A. Figure illustration for Meaningful Initialization and Inter-class Discrimination

As described in the main text, we initialize each label embedding by averaging the example embeddings connected by a True edge and apply an orthogonal loss to enhance inter-class separability. Figure 1 in the appendix illustrates this process and clarifies the rationale behind this design.

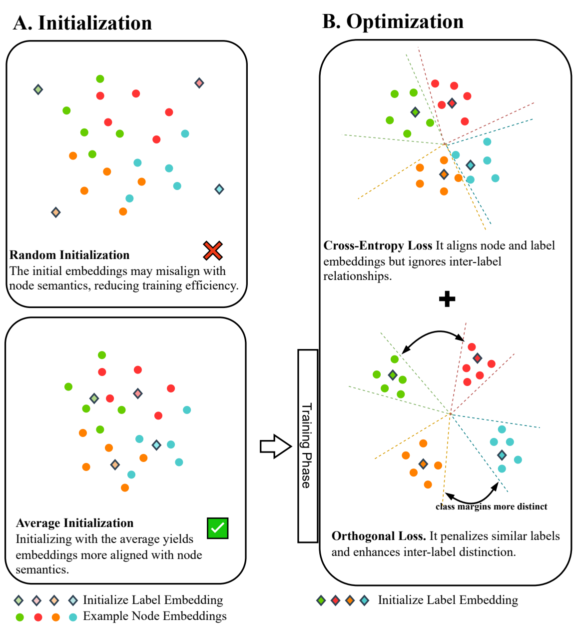

Figure 1: **Average Initialization and Orthogonal loss.** During label embedding initialization (A), random Gaussian initialization provides no contextual prior, resulting in class-agnostic embeddings. A more effective approach initializes each label with the mean embedding of its class examples, placing it closer to its target region and improving optimization—especially in few-shot settings. However, label embeddings generated by GNNs could drift toward each other (B), as cross-entropy aligns queries to their labels but does not enforce inter-label separation. To address this, we introduce an orthogonality loss that penalizes label similarity, preserving class margins while cosine-based cross-entropy maintains alignment.

B. Attribute Prediction Loss

Attribute Prediction Loss. During MASKNODE augmentation on each augmented graph $\widetilde{\mathcal{G}_x^{con}}$, we mask a subset of node attributes F_v. An MLP takes the learned node embeddings E_v to predict the masked attributes, and we add a mean squared error (MSE) reconstruction loss as an auxiliary augmentation term:

$$\mathcal{L}_{\text{attr}}(\widetilde{\mathcal{G}_x^{con}}) = \frac{1}{|\mathcal{V}_x|} \sum_{v \in \mathcal{V}_x} \text{MSE}(F_v, \text{MLP}(E_v)), \quad (1)$$

where $\mathcal{V}^D$ denotes the index set of nodes whose attributes are masked by MASKNODE.

C. Datasets Statistics

Table 1: Dataset statistics

Dataset	# Nodes	# Edges	# Classes
MAG240M	122M	1.3B	153
Wiki	4.8M	5.9M	639
arXiv	169K	1.2M	40
ConceptNet	791K	2.5M	14
FB15K-237	15K	268K	200
NELL	69K	181K	291

D. Algorithm for MAG24m Sub-sampling

This algorithm samples a subgraph from MAG240M using a BFS-based strategy. It first computes node degrees from the adjacency matrix and initializes the sampled set S with the highest-degree node. A queue is used to perform BFS, iteratively adding unvisited neighbors of nodes in S until the sampled set reaches the target size T.

Algorithm 1: BFS-Based Subgraph Sampling from MAG240M

Input: Adjacency matrix A, target size T
Output: Sampled node set $\mathcal{S}$

Compute degree for each node: $\deg(i) = \sum_j A_{ij} + \sum_j A_{ji}$
$\mathcal{S} \leftarrow \emptyset$
while $|\mathcal{S}| < T$ **do**
    $u \leftarrow \arg\max_{i \notin \mathcal{S}} \deg(i)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{u\}$; **initialize** queue $Q \leftarrow [u]$
    while $|\mathcal{S}| < T$ **and** $Q \neq \emptyset$ **do**
        $v \leftarrow Q$.pop(0)
        $N(v) \leftarrow \{ w \mid A_{vw} = 1 \lor A_{wv} = 1 \}$
        for all $w \in N(v)$ **and** $w \notin \mathcal{S}$ **do**
            $\mathcal{S} \leftarrow \mathcal{S} \cup \{w\}$; Q.append(w)
            if $|\mathcal{S}| = T$ **then**
                break
            end if
        end for
    end while
end while
return $\mathcal{S}$

E. Heatmap for different set of ways

We presented the parameter study heatmaps for the 3-way and 10-way settings in the main paper; here, we include the remaining three settings. These figures illustrate the combinations of λ and p, where each cell reports accuracy and improvement over the baseline (red indicates gains and blue indicates drops).

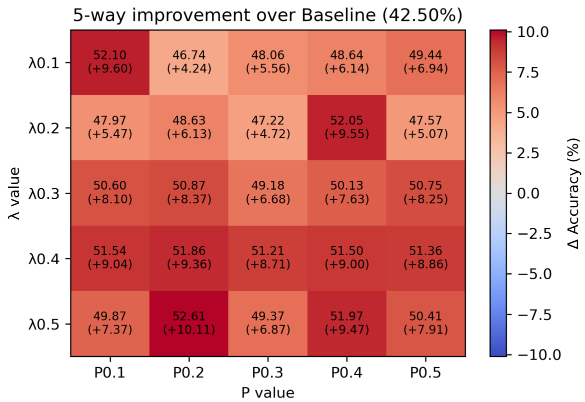

Figure 2: **Parameter study of λ and p in 5 ways**

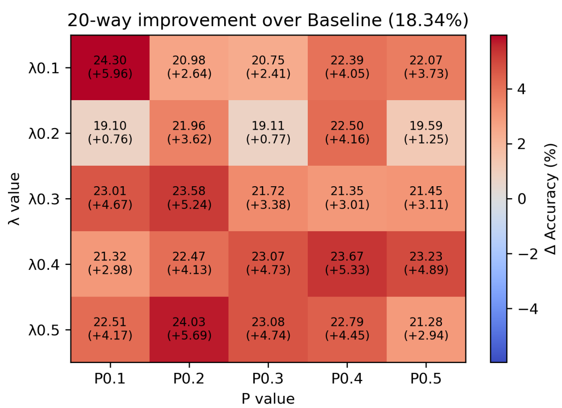

Figure 3: **Parameter study of λ and p in 20 ways**

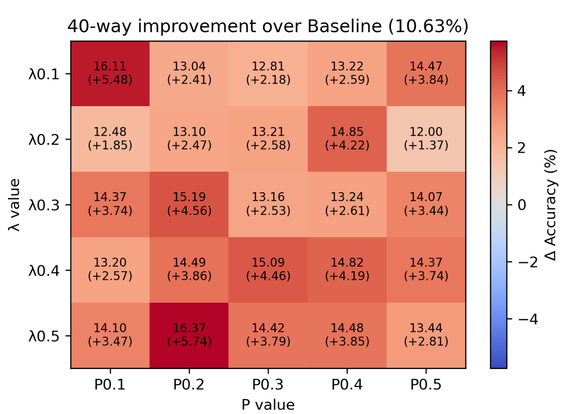

Figure 4: **Parameter study of λ and p in 40 ways**