FEDERATED FEW-SHOT GRAPH DOMAIN GENERALIZATION WITH NON-ALIGNED LABEL SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated domain generalization (FDG) on graphs aims to collaboratively learn from decentralized source domains and generalize to unseen target domains without centralizing private graph data. Existing methods typically assume a predefined prediction task with globally aligned label spaces, limiting their applicability to heterogeneous graph domains with different class cardinalities and semantics. To address this limitation, we formulate *Federated Few-Shot Graph Domain Generalization with Non-Aligned Label Spaces*, where clients learn transferable model parameters without cross-domain label correspondence and predict query nodes in an unseen graph from only a few labeled context nodes. We further propose `FedDARC`, which first characterizes graph domains through the optimization responses of a common reference encoder, learns geometry-preserving domain states, and uses them to residually calibrate node representations. It then performs label-space-agnostic relational inference by constructing task-relative labels and aggregating query-to-context relational evidence, avoiding globally shared class semantics or a fixed-output classifier. At inference, the learned shared model remains fixed, while the target domain state and task-relative labels are derived from the labeled context nodes for prediction. Extensive experiments on graph datasets demonstrate the effectiveness of `FedDARC` in generalizing to unseen graph domains with non-aligned label spaces. Our code is available at https://anonymous.4open.science/r/fedgraph-0B15.

1 INTRODUCTION

Federated learning (FL) (McMahan et al., 2017) enables multiple clients to collaboratively train a model without sharing private data. However, conventional FL mainly targets participating clients and often assumes similar training and test distributions. Federated domain generalization (FDG) (Li et al., 2023) extends FL by learning from decentralized source domains to generalize to an unseen target domain unavailable during federated training. Extending FDG to graphs is more challenging (Wu et al., 2022) because graph domains may differ in feature distributions, structures, and input feature spaces, while graph data often cannot be centralized due to privacy or ownership constraints.

Related Works: Recent studies have explored generalization under distribution shifts in federated graph learning. FedGOG (Zhou et al., 2025) improves graph-level OOD generalization by generating diverse graph samples and separating invariant from domain-specific information. FGOOD (Zhu et al., 2025) enhances node-level generalization through representation propagation and scattering, while FedLoG (Kim et al., 2025) transfers condensed local knowledge to generalize to unseen nodes, missing classes, and new clients. Despite their effectiveness, these methods generally assume a predefined prediction task or globally aligned label space, allowing a common classifier to be shared across clients. This assumption fails for heterogeneous graph datasets with different class cardinalities and unrelated semantics, where identical labels may represent different categories and target classes may be unseen during training. A fixed global classifier is therefore difficult to share across clients or transfer to an unseen target domain.

Motivation: Our key observation is that transferable knowledge across heterogeneous graph domains need not depend on specific class semantics. Even with unrelated label spaces, two capabilities remain transferable: adapting representations to domain-specific characteristics and inferring class membership from a few labeled examples. We therefore move beyond aggregating a classifier

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

over predefined categories and instead learn a label-space-independent prediction mechanism. At deployment, a few labeled target nodes define the current task and provide class evidence without requiring semantic correspondence with source labels.

Contribution: Motivated by this setting, we formulate *Federated Few-Shot Graph Domain Generalization with Non-Aligned Label Spaces*, where heterogeneous source clients collaboratively learn a transferable model without shared label semantics and predict query nodes in an unseen graph from only a few labeled context nodes. We further propose `FedDARC`, which derives domain states from the optimization responses of a common reference encoder to calibrate node representations and performs task-relative relational classification instead of using a fixed global classifier. Only shared, label-space-independent parameters are collaboratively optimized and remain fixed during target-domain inference. Our main contributions are summarized as follows:

- We formulate *Federated Few-Shot Graph Domain Generalization with Non-Aligned Label Spaces*, extending graph FDG to heterogeneous source and target graphs with different class cardinalities and semantics.
- We propose `FedDARC`, combining optimization-response-based domain characterization, domain-state-guided residual calibration, and task-relative relational inference to enable transferable prediction without globally aligned class semantics or a fixed-output classifier.
- We characterize how domain-state-guided calibration can improve task-relative relational inference and empirically demonstrate the effectiveness of the complete framework across unseen graph domains and few-shot settings.

2 PRELIMINARIES

Federated Learning across Heterogeneous Graph Domains: Let $\mathcal{K} = \{1, \ldots, K\}$ denote the source clients. Each client $k \in \mathcal{K}$ holds a raw graph $\mathcal{G}_k^{\text{raw}} = (\mathcal{V}_k, \mathcal{E}_k, \mathbf{X}_k^{\text{raw}}, \mathbf{Y}_k)$, where $\mathcal{V}_k$ and $\mathcal{E}_k$ are the node and edge sets, $\mathbf{X}_k^{\text{raw}} \in \mathbb{R}^{|\mathcal{V}_k| \times d_k}$ is the node feature matrix, and $\mathbf{Y}_k \in \mathcal{Y}_k^{|\mathcal{V}_k|}$ contains the node labels. Following the prior work (Yu et al., 2025), each client applies $\mathbf{X}_k \leftarrow \text{SVD}\left(\mathbf{X}_k^{\text{raw}}\right) \in \mathbb{R}^{|\mathcal{V}_k| \times d_0}$ to obtain a common input width d_0. Then define the dimension-unified graph as $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k, \mathbf{X}_k, \mathbf{Y}_k)$. Each graph follows a distribution $\mathcal{G}_k \sim \mathcal{D}_k$, where $\mathcal{D}_k$ characterizes its node attributes, graph structure, and local prediction task. For distinct $i, j \in \mathcal{K}$, $\mathcal{D}_i \neq \mathcal{D}_j, d_i^{\text{raw}} \neq d_j^{\text{raw}}$, capturing heterogeneity in feature distributions, graph structures, and input feature dimensions.

We further consider non-aligned label spaces. Client k has $\mathcal{Y}_k = \{c_{k,1}, \ldots, c_{k,C_k}\}$, where $C_k = |\mathcal{Y}_k|$. For distinct clients i and j, their class cardinalities and semantics need not align: $C_i \neq C_j$ and/or $\mathcal{Y}_i \not\equiv \mathcal{Y}_j, i, j \in \mathcal{K},\ i \neq j$, where $\not\equiv$ denotes the absence of semantic alignment. Thus, the same local class index may represent different categories across clients, i.e., $c_{i,r} \not\equiv c_{j,r}$ in general, precluding a globally predefined output space or conventional class-specific prediction head.

During federated training, raw graphs remain local, while clients collaboratively optimize label-space-independent parameters $\Theta^* = \arg\min_\Theta \sum_{k=1}^K \omega_k \mathcal{L}_k\left(\Theta; \mathcal{G}_k\right), \sum_{k=1}^K \omega_k = 1$, where $\mathcal{L}_k$ is the local objective and ω_k is the aggregation weight.

Federated Few-Shot Graph Domain Generalization: After federated training, the raw target features $\mathbf{X}_t^{\text{raw}}$ are independently mapped by SVD to the common input width d_0, yielding the dimension-unified target graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, \mathbf{X}_t, \mathbf{Y}_t) \sim \mathcal{D}_t$, where $\mathbf{X}_t \in \mathbb{R}^{|\mathcal{V}_t| \times d_0}$ and $\mathcal{D}_t \notin \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$. Its independently defined label space is $\mathcal{Y}_t = \{c_{t,1}, \ldots, c_{t,C_t}\}$, where $C_t = |\mathcal{Y}_t|$, and $\mathcal{Y}_t \not\equiv \mathcal{Y}_k, \forall k \in \mathcal{K}$. Although $\mathbf{Y}_t$ contains the complete ground-truth labels, only a small context node set $\mathcal{A}_t = \{v_i\}_{i=1}^{N_t} \subseteq \mathcal{V}_t$ and the corresponding labels $\{y_{t,v} : v \in \mathcal{A}_t\}$ are available at inference. Under the C_t-way S-shot setting, the context set contains S labeled nodes from each target class: $|\{v \in \mathcal{A}_t : y_{t,v} = c\}| = S, \quad \forall c \in \mathcal{Y}_t$, and hence $N_t = C_t S$. The remaining nodes form $\mathcal{Q}_t = \mathcal{V}_t \setminus \mathcal{A}_t$, whose labels are hidden from the model. Given Θ^*, $\mathcal{G}_t$, and $\mathcal{A}_t$, the objective is $\hat{y}_{t,v} = F_{\Theta^*}\left(v, \mathcal{G}_t, \mathcal{A}_t\right), v \in \mathcal{Q}_t$.

Neither the target graph nor its annotations participate in federated training, and Θ^* remains fixed during inference. The model must therefore generalize to an unseen graph and independently defined label space using only the labeled context nodes.

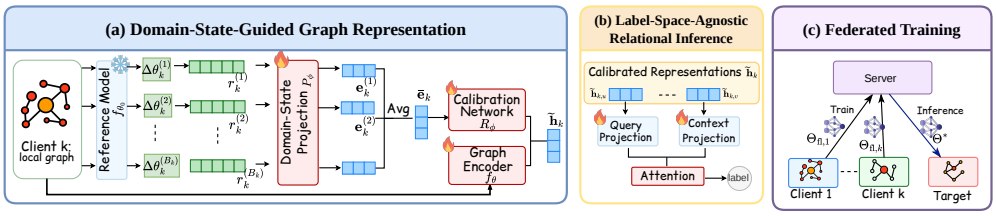


Figure 1: **Overview of the proposed framework.** (a) Normalized optimization responses are projected into domain states for representation calibration. (b) Relational inference over calibrated representations predicts task-relative labels. (c) The remaining shared parameters are optimized federatively after freezing the domain-state projection.

3 METHODOLOGY

We propose FedDARC to learn transferable prediction capabilities across heterogeneous graph domains without globally aligned class semantics. As shown in Fig. 1, FedDARC consists of 3 stages. *Domain-State-Guided Graph Representation* characterizes each domain through a common reference model's optimization response and calibrates node representations with the resulting state. *Label-Space-Agnostic Relational Inference* predicts queries from relations to labeled context nodes using task-relative labels. *Federated Training and Unseen-Domain Inference* optimizes shared parameters across source graph and transfers them to an unseen target graph while keeping them fixed.

3.1 DOMAIN-STATE-GUIDED GRAPH REPRESENTATION

Source graphs may differ in feature distributions, graph structures, and raw input dimensions. We therefore characterize each domain through the optimization response of a common reference model to its local data and use the resulting compact domain state to calibrate node representations.

Gradient-Based Domain Characterization:

Let θ_0 denote the fixed parameters of a common reference encoder f_{θ_0}. Client k partitions all local data into B_k response batches $\{\mathcal{B}_k^{(b)}\}_{b=1}^{B_k}$. For response generation, we use the task-relative relational classification objective in Eq. (14), evaluated on the uncalibrated representations produced by the reference encoder, and denote the resulting batch-wise objective by $\mathcal{L}_{\text{resp}}(\theta; \mathcal{B}_k^{(b)})$. Starting from the same reference parameters θ_0, each response batch induces one virtual gradient step. Specifically,

$$\theta_k^{(b)} = \theta_0 - \eta_d \nabla_\theta \mathcal{L}_{\text{resp}} \left(\theta; \mathcal{B}_k^{(b)}\right), \tag{1}$$

where η_d denotes the response-step learning rate. The corresponding optimization response is

$$\Delta\theta_k^{(b)} = \theta_k^{(b)} - \theta_0, \quad b = 1, \ldots, B_k. \tag{2}$$

The virtual update is used only to probe how the common reference model responds to the local graph data: it does not update θ_0 and is separate from the subsequent federated model optimization. Because all clients use the same reference initialization and response-generation procedure, differences in $\Delta\theta_k^{(b)}$ capture differences in the local optimization responses rather than differences in initialization.

Since $\Delta\theta_k^{(b)}$ contains parameter tensors from multiple layers, each displacement is vectorized and normalized to characterize its relative optimization pattern rather than its absolute update magnitude:

$$\mathbf{r}_k^{(b)} = \frac{\text{vec}(\Delta\theta_k^{(b)})}{\left\|\text{vec}(\Delta\theta_k^{(b)})\right\|_2 + \epsilon}, \tag{3}$$

where $\text{vec}(\cdot)$ concatenates the parameter tensors and $\epsilon > 0$ ensures numerical stability. Each $\mathbf{r}_k^{(b)}$ represents the normalized response of the common reference encoder to response batch $\mathcal{B}_k^{(b)}$, while $\{\mathbf{r}_k^{(b)}\}_{b=1}^{B_k}$ provides multiple observations of source domain $\mathcal{D}_k$.

Geometry-Preserving Domain-State Learning: The normalized response $\mathbf{r}_k^{(b)}$ remains high-dimensional and is therefore unsuitable for directly conditioning node representations. During domain-state initialization, each client uploads its normalized response vectors $\{\mathbf{r}_k^{(b)}\}_{b=1}^{B_k}$ to the server once. The server then learns a shared projection P_ϕ that maps each response into a compact domain state as $\mathbf{e}_k^{(b)} = P_\phi\left(\mathbf{r}_k^{(b)}\right)$. Because different response batches from client k characterize the same underlying graph domain, we summarize their response and projected-state centers as

$$\bar{\mathbf{r}}_k = \frac{1}{B_k}\sum_{b=1}^{B_k}\mathbf{r}_k^{(b)}, \qquad \bar{\mathbf{e}}_k = \frac{1}{B_k}\sum_{b=1}^{B_k}\mathbf{e}_k^{(b)}. \tag{4}$$

The client-level domain state is defined as $\mathbf{e}_k := \bar{\mathbf{e}}_k$. To retain the information contained in the optimization responses, the projected state space is encouraged to preserve inter-domain geometry, maintain distinguishability across source domains, and remain consistent across response batches from the same domain. Specifically, we first preserve inter-domain geometry by matching pairwise distances before and after projection:

$$\mathcal{L}_{\text{dist}} = \frac{1}{K(K-1)}\sum_{\substack{i,j\in\mathcal{K}\\ i\neq j}}\left[D(\mathbf{e}_i,\mathbf{e}_j) - D(\bar{\mathbf{r}}_i,\bar{\mathbf{r}}_j)\right]^2, \quad D(\mathbf{a},\mathbf{b}) = \left\|\frac{\mathbf{a}}{\|\mathbf{a}\|_2+\epsilon} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2+\epsilon}\right\|_2. \tag{5}$$

Eq. (5) encourages the compact state space to preserve the relative pairwise geometry observed in the normalized optimization-response space.

Second, to enhance domain distinguishability, we encourage different source domains to occupy distinct directions. Let $\mathbf{M}_e$ collect the normalized domain-state centers, where $[\mathbf{M}_e]_{k,:} = \frac{\mathbf{e}_k^\top}{\|\mathbf{e}_k\|_2+\epsilon}$. Then, the domain-diversity objective is

$$\mathcal{L}_{\text{div}} = -\log\det\left(\mathbf{M}_e\mathbf{M}_e^\top + \epsilon\mathbf{I}\right), \tag{6}$$

discouraging the normalized source-domain states from collapsing into highly similar directions.

Third, to ensure within-domain consistency, we penalize state variation across response batches from the same client as

$$\mathcal{L}_{\text{stab}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{B_k}\sum_{b=1}^{B_k}\left\|\mathbf{e}_k^{(b)} - \mathbf{e}_k\right\|_2^2. \tag{7}$$

Then, the complete domain-state learning objective is $\mathcal{L}_{\text{state}} = \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{stab}}\mathcal{L}_{\text{stab}}$. After receiving the one-time source response collection, the server optimizes P_ϕ using $\mathcal{L}_{\text{state}}$ and then freezes the resulting projection P_{ϕ^*}. The fixed source-domain state used during subsequent federated training is therefore

$$\mathbf{e}_k = \frac{1}{B_k}\sum_{b=1}^{B_k}P_{\phi^*}\left(\mathbf{r}_k^{(b)}\right). \tag{8}$$

Thus, the domain-state projection is learned only during domain-state initialization and is not updated during the subsequent federated optimization stage.

Domain-State-Guided Residual Calibration: The fixed domain state $\mathbf{e}_k$ provides a compact conditioning signal for calibrating the shared node representations to client k. Let f_θ denote the shared graph encoder and $\mathbf{h}_{k,v} = f_\theta(v,\mathcal{G}_k)$ the representation of node $v \in \mathcal{V}_k$. A shared calibration network R_ψ maps $\mathbf{e}_k$ to feature-wise scaling and shifting parameters: $[\boldsymbol{\gamma}(\mathbf{e}_k), \boldsymbol{\beta}(\mathbf{e}_k)] = R_\psi(\mathbf{e}_k)$.

To avoid unnecessary changes when little domain-specific adjustment is needed, we define the calibration as a residual relative to the zero state:

$$\Delta\boldsymbol{\gamma}(\mathbf{e}_k) = \boldsymbol{\gamma}(\mathbf{e}_k) - \boldsymbol{\gamma}(\mathbf{0}), \quad \Delta\boldsymbol{\beta}(\mathbf{e}_k) = \boldsymbol{\beta}(\mathbf{e}_k) - \boldsymbol{\beta}(\mathbf{0}). \tag{9}$$

The calibrated representation is then

$$\widetilde{\mathbf{h}}_{k,v} = \mathbf{h}_{k,v} + \lambda_{\text{cal}}\left[\Delta\boldsymbol{\gamma}(\mathbf{e}_k)\odot\mathbf{h}_{k,v} + \Delta\boldsymbol{\beta}(\mathbf{e}_k)\right], \tag{10}$$

where $\lambda_{\text{cal}} \geq 0$ controls the calibration strength and $\odot$ denotes element-wise multiplication. Centering at the zero state ensures no residual adjustment when $\mathbf{e}_k = \mathbf{0}$, while allowing domain-specific feature calibration otherwise.

3.2 LABEL-SPACE-AGNOSTIC RELATIONAL CLASSIFICATION

Domain-state-guided calibration handles graph-domain heterogeneity but not non-aligned label spaces. Since class cardinalities and semantics vary across clients, a fixed-output classifier cannot be shared. We therefore predict queries from their relations to labeled context nodes rather than globally predefined class identities.

Task-Relative Label Construction: For an S-shot task on client k, we use all classes in $\mathcal{Y}_k$ and sample a context node set $\mathcal{A}_k \subseteq \mathcal{V}_k$ with S labeled nodes per class, together with a disjoint query node set $\mathcal{Q}_k \subseteq \mathcal{V}_k$: $|\{v \in \mathcal{A}_k : y_{k,v} = c\}| = S, \forall c \in \mathcal{Y}_k, \mathcal{A}_k \cap \mathcal{Q}_k = \varnothing$. Thus, $|\mathcal{A}_k| = C_k S$. Since local class identities are not comparable across clients, we re-index the local classes within the current task as $\rho_k : \mathcal{Y}_k \to \{1, \ldots, C_k\}$. The mapping ρ_k is task-specific and never shared across clients. Thus, each relative index identifies a class only within the current task and implies no cross-client semantic correspondence.

Query-to-context relational modeling: Given the task-relative classes and calibrated representations $\widetilde{\mathbf{h}}_{k,v}$, we infer each query by measuring its relations to labeled context nodes. Rather than using a fixed distance metric, we learn a transferable relation space through two projections. For query node $u \in \mathcal{Q}_k$ and context node $v \in \mathcal{A}_k$,

$$\mathbf{q}_{k,u} = \mathbf{W}_Q \,\mathrm{LN}\left(\widetilde{\mathbf{h}}_{k,u}\right), \qquad \mathbf{k}_{k,v} = \mathbf{W}_K \,\mathrm{LN}\left(\widetilde{\mathbf{h}}_{k,v}\right), \tag{11}$$

where $\mathbf{W}_Q$ and $\mathbf{W}_K$ are trainable projections, $\mathrm{LN}(\cdot)$ denotes layer normalization, and $\mathbf{q}_{k,u}, \mathbf{k}_{k,v} \in \mathbb{R}^{d_a}$. Layer normalization reduces scale variation, while the projections learn dimensions informative for transferable query-to-context matching. We then measure their relation by $s_{k,u,v} = \frac{\mathbf{q}_{k,u}^\top \mathbf{k}_{k,v}}{\sqrt{d_a}\,\tau}$, where $\sqrt{d_a}$ controls the dot-product scale and $\tau > 0$ controls the sharpness of the relation distribution. Then, the scores are normalized over the context set:

$$a_{k,u,v} = \frac{\exp(s_{k,u,v})}{\sum_{v' \in \mathcal{A}_k} \exp(s_{k,u,v'})}, \qquad v \in \mathcal{A}_k, \tag{12}$$

where $a_{k,u,v}$ measures the relative evidence contributed by context node v to query node u.

Context-conditioned label prediction: To convert relational evidence into class predictions, each context node v is represented only by its task-relative one-hot label $\mathbf{y}_{k,v} = \mathrm{OneHot}\left(\rho_k(y_{k,v})\right) \in \mathbb{R}^{C_k}$. The class distribution of query node u is $\widehat{\mathbf{p}}_{k,u} = \sum_{v \in \mathcal{A}_k} a_{k,u,v} \mathbf{y}_{k,v}$. Equivalently, the probability assigned to class $c \in \{1, \ldots, C_k\}$ is

$$\widehat{p}_{k,u}(c) = \sum_{v \in \mathcal{A}_k} a_{k,u,v} \mathbb{I}\left[\rho_k(y_{k,v}) = c\right]. \tag{13}$$

Thus, a class receives high probability when the query assigns strong relational evidence to its context nodes. Since the output dimension is determined by the current task through C_k, prediction does not require a globally aligned label space.

Relational classification objective: During source-domain training, query labels supervise the transferable relational mechanism. For client k, we minimize

$$\mathcal{L}_{\mathrm{cls}}^k = -\frac{1}{|\mathcal{Q}_k|} \sum_{u \in \mathcal{Q}_k} \log \widehat{p}_{k,u}\left(\rho_k(y_{k,u})\right). \tag{14}$$

Minimizing $\mathcal{L}_{\mathrm{cls}}^k$ encourages stronger relations to same-class context nodes and weaker relations to other classes. By resampling the context and query nodes across training tasks, the model learns transferable within-task relations rather than fixed mappings to source-domain class identities.

3.3 FEDERATED TRAINING AND UNSEEN-DOMAIN INFERENCE

The preceding modules produce domain-calibrated representations $\widetilde{\mathbf{h}}_{k,v}$ and task-relative predictions $\widehat{\mathbf{p}}_{k,u}$. We next describe their federated optimization and transfer to an unseen target graph.

Federated optimization: During the first round, each client transmits its normalized response vectors $\{\mathbf{r}_k^{(b)}\}_{b=1}^{B_k}$ to the server only once. The server uses the responses from all source clients to

optimize the projection P_ϕ according to $\mathcal{L}_{\text{state}}$. After the server-side domain-state projection P_ϕ has been frozen, let $\Theta_{\text{fl}} = \{\theta, \psi, \mathbf{W}_Q, \mathbf{W}_K\}$ denote the remaining shared trainable parameters. At communication round r, the server broadcasts $\Theta_{\text{fl}}^{(r)}$ to all clients $\mathcal{K}$. Each client initializes its local model from $\Theta_{\text{fl}}^{(r)}$ and updates it by minimizing its local classification objective $\mathcal{L}_{\text{cls}}^k$ in Eq. (14). The server aggregates the updates according to local data size as $\Theta_{\text{fl}}^{(r+1)} = \sum_{k\in\mathcal{K}} \frac{n_k}{\sum_{j\in\mathcal{K}} n_j} \Theta_{\text{fl},k}^{(r+1)}$, where n_k denotes the number of local training instances. Only shared parameters are communicated, while raw graph data, labels, context–query partitions, and task-relative mappings ρ_k remain local. Repeating this process yields $\Theta_{\text{fl}}^* = \{\theta^*, \psi^*, \mathbf{W}_Q^*, \mathbf{W}_K^*\}$, which together with the frozen ϕ^* forms the final parameter set $\Theta^* = \{\phi^*, \Theta_{\text{fl}}^*\}$.

Domain-state estimation on the unseen target graph: After federated training, the fixed Θ^* is transferred to the unseen target graph $\mathcal{G}_t$. Since $\mathcal{D}_t$ is unavailable during training, its domain state is estimated at inference from $\mathcal{A}_t$ using the same reference-response procedure. Specifically, we obtain B_t normalized responses $\{\mathbf{r}_t^{(b)}\}_{b=1}^{B_t}$ and compute $\mathbf{e}_t = \frac{1}{B_t}\sum_{b=1}^{B_t} P_{\phi^*}\left(\mathbf{r}_t^{(b)}\right)$. The resulting $\mathbf{e}_t$ calibrates the target representations through Eq. (10), enabling domain-specific calibration without optimizing target-specific model parameters.

Few-shot relational inference on the target graph: As on source clients, we construct the task-relative mapping $\rho_t : \mathcal{Y}_t \to \{1, \ldots, C_t\}$. Using the calibrated target representations and fixed $\mathbf{W}_Q^*$ and $\mathbf{W}_K^*$, Eqs. (11)–(13) yield $\widehat{\mathbf{p}}_{t,u}$ for each $u \in \mathcal{Q}_t$. The predicted task-relative class is mapped back to the target label space: $\widehat{y}_{t,u} = \rho_t^{-1}\left(\arg\max_{c\in\{1,\ldots,C_t\}} \widehat{p}_{t,u}(c)\right), u \in \mathcal{Q}_t$.

The temporary reference updates used for domain-state estimation are discarded afterward. Thus, Θ^* remains fixed and no target-specific classifier is optimized; target labels are used only through $\mathcal{A}_t$ to define task-relative classes and provide relational evidence.

4 Theoretical analysis

This section characterizes how calibration-induced changes in relational separation affect prediction. For a target query node $u \in \mathcal{Q}_t$ with task-relative label $c_u = \rho_t(y_{t,u})$, let $s_{u,v}^{\text{raw}}$ and $s_{u,v}^{\text{cal}}$ denote the relation scores computed from the original and calibrated representations, respectively, using the same relational parameters. For $\xi \in \{\text{raw}, \text{cal}\}$, define the class-wise relational evidence as $Z_{u,c}^{(\xi)} = \sum_{v\in\mathcal{A}_t(c)} \exp\left(s_{u,v}^{(\xi)}\right), \widehat{p}_{t,u}^{(\xi)}(c) = \frac{Z_{u,c}^{(\xi)}}{\sum_{j=1}^{C_t} Z_{u,j}^{(\xi)}}$, where $\mathcal{A}_t(c) = \{v \in \mathcal{A}_t : \rho_t(y_{t,v}) = c\}$. The correct-versus-incorrect relational evidence margin is

$$M_u^{(\xi)} := \log \frac{Z_{u,c_u}^{(\xi)}}{\sum_{c\neq c_u} Z_{u,c}^{(\xi)}}, \quad \widehat{p}_{t,u}^{(\xi)}(c_u) = \sigma\left(M_u^{(\xi)}\right). \tag{15}$$

Proposition 1 (Calibration-induced relational improvement)**.** *For query u, if calibration increases the relational evidence margin by at least $\delta_u > 0$, i.e., $M_u^{(\text{cal})} \geq M_u^{(\text{raw})} + \delta_u$, then $\widehat{p}_{t,u}^{(\text{cal})}(c_u) \geq \frac{e^{\delta_u}\widehat{p}_{t,u}^{(\text{raw})}(c_u)}{1-\widehat{p}_{t,u}^{(\text{raw})}(c_u)+e^{\delta_u}\widehat{p}_{t,u}^{(\text{raw})}(c_u)} > \widehat{p}_{t,u}^{(\text{raw})}(c_u)$, and consequently $\mathcal{L}_{t,u}^{(\text{cal})} < \mathcal{L}_{t,u}^{(\text{raw})}$, where $\mathcal{L}_{t,u}^{(\xi)} = -\log \widehat{p}_{t,u}^{(\xi)}(c_u)$.*

We next relate this margin improvement to changes in individual relation scores. Let $\Delta s_{u,v} = s_{u,v}^{(\text{cal})} - s_{u,v}^{(\text{raw})}$. Suppose calibration changes correct-class scores by at least α_u and incorrect-class scores by at most β_u: $\Delta s_{u,v} \geq \alpha_u, v \in \mathcal{A}_t(c_u), \Delta s_{u,v} \leq \beta_u, v \notin \mathcal{A}_t(c_u)$, then, $Z_{u,c_u}^{(\text{cal})} \geq e^{\alpha_u} Z_{u,c_u}^{(\text{raw})}, \sum_{c\neq c_u} Z_{u,c}^{(\text{cal})} \leq e^{\beta_u}\sum_{c\neq c_u} Z_{u,c}^{(\text{raw})}$, and hence $M_u^{(\text{cal})} - M_u^{(\text{raw})} \geq \alpha_u - \beta_u$. Therefore, $\alpha_u > \beta_u$ is sufficient to improve the relational evidence margin.

This separation is consistent with the source-domain training objective. For a source query u, let $\ell_{k,u} = -\log \widehat{p}_{k,u}(c_u)$. For a non-degenerate multi-class task,

$$\frac{\partial \ell_{k,u}}{\partial s_{k,u,v}} = \begin{cases} -\dfrac{a_{k,u,v}\,(1-\widehat{p}_{k,u}(c_u))}{\widehat{p}_{k,u}(c_u)}, & v \in \mathcal{A}_k(c_u), \\ a_{k,u,v}, & v \notin \mathcal{A}_k(c_u). \end{cases} \tag{16}$$

Table 1: Few-shot node classification accuracy (%) on unseen target graphs under 1-, 3-, and 5-shot settings. Avg. denotes the mean over all 12 dataset-shot combinations. Bold means the best results.

Method	Cora			ogbn-Products			Computers			Physics			Avg.
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	
FedAvg	$29.95_{\pm 4.45}$	$39.62_{\pm 3.73}$	$43.46_{\pm 2.72}$	$9.79_{\pm 2.44}$	$11.45_{\pm 1.79}$	$15.67_{\pm 1.61}$	$23.00_{\pm 4.53}$	$38.00_{\pm 4.56}$	$41.52_{\pm 4.25}$	$45.95_{\pm 10.85}$	$65.11_{\pm 6.38}$	$72.53_{\pm 4.19}$	36.34
FedProx	$30.34_{\pm 4.80}$	$39.79_{\pm 3.84}$	$43.82_{\pm 2.89}$	$10.84_{\pm 2.38}$	$11.43_{\pm 1.82}$	$14.22_{\pm 1.60}$	$23.17_{\pm 4.52}$	$37.72_{\pm 4.54}$	$44.43_{\pm 4.48}$	$46.92_{\pm 10.46}$	$66.21_{\pm 6.54}$	$72.96_{\pm 4.77}$	36.82
SCAFFOLD	$29.92_{\pm 4.55}$	$41.35_{\pm 3.45}$	$45.56_{\pm 2.92}$	$15.89_{\pm 2.50}$	$20.73_{\pm 1.69}$	$22.95_{\pm 1.50}$	$22.34_{\pm 4.86}$	$37.94_{\pm 4.43}$	$44.58_{\pm 4.39}$	$48.54_{\pm 10.63}$	$67.60_{\pm 5.60}$	$74.23_{\pm 3.90}$	39.30
FedDyn	$31.30_{\pm 8.20}$	$45.79_{\pm 5.42}$	$50.36_{\pm 4.11}$	$13.06_{\pm 1.87}$	$15.35_{\pm 1.72}$	$16.31_{\pm 1.45}$	$26.66_{\pm 6.26}$	$34.79_{\pm 7.03}$	$38.72_{\pm 6.77}$	$53.55_{\pm 2.90}$	$70.80_{\pm 9.81}$	$81.04_{\pm 4.47}$	39.81
FedRoD	$29.28_{\pm 4.35}$	$38.34_{\pm 3.58}$	$42.70_{\pm 3.12}$	$15.90_{\pm 2.46}$	$20.69_{\pm 1.64}$	$22.05_{\pm 1.59}$	$24.64_{\pm 4.74}$	$38.76_{\pm 4.59}$	$45.86_{\pm 4.50}$	$48.03_{\pm 9.53}$	$65.80_{\pm 5.32}$	$73.12_{\pm 3.94}$	38.76
FedSR	$35.34_{\pm 7.00}$	$45.64_{\pm 4.64}$	$49.68_{\pm 4.35}$	$15.96_{\pm 2.07}$	$19.19_{\pm 1.75}$	$20.29_{\pm 1.56}$	$27.90_{\pm 6.42}$	$40.96_{\pm 5.42}$	$45.44_{\pm 4.96}$	$54.50_{\pm 11.62}$	$73.63_{\pm 4.36}$	$77.73_{\pm 3.21}$	42.19
FedIIR	$33.70_{\pm 5.78}$	$44.45_{\pm 4.52}$	$48.84_{\pm 3.43}$	$16.25_{\pm 2.22}$	$21.35_{\pm 1.53}$	$23.36_{\pm 1.26}$	$23.22_{\pm 4.67}$	$35.86_{\pm 4.13}$	$44.88_{\pm 4.43}$	$54.27_{\pm 11.89}$	$73.21_{\pm 4.98}$	$78.02_{\pm 5.39}$	41.45
FedADG	$34.40_{\pm 6.44}$	$46.09_{\pm 4.75}$	$50.93_{\pm 3.50}$	$16.30_{\pm 2.24}$	$20.71_{\pm 1.77}$	$21.61_{\pm 1.54}$	$23.49_{\pm 4.72}$	$36.69_{\pm 4.35}$	$42.51_{\pm 4.29}$	$56.25_{\pm 11.95}$	$74.80_{\pm 4.57}$	$79.05_{\pm 4.85}$	41.90
FedDG-GA	$31.16_{\pm 4.88}$	$40.08_{\pm 3.86}$	$44.69_{\pm 3.12}$	$16.02_{\pm 2.31}$	$20.62_{\pm 1.83}$	$21.43_{\pm 1.62}$	$23.31_{\pm 4.62}$	$37.07_{\pm 4.50}$	$43.05_{\pm 4.58}$	$47.68_{\pm 11.64}$	$66.92_{\pm 7.05}$	$73.49_{\pm 5.59}$	38.79
FGOOD	$36.81_{\pm 6.06}$	$49.14_{\pm 3.70}$	$53.56_{\pm 3.42}$	$\mathbf{22.52}_{\pm 2.78}$	$23.41_{\pm 1.88}$	$24.80_{\pm 2.12}$	$31.11_{\pm 6.18}$	$38.37_{\pm 5.38}$	$\mathbf{47.60}_{\pm 4.63}$	$56.88_{\pm 11.18}$	$74.84_{\pm 5.38}$	$78.89_{\pm 3.51}$	44.82
FedGOG	$30.97_{\pm 7.51}$	$39.82_{\pm 5.70}$	$42.77_{\pm 4.86}$	$12.85_{\pm 1.89}$	$13.88_{\pm 1.49}$	$14.46_{\pm 1.33}$	$28.64_{\pm 6.15}$	$39.93_{\pm 6.58}$	$43.79_{\pm 5.81}$	$57.49_{\pm 9.04}$	$71.42_{\pm 3.78}$	$74.69_{\pm 3.29}$	39.23
Ours	$\mathbf{45.45}_{\pm 5.86}$	$\mathbf{54.85}_{\pm 4.69}$	$\mathbf{58.20}_{\pm 2.90}$	$21.08_{\pm 3.08}$	$\mathbf{23.72}_{\pm 2.31}$	$\mathbf{24.86}_{\pm 2.05}$	$\mathbf{39.53}_{\pm 10.10}$	$\mathbf{44.56}_{\pm 8.08}$	$47.04_{\pm 7.49}$	$\mathbf{62.09}_{\pm 12.10}$	$\mathbf{78.92}_{\pm 5.90}$	$\mathbf{83.98}_{\pm 2.83}$	**48.69**

Thus, minimizing the relational classification loss increases same-class relations and suppresses different-class relations. These gradients propagate through the relational projections, graph encoder, and calibration network, providing a direct training signal for improving relational separation. Section 5.4 further compares relational margins before and after calibration; larger mean margins are consistent with Proposition 1 and the corresponding gains in correct-class confidence and accuracy. Detailed proofs are provided in Appendix A.1.

Table 2: Component ablation and classifier comparison. For the Cora ablation, $\Delta = \mathrm{Acc}_{\mathrm{variant}} - \mathrm{Acc}_{\mathrm{full}}$ in percentage points (pp). Classifier macro accuracy and average rank are computed across all evaluated target datasets.

Components: Cora (5-shot)		
Variant	Acc. (%) ↑	Δ (pp) ↑
`FedDARC` (Full Model)	58.20	–
w/o Domain Calib.	54.13	-4.07
Random Frozen Projection	54.95	-3.25
Graph-Statistics State	53.91	-4.29
w/o Label-Relative Inf.	13.04	-45.16
Local Training Only	37.31	-20.89
Classifiers: all target datasets (5-shot)		
Method	Macro Acc. (%) ↑	Avg. Rank ↓
`FedDARC`	**53.52**	**2.0043**
Cosine Prototype	50.01	2.3335
Matching Network	49.53	3.0356
Euclidean Prototype	45.30	3.9383
k-NN	44.78	4.8080
Linear Probe	43.30	4.8804

Table 3: Effect of domain-state selection on relational prediction on Cora under the 5-shot setting. Metrics are averaged over query nodes. The target-derived state uses labeled context nodes only. Evidence margins and correct-class attention masses are first computed for each query and then averaged; therefore, the reported mean attention is not equal to the sigmoid of the reported mean margin. *w/o Calib.* indicates without calibration. *Mismatched Source* indicates Mismatched Source State. *Target-Derived* indicates Target-Derived State.

Domain-State Condition	Evidence Margin ↑	Correct-Class Attention ↑	Query NLL ↓	Accuracy ↑
w/o Calib.	−1.545	0.2628	1.9211	54.13%
Mismatched Source	−1.947	0.2468	2.4393	54.09%
Target-Derived	**−1.312**	**0.2716**	**1.6929**	**58.20%**

5 Experiments

5.1 Experimental Setup and Baseline

(1) Datasets and Implementation Details. We use 4 heterogeneous source graphs for federated training: WikiCS (Mernyei & Cangea, 2020), PubMed (Yang et al., 2016), ogbn-Arxiv (Hu et al., 2020), and Amazon-ratings (Platonov et al., 2023). Each graph is assigned to one client and remains local throughout training. They differ in feature spaces, structures, class compositions, and label semantics. For unseen-domain evaluation, we use 4 target graphs: Cora (Yang et al., 2016), ogbn-Products (Hu et al., 2020) and Computers (Shchur et al., 2018), and Physics (Shchur et al., 2018), spanning citation, e-commerce, and co-authorship. None participates in federated training. Unless otherwise specified, `FedDARC` is trained for 500 communication rounds with 10 local epochs per round. Results are averaged over three random seeds: 42, 142, and 242. To further evaluate scalability with the number of clients, we partition the four source datasets up to 32 clients according to their class labels. Implementation details for the other methods are provided in Appendix C. **(2) Baselines.** We compare `FedDARC` with 3 baseline groups: conventional FL methods, including

FedAvg (McMahan et al., 2017), FedProx (Sahu et al., 2018), SCAFFOLD (Karimireddy et al., 2020), FedDyn (Acar et al., 2021), and FedRoD (Chen & Chao, 2022); federated domain generalization methods adapted to graphs, including FedSR (Nguyen et al., 2022), FedIIR (Guo et al., 2023), FedADG (Zhang et al., 2023a), and FedDG-GA (Zhang et al., 2023b); and graph-specific federated generalization methods, FGOOD (Zhu et al., 2025) and FedGOG (Zhou et al., 2025).

5.2 Main results

As shown in Table 1, FedDARC achieves the best overall performance and outperforms the baselines across most datasets and few-shot settings. Under our adaptations, conventional FL methods perform poorly, plausibly because they address optimization heterogeneity across clients while assuming a shared prediction task and aligned label semantics, so their representations and prediction mechanisms transfer poorly to unseen graphs with different structures and label spaces. General FDG methods improve robustness to domain shifts by learning more transferable representations, but most are designed for Euclidean data and fixed label spaces, limiting their ability to handle graph-specific structural variation and unseen class semantics. Graph-specific FDG methods further model graph heterogeneity and achieve stronger results, with FGOOD providing the strongest competing performance in several settings, yet they still rely on predefined prediction spaces. In contrast, FedDARC combines domain-state-guided representation calibration with task-relative relational inference, adapting representations to unseen graph domains while constructing predictions without globally aligned class semantics. This enables transfer across both graph-domain and label-space heterogeneity while keeping the learned shared parameters fixed during target inference.

5.3 Ablation Studies

We analyze the contribution of each component and the effectiveness of the relational classifier.

Components: Table 2 (Cora, 5-shot) shows that all components contribute. First, removing domain-state-guided calibration costs 4.07 pp. Replacing the learned projection with a frozen random projection weakens performance, while constructing states from the mean and standard deviation of reference node representations performs slightly worse than removing calibration. The full method performs best, supporting the joint use of gradient responses and the learned projection. Second, removing relational inference causes the largest drop of 45.16 pp, showing that a task-relative prediction mechanism is necessary under non-aligned label spaces. Local-only training also decreases accuracy by 20.89 pp, demonstrating that federated collaboration provides transferable knowledge beyond independent local training. **Classifier:** Table 2 shows that our label-space-agnostic relational classifier with the same encoder achieves the highest macro accuracy and best average rank across target datasets, outperforming prototype-, neighbor-, matching-, and probe-based classifiers. The margin over the strongest alternative, cosine prototypes, is 3.51 pp, suggesting a benefit from learned query-to-context relations over fixed metrics or target-fitted classifiers.

5.4 Empirical Analysis of Relational Calibration

Table 3 is consistent with Proposition 1 on average. The target-derived domain state enlarges the mean evidence margin, strengthens correct-class attention, and improves predictive performance, whereas a mismatched source state degrades them, suggesting that the gains arise from domain-aligned calibration rather than arbitrary modulation. On ogbn-Products, the evidence margin is also positively correlated with accuracy (Spearman rank correlation $\rho = 0.295$, $p = 0.0029$), which is consistent with a link between relational separation and inference quality.

5.5 Sensitivity and Scalability Analysis

Hyperparameter Sensitivity. Fig. 2 shows that FedDARC remains relatively stable across a broad range of relational temperatures τ and residual-calibration scales λ_{cal}. The best performance occurs with a moderate τ and relatively small λ_{cal}, while neighboring configurations remain competitive. Performance decreases mainly at more extreme settings, indicating that FedDARC is not highly sensitive to these hyperparameters and generally benefits from controlled residual calibration and moderate relational sharpness.

Figure 2: Sensitivity to the relational temperature τ and residual-calibration scale λ_{cal} on Cora under the 3-shot setting.

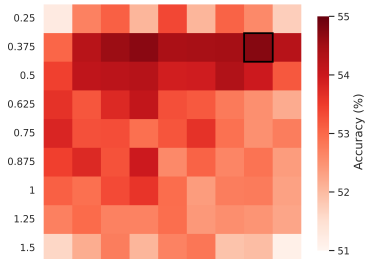
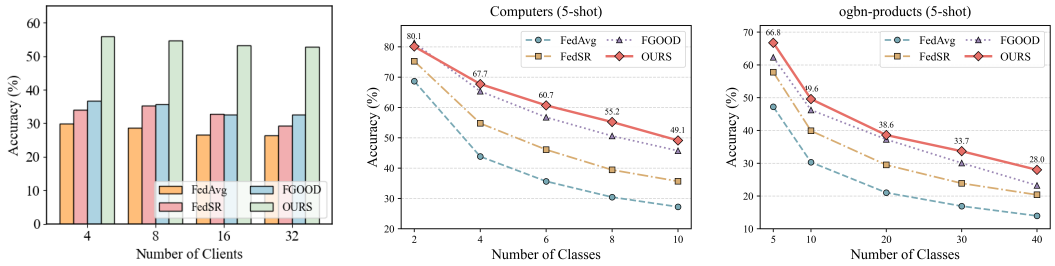


Table 4: Effect of source-domain coverage on few-shot node classification accuracy (%). Average denotes the arithmetic mean over Cora and ogbn-Products (5-shot).

Sources	Training Datasets	Cora	ogbn-Products	Average
1	PubMed	28.59	6.49	17.54
1	Amazon	25.45	10.59	18.02
2	PubMed + Arxiv	51.29	10.30	30.80
2	PubMed + Amazon	52.03	9.72	30.88
3	PubMed + Arxiv + WikiCS	57.45	17.49	37.47
3	PubMed + Arxiv + Amazon	56.82	22.59	39.71
4	PubMed + Arxiv + WikiCS + Amazon	**58.20**	**24.86**	**41.53**



(a) (b) (c)

Figure 3: Performance under different federated scales and target-task complexities. (a) Accuracy on Cora with varying numbers of participating source clients. (b–c) Accuracy of different methods as the number of target classes increases on Computers and ogbn-Products.

Effect of the Number of Clients. As shown in Fig. 3a, `FedDARC` achieves the highest accuracy across all client configurations and exhibits the smallest degradation as the federation scales up. From 4 to 32 clients, its accuracy decreases by only 3.08 pp, while the competing methods degrade more noticeably. This suggests that learning shared, label-space-independent parameters together with domain-state-guided calibration remains effective under increasing federation scale.

Effect of the Number of Target Classes. This sensitivity study varies the number of classes sampled from the target label space while keeping the number of context nodes per selected class fixed. As shown in Figs. 3b and 3c, all methods become less accurate as the number of target classes increases, reflecting the greater difficulty of distinguishing more classes from only a few labeled examples per class. Nevertheless, `FedDARC` maintains the strongest overall performance on both datasets, with a clear advantage in high-way settings. This result highlights the flexibility of task-relative relational inference, whose prediction space is determined by the current target task rather than a fixed global classifier.

Source-Domain Coverage. Table 4 shows that broader source-domain coverage generally improves unseen-domain performance, with the full four-source setting achieving the best results on both Cora and ogbn-Products. This supports the benefit of learning from diverse source-domain responses, which provides richer domain information for transferable representation calibration.

6 Conclusion

In this work, we study federated few-shot graph domain generalization with non-aligned label spaces. We propose `FedDARC`, which combines domain-state-guided representation calibration and task-relative relational inference to enable transfer across heterogeneous graph domains without globally shared class semantics. Theoretical analysis characterizes calibration-induced relational improvement. Experiments on four unseen target graphs with four source graphs and three seeds show the effectiveness of `FedDARC` across few-shot settings; broader source–target pairings remain to be studied.

AI USE STATEMENT

We used generative AI tools to assist with providing feedback on the methodology and experimental design. The authors reviewed and revised all AI-assisted material and take full responsibility for the final paper.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Jinheon Baek, Wonyong Jeong, Jiongdao Jin, Jaehong Yoon, and Sung Ju Hwang. Personalized subgraph federated learning. *arXiv preprint arXiv:2206.10206*, 2023.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=I1hQbx10Kxn.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Tao Fan, Hanlin Gu, Xuemei Cao, Chee Seng Chan, Qian Chen, Yiqiang Chen, Yihui Feng, Yang Gu, Jiaxiang Geng, Bing Luo, et al. Ten challenging problems in federated foundation models. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 37(7):4314–4337, 2025.

Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, pp. 11905–11933. PMLR, 2023.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Sungwon Kim, Yoonho Lee, Yunhak Oh, Namkyeong Lee, Sukwon Yun, Junseok Lee, Sein Kim, Carl Yang, and Chanyoung Park. Subgraph federated learning for local generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cH65nS5sOz.

Ying Li, Xingwei Wang, Rongfei Zeng, Praveen Kumar Donta, Ilir Murturi, Min Huang, and Schahram Dustdar. Federated domain generalization: A survey. *arXiv preprint arXiv:2306.01334*, 2023.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks, 2024a. URL https://arxiv.org/abs/2310.00149.

Rui Liu, Pengwei Xing, Zichao Deng, Anran Li, Cuntai Guan, and Han Yu. Federated graph neural networks: Overview, techniques, and challenges. *IEEE transactions on neural networks and learning systems*, 36(3):4279–4295, 2024b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Péter Mernyei and Catalina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *CoRR*, abs/2007.02901, 2020. URL https://arxiv.org/abs/2007.02901.

A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *CoRR*, abs/1812.06127, 2018. URL http://arxiv.org/abs/1812.06127.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International journal of machine learning and cybernetics*, 14(2):513–535, 2023.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Tao Qi, Yongfeng Huang, and Xing Xie. A federated graph neural network framework for privacy-preserving personalization. *Nature Communications*, 13(1), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30714-9. URL http://dx.doi.org/10.1038/s41467-022-30714-9.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 40–48. PMLR, 2016.

Yuhang Yao, Weizhao Jin, Srivatsan Ravi, and Carlee Joe-Wong. Fedgcn: Convergence-communication tradeoffs in federated training of graph convolutional networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuhang Yao, Yuan Li, Xinyi Fan, Junhao Li, Kay Liu, Weizhao Jin, Srivatsan Ravi, Philip S Yu, and Carlee Joe-Wong. Fedgraph: A research library and benchmark for federated graph learning. *arXiv preprint arXiv:2410.06340*, 2024.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. Samgpt: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 1142–1153, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714828. URL https://doi.org/10.1145/3696410.3714828.

Ke Zhang, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. Subgraph federated learning with missing neighbor generation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen. Federated learning for iot devices with domain generalization. *IEEE Internet of Things Journal*, 10(11):9622–9633, 2023a. doi: 10.1109/JIOT.2023.3234977.

Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3954–3963, June 2023b.

Pengyang Zhou, Chaochao Chen, Weiming Liu, Xinting Liao, Wenkai Shen, Jiahe Xu, Zhihui Fu, Jun Wang, Wu Wen, and Xiaolin Zheng. Fedgog: federated graph out-of-distribution generalization with diffusion data exploration and latent embedding decorrelation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i21.34459. URL https://doi.org/10.1609/aaai.v39i21.34459.

Yukai Zhu, He Sun, and Mingjun Xiao. Federated graph out-of-distribution generalization via representation propagation and scattering. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 993–1002, 2025. doi: 10.1109/ICDM65498.2025.00107.

A Theoretical Analysis

Section A.1 shows Calibration-Induced Relational Improvement. Section A.2 bounds the variation of the estimated domain state across sampled response batches and the resulting change in calibrated representations. Section A.3 decomposes unseen-domain risk.

A.1 Proofs for Calibration-Induced Relational Improvement

A.1.1 Proof of Proposition 1

Consider a target query node $u \in \mathcal{Q}_t$ with task-relative label $c_u = \rho_t(y_{t,u})$. For $\xi \in \{\mathrm{raw}, \mathrm{cal}\}$, define

$$\mathcal{A}_t(c) = \{v \in \mathcal{A}_t : \rho_t(y_{t,v}) = c\}, \tag{17}$$

and denote the corresponding correct-class probability by

$$p_u^{(\xi)} = \widehat{p}_{t,u}^{(\xi)}(c_u). \tag{18}$$

Recall that the class-wise relational evidence is

$$Z_{u,c}^{(\xi)} = \sum_{v \in \mathcal{A}_t(c)} \exp\left(s_{u,v}^{(\xi)}\right), \tag{19}$$

and the correct-versus-incorrect evidence margin is

$$M_u^{(\xi)} = \log \frac{Z_{u,c_u}^{(\xi)}}{\sum_{c \neq c_u} Z_{u,c}^{(\xi)}}. \tag{20}$$

By the definition of the relational classifier,

$$e^{M_u^{(\xi)}} = \frac{Z_{u,c_u}^{(\xi)}}{\sum_{c \neq c_u} Z_{u,c}^{(\xi)}} = \frac{p_u^{(\xi)}}{1 - p_u^{(\xi)}}. \tag{21}$$

Suppose calibration increases the relational evidence margin by at least $\delta_u > 0$:

$$M_u^{(\mathrm{cal})} \geq M_u^{(\mathrm{raw})} + \delta_u. \tag{22}$$

Using Eq. (21), we obtain

$$\frac{p_u^{(\mathrm{cal})}}{1 - p_u^{(\mathrm{cal})}} \geq e^{\delta_u} \frac{p_u^{(\mathrm{raw})}}{1 - p_u^{(\mathrm{raw})}}. \tag{23}$$

Rearranging gives

$$p_u^{(\mathrm{cal})} \geq \frac{e^{\delta_u} p_u^{(\mathrm{raw})}}{1 - p_u^{(\mathrm{raw})} + e^{\delta_u} p_u^{(\mathrm{raw})}}. \tag{24}$$

Since $\delta_u > 0$, the right-hand side of Eq. (24) is strictly larger than $p_u^{(\mathrm{raw})}$. Therefore,

$$p_u^{(\mathrm{cal})} > p_u^{(\mathrm{raw})}. \tag{25}$$

Since $-\log x$ is strictly decreasing on $(0, 1)$, it follows that

$$\mathcal{L}_{t,u}^{(\mathrm{cal})} < \mathcal{L}_{t,u}^{(\mathrm{raw})}. \tag{26}$$

This proves Proposition 1. □

A.1.2 Sufficient Relation-Score Condition

We next relate the margin improvement to changes in individual query-to-context relation scores. Define

$$\Delta s_{u,v} = s_{u,v}^{(\mathrm{cal})} - s_{u,v}^{(\mathrm{raw})}. \tag{27}$$

Suppose calibration increases every same-class relation score by at least α_u:

$$\Delta s_{u,v} \geq \alpha_u, \qquad v \in \mathcal{A}_t(c_u), \tag{28}$$

while every different-class relation score increases by at most β_u:

$$\Delta s_{u,v} \le \beta_u, \qquad v \notin \mathcal{A}_t(c_u). \tag{29}$$

Then,

$$\begin{aligned} Z^{(\mathrm{cal})}_{u,c_u} &= \sum_{v \in \mathcal{A}_t(c_u)} \exp\left(s^{(\mathrm{raw})}_{u,v} + \Delta s_{u,v}\right) \\ &\ge e^{\alpha_u} Z^{(\mathrm{raw})}_{u,c_u}. \end{aligned} \tag{30}$$

Similarly,

$$\sum_{c \ne c_u} Z^{(\mathrm{cal})}_{u,c} \le e^{\beta_u} \sum_{c \ne c_u} Z^{(\mathrm{raw})}_{u,c}. \tag{31}$$

Substituting Eqs. (30) and (31) into Eq. (20) yields

$$\begin{aligned} M^{(\mathrm{cal})}_u &\ge \log\left(e^{\alpha_u} Z^{(\mathrm{raw})}_{u,c_u}\right) - \log\left(e^{\beta_u} \sum_{c \ne c_u} Z^{(\mathrm{raw})}_{u,c}\right) \\ &= M^{(\mathrm{raw})}_u + \alpha_u - \beta_u. \end{aligned} \tag{32}$$

Hence,

$$M^{(\mathrm{cal})}_u - M^{(\mathrm{raw})}_u \ge \alpha_u - \beta_u. \tag{33}$$

Therefore, $\alpha_u > \beta_u$ is sufficient for a strict improvement in the relational evidence margin. Importantly, this condition concerns relative separation rather than absolute score changes: both same-class and different-class scores may increase, provided that the same-class increase is larger.

A.1.3 Optimization Signal on Source Clients

We next show that the source-domain relational objective provides a direct optimization signal for the relational separation characterized above. For a source query node $u \in \mathcal{Q}_k$, let $c_u = \rho_k(y_{k,u})$ and define

$$\mathcal{A}_k(c_u) = \{v \in \mathcal{A}_k : \rho_k(y_{k,v}) = c_u\}. \tag{34}$$

For brevity, we omit the superscript (cal) in this subsection. The query-level relational classification loss can be written as

$$\ell_{k,u} = -\log \widehat{p}_{k,u}(c_u) = \log\left(1 + e^{-M_{k,u}}\right), \tag{35}$$

where $M_{k,u}$ denotes the relational evidence margin after domain-state-guided calibration. Its derivative with respect to the margin is

$$\frac{\partial \ell_{k,u}}{\partial M_{k,u}} = -\left(1 - \widehat{p}_{k,u}(c_u)\right) < 0. \tag{36}$$

Thus, decreasing the classification loss directly favors a larger correct-versus-incorrect relational evidence margin.

More specifically, since

$$\widehat{p}_{k,u}(c_u) = \sum_{v \in \mathcal{A}_k(c_u)} a_{k,u,v}, \tag{37}$$

the derivative with respect to an individual relation score is

$$\frac{\partial \ell_{k,u}}{\partial s_{k,u,v}} = \begin{cases} -\dfrac{a_{k,u,v}\left(1 - \widehat{p}_{k,u}(c_u)\right)}{\widehat{p}_{k,u}(c_u)}, & v \in \mathcal{A}_k(c_u), \\ a_{k,u,v}, & v \notin \mathcal{A}_k(c_u). \end{cases} \tag{38}$$

Therefore, gradient descent increases the relation scores of same-class context nodes and decreases those of different-class context nodes.

Importantly, these relation scores are computed from the calibrated representations produced by the residual-calibration network R_ψ. By the chain rule,

$$\nabla_\psi \ell_{k,u} = \sum_{v \in \mathcal{A}_k} \frac{\partial \ell_{k,u}}{\partial s_{k,u,v}} \frac{\partial s_{k,u,v}}{\partial \psi} = -\left(1 - \widehat{p}_{k,u}(c_u)\right) \nabla_\psi M_{k,u}. \tag{39}$$

Hence, the relational classification objective backpropagates a margin-oriented training signal through the relation module into R_ψ, allowing the domain-state-guided calibration to be optimized toward representations with stronger task-relative relational separation.

This analysis characterizes the local optimization signal induced by each query. Under joint mini-batch optimization, gradients from different queries and model components are aggregated, so it does not require every individual query margin to increase after every parameter update.

A.1.4 Interpretation of Negative Evidence Margins

The relational evidence margin compares the correct-class evidence against the sum of all incorrect-class evidence. In particular,

$$M_u > 0 \quad \Longleftrightarrow \quad Z_{u,c_u} > \sum_{c \neq c_u} Z_{u,c}, \tag{40}$$

By contrast, correct multiclass prediction only requires

$$Z_{u,c_u} > \max_{c \neq c_u} Z_{u,c}. \tag{41}$$

The former condition is strictly stronger than the latter. Consequently, a query may be correctly classified even when its correct-versus-all-incorrect evidence margin remains negative. Our theoretical and empirical analyses therefore focus on calibration-induced margin enlargement rather than requiring the margin itself to become positive.

A.2 Stability of Domain-State-Guided Representations

We analyze how variation in sampled response batches affects the estimated domain state and the resulting calibrated representations. Consider an unseen target graph $\mathcal{G}_t$ with labeled context set $\mathcal{A}_t$. Let $\mathcal{R}_1, \ldots, \mathcal{R}_B$ be response batches sampled independently, conditional on $\mathcal{G}_t$ and $\mathcal{A}_t$, using the procedure in Section 3.1. Each batch uses only labels in $\mathcal{A}_t$.

Starting from the same reference parameters θ_0, the one-step virtual update on batch $\mathcal{R}_b$ produces

$$\Delta\theta_{\mathcal{R}_b} = -\eta_{\text{resp}} \nabla_\theta \mathcal{L}_{\text{resp}}(\theta; \mathcal{G}_t, \mathcal{R}_b)\big|_{\theta=\theta_0}. \tag{42}$$

The temporary parameters are discarded after the response is computed. Following the construction in the main method, define

$$\mathbf{r}_{\mathcal{R}_b} = \frac{\text{vec}(\Delta\theta_{\mathcal{R}_b})}{\|\text{vec}(\Delta\theta_{\mathcal{R}_b})\|_2 + \epsilon}, \qquad \widehat{\mathbf{e}}_B = \frac{1}{B}\sum_{b=1}^{B} P_{\phi^*}(\mathbf{r}_{\mathcal{R}_b}), \tag{43}$$

where P_{ϕ^*} is the projection learned from source clients and frozen before target-domain inference. The population state associated with the same response-batch sampling procedure is

$$\mathbf{e}_* = \mathbb{E}_{\mathcal{R}}[P_{\phi^*}(\mathbf{r}_{\mathcal{R}}) \mid \mathcal{G}_t, \mathcal{A}_t]\,. \tag{44}$$

Thus, both the estimated and population states apply the projection *before* averaging.

Assumption 1 (Bounded projected responses)**.** *The projection P_{ϕ^*} is fixed independently of the target response batches, and $\|P_{\phi^*}(\mathbf{r}_{\mathcal{R}})\|_2 \leq M$ for every batch generated by the specified sampling procedure.*

Theorem 1 (Stability of domain-state estimation)**.** *Under Assumption 1, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widehat{\mathbf{e}}_B - \mathbf{e}_*\|_2 \leq \frac{M}{\sqrt{B}} + M\sqrt{\frac{2\log(1/\delta)}{B}}. \tag{45}$$

Proof. Write $\mathbf{X}_b = P_{\phi^*}(\mathbf{r}_{\mathcal{R}_b})$ and $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}_b \mid \mathcal{G}_t, \mathcal{A}_t] = \mathbf{e}_*$. Conditional independence gives

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{b=1}^{B}\mathbf{X}_b - \boldsymbol{\mu}\right\|_2^2 \middle| \mathcal{G}_t, \mathcal{A}_t\right] \leq \frac{M^2}{B}. \tag{46}$$

By Jensen's inequality, the expected deviation is at most $M/\sqrt{B}$.

Replacing one response batch changes the norm of the sample mean by at most $2M/B$. Applying the bounded-differences inequality therefore shows that this deviation exceeds its expectation by at most $M\sqrt{2\log(1/\delta)/B}$ with probability at least $1-\delta$. Combining the two terms proves Eq. equation 45. □

Stability of representation calibration. We next examine the centered residual calibration in Eq. equation 10. For a domain state $\mathbf{e}$ and an uncalibrated node representation $\mathbf{h}$, write

$$T_{\mathbf{e}}(\mathbf{h}) = \mathbf{h} + \lambda_{\mathrm{cal}}\Big[\big(\boldsymbol{\gamma}(\mathbf{e}) - \boldsymbol{\gamma}(\mathbf{0})\big) \odot \mathbf{h} + \boldsymbol{\beta}(\mathbf{e}) - \boldsymbol{\beta}(\mathbf{0})\Big]. \tag{47}$$

Assumption 2 (Regularity of representation calibration)**.** *For any states* $\mathbf{e}$ *and* $\mathbf{e}'$*,*

$$\|\boldsymbol{\gamma}(\mathbf{e}) - \boldsymbol{\gamma}(\mathbf{e}')\|_2 + \|\boldsymbol{\beta}(\mathbf{e}) - \boldsymbol{\beta}(\mathbf{e}')\|_2 \le L_c\|\mathbf{e} - \mathbf{e}'\|_2. \tag{48}$$

The uncalibrated representations satisfy $\|\mathbf{h}\|_2 \le B_h$.

Proposition 2 (Stability of state-guided representations)**.** *Under Assumption 2, for any node representation* $\mathbf{h}$ *and any states* $\mathbf{e}, \mathbf{e}'$*,*

$$\|T_{\mathbf{e}}(\mathbf{h}) - T_{\mathbf{e}'}(\mathbf{h})\|_2 \le \lambda_{\mathrm{cal}} L_c \max\{B_h, 1\}\|\mathbf{e} - \mathbf{e}'\|_2. \tag{49}$$

Consequently, under the assumptions of Theorem 1, with probability at least $1-\delta$*,*

$$\|T_{\widehat{\mathbf{e}}_B}(\mathbf{h}) - T_{\mathbf{e}_*}(\mathbf{h})\|_2 \le \lambda_{\mathrm{cal}} L_c \max\{B_h, 1\} M\left(\frac{1}{\sqrt{B}} + \sqrt{\frac{2\log(1/\delta)}{B}}\right). \tag{50}$$

Proof. The shared base representation and zero-state centering terms cancel when two calibrated representations are compared. Hence,

$$\begin{aligned}&\|T_{\mathbf{e}}(\mathbf{h}) - T_{\mathbf{e}'}(\mathbf{h})\|_2 \\ &\le \lambda_{\mathrm{cal}}\Big[\|\mathbf{h}\|_\infty\|\boldsymbol{\gamma}(\mathbf{e}) - \boldsymbol{\gamma}(\mathbf{e}')\|_2 + \|\boldsymbol{\beta}(\mathbf{e}) - \boldsymbol{\beta}(\mathbf{e}')\|_2\Big] \\ &\le \lambda_{\mathrm{cal}} L_c \max\{B_h, 1\}\|\mathbf{e} - \mathbf{e}'\|_2.\end{aligned} \tag{51}$$

The second claim follows by applying Theorem 1. □

A.3 Generalization to an Unseen client Domain

Let $\mathbb{P}_i^{\mathcal{T}}$ denote the distribution of local classification tasks generated from source graph domain $\mathcal{D}_i$, and let

$$\mathbb{P}_\alpha^{\mathcal{T}} = \sum_{i=1}^{K} \alpha_i \mathbb{P}_i^{\mathcal{T}}, \qquad \sum_{i=1}^{K} \alpha_i = 1, \tag{52}$$

be the source-task mixture. Let $\mathbb{P}_T^{\mathcal{T}}$ denote the corresponding distribution on an unseen target graph domain.

Because label identities are graph-specific, task distributions are compared up to an arbitrary permutation of their local classes.

Let $\overline{R}_{\mathcal{D}}(\Theta)$ denote the ideal task risk computed using the population domain state and population class relations. We define the residual task discrepancy as

$$\mathrm{Disc}_{\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_\alpha) \quad = \sup_{\Theta \in \mathcal{H}} \left| \overline{R}_{\mathcal{D}_T}(\Theta) - \sum_{i=1}^{K} \alpha_i \overline{R}_{\mathcal{D}_i}(\Theta) \right|. \tag{53}$$

Theorem 2 (Unseen-domain generalization)**.** *Assume that the classification loss is bounded by B_ℓ and is L_ℓ-Lipschitz with respect to the predicted class distribution. Let N_{eff} be the effective number of source training tasks and let $\widehat{\Theta}$ be the learned model. Then, with probability at least $1-\delta$,*

$$\boxed{\begin{aligned} R^k_{\mathcal{D}_T}(\widehat{\Theta}) \le \widehat{R}^k_\alpha(\widehat{\Theta}) &+ 2\mathfrak{R}_{N_{\text{eff}}}(\ell \circ \mathcal{H}) \\ &+ B_\ell\sqrt{\frac{\log(4/\delta)}{2N_{\text{eff}}}} + \text{Disc}_{\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_\alpha) \\ &+ 2\epsilon_{\text{state}} + 2\epsilon_{\text{rel}}, \end{aligned}} \tag{54}$$

where $\epsilon_{\text{state}} = C_{\text{state}}\sqrt{\frac{\log(4/\delta)}{n_{\text{eff}}}}$ is an assumed bound on the target-state estimation error from a finite labeled reference set, and

$$\epsilon_{\text{rel}} = C_{\text{rel}} m \sqrt{\frac{\log(4m/\delta)}{k_{\text{eff}}}} \tag{55}$$

is an assumed bound on the class-relation estimation error from k labeled nodes per class.

Proof sketch. We first relate the practical target risk to its population counterpart:

$$R^k_{\mathcal{D}_T}(\Theta) \le \overline{R}_{\mathcal{D}_T}(\Theta) + \epsilon_{\text{state}} + \epsilon_{\text{rel}}. \tag{56}$$

By the definition of the residual task discrepancy,

$$\overline{R}_{\mathcal{D}_T}(\Theta) \le \overline{R}_{\mathcal{D}_\alpha}(\Theta) + \text{Disc}_{\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_\alpha). \tag{57}$$

The source population risk is bounded by its empirical counterpart using standard uniform convergence:

$$\overline{R}_{\mathcal{D}_\alpha}(\Theta) \le \ \widehat{R}^k_\alpha(\Theta) + 2\mathfrak{R}_{N_{\text{eff}}}(\ell \circ \mathcal{H}) + B_\ell\sqrt{\frac{\log(4/\delta)}{2N_{\text{eff}}}} + \epsilon_{\text{state}} + \epsilon_{\text{rel}}. \tag{58}$$

Combining the three inequalities proves the result. □

Theorem 2 decomposes the target risk into five interpretable terms. The first term measures the empirical performance on source classification tasks. The second and third terms characterize generalization from finitely many source tasks. The fourth term captures the residual difference between the unseen target domain and the source-domain mixture after representation calibration. The last two terms quantify the uncertainty caused by limited labeled nodes in the target graph.

Under these assumed rates, with n_{eff} and k_{eff} the effective numbers of labeled reference and per-class context nodes, more labeled target nodes shrink both estimation terms. However, when the target graph substantially differs from all source graphs, the residual task discrepancy may remain dominant. Thus, the bound does not assume that calibration fully removes domain shift.

B Related work

B.1 Federated Learning

Federated learning (Kairouz et al., 2021; Wen et al., 2023; Fan et al., 2025) enables multiple clients to collaboratively train models without centralizing their private data. FedAvg (McMahan et al., 2017) established the standard paradigm of iterative local training and server-side model averaging. Subsequent methods improve optimization under heterogeneous client distributions. FedProx (Sahu

et al., 2018) introduces a proximal objective to constrain local drift, SCAFFOLD (Karimireddy et al., 2020) employs control variates to reduce client-update variance, and FedDyn (Acar et al., 2021) uses dynamic regularization to align local and global objectives. Personalized approaches such as FedRoD (Chen & Chao, 2022) further decouple generic representations from client-specific prediction heads. Although these methods improve optimization or personalization on participating clients, they generally assume a predefined prediction task and focus primarily on performance within the observed client population.

B.2 Federated Graph Learning

Federated graph learning (Liu et al., 2024b) extends federated optimization to graph field (Wu et al., 2020) while preserving local graph ownership. Existing methods address several forms of graph heterogeneity. When clients hold subgraphs of a larger network, FedSage and FedSage+ (Zhang et al., 2021) generate missing neighbors to compensate for unobserved cross-client edges. Personalized methods such as FED-PUB (Baek et al., 2023) construct functional client representations to support similarity-aware aggregation. FedGCN (Yao et al., 2023) studies privacy-preserving collaborative training of graph convolutional networks, while FedGraph (Yao et al., 2024) augments local graph information through cross-client collaboration. These methods mainly improve collaborative or personalized learning for observed clients. Subsequently, several studies have extended this formulation to domain generalization. FedGOG (Zhou et al., 2025) improves graph-level OOD generalization by generating diverse graph samples and separating invariant from domain-specific information. FGOOD (Zhu et al., 2025) enhances node-level generalization through representation propagation and scattering, while FedLoG (Kim et al., 2025) transfers condensed local knowledge to generalize to unseen nodes, missing classes, and new clients.

C Additional Experiments

C.1 Datasets and Baseline Details

Datasets. We evaluate `FedDARC` on heterogeneous graph datasets spanning Wikipedia, citation, e-commerce, and co-authorship networks. For federated training, we use four source graphs: WikiCS (Mernyei & Cangea, 2020), PubMed (Yang et al., 2016), ogbn-Arxiv (Hu et al., 2020), and Amazon-ratings (Platonov et al., 2023). Each source graph is treated as an independent source domain and remains local throughout federated training. As summarized in Table 5, these graphs differ substantially in graph size, original feature dimensionality, graph structure, class cardinality, and label semantics.

For unseen-domain evaluation, we use Cora (Yang et al., 2016), ogbn-Products (Hu et al., 2020), Computers (Shchur et al., 2018), and Physics (Shchur et al., 2018). None of the target graphs or their annotations participates in federated training. At inference, only the sampled few-shot context nodes and their labels are available to the model, while the remaining query nodes are used exclusively for evaluation. Table 5 summarizes the statistics of all source and unseen target graphs used in the main experiments.

Baseline Evaluation Protocol. All methods use the same source-domain training data and the same target context–query splits. For baselines requiring target-specific post-training adaptation, adaptation uses only the labeled target context set $\mathcal{A}_t$, whereas the disjoint query set $\mathcal{Q}_t$ is used only for final evaluation. As these baselines assume aligned labels, Table 1 reports their adapted versions.

We further control the federated comparison by using the same round-communication budget and the same local-training protocol where applicable. Because the source graphs have independently defined label spaces, no semantic correspondence between class indices from different source domains is assumed. Detailed communication costs and the corresponding numbers of global communication rounds are reported in Section C.5.

FedLoG assumes globally shared label semantics and prior knowledge of unseen classes when constructing their representations, which conflicts with the label-identity-free setting considered in this work. We therefore do not treat FedLoG as a directly comparable accuracy baseline under the proposed problem formulation.

Table 5: Statistics of the source and unseen target graph datasets. The feature dimensions refer to the original node features before feature alignment. Edge counts follow the corresponding dataset loaders; reverse edges are included for undirected graphs when stored explicitly.

Source Graphs for Federated Training				
Statistic	WikiCS	PubMed	ogbn-Arxiv	Amazon-ratings
Domain	Wikipedia	Citation	Citation	E-commerce
# Nodes	11,701	19,717	169,343	24,492
# Edges	216,123	88,648	1,166,243	93,050
Feature Dim.	300	500	128	300
# Classes	10	3	40	5
Unseen Target Graphs				
Statistic	Cora	ogbn-Products	Computers	Physics
Domain	Citation	E-commerce	E-commerce	Co-authorship
# Nodes	2,708	2,449,029	13,752	34,493
# Edges	10,556	61,859,140	491,722	495,924
Feature Dim.	1,433	100	767	8,415
# Classes	7	47	10	5

Implementation Details. Following the preprocessing described in Section 2, we independently apply SVD to each graph to map its original node features to the common input dimension $d_0 = 64$. The implementation uses a two-layer CNN, a one-layer MLP projection, and a two-layer GNN encoder with hidden dimension 64. Unless otherwise specified, we use a learning rate of 10^{-2} and a batch size of 64. The main federated experiments allocate 500 global rounds to `FedDARC`, with 10 local training epochs per round; the communication-equivalent round allocations of the baselines are reported in Section C.5. The training and validation splits follow the OneForAll (Liu et al., 2024a) code. For 1-shot target tasks, we compute the reference-response loss using two independently augmented views of each labeled context node, allowing it to serve as both a query and a same-class context example without using target query labels or updating the trained model. For response generation, we use the same fixed, randomly initialized W_Q and W_K across all source clients and the target graph.

C.2 Extension to link classification task

We further examine cross-task generalization on two unseen knowledge graphs from distinct domains: FB15K-237 (Bordes et al., 2013), which represents encyclopedic knowledge, and WN18RR (Dettmers et al., 2018), which captures lexical relations. Unlike the node-classification tasks used during federated training, the target task is to predict relation types for graph links. Thus, both datasets introduce unseen graph domains as well as a different prediction task.

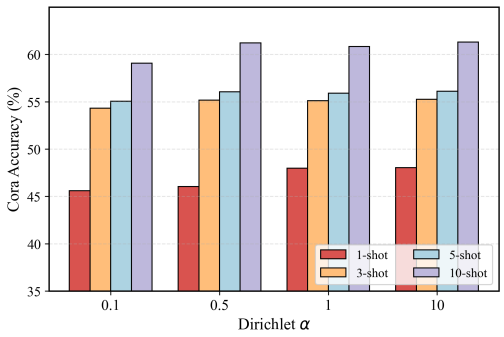


Figure 4: Sensitivity to different degrees of client heterogeneity on Cora.

As shown in Table 6, `FedDARC` achieves the highest accuracy across all few-shot settings on FB15K-237 and in the 3- and 5-shot settings on WN18RR, yielding the best performance in five of the six settings overall. The weaker 1-shot result on WN18RR also indicates that extremely limited context can remain challenging under substantial cross-domain and cross-task shifts. Overall, these results show that the learned task-relative relational inference mechanism can extend beyond node classification to relation-type prediction on unseen graphs.

Table 6: Few-shot link classification accuracy (%) on two unseen knowledge graphs. Results are reported as mean ± standard deviation. Bold values indicate the best result in each setting.

	FB15K-237			WN18RR		
Method	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
FedAvg	22.25 ± 8.69	26.51 ± 10.38	27.39 ± 9.29	17.77 ± 5.09	19.72 ± 4.83	20.25 ± 4.92
FedProx	23.07 ± 8.54	28.30 ± 9.00	30.03 ± 7.82	20.84 ± 5.10	25.64 ± 4.80	27.34 ± 4.45
FedSR	22.32 ± 8.62	26.64 ± 10.15	27.82 ± 9.14	**21.29 ± 6.06**	25.43 ± 5.14	26.63 ± 5.03
FGOOD	22.17 ± 8.72	26.46 ± 10.38	27.06 ± 8.87	21.28 ± 5.99	22.14 ± 5.09	26.29 ± 5.02
Ours	**31.79 ± 8.89**	**35.52 ± 11.48**	**37.27 ± 10.53**	19.09 ± 6.60	**26.32 ± 6.39**	**29.15 ± 6.43**

C.3 Sensitivity to Client Heterogeneity

Figure 4 evaluates the sensitivity of FedDARC to different degrees of client heterogeneity by varying the Dirichlet concentration parameter α, where a smaller α induces more skewed client label distributions. The performance remains relatively stable as α decreases, with only moderate variation even under highly non-IID partitions. This indicates that the learned domain-state-guided calibration and task-relative relational inference remain effective across different degrees of source-client label heterogeneity.

C.4 Label-renaming equivariance.

Table 7 evaluates label-renaming equivariance over 200 episodes per dataset–way setting by randomly permuting task-local labels and mapping predictions back to the original order. Results show unchanged accuracy and negligible probability discrepancies, confirming equivariance to arbitrary class re-indexing without relying on globally aligned label identities.

C.5 Communication cost and computation time

Table 7: Label-renaming equivariance under random task-local label permutations.

Dataset	Way	$\lvert\Delta\text{Acc}\rvert \downarrow$	Prob. $L_1 \downarrow$
Computers	6	0.00	2.05×10^{-8}
	8	0.00	2.33×10^{-8}
	10	0.00	2.47×10^{-7}
ogbn-Products	20	0.00	2.89×10^{-8}
	40	0.00	2.91×10^{-8}

We further compare the communication and computational efficiency of different methods under a common communication budget. Specifically, all methods are allocated about 575 MiB of round communication (500 rounds of FedDARC, excluding its one-time upload). Since their per-round communication volumes differ, the number of global rounds for each method is adjusted accordingly, as summarized in Table 8. Each global round consists of 10 local training epochs.

After federated training, baselines requiring target-specific adaptation use only the available labeled context nodes for post-training optimization. Since this procedure is performed entirely on the target graph without further client–server communication, its cost is excluded from the federated communication budget. We apply early stopping if neither classification accuracy nor loss on the labeled target context set improves for 10 consecutive epochs.

As shown in Table 8, FedDARC requires 1.15 MiB per communication round and 1.04 s per round. Although this introduces a moderate per-round overhead compared with standard FedAvg-style training, it remains substantially more computationally efficient per round than FedADG and FGOOD. Under the same round-communication budget, FedDARC therefore achieves a favorable balance between communication cost and computational efficiency.

D Privacy Analysis

FedDARC follows the standard federated-learning setting in which each client retains its graph structure, node features, and labels locally. During one-time domain-state initialization, clients transmit their normalized optimization responses $\{\mathbf{r}_k^{(b)}\}$ generated by $\Delta\theta$ to the server. Subsequent federated

Table 8: Communication and computational efficiency under an equal cumulative round-communication budget of about 575 MiB. Global rounds are adjusted according to the communication volume per round.

Method	Comm./Round (MiB)	Time/Round (s)	Global Rounds
FedAvg	0.63	0.77	913
FedProx	0.63	0.83	913
SCAFFOLD	1.25	0.97	460
FedADG	2.09	2.17	276
FGOOD	0.63	1.89	913
Ours	1.15	1.04	500

Table 9: Black-box membership inference attack results. Values closer to the random-guessing reference of 50% indicate weaker membership distinguishability under the evaluated attack.

Method	Attack AUC (%) $\rightarrow 50$	Attack Accuracy (%) $\rightarrow 50$
FedAvg	50.08	50.52
FedProx	50.35	50.69
SCAFFOLD	50.01	50.44
FedADG	50.19	50.44
FGOOD	50.37	51.52
Ours	49.74	**50.25**

training communicates updated shared model parameters. In addition, the temporary reference-model displacements used for domain-state construction are computed locally and discarded after the corresponding domain state is obtained. Therefore, the proposed domain-state construction does not require additional direct transmission of raw graph data, node-level annotations, or task-relative label identities.

This design reduces direct exposure of local graph information, but does not by itself provide a formal privacy guarantee against information inference from shared model updates. We therefore further evaluate empirical membership leakage using a black-box membership inference attack.

Membership Inference Attack Evaluation. As shown in Table 9, the attack AUC and accuracy for all methods remain close to the random-guessing reference of 50%. In particular, `FedDARC` yields an attack AUC of 49.74% and an attack accuracy of 50.25%, both close to random guessing. Compared with conventional federated and federated domain-generalization baselines, introducing domain-state-guided calibration and task-relative relational inference does not result in noticeable additional membership leakage under the evaluated attack. These results provide empirical evidence that the additional components of `FedDARC` preserve a privacy profile comparable to the considered federated baselines, while they should not be interpreted as a formal privacy guarantee.

E ALGORITHM

Algorithm 1 summarizes the training and inference procedure of `FedDARC`. It first constructs fixed source-domain states from one-time optimization responses, then performs federated training of the remaining shared parameters, and finally transfers the learned model to an unseen target graph.

Algorithm 1 Federated Training and Inference of `FedDARC`

Require: Source graphs $\{\mathcal{G}_k\}_{k=1}^{K}$, reference parameters θ_0, communication rounds T, local steps E, and initial parameters ϕ and $\Theta_{\text{fl}}^{(0)} = \{\theta, \psi, \mathbf{W}_Q, \mathbf{W}_K\}$; at inference, unseen target graph $\mathcal{G}_t$ and context set $\mathcal{A}_t$

Ensure: Final parameters Θ^* and target predictions $\{\widehat{y}_{t,u} : u \in \mathcal{Q}_t\}$

Domain-state initialization
for all $k \in \mathcal{K}$ **do**
  for $b = 1, \ldots, B_k$ **do**
    Starting from θ_0, take one virtual gradient step on $\mathcal{L}_{\text{resp}}$ using $\mathcal{B}_k^{(b)}$ to obtain $\theta_k^{(b)}$
    Compute $\Delta\theta_k^{(b)} \leftarrow \theta_k^{(b)} - \theta_0$ and normalize it as $\mathbf{r}_k^{(b)}$ by Eq. (3)
  end for
  Upload $\{\mathbf{r}_k^{(b)}\}_{b=1}^{B_k}$ to the server once
end for
Optimize P_ϕ using $\mathcal{L}_{\text{state}}$ and freeze the resulting ϕ^*
for all $k \in \mathcal{K}$ **do**
  Construct the fixed source-domain state $\mathbf{e}_k \leftarrow \frac{1}{B_k} \sum_{b=1}^{B_k} P_{\phi^*}(\mathbf{r}_k^{(b)})$
end for
Federated training
for $r = 0, \ldots, T-1$ **do**
  Server selects $\mathcal{K} \subseteq \mathcal{K}$ and broadcasts $\Theta_{\text{fl}}^{(r)}$
  for all $k \in \mathcal{K}$ **do**
    Initialize local parameters $\Theta_{\text{fl},k} \leftarrow \Theta_{\text{fl}}^{(r)}$
    for $e = 0, \ldots, E-1$ **do**
      Sample context and query sets $\mathcal{A}_k$ and $\mathcal{Q}_k$
      Construct $\rho_k : \mathcal{Y}_k \to \{1, \ldots, C_k\}$
      Compute $\mathbf{h}_{k,v} = f_\theta(v, \mathcal{G}_k)$ and calibrate it to $\widetilde{\mathbf{h}}_{k,v}$ using $\mathbf{e}_k$ via Eq. (10)
      Compute relational weights $a_{k,u,v}$ using Eqs. (11)–(12)
      Predict $\widehat{\mathbf{p}}_{k,u}$ using Eq. (13)
      Update $\Theta_{\text{fl},k}$ by minimizing $\mathcal{L}_{\text{cls}}^k$ in Eq. (14)
    end for
    Set $\Theta_{\text{fl},k}^{(r+1)} \leftarrow \Theta_{\text{fl},k}$ and upload it
  end for
  Aggregate client models: $\Theta_{\text{fl}}^{(r+1)} \leftarrow \sum_{k \in \mathcal{K}} \frac{n_k}{\sum_{j \in \mathcal{K}} n_j} \Theta_{\text{fl},k}^{(r+1)}$
end for
Set $\Theta_{\text{fl}}^* \leftarrow \Theta_{\text{fl}}^{(T)}$ and $\Theta^* \leftarrow \{\phi^*, \Theta_{\text{fl}}^*\}$
Unseen-domain inference
Obtain $\{\mathbf{r}_t^{(b)}\}_{b=1}^{B_t}$ from $\mathcal{A}_t$ using the same fixed reference parameters and response construction
Compute $\mathbf{e}_t \leftarrow \frac{1}{B_t} \sum_{b=1}^{B_t} P_{\phi^*}(\mathbf{r}_t^{(b)})$
Discard the temporary reference updates used for domain-state estimation
Construct $\rho_t : \mathcal{Y}_t \to \{1, \ldots, C_t\}$
Calibrate target representations using $\mathbf{e}_t$ with fixed θ^* and ψ^*
for all $u \in \mathcal{Q}_t$ **do**
  Compute $\widehat{\mathbf{p}}_{t,u}$ using the fixed $\mathbf{W}_Q^*$ and $\mathbf{W}_K^*$
  Predict $\widehat{y}_{t,u} \leftarrow \rho_t^{-1}\left(\arg\max_{c \in \{1,\ldots,C_t\}} \widehat{p}_{t,u}(c)\right)$
end for
return Θ^* and $\{\widehat{y}_{t,u} : u \in \mathcal{Q}_t\}$