One Backbone, Many Graphs (Without Text): Spectral-Alignment Prompts for Multi-Domain Graph Pre-Training

Liz Lemma Future Detective

January 20, 2026

Abstract

Few-shot learning on graphs increasingly requires cross-domain transfer: a single model must adapt to many graph domains (social, e-commerce, academic, biochemical) with different topologies and feature statistics, often without reliable text. Motivated by the recent taxonomy of few-shot learning on graphs (meta-learning vs. pre-training vs. prompting), we focus on the setting where labeled base tasks are scarce and domain shift is primarily structural. We propose a text-free multi-domain pre-training framework that learns (i) a shared graph encoder backbone and (ii) compact per-domain prompts (routing tokens) that explicitly correct structural shift by aligning spectral/topological statistics across domains. Our method uses self-supervised pre-training (contrastive and masked generative objectives) augmented with a spectral alignment regularizer: prompt-conditioned spectral filters are learned so that prompt-filtered representation covariances (and optionally degree/Laplacian moments) match across domains. We formalize a clean problem where domain shift arises from domain-specific Laplacian eigenvalues with an approximately shared eigenbasis, and we prove upper bounds showing prompt size m controls approximation error while few-shot labeled samples K control estimation error $\tilde{O}(\sqrt{(m/K)})$. We complement this with matching lower bounds on prompt size and impossibility results when no shared spectral structure exists. Experiments (to strengthen the contribution) would construct a multi-domain benchmark with synthetic spectral shifts and real domains, evaluating few-shot transfer, catastrophic forgetting, and prompt-size/accuracy tradeoffs against single-domain pre-training, naive multi-domain pooling, mixture-of-experts, and text-based alignment baselines.

Table of Contents

2. 2. Related Work: few-shot learning on graphs; self-supervised pre-training; prompt tuning and PEFT; multi-domain and cross-domain graph pre-training; degree bias and structure scarcity; contrast with text-attributed/LLM-based alignment.

3. 3. Problem Setup and Computational Model: domains, unlabeled pre-training access, few-shot labeled adaptation; formal objectives; what "text-free" means; evaluation protocols (transfer + forgetting).

4. 4. Spectral-Shift Model (Clean Theoretical Sandbox): commuting-Laplacian hypothesis; spectral feature tasks; prompt as spectral filter; discussion of realism and where assumptions break.

5. 5. Method: Spectral-Alignment Prompt Pre-Training: backbone architecture; prompt parameterization; spectral/statistical alignment regularizer; combined contrastive+generative objective; practical polynomial approximation of spectral filters.

6. 6. Main Theorems (Upper Bounds): approximation error vs prompt size; few-shot estimation error; robustness to approximate commuting (δ); extension to unseen domains with unlabeled adaptation.

7. 7. Lower Bounds and Impossibility: prompt-size lower bounds under shared basis; impossibility when domains do not share approximate eigenbasis; information-theoretic/sample lower bounds for learning prompts from unlabeled graphs.

8. 8. Complexity and Scalability: training and adaptation costs; memory; prompt routing vs mixture-of-experts; local subgraph vs full-graph considerations.

9. 9. Experimental Plan (Recommended): benchmark construction; baselines; ablations isolating spectral alignment; prompt size scaling; cross-domain generalization; catastrophic forgetting under sequential domain arrival.

10. 10. Discussion and Future Work: beyond commuting-Laplacian; heterophily and dynamic/heterogeneous graphs; discrete/structured prompts; connections to foundation models and interpretability.

11. 11. Conclusion.

1 Introduction

We consider the regime of *cross-domain* few-shot learning on graphs in which a practitioner has access to many unlabeled graphs drawn from several domains, yet must subsequently solve downstream prediction problems in each domain from only a small labeled set. Unlike settings where graphs come with rich textual annotations (e.g., documents, knowledge graphs with descriptions), we focus on the *text-free* case: node and edge attributes are either absent or purely numeric and non-linguistic, so there is no external semantic channel through which one may align domains. Consequently, the dominant obstacle to transfer is not primarily label shift but rather *topology shift*: domains exhibit different connectivity patterns, degree distributions, and spectral profiles, which can cause a graph encoder trained on one domain to induce a representation geometry poorly matched to another.

The standard approach to amortize representation learning across tasks is self-supervised pre-training of a single graph neural network or graph transformer. However, in multi-domain corpora, a single shared encoder tends to overfit to the most frequent structural motifs and spectral scales present in the pooled data. If domains differ mainly in their graph Laplacians, a naive pooling strategy implicitly selects a single frequency response and thereby privileges some domains at the expense of others. On the other hand, training and storing a separate backbone per domain is parameter-inefficient and offers no mechanism for adaptation to new domains with few labels. Our objective is to isolate a setting in which we can provably obtain (i) a *single* shared backbone and (ii) a *compact* per-domain adaptation mechanism that is learnable from limited labeled data, while still providing quantitative guarantees on downstream risk.

We propose to represent domain-specificity by a small vector $p_d \in \mathbb{R}^m$ that parameterizes a *spectral reweighting* of the domain Laplacian. Operationally, p_d defines a polynomial filter $g_{p_d}(L)$ (e.g., via a Chebyshev recurrence) that is inserted into the message-passing operator of a frozen backbone f_θ. Thus, the only domain-dependent component is a low-dimensional prompt controlling the frequency response of aggregation. This choice is motivated by the observation that many topological discrepancies between domains—including degree bias, mixing time, and homophily/heterophily structure—are reflected in the spectrum of the normalized Laplacian. A polynomial spectral filter is a particularly convenient class: it is expressive enough to approximate broad families of smooth transfer functions while remaining computationally compatible with sparse graph primitives, and its effective capacity is controlled by the degree m, which we treat as the prompt budget.

Prompting alone does not specify *how* different domains should be mapped into a representation space that admits a shared downstream inductive bias. Accordingly, we introduce an alignment principle that is also text-free: dur-

ing pre-training we add a regularizer that encourages *prompt-filtered* embedding statistics to match across domains. Concretely, let Σ_d denote a second-moment statistic (e.g., covariance) of embeddings produced by f_θ on domain d under prompt p_d. We penalize a distance between suitably normalized versions of Σ_d and $\Sigma_{d'}$ for sampled domain pairs (d, d'). This spectral/statistical alignment is combined with standard self-supervised losses (contrastive objectives and masked reconstruction), yielding a two-stage protocol: offline pre-training over unlabeled multi-domain graphs, followed by downstream adaptation in which θ is frozen and only p_d (and optionally a small linear head) is optimized on K labeled examples.

Our analysis is organized around a structural hypothesis that makes precise when such a prompting scheme should be sufficient. We assume an approximate *commuting-Laplacian* condition: the Laplacians L_d of different domains are simultaneously (approximately) diagonalizable, meaning there exists a shared orthogonal basis U such that

$$L_d \;=\; U\Lambda_d U^\top \;+\; \Delta_d, \qquad \|\Delta_d\|_2 \le \delta,$$

where Λ_d is diagonal with entries in $[0, 2]$. This condition does not assert that the graphs are identical; rather, it asserts that domains share latent "eigendirections" while differing primarily in the weighting of those directions. Under this hypothesis, a domain prompt that implements a polynomial g_{p_d} can approximate a domain-dependent spectral transfer ψ_d by choosing coefficients so that $g_{p_d}(\lambda) \approx \psi_d(\lambda)$ uniformly on $[0, 2]$. The role of alignment is to encourage the learned prompts to select compatible frequency responses so that embeddings across domains inhabit a common geometry, thereby making few-shot adaptation statistically efficient.

Our contributions are as follows.

(1) A parameter-efficient, text-free multi-domain pre-training objective. We formalize a pre-training procedure in which domain prompts parameterize spectral filters inside a shared backbone, and an explicit alignment regularizer matches prompt-conditioned embedding moments across domains. The method is designed so that, at downstream time, adaptation is restricted to m prompt parameters (and possibly a small linear head), with memory $O(Dm)$ across D domains. This is the graph analogue of parameter-efficient tuning, but without relying on textual side information or language-based alignment.

(2) Approximation guarantees for spectral prompting. In the idealized case $\delta = 0$, we show that if a downstream task depends on a band-limited (or Lipschitz) spectral transfer function ψ_d, then there exists a degree-m polynomial prompt filter that approximates ψ_d to error ε with

$$m \;=\; \widetilde{O}\big(\log(1/\varepsilon)\big),$$

and that the resulting operator approximation yields an $O(\varepsilon)$ contribution to excess risk. The proof reduces to Chebyshev approximation on $[0, 2]$ and a stability-to-risk argument translating uniform spectral error into prediction error for Lipschitz functionals.

(3) Few-shot excess risk bounds for prompt-only adaptation. When $\delta \geq 0$, the Laplacian perturbations Δ_d introduce an irreducible mismatch between the assumed shared basis and the actual domain Laplacians. We quantify this effect by showing an $O(\delta)$ contribution to excess risk via standard spectral perturbation inequalities. Combining approximation, perturbation, and estimation yields a bound of the form

$$\mathbb{E}\big[\mathrm{Risk}_d(\widehat{p}_d)\big] - \mathrm{Risk}_d(p_d^\star) \;\leq\; O(\varepsilon + \delta) \;+\; \widetilde{O}\left(\sqrt{\frac{m}{K}}\right),$$

for prompt-only adaptation from K labeled samples, under a linear downstream predictor in the prompt-filtered embedding space. The term $\sqrt{m/K}$ is the expected estimation error for an m-parameter family, emphasizing that prompt tuning can be statistically preferable to full fine-tuning when K is small.

(4) Matching lower bounds and structural necessity. We complement the upper bounds with two negative results that clarify what can and cannot be achieved in this setting. First, even when the shared eigenbasis U is known and $\delta = 0$, there exist Lipschitz spectral transfer families for which any degree-m polynomial prompt achieving uniform error ε requires $m = \Omega(\log(1/\varepsilon))$, matching the approximation rate above. Second, if the commuting-Laplacian hypothesis fails in a strong sense—the eigenbases are sufficiently misaligned across domains—then any method that keeps a fixed backbone and adapts only through $m = o(n)$ prompt parameters applied as spectral filters (or comparable diagonal modulations in a fixed basis) can be forced to incur constant excess risk on at least one domain. This impossibility result formalizes the intuition that low-dimensional prompting cannot simultaneously undo arbitrary cross-domain rotations of the spectral geometry.

(5) A new-domain adaptation perspective from unlabeled data. Finally, we outline how prompts for a previously unseen domain may be inferred from unlabeled graphs by minimizing an alignment loss based on estimated Laplacian moments or embedding moments. We provide sample complexity bounds showing that $N = \widetilde{O}((m+\log(1/\rho))/\varepsilon^2)$ unlabeled graphs suffice to achieve alignment error ε with probability at least $1 - \rho$, and that $N = \Omega(m/\varepsilon^2)$ is unavoidable in general.

In sum, the technical message of this work is that, in the absence of text, cross-domain transfer on graphs admits a natural spectral organization. If domains share an approximate eigenbasis, then compact polynomial prompts can reweight frequencies so that a single backbone supports multiple domains with few-shot prompt adaptation. The resulting guarantees explicitly separate approximation, structural perturbation, and finite-sample estimation effects, thereby clarifying how the prompt budget m, the few-shot size K, and the spectral mismatch δ jointly control downstream risk.

2 Related Work

Few-shot learning on graphs. Few-shot prediction on graphs has been studied under several paradigms, including meta-learning across tasks, metric learning in an embedding space, and transductive label propagation on novel classes. Graph meta-learning methods adapt a base learner using a small support set, often via gradient-based meta-learning (e.g., MAML-style objectives) or learned update rules **?**. In the graph setting, this includes approaches that meta-learn message passing parameters, pooling operators, or node-level label propagation mechanisms, typically assuming multiple supervised tasks during meta-training **??**. Metric-based formulations build class prototypes from a few labeled nodes/graphs and classify queries by distances in a learned embedding space **?**; graph-specific instantiations combine a GNN encoder with prototype aggregation or relation networks to handle neighborhood effects **??**. A recurring limitation is that many few-shot graph methods assume either (i) access to many labeled tasks during training, or (ii) a single-domain distributional match between meta-train and meta-test tasks. Our focus differs in that we treat *domains*—not merely tasks—as the primary axis of variation, and we explicitly target the regime where multi-domain data are available primarily in *unlabeled* form, with only few-shot labels per domain at downstream time.

Self-supervised and masked pre-training for graphs. Self-supervised learning (SSL) is a central mechanism for amortizing representation learning on graphs when labeled data are scarce. Contrastive objectives maximize agreement between multiple views of a graph or node neighborhood, typically created by stochastic augmentations such as edge dropping, feature masking, or subgraph sampling **???**. Generative and reconstruction-based objectives predict masked node/edge attributes, recover adjacency structure, or reconstruct latent codes, including masked autoencoding variants **????**. These methods usually target *single-domain* corpora, and their transfer guarantees (when studied) often rely on assumptions about augmentation invariances, smoothness over the graph, or stability of embeddings under perturbations. In the multi-domain setting, naive pooling of graphs across

domains can implicitly bias the learned encoder toward the dominant spectral scales and structural motifs present in the aggregated data, leading to suboptimal transfer to underrepresented domains. This motivates objectives that explicitly account for cross-domain heterogeneity during pre-training, rather than treating it as additional i.i.d. variability.

Parameter-efficient fine-tuning and prompt tuning. Parameter-efficient fine-tuning (PEFT) methods, developed primarily in the context of large language models, restrict adaptation to low-dimensional parameter subsets such as adapters, prefix/prompt vectors, or low-rank updates **???**. The guiding principle is that a shared backbone can be reused across many tasks, while task-specific vectors with small dimension suffice to steer the computation. Graph counterparts of prompting have appeared in several forms: (i) *hard prompts* that select or attach a small set of virtual nodes/edges or subgraphs to condition the encoder, (ii) *soft prompts* that introduce trainable tokens or vectors injected into node representations, and (iii) prompt-style reformulations of graph problems into masked prediction or template-based objectives **??**. Most existing graph prompt constructions are defined in the vertex/feature domain (e.g., additional node embeddings or feature offsets), whereas our interest is in prompts that act as *spectral controls* on message passing through polynomial filters. This distinction matters in the multi-domain, text-free setting: spectral prompting targets topological shift directly by reweighting frequencies of the Laplacian, and thus admits a natural capacity control via the polynomial degree.

Multi-domain, domain adaptation, and domain generalization on graphs. Domain adaptation on graphs has been explored for node classification and graph classification under covariate shift, label shift, and structural shift, using discrepancy minimization, adversarial alignment, and invariant representation learning **???**. Related work in domain generalization seeks representations that perform well on unseen domains without test-time adaptation, often by enforcing invariances across training domains or by episodic training **??**. In graphs, domain shift is complicated by the interplay between features and topology; in particular, differences in degree distributions, homophily levels, and connectivity patterns can induce substantial changes in the spectrum of the Laplacian and in the stability of message passing **??**. Several works study out-of-distribution generalization in graph neural networks and propose regularizers based on causal invariance, stable neighborhood aggregation, or subgraph-based augmentations **??**. Our formulation can be viewed as complementary: rather than insisting on strict invariance of the encoder across domains, we allow *controlled* domain-specific modulation via prompts, but restrict this modulation to be low-dimensional and structured (spectral filters), thereby enabling few-shot adaptation while

maintaining a single shared backbone.

Cross-domain graph pre-training and alignment objectives. A small but growing body of work considers pre-training on heterogeneous graph corpora spanning multiple sources. When domain labels are available, one can train domain-specific heads or use multi-task objectives; when domain labels are absent, methods often rely on clustering or implicit mixture modeling. Alignment objectives typically match embedding distributions across domains by moment matching (e.g., MMD), adversarial training, contrastive cross-domain pairing, or consistency constraints **??**. In graphs, such alignment is frequently implemented at the node-embedding level, sometimes assuming comparable feature semantics across domains. Our setting differs in two respects. First, we do not assume a shared semantic channel such as text, ontologies, or tokenized descriptions. Second, we emphasize alignment in a *prompt-conditioned* representation space, so that the alignment signal directly shapes the low-dimensional domain parameters rather than encouraging the backbone alone to absorb all cross-domain variability.

Degree bias, structural heterogeneity, and structure scarcity. A recurring empirical issue in GNNs is sensitivity to degree distributions and local topology. Standard message passing with normalized adjacency can induce degree-dependent smoothing and oversquashing effects, and the learned representation may systematically favor hubs or dense regions **??**. Across domains, shifts in degree distribution or mixing properties can therefore yield large changes in effective receptive fields and in the frequency response of aggregation operators. Moreover, in many text-free domains, node features are weak and the topology carries most of the predictive signal; conversely, in other domains, the topology may be sparse or noisy ("structure scarcity"), forcing models to rely more heavily on features or to learn robust structural priors **??**. These observations motivate adaptation mechanisms that can tune the aggregation behavior without relearning a full backbone. Spectral filtering provides a direct handle: it allows one to emphasize or suppress particular Laplacian frequencies that correspond to different smoothness regimes, which is closely related to classical analyses of graph signal processing and to recent studies of oversmoothing and heterophily in GNNs **???**.

Contrast with text-attributed and LLM-based alignment. Recent progress in graph learning increasingly leverages textual node/edge descriptions, captions, and documentation to align heterogeneous graphs or to transfer knowledge across domains using language models **??**. In such settings, the text channel provides a high-capacity, domain-agnostic semantic anchor, and prompting can be performed in language space with strong priors learned from web-scale corpora. While powerful, these approaches are inapplicable

in the strictly text-free regime we study, where attributes are numeric or absent and there is no external semantic alignment signal. Our emphasis is therefore on *structural* alignment mechanisms that operate directly on graph operators (e.g., Laplacians) and on representation statistics induced by the encoder, rather than on alignment through shared token semantics. This contrast is substantive: without text, cross-domain transfer must be achieved by exploiting regularities in topology and spectrum, which leads naturally to the structured prompt class and alignment regularizer developed in this work.

3 Problem Setup and Computational Model

We study text-free multi-domain graph representation learning with parameter-efficient downstream adaptation. There are D domains indexed by $d \in \{1, \dots, D\}$, and each domain induces a distribution $\mathcal{P}_d$ over attributed graphs

$$G = (V, E, X), \qquad X \in \mathbb{R}^{|V| \times f}, \qquad |V| \le n,$$

where X collects real-valued node attributes (possibly absent, in which case we take X to be a constant feature) and n is a global size bound used for padding in analysis. We write $G_i^{(d)} \sim \mathcal{P}_d$ for unlabeled samples and denote by N_d the number of unlabeled graphs available from domain d during pre-training. Throughout, "multi-domain" refers to heterogeneous sources of graphs whose topology and feature distributions may differ substantially across d; we do not assume shared label spaces across domains, and we allow downstream tasks to be domain-specific.

Text-free regime. By *text-free* we mean that no textual channel is available for alignment or prompting: nodes and edges do not carry natural-language descriptions, tokens, or externally pre-trained embeddings derived from text corpora. All training signals are obtained from the graph structure (e.g., adjacency, Laplacian) and numeric attributes (if present), together with self-supervised objectives on these quantities. In particular, we do not use a language model, do not require paired text–graph data, and do not assume that feature dimensions have shared semantics across domains beyond being bounded real vectors. This restriction rules out alignment mechanisms that rely on a universal semantic space, and it forces transfer to be achieved through structural and statistical regularities.

Two-stage learning and access patterns. Our computational model has two stages.

Stage I (offline pre-training). We are given an unlabeled multi-domain corpus $\{G_i^{(d)}\}_{i=1}^{N_d}$ for each d. Pre-training proceeds by mini-batch sampling

of graphs (or subgraphs) from one or more domains. Depending on the setting, the domain identity d may be observed during pre-training, or it may be latent; when latent, we treat d as an unobserved source index and allow the learner to infer a prompt assignment by unsupervised clustering or mixture modeling. In either case, the backbone encoder parameters are shared globally, and domain-specificity is constrained to enter only through low-dimensional prompt parameters.

Stage II (few-shot downstream adaptation). For each domain d and each downstream task of interest, we receive a labeled dataset S_d of size K (few-shot), drawn from a domain-specific task distribution. The task may be node-level, edge-level, or graph-level prediction; in all cases, we evaluate expected risk with respect to fresh test samples from the same domain and task. During adaptation we freeze the backbone and update only a compact set of trainable parameters: a domain prompt $p_d \in \mathbb{R}^m$ (and optionally a small linear head). This formalizes parameter-efficient fine-tuning in the multi-domain regime: a single shared representation is amortized across domains, while each domain is allowed a bounded-capacity modulation.

Model class: shared backbone and domain prompts. Let f_θ denote a shared graph encoder with parameters θ (e.g., a message-passing GNN or graph transformer). For each domain d we maintain a prompt vector $p_d \in \mathbb{R}^m$ with a fixed budget m. Operationally, p_d controls a prompt-parameterized operator $g_{p_d}(\cdot)$ that modulates message passing; in the specific instantiation studied later, $g_{p_d}(L)$ is a degree-m polynomial (or truncated Chebyshev expansion) of a graph Laplacian L. We emphasize here only the *computational interface*: given a graph G, the encoder produces embeddings

$$h \;=\; f_\theta\big(G;\, g_{p_d}(L(G))\big),$$

where h may denote node embeddings, edge embeddings, or a pooled graph embedding depending on the task. The key constraint is that domain-specificity is confined to p_d (and a small head), while θ is shared and, at downstream time, frozen.

Self-supervised pre-training objectives. Pre-training uses only unlabeled graphs and a self-supervised loss ℓ_{ssl}. We allow ℓ_{ssl} to combine contrastive and generative components, such as view-invariant agreement under stochastic augmentations and masked attribute/structure reconstruction. Since our setting is multi-domain, we additionally consider an alignment regularizer $\mathbb{R}_{\text{align}}$ that couples domains through prompt-conditioned statistics of representations. Abstractly, the pre-training objective is

$$\min_{\theta,\{p_d\}} \;\mathbb{E}_{d\sim\pi}\,\mathbb{E}_{G\sim\mathcal{P}_d}\big[\ell_{\text{ssl}}(f_\theta(G;g_{p_d}),G)\big] \;+\; \lambda\,\mathbb{R}_{\text{align}}(\theta,\{p_d\}),$$

where π is a sampling distribution over domains and $\lambda \geq 0$ is a weight. The role of $\mathbb{R}_{\text{align}}$ is not to enforce strict invariance of h across domains, but to encourage that the *prompted* representation geometry is comparable across domains in a sense that supports low-sample adaptation. A concrete instantiation, used later, matches second-order moments (e.g., covariances) of embeddings computed under domain prompts.

Downstream adaptation and parameter-efficiency. In domain d, few-shot adaptation solves a supervised empirical risk minimization problem over the prompt (and optionally a linear head). Let $\mathcal{L}_{\text{sup}}$ denote a supervised loss on labeled examples $(G, y) \in S_d$. We update

$$(p_d, w_d) \;\in\; \arg\min_{p\in\mathbb{R}^m,\, w\in\mathbb{R}^q} \; \frac{1}{K} \sum_{(G,y)\in S_d} \mathcal{L}_{\text{sup}}\Big(w^\top f_\theta(G; g_p),\, y\Big),$$

subject to a prompt budget constraint (e.g., $\|p\|_2 \leq B$ or an effective degree constraint for the polynomial filter). The backbone parameters θ are held fixed. This separation is essential in our setting: since we evaluate on many domains, full fine-tuning would scale storage and optimization cost with $D|\theta|$, whereas prompt tuning scales as $O(Dm)$ with $m \ll |\theta|$.

Formal objective in terms of risk. Let $\text{Risk}_d(\cdot)$ denote the population risk in domain d for the downstream task under the evaluation distribution (node/edge/graph as appropriate). We write $\text{Adapt}(f_\theta, p_d; S_d)$ for the adapted predictor obtained by optimizing only (p_d, w_d) on S_d. Our goal is to find a single shared backbone and prompts that minimize the average (or worst-case) post-adaptation risk:

$$\min_{\theta,\{p_d\}} \; \mathbb{E}_{d\sim\pi}\Big[\text{Risk}_d\big(\text{Adapt}(f_\theta, p_d; S_d)\big)\Big] \quad \text{s.t.} \quad \dim(p_d) = m \;\; \forall d,$$

with the understanding that S_d is a random K-sample and adaptation uses only these labels. In addition to average-case performance, we will later track excess risk relative to the best domain-specific predictor within an appropriate hypothesis class, thereby isolating the statistical cost of few-shot prompt estimation and the approximation cost induced by the prompt budget.

Evaluation protocols: transfer and forgetting. We consider two evaluation regimes.

(i) Multi-domain transfer. We pre-train once on unlabeled graphs from all training domains and then, for each domain d, adapt using K labeled examples and report test performance on held-out data from the same domain. We report metrics aggregated across domains (mean and, when relevant, worst-domain performance), and we vary K to quantify the sample-efficiency conferred by prompt-only adaptation. When domain ID is not

provided, we additionally evaluate the effect of inferred prompt assignment on downstream risk.

(ii) Continual domain arrival and forgetting. Domains may arrive sequentially, and we may add prompts over time without revisiting old labeled datasets. In this regime we freeze θ after pre-training and maintain a growing collection of prompts $\{p_d\}$. Upon receiving a new domain d', we initialize and adapt $p_{d'}$ using either few-shot labels or, when labels are unavailable, an unlabeled alignment objective. We measure (a) *forward transfer*, i.e., performance on the new domain after adaptation, and (b) *forgetting*, i.e., performance degradation on previously seen domains when introducing $p_{d'}$ and possibly updating shared components. Our default design avoids forgetting by construction, since θ is frozen and prompts are domain-local; nevertheless, we include forgetting metrics to quantify any interference introduced by shared normalization layers, shared batch statistics, or joint prompt regularization.

This section specifies only the learning interface and evaluation rules. In the next section we introduce a clean spectral-shift model under which the prompt class admits explicit approximation and sample-complexity guarantees.

4 Spectral-Shift Model: a Clean Sandbox for Prompted Transfer

We now introduce an idealized model in which domain shift is expressed primarily as a *spectral reweighting* of a shared latent graph geometry. The purpose of this section is not to claim that real multi-domain corpora exactly satisfy such a model, but to isolate a regime in which prompt-only adaptation admits explicit approximation and sample-complexity statements. The model will also make clear where and why prompt-only transfer can fail.

Normalized Laplacians and padding. For a graph $G = (V, E, X)$ with $|V| \le n$, let $L(G) \in \mathbb{R}^{n\times n}$ denote the (padded) normalized Laplacian. Concretely, if $\bar{L}(G) \in \mathbb{R}^{|V|\times|V|}$ is the usual normalized Laplacian on the observed nodes, we embed $\bar{L}(G)$ into an $n \times n$ matrix by adding isolated padded nodes; this is purely an analytical device to place all domains in a common ambient space. We emphasize that no eigen-decomposition is assumed available computationally.

Commuting-Laplacian hypothesis. The central structural assumption is that domains share (approximately) a common eigenbasis. Formally, we posit that there exists an orthogonal matrix $U \in \mathbb{R}^{n\times n}$ such that each domain

admits a representative Laplacian random variable L_d satisfying

$$L_d \;=\; U\Lambda_d U^\top + \Delta_d, \qquad \|\Delta_d\|_2 \le \delta, \qquad \Lambda_d = \mathrm{diag}(\lambda_{d,1}, \ldots, \lambda_{d,n}),\ \lambda_{d,i} \in [0,2]. \tag{1}$$

When $\delta = 0$, the family $\{L_d\}_{d=1}^D$ is simultaneously diagonalizable and hence commutes. When $\delta > 0$, the hypothesis asserts that the domain-specific eigenvectors remain close to a common basis and that the main variation across domains is in the spectrum Λ_d (up to a controlled perturbation). We interpret (1) as an abstraction of settings in which domains share latent geometry (e.g., similar motif structure or common generative mechanisms) but differ in coarse structural statistics such as degree profiles, edge densities, or homophily strength, each of which induces a shift in spectral content.

Graph signals and spectral features. Let $\phi(X) \in \mathbb{R}^{n\times r}$ be a feature lifting map applied nodewise, producing an r-dimensional signal on the n padded nodes. In the commuting regime, the coordinates

$$\widehat{S} \;:=\; U^\top \phi(X) \in \mathbb{R}^{n\times r}$$

play the role of "graph Fourier" coefficients relative to the shared basis. If $\delta = 0$, the operator L_d acts diagonally in this basis:

$$U^\top L_d U \;=\; \Lambda_d.$$

Thus, any polynomial in L_d corresponds to a pointwise transformation of eigenvalues in the spectral domain.

Downstream tasks as spectral transfer functions. We model the downstream label as depending on a low-complexity transformation of spectral features. For simplicity we describe a binary prediction setting; analogous statements can be made for regression or multiclass linear models. Let $\psi_d : [0,2] \to \mathbb{R}$ be a domain-specific *spectral transfer function*. Given a graph signal $\phi(X)$, define the transformed signal

$$S_d \;:=\; U\,\psi_d(\Lambda_d)\,U^\top \phi(X), \qquad \psi_d(\Lambda_d) = \mathrm{diag}(\psi_d(\lambda_{d,1}), \ldots, \psi_d(\lambda_{d,n})).$$

We assume the Bayes-optimal decision rule in domain d is linear in S_d:

$$y \;=\; \mathrm{sign}\big(\langle w_d^\star,\ \mathrm{vec}(S_d)\rangle + \xi\big), \tag{2}$$

where $w_d^\star$ is an unknown weight vector of bounded norm and ξ is noise. The task model (2) captures the idea that labels depend on particular frequency bands: for instance, community-level properties often correlate with low-frequency components, whereas anomaly-like signals may be high-frequency. To enable approximation results, we assume ψ_d is Lipschitz on $[0,2]$ and, in the strongest form used later, that the relevant dependence is effectively band-limited or well-approximable by low-degree polynomials.

Prompts as spectral filters. The prompt mechanism we analyze is a spectral reweighting applied to message passing. Abstractly, a prompt $p_d \in \mathbb{R}^m$ determines a filter

$$g_{p_d}(L) \;=\; \sum_{k=0}^{m} a_k(p_d)\, L^k, \tag{3}$$

or, more stably, a Chebyshev expansion on a rescaled Laplacian $\widetilde{L}$:

$$g_{p_d}(L) \;=\; \sum_{k=0}^{m} a_k(p_d)\, T_k(\widetilde{L}), \qquad \widetilde{L} := \frac{2}{2}L - I = L - I,$$

where T_k is the k-th Chebyshev polynomial. In the ideal case $\delta = 0$, we have

$$g_{p_d}(L_d) \;=\; U\, g_{p_d}(\Lambda_d)\, U^\top, \qquad g_{p_d}(\Lambda_d) = \mathrm{diag}(g_{p_d}(\lambda_{d,1}), \ldots, g_{p_d}(\lambda_{d,n})),$$

so the prompt acts as a domain-specific frequency response curve. Consequently, if the downstream transfer function ψ_d is well approximated by a degree-m polynomial on $[0, 2]$, then there exists a prompt p_d such that $g_{p_d}(\Lambda_d) \approx \psi_d(\Lambda_d)$ uniformly over eigenvalues. This is the mechanism behind logarithmic prompt-size scaling with the target approximation error.

Where the backbone enters. The backbone encoder f_θ is shared across domains and is meant to exploit generic structural inductive biases (locality, permutation invariance, pooling, etc.). In the sandbox, we conceptually separate two roles: (i) $g_{p_d}(L)$ shapes the *spectral content* of information propagation to compensate for domain shift in Λ_d; (ii) f_θ implements a task-agnostic feature extractor operating on the filtered messages. The idealized theory treats f_θ as sufficiently expressive to realize (or preserve) linear prediction in the prompted spectral features, while the prompt supplies the domain-specific frequency correction.

Realism and failure modes. We record the main points at which the model can break.

Eigenbasis mismatch. If domains do not admit a common (or approximately common) eigenbasis, then no family of diagonal spectral reweightings can simultaneously align all domains. In that case, prompts of dimension $m = o(n)$ are information-theoretically insufficient to undo an arbitrary rotation of spectral coordinates, and one should expect a constant excess risk on some domain families (formalized later as an impossibility statement).

Graph-size variability beyond padding. Padding aligns dimensions but does not create semantic correspondence between nodes across graphs. Our model does not assume node identity correspondence; rather, U is a latent basis in the ambient $\mathbb{R}^n$ used for analysis. If domains differ systematically

in size distributions and graph topologies in a way that changes the effective spectrum dramatically, the shared-basis assumption may become inaccurate (large δ).

Non-spectral shifts. Domain shift may arise from feature distribution changes (covariate shift in X), label shift, or changes in higher-order structures not well captured by the Laplacian spectrum. In such regimes, a purely spectral prompt may be too constrained; one would need prompts that modulate feature channels, attention patterns, or normalization statistics in addition to (or instead of) Laplacian polynomials.

Locality versus global spectral effects. Polynomial filters are local in the sense that L^k only propagates information within k hops, whereas true spectral projectors can be global. Thus, small m restricts the spatial range of the correction. Our approximation theorems interpret this as a tradeoff: increasing m expands the set of realizable frequency responses and increases receptive field, but it also increases few-shot estimation error through the m-parameter adaptation.

Interpretation. Within its scope, the spectral-shift model identifies a concrete axis along which domains can differ while still being amenable to parameter-efficient transfer: domain-specific eigenvalues with a shared (approximate) eigenbasis. The prompt is then a compact encoding of a domain's frequency response, and few-shot adaptation amounts to estimating this response (and a small linear head) from K labeled examples. In the next section we instantiate this principle algorithmically by coupling prompt learning to a spectral/statistical alignment objective during pre-training, thereby encouraging prompted representations to share comparable geometry across domains even when $\delta > 0$ and the ideal model holds only approximately.

5 Spectral-Alignment Prompt Pre-Training

We now describe the learning procedure that operationalizes the spectral-shift principle: we pre-train a single backbone encoder while learning a compact prompt for each domain, where prompts parameterize polynomial spectral filters, and we regularize the prompted representations to be statistically aligned across domains.

Backbone encoder and insertion point of the prompt. Let f_θ be a shared graph encoder producing node embeddings $H \in \mathbb{R}^{n \times d_h}$ (and optionally a pooled graph embedding h_G). The backbone may be instantiated as a message-passing GNN, a graph transformer with structural biases, or a hybrid architecture; our analysis only requires that the backbone is *shared* across domains and is sufficiently expressive to preserve linear predictability in the prompted features. We insert the prompt as a domain-dependent

linear operator on graph signals before (or inside) message propagation. Concretely, given a padded Laplacian $L = L(G)$ and lifted node features $\phi(X) \in \mathbb{R}^{n \times r}$, we form a prompt-conditioned message operator

$$M_{p_d}(G) \;:=\; g_{p_d}(L(G)) \in \mathbb{R}^{n \times n}, \qquad Z \;:=\; M_{p_d}(G)\, \phi(X), \tag{4}$$

and then feed (G, Z) through the backbone:

$$H \;:=\; f_\theta(G;\, Z). \tag{5}$$

In a multi-layer GNN, (4) can be applied once at the input (as a spectral "pre-emphasis" of the signal) or at each layer (as a domain-dependent propagation rule); we treat both as instances of the same abstraction, since both implement a domain-specific polynomial in L composed with shared nonlinearities.

Prompt parameterization as polynomial spectral filters. For each domain $d \in \{1, \ldots, D\}$ we maintain a prompt vector $p_d \in \mathbb{R}^m$ with a fixed budget m. We interpret p_d as coefficients of a degree-m polynomial filter

$$g_{p_d}(L) \;:=\; \sum_{k=0}^{m} a_k(p_d)\, L^k, \tag{6}$$

where $a_k(\cdot)$ is either the identity (direct parameterization) or a low-capacity map (e.g. an MLP) that enforces constraints such as bounded coefficients. To improve numerical stability and to respect the spectrum of the normalized Laplacian, we implement g_{p_d} using Chebyshev polynomials on the rescaled operator $\widetilde{L} := L - I$, whose eigenvalues lie in $[-1, 1]$:

$$g_{p_d}(L) \;:=\; \sum_{k=0}^{m} c_{d,k}\, T_k(\widetilde{L}), \qquad p_d = (c_{d,0}, \ldots, c_{d,m}). \tag{7}$$

The Chebyshev recurrence yields an $O(m|E|)$ implementation per graph (or per sampled subgraph) without eigendecomposition: for a signal $s \in \mathbb{R}^{n \times r}$ we compute

$$t_0 = s, \quad t_1 = \widetilde{L}s, \quad t_k = 2\widetilde{L}t_{k-1} - t_{k-2} \;\; (k \geq 2), \qquad g_{p_d}(L)s = \sum_{k=0}^{m} c_{d,k} t_k.$$

This realizes a receptive field of m hops and makes the prompt budget directly commensurate with both (i) the approximation power over spectral transfer functions on $[0, 2]$ and (ii) the number of trainable parameters exposed during few-shot adaptation.

Self-supervised pre-training objective. We pre-train the shared backbone and prompts using unlabeled graphs from all domains. We denote by ℓ_{ssl} a composite self-supervised loss that mixes an invariance term (contrastive) and a signal-modeling term (generative), since these encourage complementary properties of the representation. A representative instantiation is

$$\ell_{\text{ssl}} \;:=\; \ell_{\text{ctr}}(H^{(1)}, H^{(2)}) \;+\; \alpha\, \ell_{\text{gen}}(H), \tag{8}$$

where ℓ_{ctr} is an InfoNCE-style loss computed on two stochastic views of the same graph (node dropping, edge perturbation, attribute masking), and ℓ_{gen} reconstructs masked node attributes or local structural statistics from H. The key point for our setting is that the prompt enters the computation of H through $M_{p_d}(G)$; therefore, the pretext tasks shape prompts toward producing representations that are simultaneously predictive for generic graph structure *and* compatible across domains once filtered.

Spectral/statistical alignment regularizer. Self-supervision alone does not guarantee that representations produced under different domain prompts are geometrically comparable. We therefore add an alignment term that encourages cross-domain agreement of second-order statistics in the prompted representation space. For a minibatch B from domain d, let $h(G)$ denote either the pooled graph embedding or a concatenation/average of node embeddings after backbone processing. We form a batch moment estimate (centered or uncentered) such as

$$\widehat{\Sigma}_d(B) \;:=\; \frac{1}{|B|} \sum_{G \in B} (h(G) - \bar{h}_d)(h(G) - \bar{h}_d)^\top, \qquad \bar{h}_d := \frac{1}{|B|} \sum_{G \in B} h(G). \tag{9}$$

We then define an alignment map $A(\cdot)$ that optionally conditions on the prompt and normalizes scale, for example $A(p_d, \Sigma) = \operatorname{tr}(\Sigma)^{-1}\Sigma$ or a low-rank sketch of Σ. Given two domains $d \neq d'$, we penalize a discrepancy between their aligned moments:

$$\ell_{\text{align}}(d, d') \;:=\; \text{Dist}\Big(A(p_d, \widehat{\Sigma}_d(B)),\; A(p_{d'}, \widehat{\Sigma}_{d'}(B')) \Big), \tag{10}$$

where Dist can be the Frobenius norm, a Bures/Wasserstein distance between Gaussians, or a kernel MMD applied to embeddings. The intended effect is that, under the commuting-Laplacian hypothesis, prompts reweight domain spectra so that the backbone sees comparable frequency content across domains; matching moments is a tractable surrogate for aligning the full prompted representation distributions.

Joint optimization and training protocol. The pre-training objective combines (8) and (10):

$$\min_{\theta,\{p_d\}} \; \mathbb{E}_{d\sim\pi}\, \mathbb{E}_{B\sim\mathcal{P}_d}\Big[\ell_{\mathrm{ssl}}(\theta, p_d; B)\Big] \;+\; \lambda\, \mathbb{E}_{d\neq d'}\, \mathbb{E}_{B\sim\mathcal{P}_d,\; B'\sim\mathcal{P}_{d'}}\Big[\ell_{\mathrm{align}}(d, d')\Big], \tag{11}$$

with $\lambda > 0$. In practice we approximate the expectations by sampling one (or two) domains per iteration, computing embeddings under the corresponding prompts, estimating $\widehat{\Sigma}_d$ on the fly (or via an exponential moving average), and taking a gradient step in $(\theta, \{p_d\})$. The memory cost of storing prompts is $O(Dm)$; the incremental cost per forward pass is dominated by evaluating $g_{p_d}(L)$ via m sparse-matrix recurrences.

Few-shot adaptation and parameter-efficiency. After pre-training, we freeze θ and adapt only a domain prompt (and optionally a small linear head) using K labeled examples in the target domain. The prompt-only update is an m-dimensional supervised optimization problem; the method is therefore parameter-efficient by construction, and the prompt budget m governs the statistical complexity of adaptation. This design aligns with the sandbox model of Section 4: if g_{p_d} approximates the domain transfer function in the shared basis, then a simple predictor on top of f_θ suffices, and few-shot learning need only refine the spectral response rather than relearn a domain-specific encoder.

Practical remarks on polynomial degree and stability. Two implementation details are worth recording for the subsequent theory. First, the approximation properties of Chebyshev expansions imply that increasing m expands the set of realizable frequency responses on $[0, 2]$, consistent with logarithmic degree requirements for uniform approximation of Lipschitz transfer functions. Second, bounding coefficient norms (e.g. $\|p_d\|_2 \le B$) controls the operator norm of $g_{p_d}(L)$ on the spectrum of L, which stabilizes training and ensures that alignment based on second moments is well-behaved. These constraints will be invoked implicitly when translating approximation error of spectral filters into downstream excess risk bounds in the next section.

6 Main Theorems (Upper Bounds)

We now state the guarantees that motivate the design choices of Section 5. Throughout, we work under the commuting-Laplacian hypothesis: for each domain d, a representative normalized Laplacian satisfies

$$L_d \;=\; U\Lambda_d U^\top + \Delta_d, \qquad \|\Delta_d\|_2 \le \delta,$$

for a single orthogonal basis $U \in \mathbb{R}^{n\times n}$ shared across domains and diagonal Λ_d with entries in $[0, 2]$. We assume downstream labels are generated by a

domain-dependent spectral transfer function composed with a linear readout in the shared basis, i.e., (in the simplest node-level form)

$$y \;=\; \mathrm{sign}\big(\langle w_d,\, U^\top \phi(X)\rangle + \xi\big),$$

where ϕ is a fixed feature-lifting map and ξ is bounded or sub-Gaussian noise. The role of the prompt $p_d \in \mathbb{R}^m$ is to implement a polynomial spectral filter $g_{p_d}(L)$ which approximates the unknown domain transfer function in the eigenvalue variable $\lambda \in [0, 2]$.

Approximation error versus prompt size. The first theorem isolates the approximation power of degree-m polynomial prompts. Informally, if the Bayes-optimal domain-specific transformation is a sufficiently regular (e.g., Lipschitz) function of the Laplacian spectrum, then a logarithmic number of prompt coefficients suffices to approximate it uniformly on $[0, 2]$.

Theorem 6.1 (Approximation by prompt spectral filters)**.** *Assume $\delta = 0$ so that $L_d = U\Lambda_d U^\top$ exactly. Suppose the downstream Bayes-optimal predictor in domain d depends on a band-limited spectral transfer function $\psi_d(\lambda)$ which is Lipschitz on $[0, 2]$. Then for any $\varepsilon > 0$ there exists a degree-m polynomial filter g_{p_d} with*

$$m \;=\; \tilde{O}\left(\log \frac{1}{\varepsilon}\right)$$

such that

$$\big\|\psi_d(\Lambda_d) - g_{p_d}(\Lambda_d)\big\|_\infty \;\le\; \varepsilon,$$

and the induced excess Bayes risk from replacing ψ_d by g_{p_d} is $O(\varepsilon)$.

The key point is that the prompt budget m controls a *uniform* approximation guarantee over the entire spectral interval, rather than merely an L_2 approximation under a specific eigenvalue distribution. This is precisely the regime in which Chebyshev expansions are well-suited: on a compact interval, the Chebyshev truncation error for regular target functions decays essentially exponentially in the degree, giving the logarithmic scaling in ε^{-1}. From the modeling perspective, Theorem 6.1 justifies interpreting the prompt as a *domain-specific frequency equalizer* which can undo spectral shifts Λ_d while keeping the backbone fixed.

Few-shot adaptation: estimation error with prompt-only training. Approximation alone does not yield a downstream guarantee, since in a few-shot setting we must also estimate the prompt (and optionally a small linear head) from K labeled examples. The next theorem provides the standard decomposition into approximation, perturbation, and estimation errors, where only the estimation term depends on K and only through the prompt dimension m.

Theorem 6.2 (Few-shot excess risk with prompt-only adaptation)**.** *Assume the conditions of Theorem 6.1, but allow $\delta \geq 0$ so that $\|\Delta_d\|_2 \leq \delta$. Consider a downstream predictor that is linear in the prompt-filtered embeddings and has bounded norm. Let $\widehat{p}_d$ denote the prompt obtained by empirical risk minimization over K labeled samples in domain d, with the backbone parameters θ frozen. Then*

$$\mathbb{E}\big[\mathrm{Risk}_d(\widehat{p}_d)\big] - \mathrm{Risk}_d(p_d^\star) \;\leq\; O(\varepsilon + \delta) \;+\; \tilde{O}\left(\sqrt{\frac{m}{K}}\right),$$

where $p_d^\star$ is the best degree-m prompt for domain d in the hypothesis class.

Two features of Theorem 6.2 are essential for our parameter-efficient objective. First, the statistical price of adaptation scales as $\sqrt{m/K}$ (up to logarithmic factors), as one would expect from uniform stability or Rademacher complexity bounds for an m-parameter family. In particular, the backbone dimension and the number of backbone parameters do not enter the estimation term, since θ is frozen. Second, the approximation and perturbation terms decouple from K; increasing K cannot overcome an insufficient prompt degree (large ε) or a large violation of the commuting hypothesis (large δ).

Robustness to approximate commutativity. We briefly indicate how the δ-dependence arises. Because g_{p_d} is a polynomial of degree m, we may compare $g_{p_d}(L_d)$ to $g_{p_d}(U\Lambda_d U^\top)$ by a functional perturbation bound. Concretely, if the coefficients of g_{p_d} are bounded so that $\|g_{p_d}\|$ is controlled on $[0, 2]$, then repeated use of submultiplicativity yields an operator-norm stability estimate of the form

$$\big\|g_{p_d}(L_d) - g_{p_d}(U\Lambda_d U^\top)\big\|_2 \;\leq\; C(m, \|p_d\|)\,\delta,$$

with $C(m, \|p_d\|)$ polynomial in m when coefficients are bounded (and mild in practice under Chebyshev parameterizations with norm constraints). Translating this operator error into a prediction error bound uses the Lipschitzness of the downstream functional (or a margin argument for classification), which is why Theorem 6.2 incurs an additive $O(\delta)$ term. Thus, approximate commutativity is not merely a modeling convenience: it is the condition under which prompt-induced spectral reweighting remains stable and transferable.

Unseen domains: unlabeled prompt adaptation via moment alignment. We next address the setting in which a new domain d' arrives after pre-training, and we wish to produce a good prompt using *only unlabeled graphs* from d'. The alignment mechanism in (10) suggests a natural unsupervised criterion: choose $p_{d'}$ so that prompt-filtered statistics (e.g., Laplacian moments or embedding covariances) match those of previously seen domains. The following theorem formalizes the unlabeled sample complexity needed to reach a target alignment accuracy.

Theorem 6.3 (Sample complexity for learning a new-domain prompt)**.** *Assume a new domain d' satisfies the commuting-Laplacian hypothesis with the same U and unknown $\Lambda_{d'}$. Suppose we learn a prompt $p_{d'} \in \mathbb{R}^m$ by minimizing an alignment loss that depends on empirical Laplacian moments (or equivalently, statistics of prompt-filtered signals) up to order m. Then, for any $\varepsilon \in (0,1)$ and failure probability $\rho \in (0,1)$, it suffices to use*

$$N \;=\; \tilde{O}\Big(\frac{m+\log(1/\rho)}{\varepsilon^2}\Big)$$

unlabeled graphs (or sampled subgraphs) from domain d' to obtain alignment error at most ε with probability at least $1-\rho$. Moreover, in general $N = \Omega(m/\varepsilon^2)$ unlabeled samples are necessary.

Theorem 6.3 should be read as an information-theoretic statement about estimating m spectral degrees of freedom from unlabeled data: one cannot learn a degree-m spectral correction without observing enough graphs to reliably estimate the corresponding order-m moments (or any equivalent m-dimensional summary). Operationally, it means that prompt transfer to new domains can be amortized by unlabeled corpora, and the unlabeled data requirement scales linearly in the prompt budget.

Consequence: a unified excess-risk decomposition. Combining the above results yields a single guiding decomposition for prompted few-shot transfer:

$$\text{excess risk} \;\lesssim\; \underbrace{\varepsilon}_{\text{spectral approximation}} \;+\; \underbrace{\delta}_{\text{basis mismatch}} \;+\; \underbrace{\tilde{O}\Big(\sqrt{m/K}\Big)}_{\text{few-shot estimation}} .$$

The approximation term is controlled by the prompt degree (Theorem 6.1), the perturbation term is controlled by the extent to which the domains share spectral structure, and the estimation term is controlled by the number of labeled examples and the prompt dimension (Theorem 6.2). The unlabeled adaptation guarantee (Theorem 6.3) explains when and how a good initialization for $p_{d'}$ can be obtained without labels, thereby reducing the number of labeled samples needed for subsequent few-shot tuning. In the next section we show that these upper bounds are essentially tight: the logarithmic dependence of m on ε^{-1} cannot be improved in general, and without shared spectral structure prompt-only adaptation can be forced to fail.

7 Lower Bounds and Impossibility Results

We now complement the upper bounds of Section 6 with limitations that are intrinsic to prompt-only adaptation in the present spectral formalism. The

conclusions are threefold. First, even in the idealized regime where a common eigenbasis U exists and is perfectly shared, one cannot in general reduce the prompt budget below logarithmic in the target uniform approximation error. Second, when domains fail to share approximate spectral structure (in the sense of approximate commutativity), any approach that freezes a single backbone and permits only low-dimensional prompt modulation can be forced to fail on some domain family. Third, if a prompt for a new domain is learned solely from unlabeled graphs via moment/statistic matching, then the unlabeled sample complexity must scale at least linearly with the prompt dimension, matching Theorem 6.3 up to logarithmic factors.

Prompt-size lower bound under a shared eigenbasis. Theorem 6.1 shows that degree-m polynomial prompts suffice for uniform spectral approximation under regularity assumptions. The following result shows that the dependence on ε cannot, in general, be improved within the same hypothesis class (polynomial spectral filters), even if we grant the learner maximal structural help: exact commutativity and knowledge of the shared basis.

Theorem 7.1 (Prompt-size lower bound, matching Theorem 6.1)**.** *Assume $\delta = 0$ and suppose U is known. There exists a family of domains (equivalently, a family of diagonal spectra $\{\Lambda_d\}$) and a family of Lipschitz spectral transfer functions $\{\psi_d\}$ on $[0, 2]$ such that, for any $\varepsilon \in (0, 1)$, every polynomial filter g of degree at most m satisfying*

$$\left\|\psi_d(\Lambda_d) - g(\Lambda_d)\right\|_\infty \leq \varepsilon$$

must have

$$m \;=\; \Omega\!\left(\log \frac{1}{\varepsilon}\right).$$

Proof sketch. We reduce the existence of small prompts to a classical uniform approximation problem on a compact interval. Fix a domain d and consider the associated target transfer $\psi_d(\lambda)$ on $\lambda \in [0, 2]$. Any polynomial prompt produces a polynomial $g(\lambda)$ (after identifying the prompt coefficients with a basis such as Chebyshev polynomials rescaled to $[0, 2]$). Thus the requirement $\|\psi_d(\Lambda_d) - g(\Lambda_d)\|_\infty \leq \varepsilon$ entails that g uniformly approximates ψ_d on the spectral support of Λ_d; choosing Λ_d to be sufficiently dense in $[0, 2]$ forces uniform approximation on the whole interval.

The lower bound then follows by selecting ψ_d from a family whose best degree-m polynomial approximation error is bounded below by $\exp(-cm)$ for some absolute $c > 0$, a standard converse-type statement in approximation theory for appropriate regularity classes. Concretely, one may take ψ_d to have analytic continuation only to a Bernstein ellipse of fixed parameter (equivalently, to have Chebyshev coefficients that decay no faster than $\exp(-ck)$), which forces any truncation (hence any degree-m polynomial) to incur error at least $\exp(-cm)$. Setting $\exp(-cm) \leq \varepsilon$ yields

$m = \Omega(\log(1/\varepsilon))$. This establishes that, within polynomial prompt families, the logarithmic scaling in Theorem 6.1 is information-theoretically tight.

Impossibility without shared spectral structure. The commuting-Laplacian hypothesis does more than simplify analysis: it is the minimal condition under which *diagonal* (spectral) prompt modulation can align domains while leaving a single backbone fixed. If the eigenbases vary substantially across domains, then reweighting frequencies in any fixed basis is not expressive enough to compensate, regardless of pre-training quality. The following theorem formalizes this failure mode.

Theorem 7.2 (Impossibility without a shared eigenbasis)**.** *Consider a family of domains for which the normalized Laplacians take the form* $L_d = U_d \Lambda_d U_d^\top$ *with* $\delta = 0$*, but the bases* $\{U_d\}$ *do not approximately commute and are pairwise nearly orthogonal in the sense that no single orthogonal* U *jointly diagonalizes the set even approximately. Let a method output a single frozen backbone* f_θ *together with per-domain prompts* $p_d \in \mathbb{R}^m$*, where the prompts act only through spectral filters* $g_{p_d}(L)$ *(or, more generally, through any diagonal or coordinate-wise modulation in a fixed representation basis). If* $m = o(n)$*, then there exists a downstream linear prediction task in each domain such that the method incurs excess risk bounded away from* 0 *on at least one domain.*

Proof sketch. The core obstruction is geometric: spectral prompts modulate *eigenvalues* but cannot implement a domain-dependent *rotation* of the eigenvectors. When U_d varies adversarially with d, the information needed to map one domain's predictive direction to another is of dimension $\Theta(n^2)$ (an orthogonal matrix), whereas a prompt of dimension $m = o(n)$ supplies only vanishing degrees of freedom.

A packing argument makes this quantitative. One constructs a set of domains with eigenbases $\{U_d\}$ forming a large packing in $\mathrm{O}(n)$ such that, for any fixed backbone representation, the induced coordinate system is misaligned with at least one U_d by a constant angle on a constant fraction of coordinates. For each domain, one then defines a downstream task whose Bayes-optimal predictor corresponds to a coordinate (or a low-dimensional subspace) in that domain's eigenbasis; equivalently, labels depend on $\langle w_d, U_d^\top \phi(X) \rangle$ with w_d supported on a small set of coordinates. Any prompt family that only performs diagonal modulation in the backbone's fixed coordinates cannot consistently recover all these rotated predictors: diagonal scaling cannot undo a generic rotation. Consequently, at least one domain in the packing yields constant misclassification (or regression) error, giving a constant excess-risk lower bound. This provides a formal justification for treating approximate commutativity (small δ in our model) as a necessary structural assumption rather than a technical artifact.

Unlabeled lower bounds for learning prompts by moment alignment. Finally, we address the unlabeled setting of Theorem 6.3. There, the prompt for a new domain is inferred from unlabeled graphs by matching statistics that depend on Laplacian moments up to order m (or equivalent prompt-filtered embedding moments). The following statement shows that the linear dependence on m in the unlabeled sample size is unavoidable in general.

Theorem 7.3 (Unlabeled sample lower bound for prompt learning)**.** *Fix a prompt class of dimension m learned from unlabeled graphs via any estimator whose inputs are empirical statistics of order at most m (in particular, empirical Laplacian moments up to order m). For any target accuracy $\varepsilon \in (0, 1)$, there exists a pair of new domains d'_0, d'_1 satisfying the commuting-Laplacian hypothesis with the same U such that any procedure that, given N unlabeled graphs from the new domain, outputs a prompt with alignment error at most ε with probability at least $2/3$ must have*

$$N \;=\; \Omega\Big(\frac{m}{\varepsilon^2}\Big)\,.$$

Proof sketch. We use a two-point (Le Cam) argument. Construct two candidate new-domain distributions $\mathcal{P}_{d'_0}$ and $\mathcal{P}_{d'_1}$ whose induced spectra (or, more precisely, whose distributions over Laplacian moments up to order m) differ by an amount calibrated so that (i) the optimal prompts for the two domains are separated by $\Theta(\varepsilon)$ in the alignment objective, yet (ii) the total variation distance between the N-sample unlabeled observations under the two hypotheses is small unless N is of order m/ε^2. This is achieved by embedding m independent degrees of freedom into the first m moments, each perturbed at scale $\varepsilon/\sqrt{m}$, so that the aggregate alignment gap is $\Theta(\varepsilon)$ but the per-sample statistical signal is weak. Standard concentration then implies that $N = \Omega(m/\varepsilon^2)$ samples are necessary to estimate these m moment components reliably. The conclusion matches the upper bound in Theorem 6.3 up to polylogarithmic factors.

Interpretation. Taken together, Theorems 7.1–7.3 delineate the regime in which prompt-based transfer is plausible. If domains share a basis (small δ), then polynomial prompts of size $m = \Theta(\log(1/\varepsilon))$ are both sufficient and, in general, necessary for uniform spectral correction; moreover, learning such prompts from unlabeled data requires $\Theta(m/\varepsilon^2)$ graphs. Conversely, if domains do not share spectral structure, then no amount of unlabeled pre-training can circumvent the representational bottleneck imposed by low-dimensional, diagonal prompt modulation. In the next section we turn from information-theoretic limits to computational ones, quantifying the training and adaptation costs implied by our design.

8 Complexity and Scalability

We quantify the computational costs implied by prompt-parameterized spectral filtering, and we isolate the scaling bottlenecks that arise when we increase the number of domains, the graph sizes, or the prompt dimension. Our analysis is stated in the same access model as in Section 6: graphs are observed through mini-batches (or sampled subgraphs), and prompt modulation is implemented as a degree-m polynomial in a sparse graph operator derived from the normalized Laplacian.

Prompt filtering cost and its dependence on m. In our design, the prompt $p_d \in \mathbb{R}^m$ parametrizes a polynomial filter

$$g_{p_d}(L) \;=\; \sum_{k=0}^{m} a_k(p_d)\, T_k(\widetilde{L}),$$

where T_k denotes a (shifted/rescaled) Chebyshev polynomial and $\widetilde{L}$ is scaled to have spectrum in $[-1, 1]$. The salient point for scalability is that evaluation does not require eigendecomposition: the Chebyshev recurrence uses m sparse matrix–vector multiplications by $\widetilde{L}$ (or by a normalized adjacency operator), and hence the filter cost per forward pass is

$$\mathrm{Cost}_{\mathrm{filter}}(G) \;=\; O(m|E|),$$

up to feature dimension factors. When the backbone encoder f_θ is a message-passing GNN with r layers, the end-to-end cost is additively decomposable as

$$\mathrm{Cost}_{\mathrm{forward}}(G) \;\approx\; \mathrm{EncCost}_\theta(G) \;+\; O(m|E|),$$

where $\mathrm{EncCost}_\theta(G)$ is the cost of the backbone with its usual neighborhood aggregation. In typical regimes, the prompt overhead is linear in m and linear in sparsity, and thus tunable: increasing m improves approximation power (cf. Theorem 7.1) while incurring a predictable linear-time penalty.

Pre-training time with alignment regularization. Algorithm 1 uses (i) a self-supervised loss and (ii) an alignment regularizer comparing prompt-conditioned statistics across domains. The self-supervised term is standard and scales as in single-domain pre-training. The alignment term introduces two additional operations: computing batch moments and comparing them across two domains. If $h \in \mathbb{R}^{|V_B| \times d_h}$ denotes embeddings in a batch B, then an empirical covariance $\widehat{\Sigma}_d(B)$ can be formed in $O(|V_B| d_h^2)$ time and $O(d_h^2)$ memory. This can dominate when d_h is large. Two standard mitigations preserve the role of second-order alignment while reducing cost:

1. *Diagonal or block-diagonal alignment:* align only $\mathrm{diag}(\widehat{\Sigma}_d)$ (cost $O(|V_B| d_h)$) or small blocks corresponding to feature groups.

2. *Low-rank sketches:* maintain a rank-r_Σ approximation via random projections, yielding $O(|V_B| d_h r_\Sigma)$ time and $O(d_h r_\Sigma)$ memory.

In either case, the cross-domain comparison can be implemented as an ℓ_2 distance between sketches, with negligible overhead relative to the forward/backward pass. If we estimate alignment using two mini-batches per iteration (one from d and one from d'), then the asymptotic pre-training time per iteration becomes

$$O(\mathrm{EncCost}_\theta(B) + m|E_B| + \mathrm{StatCost}(B)) + O\big(\mathrm{EncCost}_\theta(B') + m|E_{B'}| + \mathrm{StatCost}(B')\big)\,,$$

where StatCost is chosen according to the moment estimator. Importantly, we do not incur any factor scaling with D per iteration: domain sampling yields constant additional work.

Few-shot adaptation cost and parameter efficiency. At downstream time we freeze θ and update only p_d (and optionally a linear head). If we perform S gradient steps on K labeled examples, then the optimization cost is essentially S forward/backward passes through the frozen backbone with respect to m prompt parameters:

$$\mathrm{Time}_{\mathrm{adapt}}(d) \;=\; O(S \cdot K \cdot (\mathrm{EncCost}_\theta + m|E|))\,, \qquad \mathrm{Params}_{\mathrm{train}} \;=\; m \;(+\,\mathrm{head}).$$

Thus, compared to full fine-tuning, adaptation eliminates the dependence on $|\theta|$ in both the number of updated parameters and optimizer state. This matters in practice for multi-domain deployments: adding a new domain increases trainable storage by $O(m)$, not $O(|\theta|)$.

Memory footprint across many domains. We store one backbone and D prompts. The prompt memory is

$$\mathrm{Mem}_{\mathrm{prompts}} \;=\; O(Dm),$$

which is negligible for small m even when D is large, and contrasts with mixture-of-experts approaches that replicate substantial fractions of the backbone. If we additionally maintain per-domain running statistics (e.g., an exponential moving average of Σ_d to stabilize alignment), the naive cost is $O(Dd_h^2)$, which can be prohibitive. In such cases we recommend the same diagonal/low-rank strategies described above; with rank-r_Σ sketches, the amortized storage becomes $O(Dd_h r_\Sigma)$.

Prompt routing versus mixture-of-experts. A separate scalability issue is *routing*: at test time, which prompt should be applied? If the domain identity is available, routing is trivial and cost-free. If it is not available, then we may learn a router $R(\cdot)$ producing either (i) a discrete prompt

index $\hat{d} = R(G)$ or (ii) a convex combination of prompts. The discrete case adds $O(D)$ logits per example (or $O(\log D)$ if implemented hierarchically), while the convex combination case replaces $p_{\hat{d}}$ by $\sum_{d=1}^{D} \alpha_d p_d$ with $\alpha = \mathrm{softmax}(R(G))$, costing $O(Dm)$ per example. This is still typically smaller than mixture-of-experts (MoE) routing when experts are full backbones: MoE pays $O(\text{top-}k \cdot \mathrm{EncCost}_\theta)$ per example and stores $O(\#\text{experts} \cdot |\theta|)$ parameters. In contrast, prompt routing keeps compute dominated by a single backbone forward pass and increases storage only linearly in Dm.

From a representational perspective, prompt routing also admits an intermediate regime between "one prompt per domain" and "one prompt per sample": we may cluster domains and share prompts among clusters, or maintain a small pool of prompts reused across domains. This yields a controllable trade-off between memory $O(Dm)$ and statistical efficiency (more sharing implies less per-domain specialization).

Local subgraph sampling versus full-graph processing. The cost expressions above depend on $|E|$ and $|V|$, and hence on whether we process full graphs or local neighborhoods. For node- and edge-level tasks on large graphs, it is standard to train on induced r-hop ego-subgraphs or sampled neighborhoods. In this regime, the polynomial filter is evaluated on the subgraph Laplacian $L(G_{\mathrm{sub}})$, yielding cost

$$O(m|E_{\mathrm{sub}}|) \quad \text{per labeled root node/edge,}$$

and the alignment term can likewise be computed on subgraph embeddings. This is consistent with our theory to the extent that the subgraph Laplacians inherit the same approximate spectral structure (in practice, the approximation error is absorbed into the δ-type perturbation term).

For graph-level tasks, full-graph processing may be necessary to capture global structure. Here the prompt overhead remains linear in $m|E|$, but memory becomes the bottleneck if $|V|$ is large due to storing intermediate activations. Standard remedies (gradient checkpointing, minibatch graph partitioning, and hierarchical pooling/coarsening) are compatible with prompt filtering, since $g_{p_d}(L)$ can be applied at each resolution with the corresponding sparse operator.

Practical guidance for choosing m under compute constraints. The prompt degree m is the principal knob controlling both approximation capacity and overhead. From the perspective of run time, increasing m increases only the filter cost, and does not multiply the backbone depth. Consequently, one may select m by fixing a target overhead fraction, e.g.,

$$\frac{m|E|}{\mathrm{EncCost}_\theta(G)} \leq \eta,$$

and then using the largest m satisfying this inequality. When the backbone is already expensive (e.g., attention-based graph transformers), moderate m adds little relative overhead; conversely, when the backbone is a light message-passing network, m should be chosen more conservatively. In all cases, the central scalability advantage remains: domain growth increases storage only by $O(m)$ per domain, and few-shot adaptation updates only $O(m)$ parameters per domain, enabling wide multi-domain coverage without replicating the encoder.

9 Experimental Plan

We outline an experimental program designed to (i) validate the claimed parameter–efficiency of prompt-only adaptation, (ii) isolate the contribution of spectral/statistical alignment during pre-training, (iii) probe the scaling of performance with prompt size m as suggested by Theorems 1–3, and (iv) assess robustness under cross-domain shift and sequential domain arrival.

Benchmark construction and domain definition. We require *text-free* multi-domain corpora of attributed graphs. We propose to build benchmarks in two complementary ways.

1. *Natural multi-domain corpora.* We form domains by a semantically meaningful partition that induces distribution shift while preserving the prediction interface (node/edge/graph labels). Examples include: (a) molecular graphs partitioned by assay/source laboratory, scaffold family, or measurement protocol; (b) citation/social graphs partitioned by time slices, communities, or platforms; (c) interaction graphs partitioned by geography or time; (d) program/AST graphs partitioned by repository or programming language, using only structural/node-type features (no identifiers or text). Each domain d yields unlabeled graphs for pre-training and a small labeled set S_d for evaluation.

2. *Controlled synthetic corpora with known spectral structure.* To stress-test the commuting-Laplacian hypothesis, we generate domains by fixing an orthogonal basis U and sampling diagonal spectra Λ_d, then constructing Laplacians $L_d = U\Lambda_d U^\top + \Delta_d$ with tunable perturbation $\|\Delta_d\|_2 \le \delta$; graphs can be obtained by projecting L_d to a sparse adjacency via thresholding or via a graphon/latent-space model whose Laplacian concentrates around L_d. We then define labels through band-limited spectral functions to align with the theoretical task class, enabling direct verification of m–versus–ε behavior and explicit sweeps over δ.

In all settings, we standardize node features to bounded ranges (e.g., $\|X\|_\infty \le 1$ after preprocessing) and avoid any textual attributes. We adopt a domain-

balanced sampling distribution π during pre-training to prevent the largest domain from dominating optimization.

Evaluation protocol and tasks. We consider node-, edge-, and graph-level prediction tasks depending on the dataset. Pre-training uses only unlabeled graphs. Downstream evaluation is *few-shot per domain*: for each d, we sample $K \in \{1, 5, 10, 20, 50\}$ labeled examples (nodes/edges/graphs) for adaptation, tune only the prompt p_d (and optionally a linear head), and report risk on a held-out test split. We report mean and worst-case performance over domains:

$$\text{AvgRisk} = \frac{1}{D}\sum_{d=1}^{D} \text{Risk}_d, \qquad \text{WorstRisk} = \max_{d \in [D]} \text{Risk}_d,$$

as well as accuracy/AUROC for classification and RMSE/MAE for regression. For stability, each (d, K) is repeated over multiple random labeled subsets.

Baselines. To interpret gains, we compare against baselines that separate (a) backbone sharing, (b) domain-specific adaptation capacity, and (c) spectral alignment.

1. *No adaptation:* frozen f_θ with a linear probe trained on S_d (no prompt).
2. *Full fine-tuning:* update all θ on S_d (upper bound on performance, not parameter-efficient).
3. *Standard parameter-efficient tuning:* adapters/LoRA-style low-rank updates inside the encoder (matched by parameter count), and feature-wise affine modulation (FiLM) layers conditioned on a domain embedding.
4. *Domain-specific backbones:* one separately pre-trained encoder per domain (computationally expensive reference point), and a multi-task shared encoder trained without prompts.
5. *Prompt variants:* (i) prompts without alignment regularization ($\lambda = 0$), (ii) alignment without prompts (shared filter, or statistics alignment applied only to embeddings), and (iii) non-spectral prompts that modulate MLP/readout layers with the same number of parameters as p_d.
6. *Mixture-of-experts comparisons:* a small MoE with k experts and a router, matched for compute where feasible, to test whether prompt routing provides a comparable benefit without replicating full encoders.

When domain identity is unavailable, we include a learned router $R(G)$ baseline and report both routing accuracy and downstream risk under predicted prompts.

Ablations isolating spectral alignment. We isolate the mechanism of the alignment regularizer by controlled ablations.

1. *Alignment weight sweep:* $\lambda \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ to test the trade-off between invariance and specialization.

2. *Choice of statistic:* match (a) full covariance $\widehat{\Sigma}_d$, (b) diagonal variance, (c) low-rank sketches, and (d) higher-order moments when computationally feasible. We additionally ablate the distance Dist (e.g., Frobenius, Wasserstein-2 on Gaussians, or CORAL-style losses).

3. *What is aligned:* align prompt-filtered embeddings h versus align pre-filter features $\phi(X)$ versus align only graph-level pooled summaries. This tests whether alignment must act *after* spectral modulation to be effective.

4. *Domain sampling strategy:* uniform over domains versus proportional to N_d, and paired sampling (d, d') chosen uniformly versus hard-negative pairing by large discrepancy in current statistics.

We report not only downstream performance but also the achieved alignment metric value and its correlation with few-shot error, to verify that alignment is not merely a regularizer but is predictive of transfer.

Prompt size scaling and approximation behavior. We test the dependence on the prompt dimension/degree m. For each benchmark we sweep

$$m \in \{0, 2, 4, 8, 16, 32, 64\},$$

keeping the backbone fixed. We report (i) performance as a function of m at fixed K, (ii) performance as a function of K at fixed m, and (iii) the empirical compute overhead relative to $m = 0$ to connect accuracy gains to the linear-time filtering cost. On synthetic corpora with known ground-truth spectral transfer ψ_d, we additionally measure the uniform approximation error $\|\psi_d - g_{p_d}\|_\infty$ (estimated on eigenvalues) to directly relate m to an ε-proxy, thereby operationalizing Theorem 1 and checking for the expected logarithmic trend until finite-sample effects dominate.

Cross-domain generalization and new-domain adaptation. We evaluate generalization to unseen domains by holding out a subset of domains during pre-training. At test time for a new domain d' we compare:

1. *Zero-shot:* reuse a pooled prompt (e.g., the average $\bar{p}$) or choose the nearest existing prompt in statistics space.
2. *Unlabeled prompt learning:* initialize $p_{d'}$ randomly and optimize only $p_{d'}$ on unlabeled graphs to minimize the same alignment objective (no labels), then perform few-shot supervised adaptation. We vary the number N of unlabeled graphs available from d' and report downstream risk as a function of N, testing the qualitative scaling suggested by Theorem 5.

We also test robustness to domain-mismatch in graph sizes and degree distributions, controlling for trivial covariate shift via normalization where appropriate.

Sequential domain arrival and catastrophic forgetting. To test continual deployment, we simulate a sequence of domains $(d_1, \ldots, d_T)$ arriving over time. We pre-train on unlabeled data from all domains or from the prefix only (two regimes), but we *adapt* prompts sequentially using few-shot labels without revisiting past labeled sets. We consider two systems: (i) one prompt per domain learned independently with the backbone frozen, and (ii) a shared pool of $M \ll D$ prompts updated over time with routing. Forgetting is quantified by backward transfer:

$$\mathrm{BWT} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{1}{t-1} \sum_{j=1}^{t-1} \left(\mathrm{Perf}_{d_j}^{(t)} - \mathrm{Perf}_{d_j}^{(j)} \right),$$

where $\mathrm{Perf}_{d_j}^{(t)}$ is performance on domain d_j after adapting to domains up to time t. Since the backbone is frozen and prompts are disjoint across domains in the simplest setting, we expect BWT to be near zero; any deviation indicates interference via shared components (e.g., routing, shared prompt pools, or shared running statistics). We additionally measure forward transfer to newly arriving domains relative to training prompts from scratch.

Reporting and reproducibility. We report parameter counts (trainable and total), adaptation wall-clock time, and memory overhead per domain. All results include confidence intervals over seeds and labeled-set draws. Together, these experiments are intended to connect the theoretical knobs (m, K, δ) to empirical behavior, and to delimit the regimes where prompt-only spectral modulation with alignment is competitive with heavier domain-specific adaptation.

10 Discussion and Future Work

Our development is intentionally built around a strong but analyzable structural hypothesis, namely that the dominant cross-domain variation is cap-

tured by commuting (or nearly commuting) Laplacians so that a single latent eigenbasis U suffices up to perturbation. This yields a clean separation between *shared geometry* (encoded by the frozen backbone f_θ) and *domain-specific spectral reweighting* (encoded by a compact prompt p_d). The empirical program in Section 9 is meant to test how often this separation is effective in practice; here we highlight several directions in which the hypothesis, the prompt class, and the alignment mechanism can be extended.

Beyond commuting Laplacians: approximate joint structure. The impossibility statement (Theorem 4) indicates that without additional structure one cannot hope to align arbitrarily rotated eigenbases using $m = o(n)$ prompt parameters. A natural next step is to identify intermediate assumptions between exact commutation and complete adversarial rotation. One candidate is *approximate joint diagonalizability*: there exists an orthogonal U such that $U^\top L_d U$ is *approximately* diagonal for all d, with off-diagonal energy bounded in a norm that controls downstream risk. Another candidate is *subspace commutation*: only the top-r (low-frequency) eigenspaces are shared, while high frequencies may be domain-specific. In such a regime we may replace a single global U by a decomposition $U = [U_{\text{shared}}, U_{\text{res}}^{(d)}]$ and seek prompts that act primarily on the shared subspace, possibly augmented by a small domain-specific residual module. Formalizing this would require perturbation bounds that depend on principal angles between eigenspaces rather than solely on $\|\Delta_d\|_2$, and would likely lead to risk terms scaling with these angles and with the task's effective bandwidth.

A complementary direction is to move from Laplacians to other operators whose eigenstructure might be more stable across domains, e.g., random-walk matrices, personalized PageRank operators, or diffusion kernels. In heterogeneous settings one may have multiple relation-specific Laplacians $L_d^{(r)}$; even if each family $\{L_d^{(r)}\}_d$ approximately commutes, the relations may not commute with each other. This suggests prompts that parameterize a *mixture of diffusions* across relations, with alignment losses applied per relation or to their joint moments.

Heterophily and the limits of low-frequency alignment. Spectral methods tied to Laplacian smoothing are often most effective under homophily, whereas many real graphs exhibit heterophily, where predictive signals can concentrate in higher frequencies. Our prompt class as polynomial filters is not restricted to low-pass behavior, but the alignment objective (as stated in terms of second moments) may implicitly prefer representations that are stable under smoothing. A precise future direction is to incorporate *frequency-aware* alignment: instead of matching a single covariance Σ_d, we may match covariances after applying multiple band-pass projections (or learned spectral windows) so that both low- and high-frequency components

are aligned when beneficial. Technically, this amounts to aligning statistics of $P_\ell h$ for a family of commuting projectors $\{P_\ell\}$ approximated by polynomials in L. Such a modification would also interact with Theorems 1–3: if the downstream transfer function $\psi_d(\lambda)$ is not smooth or is sharply band-pass, the degree m required for uniform approximation may increase, and one may prefer rational filters or multi-resolution bases to reduce prompt length.

Dynamic graphs: time-varying operators and prompt trajectories. Many domains are dynamic: graphs evolve over time, and so do their Laplacians. A static prompt p_d may be insufficient when the spectral content drifts. One extension is to treat prompts as *time-indexed* parameters $p_{d,t}$ governed by a low-dimensional dynamical model (e.g., an autoregressive prior), trained by minimizing a combination of self-supervised loss and temporal smoothness $\|p_{d,t} - p_{d,t-1}\|_2^2$. If $L_{d,t} = U\Lambda_{d,t}U^\top + \Delta_{d,t}$ with slowly varying $\Lambda_{d,t}$, then polynomial filters can track the drift by adjusting coefficients rather than changing the backbone. This setting raises a concrete statistical question: how many unlabeled snapshots are needed to estimate a prompt trajectory with bounded cumulative alignment error, and how does this trade off with the prompt degree m and the drift rate $\|\Lambda_{d,t} - \Lambda_{d,t-1}\|$? Theorem 5 suggests a moment-estimation route for static prompts; a dynamic analogue would require concentration for dependent data and stability analyses for online prompt updates.

Heterogeneous graphs and typed features. In many text-free corpora, node and edge types (categorical identifiers) are available even when text is not. Our formulation already allows general attributed graphs through $\phi(X)$, but heterogeneous graphs introduce additional operators beyond a single Laplacian. One approach is to define a block operator $\mathcal{L}$ on a lifted space that encodes types and relations, and to parameterize prompts as polynomials in $\mathcal{L}$. Another approach is to retain relation-specific operators and let prompts control a small set of mixing coefficients across relations, yielding a structured prompt $p_d = (p_d^{\text{spec}}, p_d^{\text{mix}})$ where p_d^{spec} sets spectral responses and p_d^{mix} controls inter-relation aggregation. The theoretical challenge is then to state an analogue of the commuting hypothesis for a *family* of operators and to characterize when a shared backbone is identifiable from unlabeled data.

Discrete and structured prompts. Our prompts are continuous vectors in $\mathbb{R}^m$ that define polynomial coefficients. For deployment, it can be advantageous to enforce additional structure: sparsity, quantization, or compositionality. A discrete prompt could be an index into a codebook of spectral

filters, or a sparse combination of a few basis filters, i.e.,

$$g_{p_d}(L) = \sum_{j=1}^{J} \alpha_{d,j} g_j(L), \qquad \|\alpha_d\|_0 \le s,$$

with $s \ll J$. This would reduce per-domain storage and may enable fast routing when domain identity is unknown. From a theoretical standpoint, such a restriction changes the approximation problem from polynomial approximation to dictionary approximation, suggesting bounds in terms of covering numbers of the filter family and the coherence between basis filters. More structured prompts also enable explicit constraints such as monotonicity (for stability) or positivity (to avoid amplifying noise), which can be expressed as convex constraints on polynomial coefficients in certain bases.

We also view rational filters (ratios of polynomials) or multi-scale wavelet constructions as promising: they can approximate sharp spectral responses with fewer parameters than a single global polynomial degree. The trade-off is numerical stability and the need for iterative solvers; one might recover parameter-efficiency by restricting denominators to a low-dimensional family that admits stable recurrences.

Connections to graph foundation models: scaling, modularity, and domain discovery. The proposed separation (frozen shared backbone plus small prompts) is aligned with the emerging practice of training large shared encoders and adapting via lightweight modules. In graph settings, however, domains are often implicit; domain labels may not be present, and distribution shift may be continuous rather than discrete. This motivates *prompt discovery* and *prompt routing* from unlabeled data: learn a small set of prompts $\{p^{(1)}, \ldots, p^{(M)}\}$ and a router $R(G)$ that selects or mixes prompts at inference. In our framework, routing can be grounded in spectral statistics: we may define R to minimize an alignment discrepancy between a graph's estimated moments and prompt-conditioned target moments. The theoretical question becomes a clustering/mixture problem in the space of operator moments, where sample complexity depends on the separation between domains in moment space and on the prompt family's expressivity.

Scaling also raises optimization questions: if D is large, storing $O(Dm)$ prompts can be burdensome. Structured prompts (shared dictionaries, low-rank prompt matrices, or hypernetworks generating p_d from domain descriptors computed from unlabeled graphs) offer a compression route while retaining per-domain specialization.

Interpretability and diagnostics. A practical advantage of spectral prompts is that they admit a direct interpretation as frequency responses. After training, we can inspect $g_{p_d}(\lambda)$ over $\lambda \in [0, 2]$ (or over empirical eigenvalues) to quantify whether a domain emphasizes low frequencies, high frequencies, or

band-pass behavior. Such diagnostics can be used to (i) detect out-of-family domains for which the learned prompt induces unstable amplification, (ii) measure how much adaptation occurs during few-shot tuning (e.g., norm changes in polynomial coefficients), and (iii) relate improvements to specific frequency bands. More formally, one may attempt to derive a stability certificate in terms of $\sup_{\lambda\in[0,2]} |g_{p_d}(\lambda)|$ and Lipschitz constants of the readout, connecting interpretability to generalization control.

Finally, while our analysis uses a linear spectral label model to obtain explicit bounds, the mechanism is not inherently tied to linear tasks. Extending the theory to nonlinear readouts and to richer self-supervised objectives (beyond moment matching) remains open; a plausible route is to treat the prompt-filtered encoder as defining a domain-conditioned representation and to analyze invariance and sufficiency properties via information-theoretic or kernel-based tools. Establishing when such representations remain transferable under weaker structural assumptions is, in our view, the central theoretical problem for prompt-based multi-domain graph learning.

11 Conclusion

We have studied the following constraint regime: a collection of graph domains $\{1, \ldots, D\}$ provides large unlabeled corpora, yet downstream supervision is scarce and domain-specific, and we require parameter-efficient adaptation in the sense that the shared encoder parameters θ are frozen at downstream time while only a compact domain prompt $p_d \in \mathbb{R}^m$ (and, optionally, a small linear head) may be optimized from K labeled examples. The setting is text-free, so adaptation cannot rely on language supervision or textual node/edge attributes, and thus must be grounded in intrinsic graph structure.

Our central modeling choice is to represent cross-domain shift at the level of graph spectra. Formally, we posit that the (normalized) Laplacians admit a shared latent eigenbasis up to perturbation, i.e.,

$$L_d \;=\; U\Lambda_d U^\top + \Delta_d, \qquad \|\Delta_d\|_2 \le \delta,$$

where U is orthogonal and Λ_d is diagonal with spectrum in $[0, 2]$. This hypothesis is deliberately strong, but it is precise enough to yield a quantitative theory for when prompt-only adaptation is feasible and when it is not. Under this hypothesis, a domain can differ from another primarily by a reweighting of the same spectral directions; accordingly, a prompt that parameterizes a spectral filter can compensate for domain-specific eigenvalue profiles without modifying the shared representation geometry encoded by the backbone.

On the algorithmic side, we introduced a two-stage procedure in which the prompt enters the encoder through a prompt-parameterized spectral operator $g_{p_d}(L)$, implemented as a degree-m polynomial (e.g., via Chebyshev

recurrences), and pre-training is driven by a combination of self-supervised learning and an explicit cross-domain alignment regularizer. The role of the self-supervised term is to learn broadly useful representations on each domain, whereas the role of the alignment term is to encourage prompt-conditioned representations to share comparable statistics across domains, thereby increasing the likelihood that a small amount of labeled supervision suffices for downstream adaptation. The resulting parameterization is explicit: domain-specificity is confined to p_d of dimension m, while the computationally expensive backbone is shared and amortized across domains.

The theoretical statements we established can be read as a three-part characterization of prompt efficiency under spectral structure. First, in the idealized commuting case $\delta = 0$, polynomial spectral filters provide a controlled approximation class for domain-specific transfer functions. Concretely, when a downstream task depends on a (band-limited, Lipschitz) spectral transfer $\psi_d(\lambda)$, we can approximate ψ_d uniformly on $[0, 2]$ by a degree-m polynomial g_{p_d} with $m = \tilde{O}(\log(1/\varepsilon))$, which translates into an $O(\varepsilon)$ degradation relative to the corresponding domain-optimal spectral transform. Second, when $\delta > 0$ and the shared eigenbasis is only approximate, classical perturbation arguments yield an additional risk contribution scaling as $O(\delta)$, reflecting that no prompt acting only through L can fully undo a rotation of eigenspaces beyond what δ permits. Third, since downstream training adjusts only m prompt parameters (and possibly a small head), estimation error is governed by the complexity of an m-dimensional hypothesis class, yielding an excess risk term of order $\tilde{O}(\sqrt{m/K})$ under standard boundedness assumptions.

Taken together, these components give an upper bound of the form

$$\mathbb{E}[\mathrm{Risk}_d(\widehat{p}_d)] - \mathrm{Risk}_d(p_d^\star) \;\le\; O(\varepsilon + \delta) \;+\; \tilde{O}\left(\sqrt{\frac{m}{K}}\right),$$

where $p_d^\star$ denotes the best prompt within the degree-m polynomial class. This bound makes the relevant tradeoffs explicit: decreasing approximation error requires increasing m; decreasing estimation error requires increasing K relative to m; and any lack of shared spectral structure manifests as an irreducible term scaling with δ (or, more generally, with the deviation from joint diagonal structure).

We also proved that these rates are essentially sharp in two distinct senses. In the shared-basis case, the logarithmic dependence of prompt size on target uniform error cannot in general be improved: there exist Lipschitz spectral transfer functions for which any degree-m polynomial approximation achieving error at most ε requires $m = \Omega(\log(1/\varepsilon))$. Thus, within the polynomial prompt family, the prompt-size requirement in the approximation theorem is not an artifact of proof technique but reflects the inherent approximation difficulty on a compact spectral interval. Separately, in the absence of shared spectral structure, we established an impossibility statement: if the

eigenbases across domains are sufficiently misaligned (for instance, nearly orthogonal), then any method that maintains a fixed backbone and adapts only through an $m = o(n)$-dimensional prompt family applied in a fixed basis can be forced to incur constant excess risk on at least one domain. This lower bound delineates the boundary of applicability of prompt-only adaptation: without a joint structure constraint, low-dimensional prompts do not have the capacity to align arbitrarily rotated spectral geometries.

A further contribution concerns adaptation to previously unseen domains using unlabeled data. If a new domain d' satisfies the same latent basis hypothesis with unknown $\Lambda_{d'}$, then moment-based alignment provides a viable mechanism to estimate a prompt from unlabeled graphs alone. We showed that, for an alignment objective based on Laplacian moments up to order m, the number of unlabeled samples required to achieve alignment error at most ε scales as $\tilde{O}((m + \log(1/\rho))/\varepsilon^2)$ with failure probability ρ, and that an $\Omega(m/\varepsilon^2)$ dependence is necessary in general. This result complements the downstream few-shot bound: it formalizes that prompt learning can be statistically efficient even without labels, provided that one is willing to estimate sufficiently many spectral statistics to identify an appropriate filter within the degree-m family.

We emphasize that the value of the present framework is not limited to the specific polynomial parameterization, but rather lies in making explicit the interface between (i) an operator-level description of domain shift, (ii) a parameter-efficient adaptation mechanism, and (iii) risk bounds that separate approximation, perturbation, and estimation effects. In particular, the prompt degree m plays a dual role: it is both a computational knob (degree-m filtering costs $O(m|E|)$ per graph under sparse multiplication) and a statistical knob (the effective dimension of the downstream adaptation problem). The upper and lower bounds jointly identify where improvement is plausible (e.g., better approximation families than polynomials, or stronger structural assumptions than approximate commutation) and where it is not (e.g., attempting to align arbitrary spectral rotations with $m = o(n)$).

In conclusion, we have provided a principled account of how compact, domain-specific prompts can support few-shot transfer in text-free multi-domain graph learning when cross-domain variation admits a shared spectral geometry. The resulting picture is logically consistent: when domains share a latent eigenbasis, prompts can reweight frequencies to bridge domain gaps with provable sample-efficiency; when domains do not share such structure, prompt-only schemes are provably insufficient in the worst case. Establishing analogous guarantees under weaker and more realistic structural assumptions, and characterizing the empirical prevalence of approximate joint spectral structure in large graph corpora, remain the primary obstacles to turning this theory into a general foundation for prompt-based graph transfer.