Do MLPs Inherit GNN Knowledge Uniformly? Class-Conditional Transfer Equity in Graph-Free Distillation

Anonymous authors
Paper under double-blind review

Abstract

Graph Neural Networks (GNNs) achieve strong performance by leveraging relational structure, but their reliance on message passing limits deployment in latency-sensitive settings. Distilling GNNs into Multilayer Perceptrons (MLPs) provides a feature-only student that approximates a graph-aware teacher without message passing at inference. Existing studies typically evaluate such transfer using average accuracy, implicitly assuming that teacher knowledge is inherited uniformly across labels. However, an MLP may match the teacher well on average while failing to inherit confident teacher knowledge for particular classes. We study this failure mode as class-conditional missed transfer in graph-free distillation. We first introduce *Valid Transfer Agreement*, a label-level metric that measures whether a graph-free student preserves teacher decisions on test samples where the teacher is correct. To mitigate this failure mode, we propose `TED` (Transfer Equity Distillation), a training-time objective augmentation for existing GNN-to-MLP distillation methods. `TED` formulates reliability-guided distillation by treating correct and confident teacher decisions as transferable knowledge and defining label-wise transfer hardness as the portion of this knowledge not inherited by the student. By optimizing a differentiable hardness-based objective, `TED` reduces excessive missed transfer for difficult labels while leaving the teacher, student architecture, and inference procedure unchanged. Across eight homophilic and heterophilic graph benchmarks, `TED` reduces class-level transfer disparity and improves worst-label teacher–student agreement while preserving competitive accuracy and graph-free inference efficiency. Source code is available at: `TED`.

1 Introduction

Graph-structured data naturally models relational dependencies in many modern applications, making representation learning on graphs a central problem in machine learning (Li et al., 2022; Barman et al., 2026; Wang et al., 2021; Wu et al., 2022). Graph Neural Networks (GNNs) have become a dominant paradigm for learning on such data. By propagating and aggregating information over graph structure, GNNs exploit both node attributes and connectivity, often outperforming feature-only models on node classification and related tasks (Kipf & Welling, 2016; Hamilton et al., 2017; Veličković et al., 2017). This representational advantage, however, comes with a substantial deployment cost. A single GNN prediction may require recursive neighborhood expansion, feature fetching, and message aggregation, making inference expensive on large graphs and difficult to serve in latency-sensitive systems (Jia et al., 2020; Min et al., 2021). Multilayer Perceptrons (MLPs) provide an appealing alternative, offering a simple, fast, and graph-free inference model. Yet, because they operate only on node features, they lack the relational inductive bias that makes GNNs effective for structured data.

Knowledge distillation offers a practical route for reconciling this trade-off between graph-aware expressivity and graph-free deployment (Hinton et al., 2015; Li et al., 2025). In GNN-to-MLP distillation, a topology-aware GNN teacher is first trained using graph structure, and its predictive knowledge is then transferred to an MLP student that uses only node features at test time (Zhang et al., 2021). Prior work has shown that such students can substantially improve over vanilla MLPs while removing inference-time graph dependency. Subsequent methods further strengthen this paradigm by injecting structural signals, refining distillation

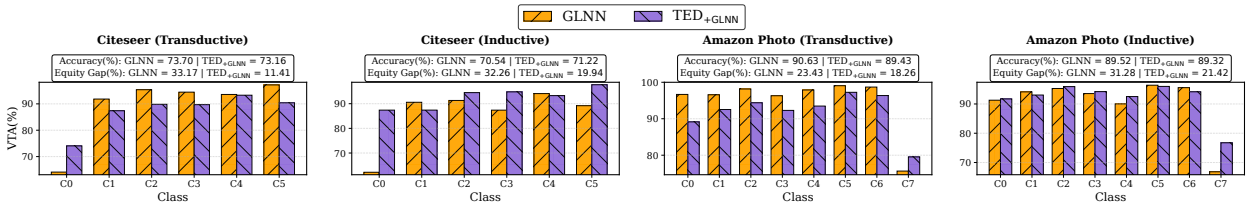


Figure 1: **Class-conditional transfer disparity in graph-free GNN-to-MLP distillation.** We report class-wise Valid Transfer Agreement (VTA) on CiteSeer and Amazon-Photo under both transductive and inductive settings. Although the base graph-free student achieves competitive average accuracy, VTA reveals that correct teacher decisions are not inherited uniformly across labels. The TED-augmented student reduces this disparity by improving transfer on under-transferred classes, leading to lower Equity Gap while preserving competitive average predictive utility.

objectives, and analyzing the nature of transferable graph knowledge (Tian et al., 2022; Wu et al., 2023; Singh et al., 2026; 2025). Together, these approaches have significantly advanced graph-free learning from GNN teachers.

Despite this progress, existing evaluations of GNN-to-MLP distillation remain dominated by average accuracy. Such aggregate metrics can obscure whether the student inherits teacher knowledge uniformly across labels. This is a consequential omission because graph-free distillation changes the information available at inference time. The GNN teacher predicts using both node features and graph structure, whereas the deployed MLP student must approximate the teacher using node features alone. This loss of structural access need not affect all labels equally. Some classes may be largely recoverable from node attributes, while others may depend more strongly on neighborhood context, structural position, high-frequency graph signals, or heterophilic evidence. As a result, an MLP can match the teacher well on average while failing to preserve valid teacher knowledge for particular classes.

In this paper, we study this phenomenon as *class-conditional transfer disparity*. Importantly, transfer disparity refers to the *outcome* of distillation, not to identical treatment of labels during optimization. Our concern is whether the final graph-free student inherits valid GNN teacher knowledge uniformly across labels after training. To make this notion measurable, we consider teacher decisions that are correct and ask whether the student preserves them label by label. We call this label-wise agreement with correct teacher decisions *Valid Transfer Agreement* (VTA), and later formalize it in Section 3. A large spread in VTA across labels indicates that some classes lose substantially more valid teacher knowledge than others.

Figure 1 illustrates this failure mode on CiteSeer and Amazon-Photo under both transductive and inductive settings. Although the graph-free GLNN student achieves competitive average accuracy, its class-wise VTA varies substantially across labels. Some classes exhibit substantially lower VTA, meaning that the MLP fails to preserve correct teacher decisions for those labels despite strong aggregate performance. This suggests that graph-free deployment can introduce a label-dependent loss of GNN knowledge that is invisible to average accuracy.

This reliability issue is especially relevant in deployed graph-based systems, where a fast graph-free student may appear accurate on average but behave inconsistently across semantic categories. Such uneven inheritance of teacher knowledge can be problematic in applications such as recommendation, fraud detection, misinformation detection, and document classification. Our concern is distinct from demographic fairness, since we do not assume that class labels are sensitive attributes. Instead, we study a reliability problem in deployed distillation: a graph-free student should not systematically lose valid GNN teacher knowledge for a subset of labels.

Research question. This paper asks: *Do graph-free MLP students inherit reliable GNN teacher knowledge uniformly across labels, and can distillation be constrained to reduce class-conditional transfer disparity without sacrificing predictive utility?*

We answer this question by using VTA as the evaluation lens and designing TED (***T**ransfer **E**quity **D**istillation*), a training-time objective augmentation for graph-free GNN-to-MLP distillation. For each class, VTA considers the samples on which the GNN teacher predicts correctly and measures how often the MLP student predicts the same label as the teacher. By conditioning on teacher correctness, VTA separates missed transfer from cases where the teacher itself is wrong, enabling a more precise diagnosis of whether valid teacher knowledge is inherited uniformly across labels. We summarize this class-conditional disparity using the Equity Gap, defined as the range between the highest and lowest class-wise VTA values.

The key challenge is to reduce this disparity without forcing the student to imitate unreliable teacher outputs. If the teacher is incorrect or uncertain, enforcing agreement may propagate noise rather than useful graph knowledge, thereby degrading predictive utility. We therefore treat teacher decisions as reliable only when they are both correct and confident. Building on this reliability view, TED defines the transferable hardness of each label as the amount of reliable teacher knowledge that the student fails to inherit. TED penalizes labels whose transferable hardness exceeds the average by more than a tolerance budget.

TED augments existing GNN-to-MLP distillation objectives with a class-level penalty that targets disproportionate missed transfer across labels. Since the constraint is used only during training, TED does not alter the student architecture, inference procedure, or graph-free deployment setting. Experiments across graph benchmarks show that TED reduces class-level transfer gaps and improves worst-label teacher–student agreement while maintaining competitive average accuracy.

Our contributions are summarized as follows:

1. We identify *class-conditional transfer disparity* as an overlooked reliability failure in graph-free GNN-to-MLP distillation.
2. We introduce *Valid Transfer Agreement* (VTA), a label-level metric that measures how faithfully the student preserves correct teacher decisions.
3. We formulate class-conditional missed-transfer control through reliable teacher knowledge. Specifically, we define label-level transferable hardness as the amount of correct and confident teacher knowledge that the student fails to inherit, and penalize excessive hardness across labels.
4. We propose TED (Transfer Equity Distillation), a training-time objective augmentation for existing graph-free distillation methods. TED reduces class-level transfer disparity and improves worst-label teacher–student agreement while preserving the student architecture, graph-free inference, and competitive average accuracy.

2 Related Works

GNN-to-MLP knowledge distillation has emerged as a practical approach for deploying graph models without inference-time message passing. Graph-less Neural Networks (GLNN) (Zhang et al., 2021) established this paradigm by training an MLP student with soft-label supervision from a GNN teacher, allowing the student to use only node features at inference. Subsequent methods improve MLP students by exposing them to additional structural signals. NOSMOG (Tian et al., 2022) augments node attributes with DeepWalk-based positional embeddings (Perozzi et al., 2014), SA-MLP (Chen et al., 2024) encodes the adjacency matrix through a linear layer, and VQGraph (Yang et al., 2024) learns a structure-aware tokenizer over local graph patterns to construct richer distillation targets. While effective, these methods rely on graph-derived representations or preprocessing during inference, and therefore differ from strictly graph-free deployment. We focus on the stricter regime where the deployed student consumes only the original node attributes. Class-view graph distillation changes the distillation unit from sample-wise prediction matching to class-wise prediction matching to improve accuracy, whereas our work measures and regularizes whether correct teacher decisions are inherited uniformly across classes (Tian et al., 2025).

More recently, trustworthiness in GNN-to-MLP distillation has been studied (Capetz et al., 2025). FAITH (Singh et al., 2026) improves group and individual fairness with respect to sensitive attributes during graph-free distillation. Furthermore, few methods have used the reliability of GNN for distillation.

KRD (Wu et al., 2023), HGMD (Wu et al., 2024), and InfGraND (Eskandari et al., 2026) use reliability and hardness estimates in GNN-to-MLP distillation. KRD measures the stability of teacher predictions under perturbations and emphasizes reliable nodes during training. HGMD uses entropy-based knowledge hardness and hardness-aware subgraphs to supervise difficult samples. InfGraND uses graph influence to identify structurally important nodes for stronger distillation. Their objective is to improve the fidelity of instance-level distillation signals. These approaches aim to improve the informativeness of the distillation signal at the instance or subgraph level. Our use of reliability is conceptually different. Rather than asking which samples should receive stronger supervision, we use reliability to define which teacher decisions constitute valid transferable knowledge, and then control how this knowledge is inherited across classes.

To the best of our knowledge, prior GNN-to-MLP distillation methods have not explicitly examined whether reliable teacher knowledge is inherited uniformly across labels. Our work fills this gap by introducing a label-level transfer metric and a training-time objective that controls class-conditional missed transfer while preserving graph-free inference.

3 Preliminaries and Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with $N = |\mathcal{V}|$ nodes, node-feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, and adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$. We denote the feature vector of node v by $\mathbf{x}_v$. Each node has a label $y_v \in \{1, \ldots, C\}$, and $\mathcal{V}_L \subseteq \mathcal{V}$ denotes the labeled node set used for training and evaluation. For class c, let $\mathcal{V}_c = \{v \in \mathcal{V}_L : y_v = c\}$ denote the labeled nodes of class c. We study node classification under GNN-to-MLP distillation, where a graph-aware teacher is used to supervise a graph-free student.

Teacher GNN. The teacher is a pre-trained and frozen GNN f_θ. For each node v, the teacher predicts using both node features and graph structure:

$$\mathbf{p}_v^t = f_\theta(\mathbf{X}, \mathbf{A})_v \in \Delta^{C-1}, \qquad \hat{y}_v^t = \arg\max_{k \in \{1,\ldots,C\}} \mathbf{p}_{v,k}^t. \tag{1}$$

Here, Δ^{C-1} denotes the probability simplex over C classes, and $\mathbf{p}_{v,k}^t$ denotes the teacher-assigned probability that node v belongs to class k. Because f_θ performs message passing over $\mathbf{A}$, its prediction may use neighborhood information, structural position, and other relational signals that are unavailable to a feature-only model.

Student MLP. The student is an MLP g_ϕ that receives only the node feature $\mathbf{x}_v$:

$$\mathbf{p}_v^s = g_\phi(\mathbf{x}_v) \in \Delta^{C-1}, \qquad \hat{y}_v^s = \arg\max_{k \in \{1,\ldots,C\}} \mathbf{p}_{v,k}^s. \tag{2}$$

At inference time, the student does not access $\mathbf{A}$, node neighborhoods, structural encodings, or the teacher. The deployed predictor is therefore strictly graph-free and depends only on the original node attributes.

Valid teacher knowledge. Our goal is not to force the student to imitate every teacher prediction. A teacher prediction is useful for evaluating transfer only when it is correct. We therefore define the set of valid teacher decisions for class c as

$$\mathcal{T}_c = \{v \in \mathcal{V}_c : \hat{y}_v^t = y_v\}. \tag{3}$$

These are the labelled nodes of class c on which the teacher predicts the ground-truth label. Conditioning on $\mathcal{T}_c$ allows us to separate missed transfer from cases where the teacher itself is incorrect.

3.1 Class-Conditional Transfer Metrics

Average accuracy and aggregate teacher–student agreement do not reveal whether teacher knowledge is inherited uniformly across labels. We therefore evaluate GNN-to-MLP distillation at the class level, focusing on teacher decisions that are correct and hence valid for measuring knowledge transfer.

For each class c, let $\mathcal{T}_c = \{v \in \mathcal{V}_c : \hat{y}_v^t = y_v\}$ denote the set of test nodes on which the teacher predicts correctly, and let $\mathcal{C}_{\text{valid}} = \{c : |\mathcal{T}_c| > 0\}$ denote the set of classes with at least one correct teacher decision. Classes outside $\mathcal{C}_{\text{valid}}$ are excluded from class-level transfer analysis.

Definition 1 (Valid Transfer Agreement)**.** *For class $c \in \mathcal{C}_{\text{valid}}$, the* Valid Transfer Agreement *is*

$$\text{VTA}_c = \frac{1}{|\mathcal{T}_c|} \sum_{v \in \mathcal{T}_c} \mathbf{1}[\hat{y}_v^s = \hat{y}_v^t]. \tag{4}$$

It measures the fraction of correct teacher decisions for class c that are preserved by the graph-free student.

Definition 2 (Equity Gap)**.** *The* Equity Gap *summarizes class-conditional transfer disparity as*

$$\text{EG} = \max_{c \in \mathcal{C}_{\text{valid}}} \text{VTA}_c - \min_{c \in \mathcal{C}_{\text{valid}}} \text{VTA}_c. \tag{5}$$

A smaller EG *indicates more uniform inheritance of correct teacher knowledge across labels.*

3.2 Problem Formulation

We study equitable graph-free knowledge transfer. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, node features $\mathbf{X}$, labels on $\mathcal{V}_L$, a pre-trained frozen GNN teacher f_θ, and an MLP student g_ϕ, our goal is to train a feature-only student that preserves the deployment efficiency of GNN-to-MLP distillation while inheriting valid teacher knowledge more uniformly across classes.

This goal entails two desiderata:

1. **Predictive utility.** The student should maintain competitive average node classification performance.

2. **Transfer equity.** The student should reduce class-level disparity in the inheritance of valid teacher knowledge, measured by the Equity Gap in Eq. (5).

The challenge is to improve worst-label transfer behavior without sacrificing average predictive utility or altering the graph-free inference interface of the deployed student.

4 Methodology

We propose `TED`, a training-time objective augmentation for graph-free GNN-to-MLP distillation. Given an existing distillation method, `TED` keeps the teacher, student architecture, and inference procedure unchanged. It only modifies the training objective so that the graph-free student does not miss reliable teacher knowledge disproportionately for particular labels.

Standard GNN-to-MLP distillation encourages the student to match the teacher on average. However, average agreement does not reveal whether teacher knowledge is inherited uniformly across labels. `TED` reframes distillation as the transfer of reliable teacher knowledge rather than the imitation of all teacher outputs. It first identifies teacher decisions that are valid transfer targets using correctness and confidence, then measures how much of this reliable knowledge the student fails to inherit for each label. The resulting objective penalizes only labels whose missed reliable transfer exceeds the average by more than a tolerance budget, directing optimization toward class-specific transfer failures that are not considered by average distillation losses. Figure 2 provides an overview of the distillation mechanism.

4.1 Reliable Teacher Decisions

Let f_θ be the frozen GNN teacher. For each labeled training node $v \in \mathcal{V}_L$, the teacher produces a predictive distribution $\mathbf{p}_v^t \in \Delta^{C-1}$ and predicted label $\hat{y}_v^t = \arg\max_{k \in \{1,\ldots,C\}} \mathbf{p}_{v,k}^t$. We do not treat every teacher prediction as equally useful for transfer. If the teacher is incorrect or uncertain, enforcing agreement may propagate noise rather than useful graph-derived knowledge. We therefore define reliable teacher knowledge using both correctness and confidence.

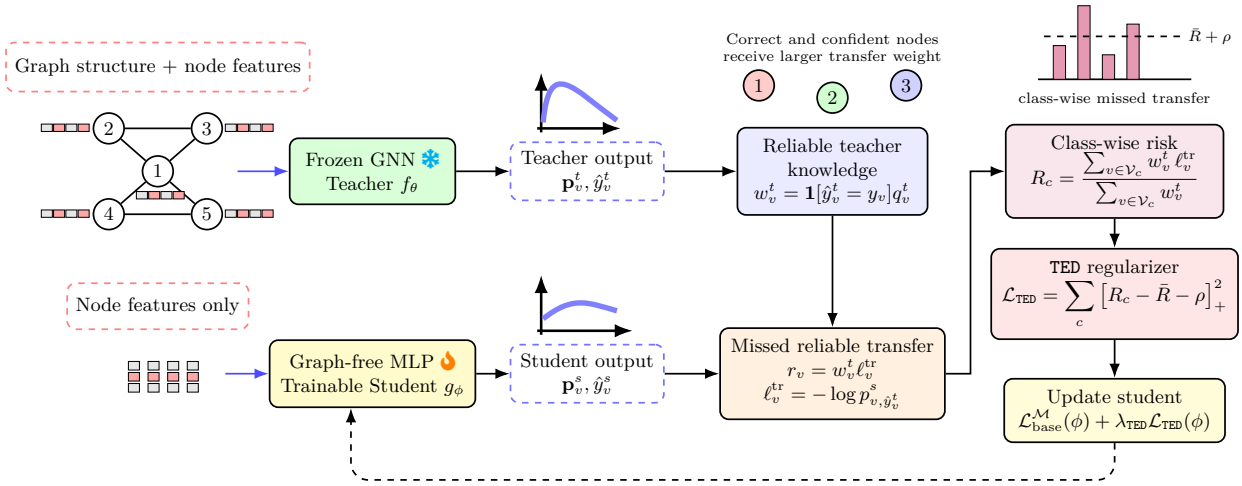


Figure 2: **Overview of TED framework**. (a) A frozen GNN teacher uses graph structure and node features, while the student MLP receives only node features. (b) TED identifies reliable teacher knowledge using correctness and confidence, and computes node-level missed reliable transfer. (c) These quantities are aggregated class-wise, and classes with excessive missed transfer are penalized through the TED regularizer.

Correctness is available on labeled training nodes through the indicator $\mathbf{1}[\hat{y}_v^t = y_v]$. Confidence is measured by normalized teacher certainty:

$$q_v^t = 1 - \frac{\mathcal{H}(\mathbf{p}_v^t)}{\log C}, \qquad \mathcal{H}(\mathbf{p}_v^t) = -\sum_{k=1}^{C} \mathbf{p}_{v,k}^t \log \mathbf{p}_{v,k}^t. \tag{6}$$

Since $0 \leq \mathcal{H}(\mathbf{p}_v^t) \leq \log C$, we have $q_v^t \in [0, 1]$. The score is large when the teacher distribution is concentrated and small when the teacher prediction is uncertain. The reliability weight of node v is

$$w_v^t = \mathbf{1}[\hat{y}_v^t = y_v]\, q_v^t. \tag{7}$$

Thus, incorrect teacher predictions receive zero weight, while correct but uncertain predictions are down-weighted. The transfer-equity objective therefore focuses on teacher decisions that are both valid and reliable.

4.2 Class-Conditional Missed Transfer Risk

Valid Transfer Agreement (Definition 1) measures whether the student preserves correct teacher decisions, but it is defined through hard predicted labels and is therefore not directly differentiable. To optimize the same transfer behavior during training, we introduce a differentiable teacher-targeted surrogate. Let $\mathbf{p}_v^s = g_\phi(\mathbf{x}_v)$ denote the student prediction. For a labeled training node v, we define

$$\ell_v^{\mathrm{tr}}(\phi) = -\log \mathbf{p}_{v,\hat{y}_v^t}^s. \tag{8}$$

This loss penalizes the student when it assigns low probability to the teacher's predicted class. Unlike entropy minimization, which can make the student confident in any class, ℓ_v^{tr} anchors the desired confidence to the teacher decision being transferred. We combine this surrogate with the reliability weight w_v^t to obtain the node-level missed reliable transfer risk

$$r_v(\phi) = w_v^t \ell_v^{\mathrm{tr}}(\phi). \tag{9}$$

This quantity is large only when the teacher decision is reliable and the student assigns insufficient probability to it. For each class c, we then define the class-conditional missed transfer risk as:

$$R_c(\phi) = \frac{\sum_{v \in \mathcal{V}_c} w_v^t \ell_v^{\mathrm{tr}}(\phi)}{\sum_{v \in \mathcal{V}_c} w_v^t}. \tag{10}$$

In implementation, we include a small constant in the denominator to prevent numerical instability. The risk $R_c(\phi)$ estimates how much reliable teacher knowledge for class c is missed by the graph-free student. High values indicate that the teacher makes correct and confident decisions for that class, yet the student fails to place sufficient probability on those decisions. This makes $R_c(\phi)$ a differentiable training proxy to the class-level missed transfer captured by VTA.

4.3 Class-Conditional Missed-Transfer Control

The class-conditional risks $\{R_c(\phi)\}_{c\in\mathcal{C}_{\text{rel}}}$ measure how much reliable teacher knowledge the graph-free student fails to inherit for each label during training. Standard GNN-to-MLP distillation objectives optimize aggregate student performance or average teacher–student agreement, but they do not control how missed transfer is distributed across labels. As a result, a student may preserve the teacher well on average while losing reliable teacher knowledge disproportionately for a subset of classes.

We formulate `TED` as an auxiliary objective for an arbitrary graph-free distillation method. Let $\mathcal{M}$ denote a base GNN-to-MLP distillation method, and let $\mathcal{L}^{\mathcal{M}}_{\text{base}}(\phi)$ denote the original objective used to train its student. This objective may include supervised classification, soft-label distillation, reliability weighting, uncertainty modeling, fairness regularization, or other method-specific components. `TED` leaves this base objective unchanged and adds a class-conditional term that penalizes excessive missed transfer of reliable teacher knowledge.

We first define the set of labels with nonzero reliable teacher mass as $\mathcal{C}_{\text{rel}} = \left\{c : \sum_{v\in\mathcal{V}_c} w_v^t > 0\right\}$. The average missed transfer risk over these labels is

$$\bar{R}(\phi) = \frac{1}{|\mathcal{C}_{\text{rel}}|} \sum_{c\in\mathcal{C}_{\text{rel}}} R_c(\phi). \tag{11}$$

For class c, the excess missed transfer risk is $R_c(\phi) - \bar{R}(\phi)$, which measures how much more reliable teacher knowledge is missed for that class relative to the average label.

The desired class-conditional transfer behavior is that no label should have excess missed transfer beyond a tolerance budget $\rho \geq 0$. This gives the bounded excess-risk condition

$$R_c(\phi) - \bar{R}(\phi) \leq \rho, \qquad \forall c \in \mathcal{C}_{\text{rel}}. \tag{12}$$

Equivalently, ρ controls how much class-level variation in missed reliable transfer is permitted. Smaller values impose stricter transfer uniformity, whereas larger values allow more heterogeneity across labels.

This condition yields the following ideal constrained formulation for a base method $\mathcal{M}$:

$$\begin{aligned} \min_{\phi} \quad & \mathcal{L}^{\mathcal{M}}_{\text{base}}(\phi) \\ \text{s.t.} \quad & R_c(\phi) - \bar{R}(\phi) \leq \rho, \qquad \forall c \in \mathcal{C}_{\text{rel}}. \end{aligned} \tag{13}$$

This formulation separates the role of the base distillation method from the class-conditional transfer requirement. The base objective preserves the original training criterion of $\mathcal{M}$, while the constraint rules out solutions in which reliable teacher knowledge is missed substantially more for some labels than for others.

Several standard constrained optimization tools could be used to address Eq. (13), including primal–dual methods, augmented Lagrangian methods, and penalty relaxations (Bertsekas, 1997; Boyd & Vandenberghe, 2004; Nocedal & Wright, 2006). We adopt a penalty relaxation because it introduces no dual variables, requires only a differentiable class-wise penalty term, and can be added to existing graph-free distillation objectives without changing their optimization pipeline. Comparison with alternative optimization approaches is beyond the scope of this work and is left for future work.

Specifically, TED uses a squared hinge penalty on violations of Eq. (12):

$$\mathcal{L}_{\texttt{TED}}(\phi) = \frac{1}{|\mathcal{C}_{\text{rel}}|} \sum_{c\in\mathcal{C}_{\text{rel}}} \left[R_c(\phi) - \bar{R}(\phi) - \rho\right]_+^2, \tag{14}$$

where $[a]_+ = \max\{a, 0\}$. The hinge form makes the penalty selective. Labels whose missed transfer risk lies within the tolerance budget do not contribute to the penalty, while labels that violate the budget contribute according to the squared magnitude of the violation.

The final objective for method $\mathcal{M}$ augmented with TED is

$$\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi) = \mathcal{L}^{\mathcal{M}}_{\text{base}}(\phi) + \lambda_{\texttt{TED}}\mathcal{L}_{\texttt{TED}}(\phi), \tag{15}$$

where $\lambda_{\texttt{TED}} \geq 0$ controls the strength of the class-conditional transfer penalty. Thus, TED is a differentiable relaxation of the bounded excess-risk formulation. Since all TED-specific quantities are used only during training, the deployed student remains identical to the graph-free student of the base method and requires no graph access at inference.

4.4 Optimization and Integration

Training with TED uses standard gradient-based optimization. Since the teacher is fixed, its predictions and reliability weights can be precomputed on the labeled nodes. The detailed training procedure is summarized in Algorithm 1.

Instantiation with existing methods. For GLNN (Zhang et al., 2021), $\mathcal{M} = \text{GLNN}$ and $\mathcal{L}^{\mathcal{M}}_{\text{base}}$ is the original supervised and soft-label distillation objective used to train the MLP student from the GNN teacher. Adding TED gives the augmented objective

$$\mathcal{L}_{\texttt{TED}_{+\text{GLNN}}}(\phi) = \alpha\mathcal{L}_{\text{CE}}(\phi) + (1-\alpha)\sum_{v\in\mathcal{V}} D_{\text{KL}}\left(\mathbf{p}^t_{v,\tau} \,\|\, \mathbf{p}^s_{v,\tau}\right) + \lambda_{\texttt{TED}}\frac{1}{|\mathcal{C}_{\text{rel}}|}\sum_{c\in\mathcal{C}_{\text{rel}}}\left[R_c(\phi) - \bar{R}(\phi) - \rho\right]^2_+ \tag{16}$$

where $\mathcal{L}_{CE}(\phi)$ is the cross-entropy loss, τ is the distillation temperature, and $D_{\text{KL}}(\cdot)$ is the Kullback–Leibler (KL) divergence. Additional instantiations follow the same augmented-objective form and are detailed in the Appendix D.

4.5 Objective Analysis

We provide a connection between the differentiable objective optimized by TED and the hard class-wise transfer metrics used for evaluation. For each class $c \in \mathcal{C}_{\text{valid}}$, recall that $\mathcal{T}_c = \{v \in \mathcal{V}_c : \hat{y}^t_v = y_v\}$ is the set of teacher-correct nodes for class c. Define the hard missed-transfer rate associated with VTA as

$$U_c = 1 - \text{VTA}_c = \frac{1}{|\mathcal{T}_c|}\sum_{v\in\mathcal{T}_c}\mathbf{1}[\hat{y}^s_v \neq \hat{y}^t_v]. \tag{17}$$

Since $U_c = 1 - \text{VTA}_c$, the Equity Gap can equivalently be written as

$$\text{EG} = \max_{c\in\mathcal{C}_{\text{valid}}} U_c - \min_{c\in\mathcal{C}_{\text{valid}}} U_c. \tag{18}$$

Let us define

$$\gamma_c = \min_{v\in\mathcal{T}_c} q^t_v, \qquad \bar{q}^t_c = \frac{1}{|\mathcal{T}_c|}\sum_{v\in\mathcal{T}_c} q^t_v. \tag{19}$$

Here, γ_c is the minimum teacher confidence among teacher-correct nodes of class c, and $\bar{q}^t_c$ is the corresponding average confidence.

Proposition 1. *For any class $c \in \mathcal{C}_{\text{valid}}$ with $\gamma_c > 0$,*

$$1 - \text{VTA}_c = U_c \leq \frac{\bar{q}^t_c}{\gamma_c \log 2} R_c(\phi). \tag{20}$$

Proof. The proof is provided in Appendix B. □

Proposition 1 establishes a connection between the differentiable missed-transfer risk $R_c(\phi)$ and the hard missed-transfer rate underlying VTA. This motivates $R_c(\phi)$ as a training-time surrogate for the class-wise transfer failures that VTA measures after optimization. All reported transfer disparity results are computed directly using the post-training VTA and Equity Gap metrics.

Convergence and Time Complexity. Under the standard smoothness conditions used in nonconvex optimization, the usual stationarity guarantee applies (Phuong & Lampert, 2019; Garrigos & Gower, 2023). A detailed discussion over convergence is provided in the Appendix C.

TED introduces only a training-time overhead on top of the base distillation method. Since the teacher is frozen, teacher predictions and reliability weights can be precomputed once before student training. During each training epoch, TED computes the teacher-targeted losses and aggregates them by class to obtain the label-wise risks, which costs $\mathcal{O}(|\mathcal{V}_L| + C)$ for $|\mathcal{V}_L|$ labeled nodes and C classes. Since node-classification benchmarks typically satisfy $|\mathcal{V}_L| \gg C$, the additional overhead is effectively dominated by $\mathcal{O}(|\mathcal{V}_L|)$ and is small relative to the forward and backward passes of the student.

5 Experiments

We evaluate TED on real-world graph benchmarks to test whether it reduces class-conditional missed transfer while preserving the predictive and deployment advantages of graph-free distillation. Specifically, we ask: **RQ1:** Does TED improve transfer equity while preserving predictive utility across different GNN teachers and both transductive and inductive settings? **RQ2:** How does TED perform on heterophilic graphs? **RQ3:** What training-time and inference-time costs are introduced by enforcing class-conditional transfer equity? **RQ4:** How does TED compare with class-balanced distillation? **RQ5:** How important is reliability-aware missed-transfer estimation in TED? **RQ6:** How does TED affect the geometry of learned student representations?

5.1 Experiment Settings and Baselines

Datasets. We evaluate TED on eight node-classification datasets spanning homophilic and heterophilic graph regimes. The homophilic datasets are Cora (McCallum et al., 2000), Citeseer (Giles et al., 1998), Pubmed (McCallum et al., 2000), Amazon-Photo, Coauthor-CS, and Coauthor-Physics (Shchur et al., 2018). The heterophilic suite includes Wisconsin and Cornell (Lim et al., 2021). Additional dataset statistics are provided in the Appendix E. Each experiment is repeated over five independent random seeds, and we report the mean and standard deviation. For Cora, Citeseer, and Pubmed, we use the standard citation-network splits introduced by (Kipf & Welling, 2016). For Coauthor-CS, Coauthor-Physics, and Amazon-Photo, we follow the protocol adopted in prior graph-free distillation studies (Zhang et al., 2021). For the heterophilic datasets Wisconsin and Cornell, we use the public splits provided by (Pei et al., 2020).

Evaluation Settings. We evaluate all methods under both transductive and inductive settings. Let the node set be partitioned as $\mathcal{V} = \mathcal{V}_L \sqcup \mathcal{V}_U^{\text{obs}} \sqcup \mathcal{V}_U^{\text{ind}}$, where $\mathcal{V}_L$ denotes labeled training nodes, $\mathcal{V}_U^{\text{obs}}$ denotes unlabeled nodes observed during training, and $\mathcal{V}_U^{\text{ind}}$ denotes nodes reserved for inductive evaluation. For any subset $S \subseteq \mathcal{V}$, we write $\mathbf{X}_S = \{\mathbf{x}_v : v \in S\}$ and $\mathbf{Y}_S = \{y_v : v \in S\}$.
In the transductive setting, the full graph structure and all node features are available during training, while supervision is provided only on $\mathcal{V}_L$. Models are evaluated on the unlabeled evaluation nodes $\mathcal{V}_U^{\text{obs}} \cup \mathcal{V}_U^{\text{ind}}$, depending on the dataset split. In the inductive setting, nodes in $\mathcal{V}_U^{\text{ind}}$ are held out during training. Training uses the observed subgraph induced by $\mathcal{V}_L \cup \mathcal{V}_U^{\text{obs}}$, together with labels only from $\mathcal{V}_L$. Evaluation is performed on $\mathcal{V}_U^{\text{ind}}$. Both evaluation protocols follow prior GNN-to-MLP distillation studies (Singh et al., 2025; Wu et al., 2023; Zhang et al., 2021).

Baselines. We compare TED-augmented methods with representative graph-free distillation baselines, including the GNN teacher, a standalone MLP, GLNN (Zhang et al., 2021), KRD (Wu et al., 2023), and FAITH (Singh et al., 2026). Following, Zhang et al. (2021), we use GraphSAGE (Hamilton et al., 2017) as the default teacher architecture. We further evaluate GCN (Kipf & Welling, 2016) to test robustness to the teacher architecture. For heterophilic datasets, we use ACM-GCN (Luan et al., 2022) as the teacher. Together, these settings evaluate TED across three teacher architectures spanning homophilic and heterophilic graph regimes.

Evaluation Metrics. We evaluate predictive utility using average classification accuracy and transfer disparity using the Equity Gap in Definition 2. In the main homophilic results, we additionally report macro-averages over all datasets within each evaluation setting. We complement scalar Equity Gap with representative class-wise VTA profiles and additional discussion in Appendix J.

Table 1: Node classification accuracy and Equity Gap transductive and inductive settings, using GraphSAGE as the teacher. MLP and GraphSAGE are included as reference accuracy baselines. Parenthesized values in red report the relative Equity Gap reduction of each TED-augmented method with respect to its corresponding base method. Average rows report macro-averages over the datasets.

Dataset	Reference Accuracy		Metric	Graph-free Knowledge Distillation					
	MLP	SAGE		GLNN	$\text{TED}_{+\text{GLNN}}$	KRD	$\text{TED}_{+\text{KRD}}$	FAITH	$\text{TED}_{+\text{FAITH}}$
Transductive Setting									
Cora	60.28±0.25	82.28±0.96	Accuracy ↑	83.68±0.91	83.98±0.71	83.20±1.23	83.62±1.20	83.64±0.69	83.28±0.32
			Equity Gap ↓	7.41±3.49	6.09±0.69 (17.81%)	8.39±2.25	6.95±2.03 (17.16%)	9.24±4.39	7.32±2.70 (20.78%)
Citeseer	60.18±0.93	70.88±0.44	Accuracy ↑	73.70±0.74	73.16±0.22	73.10±0.46	73.28±2.85	73.82±0.16	74.00±0.44
			Equity Gap ↓	33.17±21.14	11.41±1.39 (65.60%)	29.95±22.34	10.72±2.32 (64.21%)	27.95±7.55	15.74±5.63 (43.69%)
Pubmed	72.46±1.32	77.32±0.60	Accuracy ↑	79.86±0.70	81.08±0.32	80.50±0.58	80.74±0.51	81.36±0.36	81.98±0.27
			Equity Gap ↓	5.16±1.92	4.39±2.33 (14.92%)	6.07±1.43	4.48±2.11 (26.19%)	8.80±1.59	6.19±1.64 (29.66%)
Photo	77.86±1.68	90.33±1.68	Accuracy ↑	90.63±1.66	89.43±2.40	92.17±1.48	91.78±1.58	92.59±1.59	92.59±1.50
			Equity Gap ↓	23.43±7.78	18.26±5.03 (25.26%)	20.42±5.85	17.91±4.23 (12.29%)	27.63±15.87	17.57±7.05 (36.41%)
CS	89.62±1.10	89.27±0.50	Accuracy ↑	93.94±0.30	93.66±0.26	93.63±0.32	93.45±0.29	92.64±0.39	92.96±0.25
			Equity Gap ↓	9.63±2.14	8.51±1.53 (11.63%)	10.15±2.00	8.82±1.41 (13.10%)	18.32±3.45	13.42±2.59 (26.75%)
Physics	87.45±2.65	91.09±0.88	Accuracy ↑	93.33±0.89	92.86±0.93	93.41±0.92	93.36±1.27	88.51±9.08	92.98±0.90
			Equity Gap ↓	5.05±1.73	4.09±1.46 (19.01%)	4.78±2.09	2.82±1.05 (41.00%)	14.06±13.68	9.05±3.93 (35.63%)
Macro Avg.	**74.64**	**83.53**	Accuracy ↑	**85.86**	**85.69**	**86.00**	**86.04**	**85.43**	**86.30**
			Equity Gap ↓	**14.14**	**8.79** (37.84%)	**13.29**	**8.62** (35.18%)	**17.67**	**11.55** (34.63%)
Inductive Setting									
Cora	60.58±0.43	79.76±0.24	Accuracy ↑	72.54±2.38	73.16±0.71	72.14±2.52	73.04±0.78	72.42±1.56	72.12±1.91
			Equity Gap ↓	41.67±24.95	26.81±3.71 (35.66%)	40.20±24.27	27.65±3.36 (31.22%)	32.33±11.34	22.94±6.39 (29.04%)
Citeseer	60.40±0.44	70.08±0.63	Accuracy ↑	70.54±0.88	71.22±0.89	70.76±0.62	70.40±0.58	71.02±1.09	71.24±0.68
			Equity Gap ↓	32.26±9.98	19.94±3.17 (38.19%)	25.66±2.57	20.76±4.08 (19.10%)	26.08±5.20	16.52±1.68 (36.66%)
Pubmed	73.32±1.16	78.16±0.42	Accuracy ↑	80.94±0.90	80.80±0.37	80.84±0.48	81.03±0.31	80.32±1.00	81.00±0.59
			Equity Gap ↓	7.21±1.36	5.73±2.43 (20.53%)	6.25±1.67	5.09±2.10 (18.56%)	5.22±2.04	2.03±1.53 (61.11%)
Photo	78.33±1.95	88.70±2.08	Accuracy ↑	89.52±2.34	89.32±2.37	89.01±2.00	89.68±2.14	89.01±1.88	88.70±1.69
			Equity Gap ↓	31.28±9.00	21.42±7.18 (31.52%)	26.21±5.11	22.93±2.21 (12.51%)	27.48±11.87	24.22±8.26 (11.86%)
CS	89.93±1.06	89.16±0.37	Accuracy ↑	92.57±0.41	92.43±0.24	92.51±0.39	92.77±0.42	92.32±0.57	92.27±0.49
			Equity Gap ↓	20.25±2.62	17.07±5.22 (15.70%)	22.47±1.79	17.77±1.11 (20.92%)	19.76±4.51	15.77±2.22 (20.19%)
Physics	87.60±1.75	90.89±2.15	Accuracy ↑	93.04±1.48	92.75±1.64	93.10±1.44	92.82±1.80	91.74±2.56	92.52±1.42
			Equity Gap ↓	7.19±3.05	4.40±0.85 (38.80%)	6.55±2.88	3.95±0.90 (39.69%)	6.59±0.62	4.14±0.98 (37.18%)
Macro Avg.	**75.03**	**82.79**	Accuracy ↑	**83.19**	**83.28**	**83.06**	**83.29**	**82.81**	**82.98**
			Equity Gap ↓	**23.31**	**15.90** (31.81%)	**21.22**	**16.36** (22.90%)	**19.58**	**14.27** (27.11%)

5.2 Results

5.2.1 Effect on Node Classification (RQ1)

Table 1 reports transductive and inductive results on homophilic datasets using GraphSAGE as the teacher. The results show that strong average accuracy does not necessarily imply uniform transfer of teacher knowledge. Across GLNN, KRD, and FAITH, the base graph-free students achieve competitive accuracy, yet their Equity Gap reveals substantial class-conditional missed transfer. Adding TED consistently reduces Equity Gap across all three base distillation approaches while preserving competitive predictive utility. For each TED variant, we use the training-time tolerance budget $\rho = 10\%$. In the transductive setting, $\text{TED}_{+\text{GLNN}}$ reduces the macro-average Equity Gap from 14.14% to 8.79%, $\text{TED}_{+\text{KRD}}$ reduces it from 13.29% to 8.62%, and $\text{TED}_{+\text{FAITH}}$ reduces it from 17.67% to 11.55%. The corresponding macro-average accuracies remain comparable, changing from 85.86% to 85.69% for GLNN, from 86.00% to 86.04% for KRD, and from 85.43% to 86.30% for FAITH. The same trend holds in the inductive setting. $\text{TED}_{+\text{GLNN}}$ reduces the macro-average Equity Gap from 23.31% to 15.90%, $\text{TED}_{+\text{KRD}}$ reduces it from 21.22% to 16.36%, and $\text{TED}_{+\text{FAITH}}$ reduces it from 19.58% to 14.27%. Accuracy is again largely retained, changing from 83.19% to 83.28% for GLNN, from 83.06% to 83.29% for KRD, and from 82.81% to 82.98% for FAITH. These results indicate that class-conditional missed transfer is not specific to a single distillation objective. It appears in standard graph-free KD methods as well as in FAITH, a fairness-oriented distillation baseline. TED provides a method-agnostic training-time correction that reduces disparity in the inheritance of correct teacher decisions while keeping the deployed model as the same graph-free MLP student. Results with additional teacher architecture is reported in Appendix G.

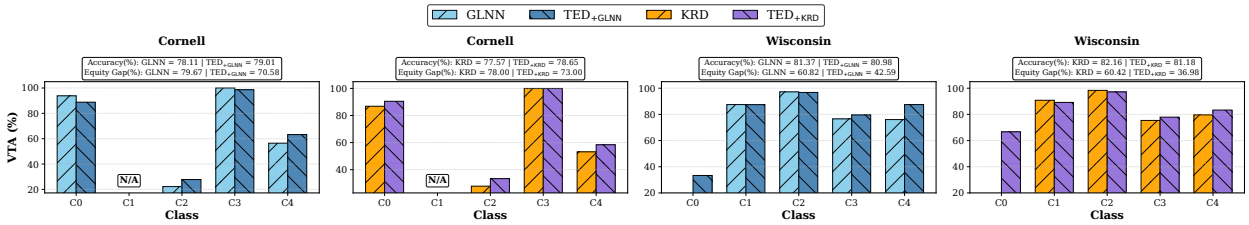


Figure 3: Class-wise Valid Transfer Agreement and Equity gap evaluation on heterophilic datasets. We compare the base GLNN and KRD distillation methods with their corresponding TED-augmented variants.

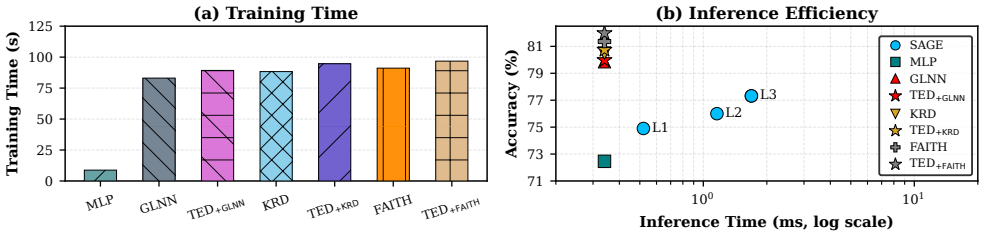


Figure 4: (a) Training time comparison of base distillation methods with the TED-augmented. (b) Inference time and accuracy across different methods. Labels L1, L2 and L3 indicate the number of layers.

5.2.2 Performance on Heterophilic Datasets (RQ2)

We further evaluate TED on heterophilic datasets, where graph neighborhoods are less aligned with class labels. For these datasets, we use ACM-GCN as the GNN teacher. Figure 3 shows that transfer disparity also arises on heterophilic graphs. The base graph-free distillation methods achieve competitive accuracy, but their Equity Gap and class-wise VTA reveal that correct teacher decisions are not inherited uniformly across labels. Adding TED reduces this disparity for the corresponding base methods while maintaining comparable predictive utility. On the Wisconsin dataset, the TED-augmented variants show a clear improvement in worst-class VTA. The improvement is reflected not only in lower Equity Gap, but also in improved worst-class VTA across the reported settings. In the public Cornell split, class C1 contains only one sample overall and no test instance. Because VTA is computed on test samples for which the teacher is correct, C1 has no evaluable transfer support. We therefore mark it as N/A in Figure 3 and exclude it from the Equity Gap calculation. Consistent with Definition 2, we compute Equity Gap only over valid classes. Under this evaluation, TED improves the worst-class VTA on Cornell as well.

These results support the broader claim that transfer equity is a deployment reliability issue in graph-free distillation. When graph structure is removed at test time, the loss of relational information can affect labels unevenly under both homophily and heterophily. TED mitigates this class-conditional missed transfer without changing the deployed MLP student.

5.2.3 Computational Time (RQ3)

We report runtime on PubMed dataset to verify that the gains in transfer equity do not change the deployment profile of graph-free distillation. Figure 4 shows that TED introduces only a small training-time overhead due to its class-wise aggregation and regularization term. At inference time, the runtime matches the corresponding graph-free base method. Together with the accuracy and Equity Gap results, this shows that TED improves transfer reliability while preserving the inference-time efficiency of graph-free distillation.

5.2.4 Ablation Study (RQ4&RQ5)

Prior distillation methods often improve tail classes by reweighting the KD loss or balancing class-level supervision (Zhang et al., 2023). We introduce a representative class-balanced GLNN to test whether such

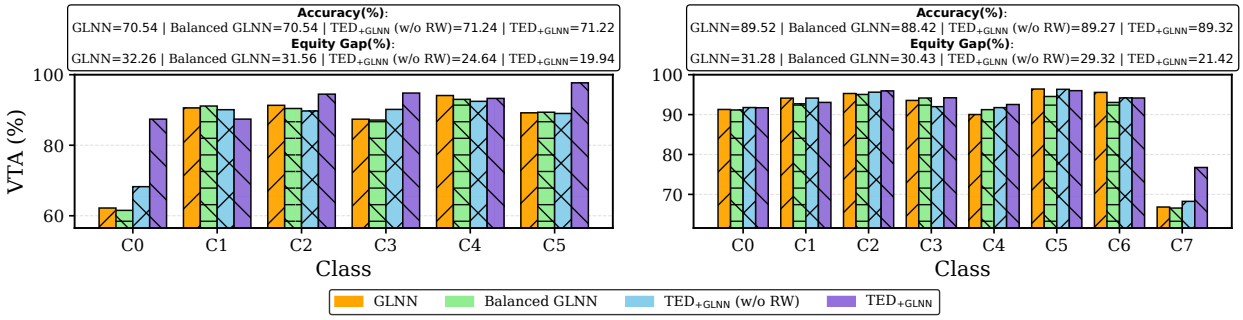
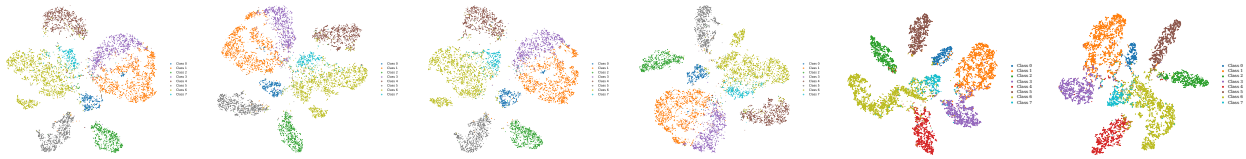


Figure 5: Ablation of class balancing and teacher reliability weighting (RW) under inductive setting on CiteSeer (left) and Amazon-Photo (right).



(a) GLNN (b) $\texttt{TED}_{+\text{GLNN}}$ (c) KRD (d) $\texttt{TED}_{+\text{KRD}}$ (e) FAITH (f) $\texttt{TED}_{+\text{FAITH}}$

Figure 6: t-SNE visualization of learned student representations on Amazon-Photo.

reweighting is sufficient for reducing transfer disparity in graph-free distillation. We also remove teacher Reliability Weighting (RW) from TED to evaluate the role of reliability-aware missed transfer. Figure 5 reports the results on GLNN under the inductive setting. Class-balanced GLNN slightly reduces Equity Gap compared with base GLNN, but it remains worse than TED. Removing reliability weighting from TED increases Equity Gap more substantially, showing that directly regularizing class-wise missed transfer is useful. The full TED objective achieves the best tradeoff between Equity Gap and accuracy, indicating that teacher reliability weighting further improves transfer equity by focusing the constraint on reliable teacher decisions. Similar trends are observed in the transductive setting, as reported in the Appendix H. We also include additional sensitivity analyses for the tolerance budget ρ and the regularization strength $\lambda_{\texttt{TED}}$ in the Appendix I.

5.2.5 Representation Visualization (RQ6)

Beyond accuracy and Equity Gap, we visualize the penultimate-layer student embeddings with t-SNE (Van der Maaten & Hinton, 2008) to qualitatively examine the learned representation geometry. Figure 6 shows results on Amazon-Photo for GLNN, KRD, FAITH, and their TED-augmented variants. With TED, several labels appear more compact and less overlapping. Additional visualizations are provided in Appendix K.

6 Conclusion

We investigated class-conditional transfer disparity in graph-free GNN-to-MLP distillation. Although distilled MLPs are commonly evaluated by average accuracy, we showed that they can fail to inherit valid GNN teacher knowledge uniformly across labels. We propose TED, a training-time objective augmentation that penalizes labels with excessive missed reliable transfer. TED preserves the student architecture and graph-free inference procedure, making it compatible with existing GNN-to-MLP distillation methods. Across graph benchmarks, TED reduces Equity Gap and improves worst-label transfer while maintaining competitive accuracy and efficient inference. These findings highlight the importance of evaluating graph-free distillation not only by aggregate utility, but also by the uniformity with which valid teacher knowledge is transferred across labels. A limitation of our approach is that reliability is estimated from labeled nodes, which may be restrictive in scarcely labeled settings.

References

Jyotismita Barman, Mohammad Yusuf, Sandeep Kumar, and Tapan Kumar Gandhi. Enhancing depression diagnosis with augmented brain signal driven decorrelated graph neural networks. *Communications Medicine*, 2026.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Margaret Capetz, Yizhou Sun, and Arjun Subramonian. Fairness implications of GNN-to-MLP knowledge distillation. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*, 2025. URL https://openreview.net/forum?id=6LPz8LlfeK.

Jie Chen, Mingyuan Bai, Shouzhen Chen, Junbin Gao, Junping Zhang, and Jian Pu. SA-MLP: Distilling graph knowledge from GNNs into structure-aware MLP. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=MZ2kKZc8m7.

Amir Eskandari, Aman Anand, Elyas Rashno, and Farhana Zulkernine. InfgraND: An influence-guided GNN-to-MLP knowledge distillation. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=lfzHR3YwlD.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pp. 89–98, 1998.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Zhihao Jia, Sina Lin, Rex Ying, Jiaxuan You, Jure Leskovec, and Alex Aiken. Redundancy-free computation for graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 997–1005, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Michelle M Li, Kexin Huang, and Marinka Zitnik. Graph representation learning in biomedicine and healthcare. *Nature biomedical engineering*, 6(12):1353–1369, 2022.

Xiang Li, Like Li, Minglei Li, Pengfei Yan, Ting Feng, Hao Luo, Yong Zhao, and Shen Yin. Knowledge distillation and teacher-student learning in medical imaging: Comprehensive overview, pivotal role, and future directions. *Medical Image Analysis*, pp. 103819, 2025.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in neural information processing systems*, 34:20887–20902, 2021.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. *Advances in neural information processing systems*, 35:1362–1375, 2022.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Seung Won Min, Kun Wu, Sitao Huang, Mert Hidayetoğlu, Jinjun Xiong, Eiman Ebrahimi, Deming Chen, and Wen-mei Hwu. Large graph convolutional network training with gpu-oriented data communication architecture. *Proceedings of the VLDB Endowment*, 14(11):2087–2100, 2021.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*, 2020.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International conference on machine learning*, pp. 5142–5151. PMLR, 2019.

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

Vipul Kumar Singh, Jyotismita Barman, Sandeep Kumar, et al. Refine: Enabling efficient and trustworthy modeling of financial networks via gnn-to-mlp knowledge distillation. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–5. IEEE, 2025.

Vipul Kumar Singh, Jyotismita Barman, Sandeep Kumar, Tapan K Gandhi, et al. Leap of faith from gnn-to-mlp: Fairness aware inference via distillation of graph knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25490–25498, 2026.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh Chawla. Learning mlps on graphs: A unified view of effectiveness, robustness, and efficiency. In *The eleventh international conference on learning representations*, 2022.

Yingjie Tian, Shaokai Xu, and Muyang Li. Class-view graph knowledge distillation: A new idea for learning mlps on graphs. *Neurocomputing*, 637:130035, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. A review on graph neural network methods in financial applications. *arXiv preprint arXiv:2111.15367*, 2021.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

Lirong Wu, Haitao Lin, Yufei Huang, and Stan Z Li. Quantifying the knowledge in gnns for reliable distillation into mlps. In *International conference on machine learning*, pp. 37571–37581. PMLR, 2023.

Lirong Wu, Yunfan Liu, Haitao Lin, Yufei Huang, and Stan Z Li. Teach harder, learn poorer: Rethinking hard sample distillation for gnn-to-mlp knowledge distillation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2554–2563, 2024.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM computing surveys*, 55(5):1–37, 2022.

Ling Yang, Ye Tian, Minkai Xu, Zhongyi Liu, Shenda Hong, Wei Qu, Wentao Zhang, Bin Cui, Muhan Zhang, and Jure Leskovec. Vqgraph: Rethinking graph representation space for bridging gnns and mlps. In *International Conference on Learning Representations*, volume 2024, pp. 16490–16509, 2024.

Shaoyu Zhang, Chen Chen, Xiyuan Hu, and Silong Peng. Balanced knowledge distillation for long-tailed learning. *Neurocomputing*, 527:36–46, 2023.

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old mlps new tricks via distillation. *arXiv preprint arXiv:2110.08727*, 2021.

Appendix

A TED Algorithm

Algorithm 1 presents the complete training procedure of the proposed TED framework.

Algorithm 1 TED Training

Require: Frozen teacher f_θ, student g_ϕ, base method $\mathcal{M}$, tolerance ρ, regularization weight $\lambda_{\texttt{TED}}$

Precompute teacher labels $\hat{y}_v^t$ and reliability weights for labeled nodes $v \in \mathcal{V}_L$

$$w_v^t = \mathbf{1}[\hat{y}_v^t = y_v]\left(1 - \frac{\mathcal{H}(\mathbf{p}_v^t)}{\log C}\right)$$

for each training step **do**
  Compute the base objective $\mathcal{L}_{\text{base}}^{\mathcal{M}}(\phi)$
  Estimate $R_c(\phi)$ for each $c \in \mathcal{C}_{\text{rel}}$ using Eq. (10)
  Compute $\bar{R}(\phi) = |\mathcal{C}_{\text{rel}}|^{-1} \sum_{c \in \mathcal{C}_{\text{rel}}} R_c(\phi)$
  Compute $\mathcal{L}_{\texttt{TED}}(\phi)$ using Eq. (14)
  Update ϕ by gradient descent on $\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi)$ from Eq. (15)
end for
return g_ϕ

B Proof of Proposition 1

Proof. Fix a class $c \in \mathcal{C}_{\text{valid}}$ with $\gamma_c > 0$. Recall that $\mathcal{T}_c = \{v \in \mathcal{V}_c : \hat{y}_v^t = y_v\}$ is the set of teacher-correct nodes for class c. For every $v \in \mathcal{T}_c$, the teacher is correct, and therefore the reliability weight satisfies $w_v^t = q_v^t$. The teacher-targeted transfer loss is

$$\ell_v^{\text{tr}}(\phi) = -\log \mathbf{p}_{v,\hat{y}_v^t}^s. \tag{21}$$

We first establish a pointwise relation between hard teacher–student disagreement and the teacher-targeted loss. If $\hat{y}_v^s = \hat{y}_v^t$, then $\mathbf{1}[\hat{y}_v^s \neq \hat{y}_v^t] = 0$, and the desired pointwise bound holds trivially. If $\hat{y}_v^s \neq \hat{y}_v^t$, then the class predicted by the student has probability at least as large as the probability assigned to the teacher class. That is, for some $k \neq \hat{y}_v^t$,

$$\mathbf{p}_{v,k}^s \geq \mathbf{p}_{v,\hat{y}_v^t}^s. \tag{22}$$

Since the probabilities sum to one, this implies

$$\mathbf{p}_{v,\hat{y}_v^t}^s \leq \frac{1}{2}. \tag{23}$$

Hence,

$$-\log \mathbf{p}_{v,\hat{y}_v^t}^s \geq \log 2. \tag{24}$$

Therefore, for every $v \in \mathcal{T}_c$,

$$\mathbf{1}[\hat{y}_v^s \neq \hat{y}_v^t] \leq \frac{\ell_v^{\text{tr}}(\phi)}{\log 2}. \tag{25}$$

By definition, $\gamma_c = \min_{v \in \mathcal{T}_c} q_v^t$. Since $\gamma_c > 0$, we have $q_v^t \geq \gamma_c$ for all $v \in \mathcal{T}_c$. Because $\ell_v^{\text{tr}}(\phi) \geq 0$, it follows that

$$\ell_v^{\text{tr}}(\phi) \leq \frac{q_v^t}{\gamma_c} \ell_v^{\text{tr}}(\phi). \tag{26}$$

Combining Eq. (25) and Eq. (26), we obtain

$$\mathbf{1}[\hat{y}_v^s \neq \hat{y}_v^t] \leq \frac{q_v^t \ell_v^{\text{tr}}(\phi)}{\gamma_c \log 2}. \tag{27}$$

Averaging over $v \in \mathcal{T}_c$ gives

$$U_c = \frac{1}{|\mathcal{T}_c|} \sum_{v \in \mathcal{T}_c} \mathbf{1}[\hat{y}_v^s \neq \hat{y}_v^t] \leq \frac{1}{\gamma_c |\mathcal{T}_c| \log 2} \sum_{v \in \mathcal{T}_c} q_v^t \ell_v^{\mathrm{tr}}(\phi). \tag{28}$$

Since nodes outside $\mathcal{T}_c$ have zero reliability weight, the class-wise TED risk can be written as

$$R_c(\phi) = \frac{\sum_{v \in \mathcal{T}_c} q_v^t \ell_v^{\mathrm{tr}}(\phi)}{\sum_{v \in \mathcal{T}_c} q_v^t}. \tag{29}$$

Therefore,

$$\sum_{v \in \mathcal{T}_c} q_v^t \ell_v^{\mathrm{tr}}(\phi) = R_c(\phi) \sum_{v \in \mathcal{T}_c} q_v^t. \tag{30}$$

Substituting this identity yields

$$U_c \leq \frac{R_c(\phi) \sum_{v \in \mathcal{T}_c} q_v^t}{\gamma_c |\mathcal{T}_c| \log 2} = \frac{\bar{q}_c^t}{\gamma_c \log 2} R_c(\phi), \tag{31}$$

where

$$\bar{q}_c^t = \frac{1}{|\mathcal{T}_c|} \sum_{v \in \mathcal{T}_c} q_v^t. \tag{32}$$

Since $U_c = 1 - \mathrm{VTA}_c$, this proves

$$1 - \mathrm{VTA}_c = U_c \leq \frac{\bar{q}_c^t}{\gamma_c \log 2} R_c(\phi). \tag{33}$$

□

C Convergence Analysis

The TED term preserves the standard optimization structure of the base distillation objective. Most graph-free distillation losses such as cross-entropy and KL divergence are differentiable functions of the student logits. The TED penalty uses the squared hinge $[z]_+^2$, which is differentiable, composed with the class-wise risks $R_c(\phi)$. Thus, $\mathcal{L}_{\mathtt{TED}}(\phi)$ remains differentiable under the same regularity conditions as the teacher-targeted transfer loss. Under the standard smoothness and lower-boundedness assumptions used in nonconvex optimization, gradient descent on the augmented objective satisfies the usual stationarity guarantee. For completeness, we state the standard stationarity result for gradient descent on a smooth lower-bounded objective.

Proposition 2. *Suppose $\mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi)$ is lower bounded and has L-Lipschitz continuous gradients along the optimization trajectory. If gradient descent is run with step size $0 < \eta \leq 1/L$, then after T iterations,*

$$\min_{0 \leq t < T} \|\nabla \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t)\|_2^2 \leq \frac{2\left(\mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_0) - \mathcal{L}_\star\right)}{\eta T}, \tag{34}$$

where $\mathcal{L}_\star$ is a lower bound on $\mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi)$.

Proof. Since $\mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi)$ has L-Lipschitz continuous gradients along the optimization trajectory, the standard descent lemma implies that, for any two iterates ϕ_t and ϕ_{t+1},

$$\mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_{t+1}) \leq \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t) + \left\langle \nabla \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t), \phi_{t+1} - \phi_t \right\rangle + \frac{L}{2} \|\phi_{t+1} - \phi_t\|_2^2. \tag{35}$$

Gradient descent updates the parameters as $\phi_{t+1} = \phi_t - \eta \nabla \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t)$. Substituting this update into the descent inequality gives

$$\begin{aligned} \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_{t+1}) &\leq \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t) - \eta \left\|\nabla \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 + \frac{L\eta^2}{2} \left\|\nabla \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 \\ &= \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t) - \eta \left(1 - \frac{L\eta}{2}\right) \left\|\nabla \mathcal{L}_{\mathtt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2. \end{aligned} \tag{36}$$

Since $0 < \eta \leq 1/L$, we have $1 - \frac{L\eta}{2} \geq \frac{1}{2}$. Therefore,

$$\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_{t+1}) \leq \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t) - \frac{\eta}{2} \left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2. \tag{37}$$

Rearranging yields

$$\left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 \leq \frac{2}{\eta} \left(\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t) - \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_{t+1})\right). \tag{38}$$

Summing over $t = 0, \ldots, T-1$ gives

$$\begin{aligned}\sum_{t=0}^{T-1} \left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 &\leq \frac{2}{\eta} \sum_{t=0}^{T-1} \left(\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t) - \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_{t+1})\right) \\ &= \frac{2}{\eta} \left(\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_0) - \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_T)\right). \end{aligned} \tag{39}$$

Because $\mathcal{L}_\star$ is a lower bound on $\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi)$, we have $\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_T) \geq \mathcal{L}_\star$. Hence,

$$\sum_{t=0}^{T-1} \left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 \leq \frac{2}{\eta} \left(\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_0) - \mathcal{L}_\star\right). \tag{40}$$

Dividing by T gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 \leq \frac{2\left(\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_0) - \mathcal{L}_\star\right)}{\eta T}. \tag{41}$$

Finally, the minimum is bounded by the average:

$$\min_{0 \leq t < T} \left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \left\|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\right\|_2^2. \tag{42}$$

Combining the last two inequalities proves

$$\min_{0 \leq t < T} \|\nabla \mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_t)\|_2^2 \leq \frac{2\left(\mathcal{L}_{\texttt{TED}_{+\mathcal{M}}}(\phi_0) - \mathcal{L}_\star\right)}{\eta T}. \tag{43}$$

□

D Integration with Existing GNN-to-MLP Methods

Integration with KRD. Following KRD (Wu et al., 2023), let $\mathbf{h}_i^{(L)}$ denote the teacher GNN logits for node i and let $\mathbf{z}_i^{(L)}$ denote the student MLP logits. KRD samples reliable knowledge points $j \in \mathcal{N}_i$ according to the reliability-based sampling distribution $p(s_j \mid \rho_i, \alpha^{(t)})$ and defines the reliable distillation loss as

$$\mathcal{L}_{\text{KRD}} = \mathbb{E}_i \, \mathbb{E}_{\substack{j \in \mathcal{N}_i \\ j \sim p(s_j \mid \rho_i, \alpha^{(t)})}} D_{\text{KL}} \left(\sigma(\mathbf{z}_j^{(L)}/\tau), \sigma(\mathbf{h}_i^{(L)}/\tau)\right), \tag{44}$$

where τ is the distillation temperature. The original KRD training objective is

$$\mathcal{L}_{\text{total}}^{\text{KRD}} = \frac{\lambda}{|\mathcal{V}_L|} \sum_{i \in \mathcal{V}_L} \mathcal{H}\left(y_i, \sigma(\mathbf{z}_i^{(L)})\right) + (1-\lambda)\left(\mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{KRD}}\right). \tag{45}$$

We instantiate $\texttt{TED}_{+\text{KRD}}$ by adding the class-level transfer-equity regularizer to the original KRD objective:

$$\mathcal{L}_{\texttt{TED}_{+\text{KRD}}} = \frac{\lambda}{|\mathcal{V}_L|} \sum_{i \in \mathcal{V}_L} \mathcal{H}\left(y_i, \sigma(\mathbf{z}_i^{(L)})\right) + (1-\lambda)\left(\mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{KRD}}\right) + \lambda_{\texttt{TED}} \frac{1}{|\mathcal{C}_{\text{rel}}|} \sum_{c \in \mathcal{C}_{\text{rel}}} \left[R_c(\phi) - \bar{R}(\phi) - \rho\right]_+^2. \tag{46}$$

Table 2: Dataset statistics used to evaluate TED.

Dataset	Type	# Nodes	# Edges	# Features	# Classes
Cora	Homophily	2,708	5,278	1,433	7
Citeseer	Homophily	3,327	4,614	3,703	6
Pubmed	Homophily	19,717	44,324	500	3
Photo	Homophily	7,650	119,081	745	8
CS	Homophily	18,333	81,894	6,805	15
Physics	Homophily	34,493	247,962	8,415	5
Wisconsin	Heterophily	251	499	1,703	5
Cornell	Heterophily	183	295	1,703	5

Integration with FAITH. We instantiate TED on top of FAITH (Singh et al., 2026) by adding the transfer-equity regularizer to the original FAITH training objective. We use FAITH as the base GNN-to-MLP distillation framework. Let $\mathbf{z}_v^s$ denote the student MLP logits for node v and $\mathbf{z}_v^t$ denote the teacher GNN logits. FAITH optimizes a supervised classification loss, a soft-label distillation loss, a neighborhood-guided Dirichlet energy alignment loss, and an oracle similarity alignment loss:

$$\mathcal{L}_{\text{FAITH}}(\phi) = \alpha \sum_{v \in \mathcal{V}_L} \mathcal{L}_{\text{CE}}\left(\sigma(\mathbf{z}_v^s), y_v\right) + (1-\alpha) \sum_{v \in \mathcal{V}} D_{\text{KL}}\left(\sigma(\mathbf{z}_v^s/\tau) \,\|\, \sigma(\mathbf{z}_v^t/\tau)\right) + \beta \mathcal{L}_{\text{DE}} + \gamma \mathcal{L}_{\text{IF}}. \tag{47}$$

Here, $\mathcal{L}_{\text{DE}}$ denotes the neighborhood-guided energy alignment loss, $\mathcal{L}_{\text{IF}}$ denotes the oracle similarity alignment loss, τ is the distillation temperature, and α, β, γ are the FAITH loss weights.

To obtain TED+$_{\text{FAITH}}$, we keep the FAITH objective unchanged and add the class-wise transfer-equity regularizer:

$$\mathcal{L}_{\text{TED+FAITH}}(\phi) = \mathcal{L}_{\text{FAITH}}(\phi) + \lambda_{\text{TED}} \mathcal{L}_{\text{TED}}(\phi) \tag{48}$$

where

$$\mathcal{L}_{\text{TED}}(\phi) = \frac{1}{|\mathcal{C}_{\text{rel}}|} \sum_{c \in \mathcal{C}_{\text{rel}}} \left[R_c(\phi) - \bar{R}(\phi) - \rho\right]_+^2. \tag{49}$$

Thus, TED+$_{\text{FAITH}}$ preserves the structural and fairness-aware objectives of FAITH while adding a constraint on class-wise missed transfer of reliable teacher knowledge. The added term is used only during training. At inference time, the deployed model remains the same graph-free MLP student used by FAITH.

E Dataset Statistics

We utilize eight open graph datasets to validate the proposed TED framework. An overview summary of statistics of the datasets is given in Table 2.

F Hyperparameter Details and System Configuration

For each baseline, we use the official implementation provided by the authors. For each TED-augmented variant, we retain the original implementation of the base method and tune its hyperparameters together with the TED-specific parameter using the validation split. The only additional hyperparameter introduced by TED is the regularization strength $\lambda_{\texttt{TED}}$, which controls the contribution of the transfer-equity regularizer. We select $\lambda_{\texttt{TED}}$ on the validation set from the range $[0, 1.5]$. We summarize the architectural and optimization hyperparameter search spaces in Table 3. All baseline methods and TED-augmented variants are implemented in PyTorch 1.6.0 (Paszke et al., 2019) using the DGL library (Wang et al., 2019). Experiments are run on a Linux machine equipped with an Intel Platinum 8360Y CPU and an NVIDIA A6000 GPU.

Table 3: Hyperparameter search space used in our experiments.

Hyperparameter	**Search space**
Number of layers	$\{2, 3\}$
Hidden dimension	$\{128, 256, 512, 1024, 2048\}$
Learning rate	$\{0.001, 0.005, 0.01\}$
Weight decay	$\{0, 0.0005, 0.001\}$

Table 4: Node classification accuracy and Equity Gap transductive and inductive settings, using GCN as the teacher. MLP and GCN are included as reference accuracy baselines. Parenthesized values in red report the relative Equity Gap reduction of each TED-augmented method with respect to its corresponding base method. Average rows report macro-averages over the datasets.

Dataset	**Reference Accuracy**		**Metric**	**Graph-free Knowledge Distillation**					
	MLP	**GCN**		**GLNN**	**TED$_{+\text{GLNN}}$**	**KRD**	**TED$_{+\text{KRD}}$**	**FAITH**	**TED$_{+\text{FAITH}}$**
				Transductive Setting					
Cora	58.88±0.95	81.48±1.64	Accuracy ↑	83.10±1.11	83.30±2.02	83.34±0.65	83.24±1.57	83.34±0.94	83.66±1.03
			Equity Gap ↓	9.87±2.01	7.78±3.44 (21.18%)	10.08±2.22	7.59±2.06 (24.70%)	7.17±2.55	5.03±1.12 (29.85%)
Citeseer	60.72±0.52	71.60±0.23	Accuracy ↑	73.32±0.86	73.62±0.42	73.12±0.69	73.54±0.57	74.68±0.52	74.88±0.37
			Equity Gap ↓	33.38±28.26	14.38±5.70 (56.92%)	41.26±27.05	13.49±6.90 (67.30%)	17.84±6.55	13.26±4.76 (25.67%)
Pubmed	72.52±0.65	77.16±0.45	Accuracy ↑	81.38±0.71	81.06±0.46	80.82±1.02	81.16±0.45	82.14±0.55	81.88±0.64
			Equity Gap ↓	6.40±2.32	5.04±3.13 (21.25%)	5.78±2.24	4.79±2.27 (17.13%)	3.81±1.90	2.64±1.36 (30.71%)
CS	89.37±1.23	89.85±0.62	Accuracy ↑	92.78±0.27	93.03±0.53	93.60±0.29	93.11±1.27	93.88±0.36	93.83±0.36
			Equity Gap ↓	9.87±1.99	8.62±2.51 (12.66%)	9.92±2.08	9.83±3.01 (0.91%)	5.24±0.65	4.38±0.81 (16.41%)
Physics	87.69±3.79	92.41±0.40	Accuracy ↑	93.68±1.23	94.03±0.38	93.86±1.29	94.00±0.37	94.59±0.37	94.65±0.21
			Equity Gap ↓	3.81±2.41	3.54±0.08 (7.09%)	3.84±2.38	3.48±0.29 (9.37%)	3.57±1.26	1.82±1.42 (49.02%)
Photo	77.66±2.59	91.02±1.46	Accuracy ↑	92.04±1.35	90.01±2.04	92.18±1.51	91.93±1.47	93.46±1.53	93.42±1.42
			Equity Gap ↓	20.43±2.31	16.75±2.43 (18.01%)	18.82±3.35	17.12±3.73 (9.03%)	4.69±2.36	3.24±0.32 (30.92%)
Macro Avg.	**74.47**	**83.92**	Accuracy ↑	**86.05**	**85.84**	**86.15**	**86.16**	**87.02**	**87.05**
			Equity Gap ↓	**13.96**	**9.35** (33.01%)	**14.95**	**9.38** (37.24%)	**7.05**	**5.06** (28.24%)
				Inductive Setting					
Cora	60.78±0.63	79.24±0.59	Accuracy ↑	73.64±0.48	73.00±1.01	73.64±0.14	73.10±0.30	73.24±0.35	73.06±0.26
			Equity Gap ↓	30.96±2.58	24.59±2.16 (20.57%)	29.40±2.83	26.04±1.33 (11.43%)	28.24±9.26	26.75±3.09 (5.28%)
Citeseer	60.38±0.58	72.20±0.24	Accuracy ↑	71.62±0.74	71.70±0.57	71.50±0.52	71.26±0.16	70.72±0.57	70.34±0.76
			Equity Gap ↓	53.31±21.07	23.16±3.96 (56.56%)	53.36±21.35	22.15±3.88 (58.49%)	61.00±22.07	19.89±5.04 (67.39%)
Pubmed	72.44±0.65	77.58±0.28	Accuracy ↑	81.02±0.39	80.98±0.31	81.06±0.62	80.94±0.28	80.20±0.20	80.32±0.72
			Equity Gap ↓	5.66±1.80	2.80±1.40 (50.53%)	4.46±2.54	3.95±2.44 (11.43%)	5.29±2.69	3.30±3.35 (37.62%)
CS	89.59±1.28	90.31±0.21	Accuracy ↑	92.01±0.36	92.38±0.33	92.91±0.46	92.71±0.48	92.75±0.21	92.86±0.14
			Equity Gap ↓	19.92±3.99	17.64±6.60 (11.45%)	22.60±3.44	18.65±6.62 (17.48%)	16.87±3.77	14.38±2.65 (14.76%)
Physics	87.51±2.14	92.68±0.74	Accuracy ↑	92.40±1.96	93.74±0.66	93.34±1.89	93.71±0.88	93.55±0.85	93.51±0.71
			Equity Gap ↓	6.91±1.85	6.53±1.16 (5.50%)	6.81±1.11	6.21±0.79 (8.81%)	4.91±2.07	2.89±0.75 (41.14%)
Photo	76.21±3.00	88.42±2.60	Accuracy ↑	87.94±2.42	89.86±1.89	89.32±3.24	89.29±1.77	89.46±1.86	89.94±2.07
			Equity Gap ↓	47.34±11.83	26.81±7.75 (43.37%)	42.54±17.55	28.86±6.35 (32.16%)	21.08±6.11	18.76±2.70 (11.01%)
Macro Avg.	**74.49**	**83.41**	Accuracy ↑	**83.11**	**83.61**	**83.63**	**83.50**	**83.32**	**83.34**
			Equity Gap ↓	**27.35**	**16.92** (38.13%)	**26.53**	**17.64** (33.49%)	**22.90**	**14.33** (37.43%)

G Performance on GCN Teacher

To show the generalizability of TED with different teacher networks, we report accuracy and Equity Gap using GCN as teacher in both GLNN and KRD. Results in Table 4 highlight that student MLPs augmented with TED consistently provide better tradeoff in utility and equity.

H Ablation under Transductive Setting

We further ablate the proposed objective under the transductive setting. We compare GLNN with a class-balanced GLNN variant and with TED$_{+\text{GLNN}}$ without teacher reliability weighting. The goal is to test whether class reweighting alone is sufficient and whether reliability-aware missed transfer improves the transfer-equity constraint. Figure 7 shows that class balancing gives only limited Equity Gap reduction, while the full TED

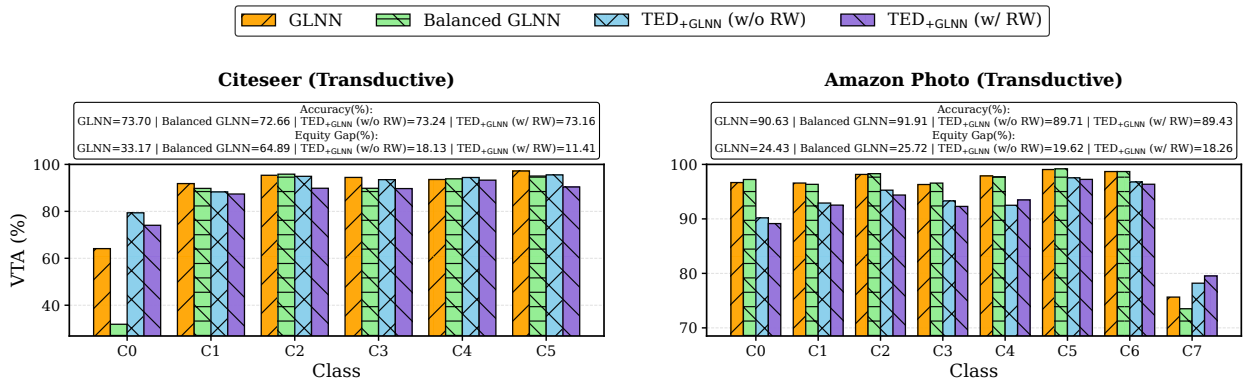


Figure 7: Ablation of class balancing and teacher reliability weighting (RW) under transductive setting on CiteSeer (left) and Amazon-Photo (right).

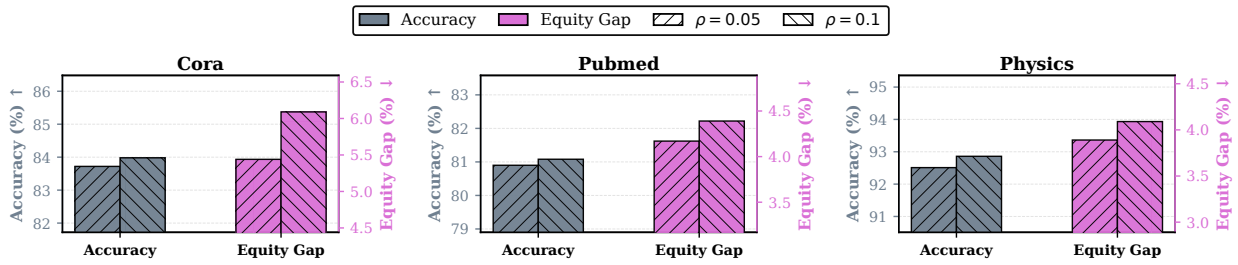


Figure 8: Sensitivity to the tolerance budget ρ under the transductive setting.

objective achieves the strongest accuracy–equity tradeoff. This indicates that teacher reliability weighting provides useful guidance for reducing class-conditional missed transfer.

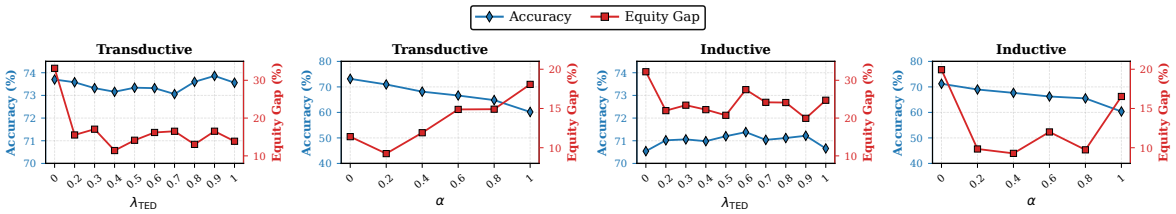
I Impact of Tolerance budget ρ and Hyperparameter $\lambda_{\texttt{TED}}$

We analyze the sensitivity of TED to its main hyperparameters using GLNN as the base distillation method. The tolerance budget ρ determines how much class-wise missed-transfer risk is allowed, while $\lambda_{\texttt{TED}}$ controls the strength of the transfer-equity regularizer. Since GLNN also includes the distillation weight α, we additionally vary α to examine accuracy–equity tradeoff based on the base method's own hyperparameters.

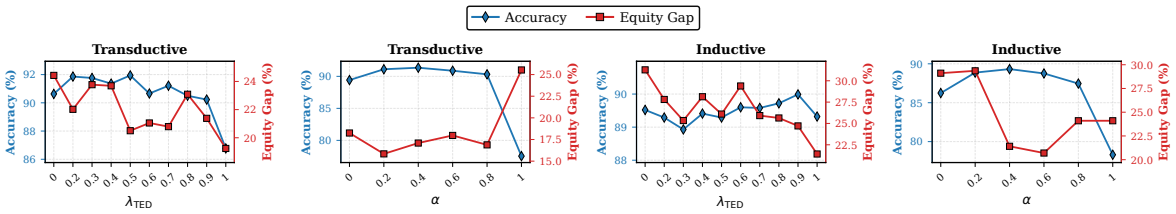
Figure 8 compares $\rho = 0.05$ and $\rho = 0.1$ under the transductive setting. The results show that both choices maintain competitive accuracy, while the stricter tolerance $\rho = 0.05$ generally yields a lower Equity Gap. This is expected, since a smaller tolerance penalizes deviations in class-wise missed transfer more strongly. We use $\rho = 0.1$ in the main experiments as a conservative setting that reduces transfer disparity while avoiding an overly restrictive constraint.

Figure 9 reports sensitivity to $\lambda_{\texttt{TED}}$ and α on CiteSeer and Amazon-Photo under both transductive and inductive settings. The results show that nonzero values of $\lambda_{\texttt{TED}}$ consistently improve the accuracy–equity tradeoff relative to disabling the TED regularizer. Across datasets and settings, moderate values provide stable reductions in Equity Gap with limited changes in accuracy.

The sensitivity to α shows a similar pattern. Moderate reliability weighting improves Equity Gap while preserving predictive utility, whereas extreme values can degrade the tradeoff. This supports the design of TED: teacher reliability is useful for identifying valid transferable knowledge, but it should complement rather than dominate the distillation loss. Overall, the sensitivity analysis indicates that TED does not depend on a narrowly tuned hyperparameter choice and remains effective across a reasonable range of settings.



(a) CiteSeer



(b) Amazon-Photo

Figure 9: Impact of the TED regularization strength λ_{TED} and the GLNN distillation weight α on accuracy and Equity Gap. Results are shown for CiteSeer and Amazon-Photo under both transductive and inductive settings. Moderate nonzero values of λ_{TED} generally reduce Equity Gap while maintaining comparable accuracy, indicating that the transfer-equity regularizer improves class-wise transfer behavior without requiring a narrowly tuned value. The sensitivity to α shows the expected accuracy–equity tradeoff of the base distillation objective.

Table 5: Transductive class-wise VTA diagnostics for GLNN. MeanVTA, WorstVTA, and StdVTA are computed over valid classes $\mathcal{C}_{\text{valid}}$. EG denotes the Equity Gap. All values are reported in percentage points.

Dataset	Method	Acc. ↑	MeanVTA ↑	WorstVTA ↑	StdVTA ↓	EG ↓
CiteSeer	GLNN	73.70	89.42	64.12	11.43	33.17
	$\text{TED}_{+\text{GLNN}}$	73.16	87.44	74.06	6.23	11.41
Amazon-Photo	GLNN	90.63	94.88	75.65	7.33	23.43
	$\text{TED}_{+\text{GLNN}}$	89.43	91.87	79.53	5.22	18.26
Wisconsin	GLNN	81.37	67.46	0.00	34.62	60.82
	$\text{TED}_{+\text{GLNN}}$	80.98	76.92	33.33	22.47	42.59

J Class-wise Valid Transfer Agreement Profile

We use Equity Gap as the primary transfer-equity metric because it directly captures the largest class-conditional disparity in inherited correct teacher knowledge. Since Equity Gap is a range-based summary, we complement it with class-wise VTA profiles to evaluate the underlying per-label transfer behavior.

Table 5 provides a diagnostic view of the scalar Equity Gap under the transductive setting. Across CiteSeer, Amazon-Photo, and Wisconsin, TED improves WorstVTA and reduces both StdVTA and EG. These joint trends indicate that the observed EG reduction is not an isolated endpoint effect: it is accompanied by stronger transfer on the most under-transferred class and lower dispersion across valid classes. The effect is most pronounced on Wisconsin, where the base GLNN student has a severe worst-class transfer failure and TED substantially improves that class. Accuracy remains comparable, showing that the improvement in transfer equity does not require changing the deployed graph-free student. Overall, the diagnostics show that TED improves the class-conditional transfer profile rather than merely producing a favorable change in a EG scalar summary.

K Visualization of Representations

Figures 10 and 11 show t-SNE visualizations of the penultimate-layer student representations on Cora and CiteSeer, respectively. These plots provide a qualitative view of the representation geometry learned by the base graph-free students and their TED-augmented variants. Across both datasets, the TED variants show slightly more class-consistent structure, with several classes appearing more compact.

(a) GLNN (b) $\texttt{TED}_{+\text{GLNN}}$ (c) KRD (d) $\texttt{TED}_{+\text{KRD}}$ (e) FAITH (f) $\texttt{TED}_{+\text{FAITH}}$

Figure 10: t-SNE visualization of learned representations on Cora; colors denote different classes.

(a) GLNN (b) $\texttt{TED}_{+\text{GLNN}}$ (c) KRD (d) $\texttt{TED}_{+\text{KRD}}$ (e) FAITH (f) $\texttt{TED}_{+\text{FAITH}}$

Figure 11: t-SNE visualization of learned representations on Citeseer; colors denote different classes.