, we may write schematically

$$\mathcal{L}_{\mathrm{gen}} = \mathbb{E}\Big[\sum_{(u,j)\in\mathcal{M}} \ell_X\big(d_\phi(z, \mathrm{id}(u), j),\, X_{u,j}\big) + \sum_{e\in\mathcal{E}^-} \ell_E\big(d_\phi(z, \mathrm{id}(e)),\, 1\big) + \sum_{e\in\mathcal{E}^+} \ell_E\big(d_\phi(z, \mathrm{id}(e)),\, 0\big)\Big],$$

where $\mathcal{E}^+$ is a sampled set of non-edges for negative edge prediction. The precise parameterization is not essential; what matters is that reconstruction is performed *from corrupted views* so that the encoder must propagate stable information across sparsification patterns rather than overfitting to a single topology.

Why the objective targets topology-shift stability. The stability term $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ measures the expected embedding perturbation induced by a test-time sparsification. By construction, the contrastive component reduces (in expectation) $\|z^{(1)} - z^{(2)}\|$ when $\tilde{G}^{(1)}$ and $\tilde{G}^{(2)}$ differ by deletions and maskings typical of Sim_η. When Shift is dominated by Sim_η, the simulator generates perturbations at least as severe as those induced at test time, and the learned invariances transfer to the shifted regime. The generative component prevents trivial solutions (e.g. mapping every neighborhood to a constant) and encourages that invariance be achieved through *predictive compression* rather than information erasure, which is crucial for downstream discrimination with a small prompt.

Implementation notes under local-access constraints. We rely on in-batch negatives for computational efficiency; this yields $O(m^2 d)$ similarity computation but avoids maintaining a memory bank, and is compatible with sampling neighborhoods independently from $\mathcal{D}_{pre}$. In practice, one may further stabilize training by (i) normalizing embeddings, (ii) using a predictor

head and stop-gradient on one branch (BYOL-style) to reduce sensitivity to batch size, or (iii) replacing InfoNCE with variance–covariance regularization (VICReg-style) to mitigate collapse; these choices preserve the intended invariance effect and are interchangeable for our purposes. For masking, we apply simulator edge deletions before feature corruption so that masked nodes do not artificially influence message passing through edges that would be absent under scarcity. When node-level masking is used, we prefer a learned mask token over all-zeros to reduce distribution shift between "missing" and "truly zero" attributes. For edge reconstruction, we sub-sample candidate pairs within Sub_r to avoid quadratic enumeration of all node pairs.

Experimental validation points. The remaining design choices—temperature τ, the balance λ, and the specific corruption policies within Sim_η—are treated as tunable hyperparameters. Empirically, we will validate (i) whether increasing simulator strength decreases $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ as predicted, (ii) whether the contrastive–generative combination outperforms either term alone under tail evaluation, and (iii) whether the learned invariances preserve label information sufficiently to be exploited by prompt-only adaptation.

6 Prompt-Only Adaptation: Unified Matching Templates and Constrained K-Shot ERM

We now specify the downstream adaptation stage under the parameter-efficiency constraint that the encoder f_θ is frozen and only a prompt object $p \in \mathbb{R}^P$ (or an equivalent prompt-graph parameterization of total dimension P) may be updated. The input to adaptation is a labeled support set $S = \{(v_i, y_i)\}_{i=1}^K$ drawn from head nodes of the downstream graph G^*, together with unlabeled query nodes from a tail/cold-start regime. For each support node we compute the local embedding

$$z_i \;=\; f_\theta\big(\mathsf{Sub}_r(G^*, v_i)\big) \in \mathbb{R}^d,$$

and we seek a prompted predictor of the form $v \mapsto h_p(z)$ that can be tuned with K labels and then evaluated on tail neighborhoods (potentially after applying Shift).

A unified similarity/matching view of h_p. To make the role of prompts explicit while keeping the hypothesis class analyzable, we instantiate h_p as a similarity-based template. We write $g_p : \mathbb{R}^d \to \mathbb{R}^{d'}$ for a prompt-conditioned transformation of the frozen embedding and define class scores by comparing $g_p(z)$ to prompt-conditioned reference vectors. For a C-class task, a general form is

$$\mathrm{score}_c(z; p) \;=\; \mathsf{Sim}\big(g_p(z),\, r_c(p, S)\big), \qquad \hat{y} \;=\; \arg\max_{c \in [C]} \mathrm{score}_c(z; p),$$

where Sim is a fixed similarity (e.g. dot product or negative squared distance), and $r_c(p, S)$ denotes a class-specific reference computed from the support set (possibly also depending on p). This umbrella includes (i) linear probing ($\mathrm{score}_c(z; p) = \langle w_c(p), z \rangle$), (ii) prototypical classification (r_c is a prototype), and (iii) nearest-neighbor matching with a prompt-induced metric. The analysis in the sequel uses only Lipschitzness of h_p as a map of the embedding argument, hence we do not commit to a single Sim.

Prototype-based instantiation (metric prompting). A particularly convenient specialization for K-shot learning is a prototypical head. Let $S_c = \{i : y_i = c\}$ and define a prompt-conditioned prototype

$$\mu_c(p) \;=\; \frac{1}{|S_c|} \sum_{i \in S_c} g_p(z_i), \qquad \mathrm{score}_c(z; p) \;=\; -\big\| g_p(z) - \mu_c(p) \big\|_2^2.$$

If g_p is close to identity, then prompts primarily modulate the geometry of the embedding space and can correct systematic bias between the pre-trained representation and the downstream label geometry using only a small number of parameters. When C is large and K is small, prototypes provide an implicit regularization: parameters are shared across classes through g_p rather than through C independent weight vectors.

Vector prompt parameterizations (post-encoder). In the strictest local-access setting, we may keep the encoder calls independent of p and implement prompting purely in embedding space. Typical choices include

$$g_p(z) = z{+}p \quad \text{(additive prompt)}, \qquad g_p(z) = \gamma(p) \odot z \quad \text{(feature-wise gating)}, \qquad g_p(z) = z{+}U(p)\,V($$

Here $\gamma(p) \in \mathbb{R}^d$ is a bounded gating vector and $U(p), V(p)$ have small rank so that the total number of free parameters is $P \ll |\theta|$. These vector prompts are computationally attractive: once z_i are cached for the support set, each prompt update step avoids re-encoding neighborhoods, reducing adaptation to inexpensive operations in $\mathbb{R}^d$.

Prompt-graphs (pre-encoder) under a frozen backbone. An alternative is to represent p as a small prompt graph Π_p consisting of n_p virtual nodes with learnable features and a fixed, bounded pattern of edges connecting Π_p to the ego-subgraph. For a node v, we form an augmented neighborhood $\mathsf{Sub}_r(G^*, v) \oplus \Pi_p$ (with attachment rules that depend only on hop index or a designated anchor) and define

$$z_v(p) \;=\; f_\theta\big(\mathsf{Sub}_r(G^*, v) \oplus \Pi_p\big), \qquad \hat{y} \;=\; h\big(z_v(p)\big),$$

where h is a fixed readout (e.g. a cosine-to-prototype map). This realizes adaptation by altering the message-passing context while keeping θ unchanged. The cost is that embeddings must typically be recomputed when

p changes; nevertheless, n_p can be made small and the attachment local so that local-access constraints remain satisfied.

Constrained ERM for K-shot adaptation. We perform downstream learning by empirical risk minimization over prompts with a norm budget:

$$p_{\mathrm{ERM}} \in \arg\min_{\|p\|\le B} \hat{\mathcal{R}}_S(p) = \arg\min_{\|p\|\le B} \frac{1}{K}\sum_{i=1}^{K} \ell\big(h_p(z_i), y_i\big).$$

In practice we implement this constraint either by projected gradient descent $p \leftarrow \Pi_{\{\|p\|\le B\}}(p - \eta\nabla\hat{\mathcal{R}}_S(p))$ or by adding a penalty $\rho\max\{0, \|p\| - B\}^2$. The constraint is not merely cosmetic: it controls the capacity of the prompt class and is the parameter-level analogue of standard regularization in linear probing.

Relation to parameter-efficient fine-tuning. Prompt-only adaptation is a graph-local analogue of parameter-efficient fine-tuning (PEFT): we restrict task-specific learning to P degrees of freedom while retaining a task-agnostic representation learned offline. Vector prompts correspond to learned offsets/gates in representation space (comparable to prompt tuning or lightweight adapters), while low-rank constructions parallel LoRA-style updates but placed after the encoder rather than inside it. Prompt-graphs, by contrast, mimic soft prompt tokens in sequence models: they modify the computation by injecting learnable context rather than changing backbone weights. The subsequent theory treats all these options through a single lens: a constrained hypothesis class indexed by p and evaluated under topology shift through the stability of the frozen encoder.

7 Theory I — Stability-Based Transfer Bound

We formalize the effect of topology shift by isolating the portion of the risk that is attributable to distribution shift in the *observed neighborhoods* rather than to statistical error in fitting the prompt. Let $\mathcal{R}_{\mathrm{head}}(p) \triangleq \mathbb{E}_{(v,y)\sim\mathcal{D}_{\mathrm{head}}}\big[\ell(h_p(f_\theta(\mathsf{Sub}_r(G^*, v))), y)\big]$ denote the population risk on head neighborhoods, and recall $\mathcal{R}_{\mathrm{test}}(p)$ is defined analogously under $\mathcal{D}_{\mathrm{tail}}$ (possibly induced by applying Shift). Our bounds proceed by (i) controlling the statistical error of prompt ERM on $\mathcal{D}_{\mathrm{head}}$ and (ii) controlling the transfer error from $\mathcal{D}_{\mathrm{head}}$ to $\mathcal{D}_{\mathrm{tail}}$ through a stability modulus of the frozen encoder.

Representation stability under topology shift. We quantify how much the representation of a node can change when its neighborhood is sparsified. Writing the shifted graph as $\mathsf{Shift}(G)$ and coupling $(G, v) \sim \mathcal{D}_{\mathrm{head}}$ with its

shifted counterpart, we define the topology-shift stability of f_θ by

$$\mathrm{Stab}(f_\theta; \mathsf{Shift}) \triangleq \mathbb{E}_{(G,v)\sim\mathcal{D}_{\mathrm{head}}} \Big\| f_\theta\big(\mathsf{Sub}_r(G,v)\big) - f_\theta\big(\mathsf{Sub}_r(\mathsf{Shift}(G),v)\big) \Big\|_2.$$

This quantity is a property of the frozen encoder and the shift operator, independent of the downstream labels. In particular, it is well-defined under the local-access model because it depends only on pairs of r-hop neighborhoods.

Step 1: prompt ERM generalization on head nodes. We first bound the excess risk incurred by fitting the prompt on K labeled head examples. Under bounded embeddings $\|f_\theta(\cdot)\| \le R$, a norm constraint $\|p\| \le B$, and Lipschitzness of the prompt class in its parameters and in the embedding input, standard Rademacher complexity arguments yield that with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{head}}(p_{\mathrm{ERM}}) - \inf_{\|p\|\le B} \mathcal{R}_{\mathrm{head}}(p) \;\le\; O\left(\frac{\alpha B R\sqrt{P} + \sqrt{\log(1/\delta)}}{\sqrt{K}} \right),$$

where P is the number of tunable prompt parameters and α is the Lipschitz constant of h_p with respect to p. The salient point is that the statistical error scales as $\sqrt{P/K}$: holding K fixed, increasing prompt capacity increases the price of adaptation, while freezing θ removes any dependence on $|\theta|$.

Step 2: transfer from head to tail via Lipschitz stability. To relate the head and tail risks, we use the fact that the downstream shift acts *through* the representation $z = f_\theta(\mathsf{Sub}_r(\cdot))$. Suppose the composed loss is L-Lipschitz in the embedding argument, i.e.,

$$\big|\ell(h_p(z),y) - \ell(h_p(z'),y)\big| \;\le\; L\,\|z-z'\|_2 \qquad \text{for all } (z,z',y,p).$$

Then, under the coupling that generates $\mathcal{D}_{\mathrm{tail}}$ by applying Shift to neighborhoods while preserving labels, we obtain the pointwise bound

$$\big|\mathcal{R}_{\mathrm{test}}(p) - \mathcal{R}_{\mathrm{head}}(p)\big| \;\le\; L \cdot \mathrm{Stab}(f_\theta; \mathsf{Shift}) \qquad \text{for all } p.$$

Indeed, we write both risks as expectations over the same draw $(G,v,y) \sim \mathcal{D}_{\mathrm{head}}$ and apply the Lipschitz property to the difference between the losses evaluated at $f_\theta(\mathsf{Sub}_r(G,v))$ and $f_\theta(\mathsf{Sub}_r(\mathsf{Shift}(G),v))$.

Excess test risk decomposition. Combining the head generalization inequality with the transfer inequality yields an excess risk bound on the tail distribution that separates *prompt complexity* from the *shift penalty*. Concretely, by adding and subtracting $\mathcal{R}_{\mathrm{head}}$ and using triangle inequalities, we obtain

$$\mathcal{R}_{\mathrm{test}}(p_{\mathrm{ERM}}) - \inf_{\|p\|\le B} \mathcal{R}_{\mathrm{test}}(p) \;\le\; O\left(\frac{\alpha B R\sqrt{P}}{\sqrt{K}} \right) + 2L\cdot\mathrm{Stab}(f_\theta; \mathsf{Shift}) \quad \text{(up to logarithmic terms)}.$$

The first term is the usual estimation error for prompt ERM on K labels; it is the price of parameter-efficiency. The second term is the irreducible cost of evaluating on neighborhoods that differ from those used to fit the prompt. Importantly, the second term is *prompt-independent* in the sense that it depends on p only through the Lipschitz constant L of the template; consequently, improving robustness to structure scarcity is primarily a representation learning problem (reducing $\mathrm{Stab}(f_\theta; \mathsf{Shift})$) rather than a downstream optimization trick.

Interpretation and conditions for tightness. The decomposition is informative precisely because the two terms correspond to distinct failure modes. If $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ is small, then tail performance is limited mainly by the few-shot estimation rate $\sqrt{P/K}$, and increasing K or decreasing P improves test risk in the expected manner. Conversely, if $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ is large, then no amount of prompt tuning on head labels can guarantee good tail performance without changing the representation: even the best prompt for $\mathcal{D}_{\mathrm{head}}$ may not transfer.

Both terms are essentially tight. For the prompt term, if the prompt class contains a P-dimensional linear family, then there exist tasks for which any method using K labeled examples suffers expected excess risk at least $\Omega(\sqrt{P/K})$, matching the upper bound scaling in P and K. For the shift term, the Lipschitz inequality is tight whenever the loss changes at rate L along the direction of the embedding perturbation induced by Shift; in particular, if Shift systematically deletes informative edges for tail nodes, the induced representation drift can be aligned with the decision boundary, and the resulting risk gap scales proportionally to $\mathbb{E}\|f_\theta(\cdot) - f_\theta(\mathsf{Shift}(\cdot))\|_2$. The next section addresses how simulator-augmented pre-training reduces this stability term in a controlled manner.

8 Theory II — Controlling Stability via Synthetic Sparsity

We now explain why simulator-augmented pre-training can *control* the topology-shift stability term. Our goal is to connect the stability modulus

$$\mathrm{Stab}(f_\theta; \mathsf{Shift}) = \mathbb{E}\Big\| f_\theta\big(\mathsf{Sub}_r(G, v)\big) - f_\theta\big(\mathsf{Sub}_r(\mathsf{Shift}(G), v)\big) \Big\|_2$$

to an observable pre-training objective in which we enforce representation consistency under synthetic structure loss.

Linearized message passing / NTK-like regime. We adopt a standard linearization argument: for a message-passing encoder with parameters θ

near an initialization θ_0, we approximate the embedding by its first-order expansion

$$f_\theta(\mathsf{Sub}_r(G,v)) \approx f_{\theta_0}(\mathsf{Sub}_r(G,v)) + J_{\theta_0}(\mathsf{Sub}_r(G,v))(\theta-\theta_0),$$

where J_{θ_0} is the Jacobian with respect to θ. For the purpose of stability under edge/feature deletions, the constant offset f_{θ_0} cancels in differences, and we may view the learned representation as a linear map applied to a (fixed) neighborhood-dependent feature vector. Concretely, we assume the simplified model

$$f_\theta(\mathsf{Sub}_r(G,v)) \;=\; W\,\psi(G,v), \tag{1}$$

where $W \in \mathbb{R}^{d\times D}$ is learned and $\psi(G,v) \in \mathbb{R}^D$ is a bounded feature map summarizing the r-hop neighborhood (e.g., aggregated node attributes and adjacency patterns). We assume $\|\psi(G,v)\|_2 \le S$ almost surely, which is natural under bounded node features and finite r.

Simulator consistency as a surrogate for stability. Let $\tilde{G}^{(1)}, \tilde{G}^{(2)} \sim \mathsf{Sim}_\eta(\cdot \mid G)$ be two independent simulator views of the same underlying neighborhood. A core term in our pre-training objective enforces invariance by penalizing the embedding discrepancy

$$\big\| f_\theta(\mathsf{Sub}_r(\tilde{G}^{(1)},v)) - f_\theta(\mathsf{Sub}_r(\tilde{G}^{(2)},v))\big\|_2^2.$$

Under (1), this becomes $\|W(\psi(\tilde{G}^{(1)},v) - \psi(\tilde{G}^{(2)},v))\|_2^2$, which is precisely the quadratic form governing sensitivity to simulator-induced deletions and masking. We therefore study the regularized objective

$$\mathcal{L}_{pre}(W) \;=\; \mathbb{E}_{(G,v)}\mathbb{E}_{\tilde{G}^{(1)},\tilde{G}^{(2)}\sim\mathsf{Sim}_\eta(\cdot|G)}\big\|W\Delta\psi_{\mathsf{Sim}}\big\|_2^2 + \lambda\|W\|_F^2, \qquad \Delta\psi_{\mathsf{Sim}} \triangleq \psi(\tilde{G}^{(1)},v) - \psi(\tilde{G}^{(2)},v). \tag{2}$$

The ridge term captures either explicit weight decay or implicit norm control from optimization dynamics, and it will be useful to ensure a bounded operator norm.

Dominated shifts and explicit dependence on simulator strength. To relate Sim_η to the downstream operator Shift, we require that Shift be *dominated* by the simulator in the sense that every deletion/masking event that can occur under Shift is at most as likely under Sim_η, up to a multiplicative factor. One convenient formulation is the following: for each local structural/feature element e (e.g., an edge incident to a node in $\mathsf{Sub}_r(G,v)$, or a feature coordinate of a node in the neighborhood), let $p_e(G,v)$ denote the probability that Sim_η removes/masks e, and let $q_e(G,v)$ denote the corresponding probability for Shift. We assume there exists $\rho \ge 1$ such that

$$q_e(G,v) \;\le\; \rho\, p_e(G,v) \qquad \text{for all relevant } (G,v,e). \tag{3}$$

In our degree-aware setting, $p_e(G, v)$ is larger for elements attached to low-degree nodes, so (3) is plausible when the test-time shift increases sparsification primarily on tail neighborhoods (possibly by increasing deletion rates in the same degree-dependent family). The parameter ρ is an explicit measure of "how much stronger" the test shift is compared to the simulator: taking η stronger (more deletions/masks, especially on low degree) increases p_e and can reduce ρ.

A stability bound from the pre-training objective. Let W^* be any minimizer of (2). We claim that small $\mathcal{L}_{pre}(W^*)$ implies small stability under any dominated shift. The key observation is that, for bounded ψ, the perturbation induced by Shift can be controlled by the perturbations induced by two independent simulator draws. Intuitively, if Sim_η deletes/masks each element with probability at least that of Shift (up to ρ), then the random difference between two simulator views stochastically covers the difference between the original neighborhood and its shifted version. In particular, one can show (by element-wise coupling and a variance comparison) that there exists a constant c_0 depending only on the boundedness and additivity properties of ψ such that

$$\mathbb{E}\big\|\psi(G, v) - \psi(\mathsf{Shift}(G), v)\big\|_2^2 \;\le\; c_0\,\rho \cdot \mathbb{E}\big\|\psi(\tilde{G}^{(1)}, v) - \psi(\tilde{G}^{(2)}, v)\big\|_2^2. \quad (4)$$

Multiplying by W^* and applying Jensen's inequality gives

$$\mathrm{Stab}(f_{W^*}; \mathsf{Shift}) = \mathbb{E}\big\|W^*(\psi(G, v) - \psi(\mathsf{Shift}(G), v))\big\|_2 \;\le\; \sqrt{\mathbb{E}\big\|W^*(\psi(G, v) - \psi(\mathsf{Shift}(G), v))\big\|_2^2}.$$

Combining this with (4) yields

$$\mathrm{Stab}(f_{W^*}; \mathsf{Shift}) \;\le\; \sqrt{c_0\rho} \cdot \sqrt{\mathbb{E}\big\|W^*(\psi(\tilde{G}^{(1)}, v) - \psi(\tilde{G}^{(2)}, v))\big\|_2^2}. \quad (5)$$

Finally, since the first term in (2) is exactly the expectation under the square root in (5), we obtain an explicit upper bound in terms of the optimized objective:

$$\mathrm{Stab}(f_{W^*}; \mathsf{Shift}) \;\le\; \sqrt{c_0\rho} \cdot \sqrt{\mathcal{L}_{pre}(W^*)}. \quad (6)$$

The dependence on simulator strength is entirely contained in ρ: stronger simulator corruption (larger p_e, especially on low degree) decreases ρ for a fixed test shift and tightens (6). The ridge parameter λ further prevents degenerate solutions by controlling $\|W^*\|_F$, ensuring that the model cannot amplify small neighborhood perturbations without incurring large regularization cost. In summary, in this linearized regime, the synthetic sparsity objective provides a certificate that the learned representation is insensitive to the same kinds of structural losses that characterize tail and cold-start evaluation.

9 Lower Bounds and Hardness

We complement the preceding upper bounds with limitations that are intrinsic to (i) prompt-only adaptation with a P-dimensional hypothesis class, (ii) tail/cold-start regimes where the observable neighborhood carries no label information, and (iii) worst-case robustness requirements under adversarial topology perturbations. These results justify why our guarantees necessarily scale with P, why any nontrivial performance on fully isolated nodes must rely on informative attributes, and why we restrict attention to stochastic (distributional) shift models rather than adversarial ones.

A matching labeled-sample lower bound in the prompt dimension. Fix a frozen encoder f_θ and consider any prompt family $\{h_p : \|p\| \le B\}$ whose induced predictors contain a P-dimensional linear subfamily. A canonical example is a template of the form

$$h_p(z) = \langle p, \Phi(z) \rangle, \qquad p \in \mathbb{R}^P, \quad \|\Phi(z)\|_2 \le 1,$$

followed by a 1-Lipschitz loss (e.g., hinge or logistic with appropriate scaling). We claim that the $\sqrt{P/K}$ dependence in prompt ERM bounds cannot be improved in general: there exists a distribution over labeled embeddings (Z, Y) with $Z \in \mathbb{R}^d$ (equivalently, over neighborhoods (G, v) pushed forward by f_θ) such that any learning algorithm that outputs a prompt $\hat{p}$ given K labeled samples satisfies the minimax excess risk lower bound

$$\mathbb{E}\big[\mathcal{R}(\hat{p})\big] - \inf_{\|p\| \le B} \mathcal{R}(p) \;\ge\; c\sqrt{\frac{P}{K}},$$

for an absolute constant $c > 0$ (up to benign dependence on B and Lipschitz constants). Equivalently, guaranteeing excess risk at most ε requires

$$K = \Omega\left(\frac{P}{\varepsilon^2}\right).$$

The proof is standard in statistical learning theory: we choose a packing of prompts $\{p^{(1)}, \ldots, p^{(M)}\}$ with $M \ge \exp(\Omega(P))$ and define distributions $\{\mathbb{P}_j\}_{j=1}^M$ over (Z, Y) such that each $p^{(j)}$ is (nearly) optimal for $\mathbb{P}_j$ while the induced sample distributions have small pairwise KL divergence. An application of Fano's inequality (or Assouad's lemma on the hypercube) shows that no algorithm can reliably identify which $\mathbb{P}_j$ generated the data when $K \ll P$, forcing a constant probability of outputting a prompt whose risk is separated from the optimum by $\Omega(\sqrt{P/K})$. Importantly, this argument already holds when f_θ is "perfect" in the sense that $Z = f_\theta(\mathsf{Sub}_r(G, v))$ is directly observed; thus the lower bound is entirely about the limited labeled-sample regime and the prompt class capacity. This establishes that our prompt-only rates are minimax-optimal (up to constants/log factors) among all methods constrained to update only P real-valued parameters.

Cold-start impossibility without informative attributes. Next we isolate a failure mode that is not a matter of finite-sample complexity but of identifiability. Consider a tail node v with $\deg(v) = 0$ after the shift (or, more generally, whose r-hop neighborhood contains no other nodes). Then $\mathsf{Sub}_r(G, v)$ contains only the node feature vector X_v (and trivial structure). If under $\mathcal{D}_{tail}$ the attribute X_v is independent of the label Y_v, i.e.,

$$\mathbb{P}(Y = y \mid X_v = x) = \mathbb{P}(Y = y) \qquad \text{for all } x,$$

then no predictor based on $\mathsf{Sub}_r(G, v)$ can outperform the label prior. In the balanced binary case, for any (possibly randomized) predictor $\hat{Y} = \mathcal{A}(X_v)$ we have

$$\mathbb{P}(\hat{Y} \neq Y) \;\geq\; \frac{1}{2}.$$

The argument is immediate: since X_v carries no information about Y, the joint distribution factorizes, and the Bayes-optimal decision rule is constant, achieving error equal to the minority-class probability (in particular $1/2$ when balanced). This conclusion holds regardless of whether we use prompting, full fine-tuning, or any other learning rule, and it also persists if we allow access to unlabeled tail features at test time. Thus, any positive result for isolated nodes must assume informative node attributes, auxiliary modalities, or additional side information (e.g., textual profiles, temporal traces, or cross-graph identity links).

Computational hardness of worst-case robust prediction under adversarial deletions. Finally, we emphasize that robustness to arbitrary edge deletions is computationally intractable in the worst case, even when we restrict attention to local neighborhoods. We sketch a reduction from CLIQUE. Fix integers k and t. Given a graph instance H, we construct a larger graph G with a distinguished node $v^\star$ such that the label of $v^\star$ is defined as

$$Y_{v^\star} = \mathbf{1}\{\text{the } r\text{-hop neighborhood of } v^\star \text{ contains a } k\text{-clique}\}.$$

Suppose we require a predictor to be *robust* in the sense that it must output the correct label for $v^\star$ under any deletion of up to t edges in $\mathsf{Sub}_r(G, v^\star)$. Choosing t appropriately, deciding whether there exists an adversarial deletion that flips the label is equivalent to deciding whether a k-clique exists (or can be destroyed) in the encoded neighborhood, which is NP-hard. In particular, computing the robust label (and therefore producing a certifiably robust predictor) would solve CLIQUE in polynomial time. This hardness persists even if we grant the algorithm full access to $\mathsf{Sub}_r(G, v^\star)$ (so the local-access constraint is not the source of difficulty); it is the adversarial quantification over deletions that induces the combinatorial barrier.

Implications. Taken together, these limitations delineate the regime in which our theory is meaningful. The $\Omega(P/\varepsilon^2)$ lower bound explains why parameter-efficiency must be paired with small-P prompt designs when K is tiny. The cold-start impossibility shows that topology-shift robustness cannot be unconditional: if the shift removes all informative structure and attributes are uninformative, no method can succeed. The NP-hardness sketch motivates our use of *dominated stochastic shifts* and stability-based bounds: we target robustness to plausible distributional sparsification, not worst-case adversaries.

10 Experimental Plan

We design experiments to validate the two central claims suggested by the theory: (i) synthetic degree-aware sparsification during self-supervised pre-training reduces sensitivity of frozen embeddings to downstream topology loss, and (ii) under a fixed frozen encoder, prompt-only adaptation exhibits the expected tradeoff between labeled sample size K, prompt dimension P, and performance under head-to-tail shift.

Controlled sparsity-shift benchmarks. We will instantiate a family of semisynthetic benchmarks in which the *same* underlying labeled graph is evaluated under progressively stronger topology loss. Concretely, starting from a standard transductive node classification dataset, we define head nodes as the top degree quantile (structure-rich) and tail nodes as the bottom degree quantile (structure-scarce). We then generate evaluation-time neighborhoods by applying a parametrized shift operator Shift_γ that performs additional edge deletions (and, optionally, feature masking) with strength γ, with higher deletion probability for smaller $\deg(v)$ to emulate cold-start and long-tail effects. For each γ, we report accuracy (or macro-F1 under imbalance) on tail queries while the support set S remains sampled from head nodes. This isolates the shift penalty while keeping label semantics fixed, matching the setting of Theorem 2.

Real-world long-tail graphs and tail evaluation protocol. To ensure external validity, we will evaluate on graphs where the degree distribution is heavy-tailed and where tail nodes are known to be challenging (e.g., product co-purchase/recommendation graphs, citation graphs with new papers, and social graphs with sparse user histories). For each dataset we will define a tail split by degree and/or by temporal cold-start (when timestamps are available), and we will additionally test robustness under explicit post-processing shifts Shift_γ applied only at evaluation time. We will ensure the local-access constraint is respected by computing predictions for each node using only

$\mathsf{Sub}_r(G, v)$, and we will report the dependence on r to quantify the extent to which performance derives from local versus broader structure.

Pre-training ablations: simulator design and schedules. We will ablate the structure-scarcity simulator Sim_η along three axes. First, we compare degree-aware deletion/masking against degree-agnostic deletion, holding the expected sparsity constant, to test whether targeting low-degree regions is necessary to improve tail performance. Second, we vary the corruption family: edge deletions only, feature masking only, and combined deletions+masking, in order to disentangle whether robustness arises primarily from structural invariance or attribute denoising. Third, we study *simulator schedules*: (a) fixed-strength corruption, (b) a curriculum that increases sparsity over training, and (c) a mixture schedule that samples η from a distribution (exposing the encoder to a range of tail severities). These experiments are designed to probe the linearized intuition in Theorem 3: stronger and appropriately dominated synthetic sparsification should reduce measured instability, but overly strong corruption may degrade semantic fidelity and hurt head performance.

Adaptation baselines under equal budgets. We will compare prompt-only adaptation to competing parameter-efficient and full-capacity alternatives, controlling for parameter count and optimization budget. Baselines will include: (i) full fine-tuning of the encoder (upper bound on adaptation capacity), (ii) linear probing on frozen embeddings, (iii) adapters or LoRA-style low-rank updates restricted to a comparable number of trainable parameters, and (iv) meta-learning style baselines that explicitly optimize for fast adaptation across pre-training tasks (where feasible). For prompt-only methods, we will sweep the prompt budget P and the norm constraint (or an equivalent regularizer) to empirically trace the capacity–generalization curve predicted by the $\sqrt{P/K}$ scaling. We will additionally evaluate sensitivity to the support set size K by varying K over a range that includes the genuinely few-shot regime.

Stability proxy measurement and correlation with accuracy under shift. To connect practice to the stability-based bound, we will measure an empirical proxy for $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ without requiring labels: for a sample of nodes (G, v), we compute

$$\widehat{\mathrm{Stab}}_\gamma(f_\theta) \;=\; \mathbb{E}\big[\|f_\theta(\mathsf{Sub}_r(G, v)) - f_\theta(\mathsf{Sub}_r(\mathsf{Shift}_\gamma(G), v))\|_2\big],$$

estimated by Monte Carlo over nodes and shift randomness. We will then correlate $\widehat{\mathrm{Stab}}_\gamma(f_\theta)$ with tail accuracy across (a) different pre-training schemes, (b) different simulator schedules, and (c) different shift strengths γ. The

intended test is not merely that lower instability coincides with higher accuracy, but that the *relative* ranking of methods under increasing γ is explained by their stability curves, as suggested by Theorem 2. We will also verify that improvements in stability induced by Sim_η are concentrated on low-degree nodes, consistent with the degree-aware design.

Scaling checks implied by theory. Finally, we will conduct targeted scaling experiments to test whether excess risk empirically follows the predicted dependence on K and P. Holding the frozen encoder fixed, we will fit a simple model of performance as a function of $\sqrt{P/K}$ (with dataset-dependent constants) and check whether prompt-only methods saturate at small K in a manner consistent with the lower bound. We will report both head and tail performance to expose the tradeoff between invariance and expressivity, and we will include calibration and robustness metrics (e.g., ECE under shift) to capture failure modes not visible in accuracy alone.

11 Discussion and Limitations

Our analysis and experimental plan are organized around a specific mechanism: by exposing the encoder to degree-aware synthetic topology loss through Sim_η during self-supervised pre-training, we aim to decrease $\mathrm{Stab}(f_\theta; \mathsf{Shift})$, thereby reducing the head-to-tail shift penalty in Theorem 2 while retaining sufficient semantic fidelity for prompt-only adaptation. This suggests a concrete set of conditions under which synthetic sparsity is expected to help, as well as regimes where it may hurt.

When synthetic sparsity helps. Synthetic sparsification is most beneficial when the downstream shift is well-modeled by the simulator family, in the sense that Shift is *dominated* by (or at least statistically close to) Sim_η on the relevant neighborhoods. In such cases, the contrastive consistency term directly penalizes representation changes induced by the same kinds of edge deletions and feature masking that appear at evaluation time, and the generative component provides an auxiliary pressure to preserve information that can be recovered from partially observed neighborhoods. Intuitively, we obtain an encoder whose embeddings vary smoothly with respect to edge/feature removal in low-degree regions, and prompt-only adaptation can exploit this smoothness because it is not required to relearn invariances under severe label scarcity. This regime corresponds to the motivating long-tail setting in which tail nodes are not semantically different, but merely structurally under-observed.

When synthetic sparsity hurts. The same mechanism can degrade performance if the simulator pushes beyond the invariances we actually desire.

Overly strong corruption may erase label-relevant signals that are *not* redundant within $\mathsf{Sub}_r(G, v)$, especially in tasks whose Bayes-optimal rule depends on specific motifs, rare edges, or high-order interactions that are fragile under deletion. In our bound, such degradation is not visible through $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ alone: a representation can be stable yet uninformative. Practically, this manifests as a head-accuracy drop (semantic underfitting) or as tail degradation when the remaining attributes are weak (cf. Corollary 5). We therefore view simulator design as a bias–variance tradeoff: increasing synthetic sparsity typically reduces the shift term, but can increase approximation error by collapsing distinct neighborhoods. This also clarifies a limitation of the linearized argument in Theorem 3: minimizing a squared perturbation objective can drive W toward low-norm, low-sensitivity solutions that may discard discriminative directions unless counterbalanced by objectives that preserve task-relevant information.

Mismatch beyond topology loss. Our theory is explicitly tailored to topology and attribute *missingness* (edge deletions and feature masking). Many real shifts are not well-approximated by such operators: edges may be *rewired* rather than removed; the semantics of edges may change across domains; node/edge types may appear or disappear; and label definitions may drift. In these cases, the stability term can be small while the risk shift is large because the mapping from neighborhoods to labels has changed. Conversely, stability can be large even when the downstream task is robust, if the encoder is sensitive to nuisance variations that the prompt template can easily ignore. Thus, $\mathrm{Stab}(f_\theta; \mathsf{Shift})$ should be interpreted as a sufficient, not necessary, control term, and our empirical correlation tests are intended to delineate the range in which it is predictive.

Cross-domain graphs. When transferring across graphs (e.g., from one platform or domain to another), the primary challenge is often a combination of topology shift, feature shift, and distribution shift in subgraph patterns. Degree-aware deletion is a plausible component of such shifts, but may be insufficient. A natural extension is to augment Sim_η with domain-inspired perturbations (feature normalizations, type masking, subgraph resampling, or edge-type dropout) and to measure stability with respect to a richer family of shifts. However, stronger augmentation families also increase the risk of learning representations invariant to domain-specific but label-relevant signals. From the perspective of Theorem 2, we would like a simulator family that is broad enough to upper-bound the anticipated deployment shifts while remaining narrow enough to avoid collapsing the label information needed by any plausible downstream prompt.

Complex graphs: heterogeneous and dynamic settings. Our notation suppresses edge features, types, and time; consequently, the current stability notion is incomplete for heterogeneous graphs (multiple node/edge types) and dynamic graphs (temporal edges, evolving attributes). In heterogeneous graphs, degree alone may not capture structural scarcity: a node may have high total degree but sparse degree in the *relevant* relation type. Degree-aware simulators should therefore be refined to type-aware sparsification, and the prompt template may need to encode type-conditional prototypes or relation-specific prompts. In dynamic graphs, the relevant local view is time-indexed, and missingness can be coupled with recency effects; here, a simulator that deletes edges uniformly over time may be miscalibrated. Extending our framework would require a temporally conditioned Sim_η and a stability definition that couples (G_t, v) with shifted histories. These extensions appear conceptually straightforward but introduce practical complications in pre-training objectives and in local-access inference, since time windows and neighborhood sampling must be carefully specified.

Local-access deployment considerations. We assume a strict local-access model at test time. This is appropriate for settings where full-graph passes are infeasible (latency, privacy, or streaming constraints), but it limits the class of computable decision rules, and it can exacerbate cold-start impossibility when attributes are weak. Moreover, local access induces variance through neighborhood sampling and can interact with Shift in subtle ways (e.g., a sampler that under-represents rare neighbors behaves like an additional shift). A deployment-oriented direction is to study resource-aware variants of stability that account for sampling noise and caching strategies, and to characterize when a small number of additional queries (e.g., adaptive expansion beyond radius r) yields disproportionate gains for tail nodes.

Interpretable prompt-graphs and open directions. Finally, prompt-only adaptation is appealing not only for parameter efficiency but also for interpretability if prompts are structured as small graphs or prototype sets. A concrete open problem is to design prompt-graphs whose nodes/edges admit semantic alignment to substructures in $\mathsf{Sub}_r(G, v)$, enabling explanations in terms of matched motifs or relation-specific evidence. This suggests combining (i) constrained prompt parameterizations (e.g., sparse prototypes, typed prompt nodes) with (ii) stability-aware pre-training so that the same prompt remains meaningful under topology loss. More broadly, we regard the central question as one of *calibrating invariance*: how to choose Sim_η so that it matches the deployment shifts we care about, while preserving the fine-grained information that prompts must leverage when only K labels are available.