Towards Automatic Evolution Tree Generation from Citation Graphs

Zexing Zhao[1], **Yuntong Hu**[2], **Liang Zhao**[2†]
[1]School of Mechanical Engineering, Georgia Institute of Technology
[2]Department of Computer Science, Emory University
alfred.zhao@gatech.edu, {yuntong.hu, liang.zhao}@emory.edu

Abstract

Surveys remain the primary way researchers grasp the lineage of methods within an AI subfield, but they scale poorly against the current rate of publication. Existing taxonomy-induction methods are largely leaf-bound and time-agnostic; they tend to force transitional papers into mature leaves and can create topological inversions between ancestors and descendants. We propose EvoTree, a staged framework that decouples conceptual backbone learning from temporal refinement: a graph-aware encoder with distribution-based hierarchical clustering yields a stable taxonomy backbone; temporal fine-tuning then re-attaches marginal papers to internal nodes under monotonic-path constraints; a final LLM pass labels concepts without altering the topology. We release the first annotated benchmark for this task across 11 AI subfields. EvoTree attains the highest NMI and citation-direction accuracy among all baselines and the best concept purity on the annotated benchmark, and is the only method with non-trivial marginal-paper detection on the annotated set.

1 Introduction

The rapid growth of scientific literature has fundamentally challenged human-centered knowledge organization and field understanding. Open scholarly infrastructures such as OpenAlex (Priem et al., 2022) now index over 470 million scientific works together with continuously expanding citation networks. As research fields evolve at an unprecedented pace, researchers increasingly struggle to efficiently understand the historical evolution, conceptual lineage, and paradigm transitions of rapidly developing domains (Fortunato et al., 2018), creating an increasing need for AI-assisted methods to organize, synthesize, and understand scientific progress (Le et al., 2026; Hu et al., 2025).

Among various forms of scientific organization, evolution trees and hierarchical lineage diagrams are particularly effective because they explicitly reveal conceptual inheritance, branching, and progressive refinement across generations of work. Such structures are widely adopted in modern survey papers to summarize the evolution of research paradigms, model families, and emerging subfields (Zhang et al., 2018; Wang et al., 2024; Hu et al., 2024; Liang et al., 2025). However, constructing these evolution trees remains largely manual, requiring substantial domain expertise and intensive literature analysis, while quickly becoming outdated as new papers continuously emerge (Wang et al., 2024; Liang et al., 2025). This motivates a fundamental research question:

> ***Can we automatically generate scientific evolution trees from citation graphs?***

Addressing this problem is highly challenging due to several underexplored issues. First, scientific evolution trees are hierarchical, temporal, and semantically structured objects, whereas citation graphs are large-scale sparse relational networks. Bridging these two fundamentally different structures requires learning evolution-aware representations that preserve both citation dependencies and temporal scientific progression, which is not directly addressed by existing taxonomy induction methods (Zhang et al., 2018; Shen et al., 2020; Yu et al., 2020) or temporal topic modeling approaches (Blei and Lafferty, 2006; Wang and McCallum, 2006; Wang et al., 2008). Second, scientific evolution is inherently dynamic: research paradigms continuously branch, merge, and refine over time, making it difficult to construct stable and interpretable lineage structures from evolving citation graphs. Existing citation and co-citation analy-

†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 28680–28703
October 24-29, 2026 ©2026 Association for Computational Linguistics

ses in science-of-science research primarily characterize global scientific trends rather than inducing explicit hierarchical evolutionary structures (Fortunato et al., 2018). Third, existing survey papers provide only limited supervision signals, requiring models to generalize from scarce high-quality evolution annotations. Moreover, marginal or transitional papers often lie between major paradigms, making them difficult to position within conventional taxonomy induction frameworks without introducing chronological inconsistencies.

To address those challenges, we propose a structurally staged and temporally constrained framework that unifies taxonomy induction, temporal admissibility, and marginal-paper re-attachment into a cohesive pipeline. Specifically, our framework first constructs a stable, time-agnostic core taxonomy via distribution-based hierarchical clustering on graph-aware node embeddings. Subsequently, it performs temporal refinement to model evolutionary dynamics and handle marginal papers by attaching them to broader internal concepts, while enforcing temporal consistency and structural legality constraints (e.g., path monotonicity). This staged approach ensures that the conceptual backbone remains robust while temporal dynamics are accurately and logically mapped.

Our contributions are summarized as follows:

- **Novel Problem Formulation.** We formulate evolutionary tree construction from citation graphs, introducing a dynamic perspective for scientific literature mining.
- **Staged & Constrained Framework.** We propose a staged framework that separates taxonomy initialization from temporal refinement, preserving citation-aware structures while filtering peripheral papers from core trajectories.
- **Evolutionary Dynamics Modeling.** We incorporate temporal and structural constraints with marginal-paper detection and internal-node reattachment, alleviating chronological distortions in traditional taxonomy generation.
- **Comprehensive Benchmark.** We construct the first benchmark for this task across 11 AI subfields, where extensive experiments show consistent gains over taxonomy, citation-based, and LLM baselines in evolution-structure quality and annotated concept purity, together with the only non-trivial marginal-paper detection.

2 Related Work

2.1 Automatic Taxonomy Induction

Automatic taxonomy induction studies how to organize terms, entities, or topics into hierarchical structures such as trees or DAGs. Prior work has shown that taxonomy quality requires more than isolated pairwise hypernym prediction (Snow et al., 2006; Bansal et al., 2014; Mao et al., 2018). Most existing approaches focus on taxonomy expansion or completion given an existing hierarchy (Shen et al., 2020; Yu et al., 2020; Mishra et al., 2025; Jiang et al., 2022), or extend the hierarchy formalism itself (Lee et al., 2022; Lu et al., 2024; Kargupta et al., 2025). A smaller line of work constructs topic or concept hierarchies directly from corpora. TaxoGen (Zhang et al., 2018) recursively builds topic taxonomies through adaptive spherical clustering and local embeddings, while HiExpan (Shen et al., 2018) expands user-guided seed hierarchies using weakly supervised relation extraction. A recent line of work applies language models to scholarly taxonomy construction. TaxoAlign (Lahiri et al., 2025) generates taxonomies for scientific domains by aligning LM-induced hierarchies to reference structures, and Zhu et al. (2025) induce hierarchical taxonomies for scientific papers via LLM-guided multi-aspect clustering. However, these methods are primarily designed for text corpora and cannot effectively model the structural dependencies. Moreover, they generally assume static taxonomies and therefore fail to capture temporally constrained scientific evolution, including paradigm transitions and lineage progression across generations of work.

2.2 Knowledge Evolution Modeling

Knowledge evolution modeling studies how scientific concepts, research topics, and paradigms emerge, evolve, and transition over time. Early work primarily focused on temporal topic modeling, including Dynamic Topic Models (Blei and Lafferty, 2006; Wang et al., 2008; Bhadury et al., 2016) and Topics over Time (Wang and McCallum, 2006), which capture topic drift and temporal word distributions across document collections. Another line of research investigates scientific evolution through citation and co-citation analysis within the science-of-science community (Fortunato et al., 2018). More recently, LLM-based survey systems such as AutoSurvey (Wang et al., 2024) and SurveyX (Liang et al., 2025) retrieve papers, construct

outlines, and generate survey-style text, but primarily optimize textual synthesis rather than explicit structural evolution modeling. Although several recent works summarize papers into taxonomy-like structures (Hu et al., 2024; Zhu et al., 2025), none directly addresses scientific evolution tree generation. Existing methods lack temporal constraints over evolutionary paths, and mechanisms for handling marginal papers or constructing survey-style evolution structures.

3 Problem Formulation

Citation graph. Given a survey paper, let $\mathcal{G} = (\mathcal{V}, \mathcal{E}_{\mathcal{G}})$ denote the citation graph constructed from its reference set, where each node $v_i \in \mathcal{V}$ is a referenced paper with text content s_i (title and abstract) and publication year $\tau(v_i)$. A directed edge $(v_i, v_j) \in \mathcal{E}_{\mathcal{G}}$ indicates that paper v_i cites paper v_j. For the main evaluation, the target survey itself is excluded from $\mathcal{G}$.

Taxonomy tree. A taxonomy tree organizes papers by conceptual inclusion. We denote it as $\mathcal{T}^{\text{tax}} = (\mathcal{C}^{\text{tax}}, \mathcal{V}, \mathcal{E}^{\text{tax}}, \phi^{\text{tax}})$, where $\mathcal{C}^{\text{tax}}$ is the set of topic nodes, $\mathcal{V}$ is the set of paper nodes, and $\mathcal{E}^{\text{tax}} \subseteq \mathcal{C}^{\text{tax}} \times (\mathcal{C}^{\text{tax}} \cup \mathcal{V})$ forms a rooted tree. Each topic node $c \in \mathcal{C}^{\text{tax}}$ represents a semantic concept with label $\phi^{\text{tax}}(c)$, and each paper node $v_i \in \mathcal{V}$ appears only as a terminal leaf (i.e., for every paper node, its parent must be a concept node rather than another paper node, and this parent corresponds to the lowest-level concept cluster that directly contains the paper). Thus, edges in $\mathcal{T}^{\text{tax}}$ encode conceptual refinement only, with no temporal semantics.

Evolutionary tree. An evolutionary tree $\mathcal{T} = (\mathcal{C}, \mathcal{V}, \mathcal{E}_C, \mathcal{E}_P, \phi)$ extends this organization with temporally admissible development relations. $\mathcal{E}_C$ forms a rooted tree over concepts $\mathcal{C}$ with labels ϕ, and $\mathcal{E}_P = \{(a(v_i), v_i)\}$ attaches each paper to a concept via the attachment map $a : \mathcal{V} \to \mathcal{C}$. Unlike a taxonomy tree, papers may attach to internal concept nodes when better explained at a coarser abstraction level.

For each concept c, let $\mathcal{V}_{\text{att}}(c) = \{v_i : a(v_i) = c\}$ denote the papers attached directly to c, and $\mathcal{V}(c)$ the papers in its full subtree. We define two temporal statistics over publication years $\tau(\cdot)$:

$$\tau^{\text{att}}(c) = \frac{1}{|\mathcal{V}_{\text{att}}(c)|} \sum_{v \in \mathcal{V}_{\text{att}}(c)} \tau(v), \quad \bar{\tau}(c) = \frac{1}{|\mathcal{V}(c)|} \sum_{v \in \mathcal{V}(c)} \tau(v). \tag{1}$$

τ^{att} supports structural constraints: because the attachment sets of a parent and its children are disjoint, the constraint it induces is non-trivial. $\bar{\tau}$ summarizes the full concept body and is used for path-level evaluation.

The tree must satisfy (i) **Rooted Tree**; (ii) **Temporal Consistency**, for each concept edge $(c_p, c_c) \in \mathcal{E}_C$ from parent c_p to child c_c,

$$\tau^{\text{att}}(c_p) \leq \tau^{\text{att}}(c_c) + \epsilon; \tag{2}$$

and (iii) **Sibling Ordering** by non-decreasing τ^{att}. Eq. 2 requires each root-to-leaf path to proceed from earlier, broader concepts to later, more specialized ones, with the slack ϵ absorbing noise in publication dates so that a single mis-dated paper does not invalidate an otherwise valid edge.

Marginal papers. A marginal paper has weak compatibility with all leaf-level concepts. Given a continuous compatibility score $m(v_i, c) \in [0, 1]$ (instantiated in §4.4.3), the marginal set is

$$\mathcal{M} = \{v_i \in \mathcal{V} : \max_{\ell \in \text{Leaves}(\mathcal{T})} m(v_i, \ell) < \eta\}, \tag{3}$$

where η is a threshold. Marginal papers are not discarded; they may attach to internal nodes to represent isolated, transitional, or weakly continued contributions.

Challenges and design overview. This formalism poses four challenges that shape our method. (C1) The discrete structure $(\mathcal{C}, \mathcal{E}_C, a)$ and continuous embeddings $\{\mathbf{e}_i\}$ cannot be jointly optimized end-to-end, motivating *alternating refinement* between structure and representation (§4.4.1). (C2) Temporal consistency (Eq. 2) cannot emerge from a time-agnostic clustering loss, motivating a *temporal merge filter* on τ^{att} (§4.4.1). (C3) Marginal papers fall outside leaf-only partitions, motivating explicit *internal-node re-attachment* (§4.4.3). (C4) Temporal refinement risks collapsing the conceptual organization, motivating a *staged design* that establishes a stable backbone $\mathcal{T}^{\text{tax}}$ (§4.3) before applying temporal pressure (§4.4).

4 Methodology

4.1 Overview

EvoTree (Fig. 1) addresses challenges C1–C4 through three stages over a shared backbone. The input of EvoTree (Fig. 1A) is a citation graph in which each node carries paper text and a timestamp,

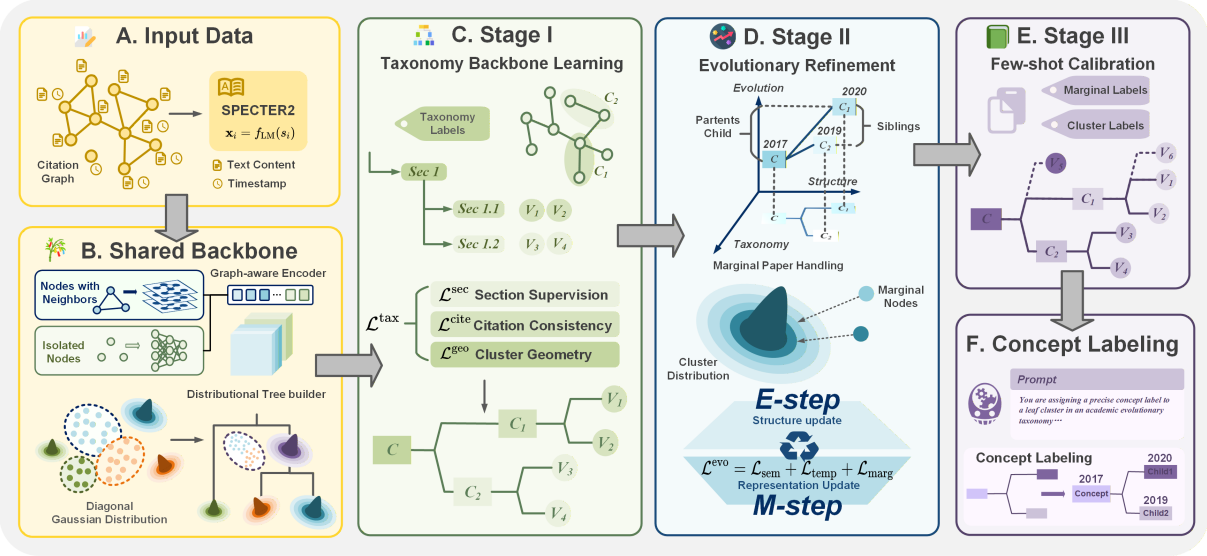


Figure 1: Overview of the EvoTree pipeline. **(A)** Input citation graph encoded by SPECTER2. **(B)** Shared backbone with graph-aware encoder and distributional tree builder. **(C)** Stage I taxonomy backbone learning under $\mathcal{L}^{\text{tax}}$. **(D)** Stage II evolutionary refinement via alternating updates under $\mathcal{L}^{\text{evo}}$. **(E)** Stage III few-shot calibration. **(F)** LLM concept labeling.

encoded by SPECTER2. The backbone (§4.2) couples a graph-aware encoder with a distributional tree builder that summarizes each cluster as a diagonal Gaussian. Stage I (§4.3) learns a time-agnostic taxonomy backbone $\mathcal{T}^{\text{tax}}$ under section-derived weak supervision. Stage II (§4.4) refines $\mathcal{T}^{\text{tax}}$ into the evolutionary tree $\mathcal{T}$ through alternating structure–representation updates with temporal admissibility and marginal-paper handling. Stage III (§4.5) calibrates against reference trees via few-shot metric learning, and a final LLM pass (§4.6) verbalizes concept labels without altering topology. The main text is self-contained for the design rationale of each component; appendices provide derivations, loss formulations, and implementation details.

4.2 Shared Backbone

The pipeline rests on two reusable modules: a *graph-aware paper encoder* and a *distributional tree builder*.

4.2.1 Graph-aware Encoder

The encoder maps each paper v_i to a clustering embedding $\mathbf{e}_i \in \mathbb{R}^d$ from text and citation context. A pretrained scientific encoder SPECTER2 (Singh et al., 2022) produces a time-agnostic vector $\mathbf{x}_i$, from which a semantic stream $\mathbf{z}_i^{\text{sem}} = \text{MLP}_{\text{sem}}(\mathbf{x}_i)$ and a graph stream $\mathbf{h}_i^{(0)} = \text{MLP}_{\text{graph}}(\mathbf{x}_i)$ are derived. The graph stream propagates through L layers with anchored residuals and PairNorm (Zhao and Akoglu, 2020):

$$\mathbf{h}_i^{(l+1)} = \text{PairNorm}\Big(\alpha\mathbf{x}_i + (1-\alpha)\big[\beta\mathbf{h}_i^{(l)} + (1-\beta)\mathbf{m}_i^{(l)}\big]\Big), \tag{4}$$

where $\mathbf{m}_i^{(l)}$ aggregates messages from citation neighbors (set to $\mathbf{0}$ for isolated nodes). The anchored term mitigates over-smoothing while PairNorm stabilizes propagation under uneven density. The two streams fuse via a connectivity-biased gate:

$$\begin{aligned}\mathbf{z}_i^{\text{fused}} &= \lambda_i\mathbf{h}_i^{(L)} + (1-\lambda_i)\mathbf{z}_i^{\text{sem}},\\ \lambda_i &= \sigma\big(\mathbf{W}_g[\cdot] + \delta\mathbb{1}[\text{iso}(v_i)]\big),\end{aligned} \tag{5}$$

with $\delta < 0$ biasing isolated nodes toward the semantic stream. The final embedding is $\mathbf{e}_i = \text{L}_2\text{Norm}(\mathbf{z}_i^{\text{fused}})$; a separate projection $\mathbf{p}_i$ is reserved for contrastive learning (Chen et al., 2020). Full details in Appendix F.

4.2.2 Distributional Tree Builder

Given embeddings $\{\mathbf{e}_i\}$, the builder constructs a hierarchical tree through distributional clustering. HDBSCAN with soft membership (Campello et al., 2013) yields base clusters $\{C_k\}_{k=1}^K$ and weights w_{ik}, each summarized as $C_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \text{diag}(\boldsymbol{\sigma}_k^2))$. Higher-level concepts arise from bottom-up agglomeration under the closed-form 2-Wasserstein distance:

$$d^2(C_p, C_q) = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\|_2^2 + \|\boldsymbol{\sigma}_p - \boldsymbol{\sigma}_q\|_2^2. \tag{6}$$

The parent Gaussian is re-estimated from union members after each merge, reducing level-wise drift. The same builder is used in Stages I and II; what changes between them is the supervision and whether temporal admissibility is enforced.

4.3 Stage I: Taxonomy Backbone Learning

Stage I learns a time-agnostic backbone $\mathcal{T}^{\text{tax}}$ that captures conceptual inclusion. Inputs are the citation graph, paper texts, and weak hierarchical labels derived from the survey section structure (each paper inherits its section path). We pretrain the graph-aware paper encoder (§4.2.1) under three complementary signals: **section supervision** (multi-label classification and level prediction over the section hierarchy), **citation consistency** (link prediction and contrastive ranking on citation edges), and **cluster geometry** (centroid pulling for labeled papers):

$$\mathcal{L}^{\text{tax}} = \mathcal{L}_{\text{sec}} + \mathcal{L}_{\text{cite}} + \mathcal{L}_{\text{geo}}. \tag{7}$$

Applying the distributional tree builder to the pretrained embeddings yields $\mathcal{T}^{\text{tax}}$, which serves as the structural prior for Stage II. Component formulations are given in Appendix H.

We emphasize that section structures serve as weak supervision rather than ground-truth concept hierarchies: headings reflect expository choices as much as conceptual organization, and their granularity varies across surveys. Their influence is correspondingly bounded. They contribute only one of the three terms in Eq. 7, shaping the initialization jointly with the citation and geometry objectives rather than acting as a direct optimization target, and are unavailable at inference.

4.4 Stage II: Evolutionary Refinement

Stage II transforms $\mathcal{T}^{\text{tax}}$ into the evolutionary tree $\mathcal{T}$, driven by three signals absent from Stage I: **temporal admissibility** (Eq. 2), **citation directionality** (citing papers attach deeper than cited ones), and **marginal-paper re-attachment**. The heavy components (SPECTER2, GNN weights) remain frozen; only lightweight adapters (projection MLP, fusion gate, marginal head, cluster Gaussians) are fine-tuned.

4.4.1 EM-style Refinement

Because discrete tree construction and continuous adapter updates cannot be jointly optimized end-to-end (C1), we adopt an alternating procedure inspired by EM. We label the two phases as E-step and M-step for brevity, without claiming a strict EM optimization of a probabilistic likelihood.

E-step — structure update. With adapters fixed, we re-apply the distributional tree builder to the current embeddings. Unlike Stage I, agglomeration is filtered by temporal admissibility: a candidate merge producing parent c_p from children c_a, c_b is rejected if

$$\tau^{\text{att}}(c_p) > \min\big(\tau^{\text{att}}(c_a),\, \tau^{\text{att}}(c_b)\big) + \epsilon, \tag{8}$$

which uses the same τ^{att} statistic as the formal admissibility condition (Eq. 2). This step yields an updated tree $\mathcal{T}^{(t)}$ and a set of low-membership marginal candidates.

M-step — representation update. With $\mathcal{T}^{(t)}$ fixed, we update the adapters under three self-supervised objectives:

$$\mathcal{L}^{\text{evo}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{marg}}, \tag{9}$$

where $\mathcal{L}_{\text{sem}}$ preserves parent–child semantic coherence, $\mathcal{L}_{\text{temp}}$ encourages temporal and citation-direction consistency, and $\mathcal{L}_{\text{marg}}$ trains the marginal detector (§4.4.3). Detailed formulations in Appendix I.

4.4.2 Structural Admissibility Constraints

In addition to the differentiable objectives, the E-step filters candidate trees using admissibility constraints. These constraints prevent the evolutionary refinement from producing structures that are temporally invalid or that destroy the semantic backbone learned in Stage I.

Temporal consistency. This constraint follows Eq. 2, directly rejecting violating merges.

Backbone preservation. This retains the major branches of $\mathcal{T}^{\text{tax}}$ via the Major Branch Count:

$$\text{MBC}(\mathcal{T}) = \big|\{c \in \text{children}(r) : |\mathcal{V}(c)| \geq \rho|\mathcal{V}|\}\big|, \tag{10}$$

requiring $\text{MBC}(\mathcal{T}) \geq \kappa\, \text{MBC}(\mathcal{T}^{\text{tax}})$.

Tree legality. The output must remain a valid rooted tree, with a single root, no cycles, and a unique parent for each non-root concept node. Detailed constraints are described in Appendix J.

4.4.3 Marginal Paper Handling

Stage II explicitly handles *marginal papers*: papers poorly explained by all fine-grained leaf concepts, which would blur concept boundaries and weaken

temporal coherence if forced into leaves. We separate handling into leaf-level detection and all-node re-attachment.

Detection. For each paper v_i, we compute its best leaf compatibility $A_i^{\text{leaf}} = \max_{c \in \text{Leaves}(\mathcal{T}^{(t)})} m(v_i, c)$, where $m(v_i, c) \in [0, 1]$ combines semantic similarity, citation-neighbor overlap, and temporal compatibility (Appendix K). A paper is marginal when $A_i^{\text{leaf}} < \eta$, providing pseudo-labels for a marginal classifier (optionally reinforced by HDBSCAN low-membership outliers). Intuitively, a paper is marginal when it is not well explained by any single leaf: its compatibility with even its best-matching concept falls below η, typically because its citations span multiple sibling branches rather than concentrating in one.

Re-attachment. Each flagged marginal paper is re-attached to the concept whose Gaussian best explains its embedding:

$$a(v_m) = \arg\max_{c \in \mathcal{C}^{(t)}} \log \mathcal{N}\big(\mathbf{e}_m;\, \boldsymbol{\mu}_c, \text{diag}(\boldsymbol{\sigma}_c^2)\big)\,. \tag{11}$$

Unlike detection, re-attachment searches over all concept nodes, allowing papers that fit no fine-grained leaf to attach to broader internal concepts.

4.5 Stage III: Few-shot Metric Learning

To inject structural priors that self-supervised signals alone cannot recover, we calibrate the model on a small set of reference evolutionary trees via few-shot metric learning. Only a lightweight adaptation layer (∼17% of parameters) is updated under a small learning rate (5e-6) for 20 epochs over the reference trees (10 per leave-one-domain-out fold) — orders of magnitude smaller than standard supervised fine-tuning. The calibration losses are purely metric-learning: they shape embedding geometry to reflect concept clusters and evolution directions, without predicting discrete labels.

$$\mathcal{L}^{\text{cal}} = \mathcal{L}^{\text{cal}}_{\text{marg}} + \mathcal{L}^{\text{cal}}_{\text{concept}} + \mathcal{L}^{\text{cal}}_{\text{edge}}. \tag{12}$$

Weight and learning-rate choices ensure the reference signal refines rather than overrides earlier structure. Full formulations in Appendix L.

4.6 Concept Labeling

The tree $\mathcal{T}$ encodes structure but no human-readable labels. A single LLM pass traverses $\mathcal{T}$ bottom-up:

$$\phi(c) = \text{LLM}\big(\Pi(c,\, \mathcal{V}(c),\, \{\phi(c') : c' \in \text{children}(c)\})\big), \tag{13}$$

where Π is a structured prompt. Leaf prompts request a precise method name; internal-node prompts require a more abstract noun phrase that generalizes over (and does not repeat) child labels. **The LLM is used only for concept verbalization and does not modify topology or paper attachments.** Prompt templates in Appendix C.

5 Experiments

5.1 Datasets

Survey-reference graph dataset. We construct 411 ego-graphs from arXiv survey papers. Hierarchical weak labels are derived from the survey section structure: a referenced paper inherits the section path in which it appears, e.g., *Chapter 2 → 2.2 → 2.2.3*. We split graphs, rather than individual nodes, into train/validation/test sets with a 70/15/15 ratio. Detailed statistics are shown in Table 6.

Few-shot (FS) calibration dataset. For direct evaluation against annotated labels, we curate 11 reference evolutionary trees spanning major AI subfields, including computer vision, natural language processing, graph learning, reinforcement learning, and large language models. Each graph contains manual annotations for concept membership (`FS_concept`), marginal-paper labels (`FS_is_marginal`), evolution edges (`FS_tree_edges`), and node depth (`FS_depth`). Detailed statistics are shown in Table 7. Calibration follows an 11-fold leave-one-domain-out protocol (Appendix B.3).

5.2 Clustering Semantic Quality and Evolution Structure Quality

5.2.1 Metrics & Baselines

We evaluate two aspects of tree quality. For semantic clustering, we report **Leaf Purity (LP)** and **Path Purity (PP)** to measure topic consistency at the leaf and root-to-leaf path levels, respectively, and **NMI** to measure agreement between the tree-induced partition and topic labels. For evolution structure, we report **Citation Direction Accuracy (CDA)**, which checks whether cited papers are placed at shallower depths than citing papers, and **Path Monotonicity Rate (PMR)**, which measures whether publication years are non-decreasing along root-to-leaf paths.

We compare EvoTree with representative flat, hierarchical, citation-aware, and taxonomy-induction baselines, including HAC-SPECTER2 (Murtagh and Contreras, 2012), KMeans-flat (Ahmed et al., 2020), HAC-4level (Murtagh and Contreras, 2012), CitRank+HAC (Woods, 2024), TaxoGen-style (Zhang et al., 2018), Hu-CiteTaxo (Hu et al., 2024), TaxoAlign (Lahiri et al., 2025), and Context-Aware (Zhu et al., 2025). We also report **EvoTree (Pre-train)**, which disables temporal fine-tuning, to isolate the effect of representation–structure refinement.

5.2.2 Results

Semantic clustering quality. Taxonomy-generation methods remain competitive here: Context-Aware leads on LP (0.713) and PP (0.742), exceeding EvoTree by 6.0 and 3.9 points. EvoTree attains the highest NMI (0.526). These metrics reward flat leaf coherence alone; EvoTree additionally constrains its hierarchy to be temporally admissible, trading some semantic compactness for evolution-consistent structure.

Citation- and time-consistent evolution structure. Hu-CiteTaxo attains the highest PMR (0.322); its CDA is not comparable, as its output does not expose the directional edges the metric evaluates. The more informative distinction is structural: none of the three taxonomy-generation baselines emits evolution edges or marginal scores, so they are not evaluated on the FS-aligned metrics (§5.3). The indirect metrics above cover only the semantic objective that taxonomy generation shares with evolution-tree induction.

Effect of temporal fine-tuning. Comparing EvoTree with EvoTree (Pre-train) isolates the effect of the temporal fine-tuning stage. Full EvoTree improves LP and CDA, showing that joint representation–structure refinement improves both semantic coherence and citation-direction consistency.

5.3 Evaluation on Few-shot Test

5.3.1 Metrics & Baselines

We evaluate against the annotated few-shot dataset using three metrics: **FS-CP**, which measures concept purity within predicted leaf nodes; **FS-EDA**, which measures whether annotated evolutionary descendants are placed deeper than their predecessors; and **FS-Mar**, which evaluates marginal-paper detection by AUROC.

We compare EvoTree with clustering baselines from §5.2 and LLM-based baselines that prompt GPT-4o to generate hierarchies from titles, titles plus years, or titles plus citation-graph context. We again include EvoTree (Pre-train) to measure the effect of temporal fine-tuning.

5.3.2 Results

Alignment with annotated concepts. On the few-shot dataset, EvoTree achieves the highest FS-CP, though the gap is narrow. Its improvement over the strongest LLM baseline is modest, but it consistently outperforms clustering-based baselines and substantially improves over EvoTree (Pre-train), indicating that temporal fine-tuning and tree refinement help align predicted clusters with ground-truth concepts.

Evolution-direction accuracy. EvoTree is on par with the strongest LLM baseline on FS-EDA. The strong performance of LLM-title suggests that large language models can often infer local predecessor–successor relations from method names alone. However, LLM-only methods do not provide explicit structure-level constraints or marginality estimates, which are required for controlled, audit-friendly tree construction.

Marginal-paper detection. FS-Mar reflects a structural capability that, to our knowledge, no other method evaluated in this work exposes: standard clustering and LLM baselines produce only hard assignments and therefore admit no continuous marginality score, while EvoTree directly outputs marginality scores from its hierarchical structure. EvoTree obtains a non-trivial FS-Mar AUROC and improves over its pre-training-only variant, supporting the value of modeling boundary papers rather than forcing every paper into a fine-grained leaf cluster.

5.4 Ablation Studies

5.4.1 Contribution of Each Training Stage

EvoTree is trained in three stages: **Pre-train only** (Stage I) uses multi-task supervision with section structure as weak labels; **+ Post-train** (Stage II) adds self-supervised evolutionary losses; **+ FS-correction** (Stage III few-shot calibration; our full model) injects structural priors via few-shot metric learning on the reference trees. We evaluate on test graphs and on the FS dataset (Table 3) to verify that gains generalize beyond the FS-correction training set.

Table 1: Clustering semantic quality and evolution structure quality. Dashes indicate that the method does not produce the required output. Results are mean ± standard deviation over 5 runs. Best results are in **bold**.

Method	Clustering			Evolution	
	LP ↑	PP ↑	NMI ↑	CDA ↑	PMR ↑
HAC-SPECTER2	0.526±0.008	0.607±0.009	0.399±0.011	–	–
KMeans-flat	0.524±0.010	0.603±0.011	0.392±0.012	–	–
HAC-4level	0.586±0.007	0.665±0.008	0.437±0.010	1.000	0.293±0.014
CitRank+HAC	0.498±0.011	0.574±0.012	0.346±0.013	1.000	0.250±0.015
TaxoGen-style	0.601±0.009	0.674±0.008	0.511±0.010	0.892±0.013	0.265±0.013
Hu-CiteTaxo	0.664±0.006	0.718±0.008	0.519±0.005	–	**0.322±0.007**
TaxoAlign	0.650±0.007	0.694±0.006	0.480±0.009	0.883±0.005	0.305±0.008
Context-Aware	**0.713±0.005**	**0.742±0.007**	0.512±0.006	0.821±0.009	0.318±0.004
EvoTree (Pre-train)	0.591±0.008	0.689±0.009	0.494±0.011	0.829±0.012	0.284±0.013
EvoTree	0.653±0.007	0.703±0.008	**0.526±0.009**	**0.893±0.011**	0.312±0.012

Table 2: Evaluation on the few-shot benchmark. Results are mean ± standard deviation over 5 runs. Best results are in **bold**.

Method	FS-CP ↑	FS-EDA ↑	FS-Mar ↑
HAC-SPECTER2	0.600±0.011	0.838±0.013	–
CitRank+HAC	0.739±0.010	1.000	–
TaxoGen-style	0.661±0.011	0.929±0.012	–
LLM (title)	0.767±0.009	0.964±0.010	–
LLM (title + year)	0.759±0.010	0.937±0.011	–
LLM (citation graphs)	0.769±0.009	0.952±0.010	–
EvoTree (Pre-train)	0.653±0.011	0.904±0.012	0.662±0.013
EvoTree	**0.775±0.008**	0.949±0.010	**0.726±0.011**

Both later stages deliver substantial gains in leaf purity and concept-level alignment (FS-CP $+0.104$ after post-training and a further $+0.118$ after FS-correction), consistent with the self-supervised evolutionary losses injecting temporal structure that label supervision cannot capture, and with the few-shot priors sharpening concept boundaries. Post-training yields the largest gain in marginal-paper recognition (FS-Mar $0.662 \rightarrow 0.750$), most of which FS-correction preserves (0.726). These gains also transfer to test graphs unseen during FS-correction training, mitigating concerns about overfitting to the reference trees. The dip in FS-EDA after post-training reflects the tension between self-supervised temporal refinement and strict depth ordering; FS-correction recovers most of it ($0.880 \rightarrow 0.949$).

5.4.2 Necessity of Graph Structure

To verify that the GNN exploits the citation graph rather than merely propagating SPECTER2 embeddings, we construct **Pre-train (MLP)** — identical to **Pre-train (GNN)** in hyperparameters and training objective but with the citation edge set replaced by an empty set, so any gap is attributable solely to the citation graph.

Removing the citation graph causes an even larger gap in validation score than at inference, indicating that the graph contributes signal throughout optimization. The MLP variant's slightly higher PMR is a degenerate artifact: without citation constraints, tree construction relies on the temporal distribution of embeddings alone, trivially producing monotonic paths — mirroring the CDA artifact for HAC-4level in §5.2. Together with the training-stage ablation (§5.4.1), these results confirm that EvoTree relies jointly on the citation graph as a structural backbone and on the staged training framework that progressively injects supervised, self-supervised, and correction signals.

5.5 Generation Quality

The metrics in Tables 1 and 2 measure agreement with reference partitions and citation ordering, but not whether a structure reads as a coherent account of how a field developed. We therefore evaluate the generated structures directly under two independent protocols: an LLM judge, instantiated with the same model as concept labeling (§4.6) and run multiple times per structure, and three CS

Table 3: Ablation study on the contribution of each training stage. Left: clustering and evolution-structure metrics on test graphs; right: FS-aligned metrics on the 11 annotated domains. Results are mean ± standard deviation over 5 runs. Best results are in **bold**.

Variant	Clustering			Evolution		FS-aligned		
	LP ↑	PP ↑	NMI ↑	CDA ↑	PMR ↑	FS-CP ↑	FS-EDA ↑	FS-Mar ↑
Pre-train only	0.541±0.010	0.629±0.011	**0.530±0.011**	0.829±0.013	0.240±0.014	0.553±0.011	**0.954±0.012**	0.662±0.013
+ Post-train	0.598±0.009	0.665±0.009	0.455±0.011	0.857±0.012	0.270±0.013	0.657±0.010	0.880±0.012	**0.750±0.012**
+ FS-correction	**0.653±0.007**	**0.703±0.008**	0.526±0.009	**0.893±0.011**	**0.312±0.012**	**0.775±0.008**	0.949±0.010	0.726±0.011

Table 4: Ablation study on the necessity of the citation graph. "Val. score" is the validation score during training. Best in **bold**. Mean ± std over 5 runs.

Variant	LP ↑	PP ↑	NMI ↑	PMR ↑	Val. score ↑
Pre-train (MLP, no graph)	0.450±0.012	0.536±0.013	0.363±0.013	**0.306±0.013**	0.514±0.014
Pre-train (GNN, with graph)	**0.541±0.010**	**0.629±0.011**	**0.464±0.011**	0.240±0.014	**0.780±0.010**

Table 5: Generation-quality evaluation on the FS domains under two independent protocols, LLM-as-a-Judge and Human evaluation. Scores are on a 1–10 scale (mean ± standard deviation over LLM runs or human evaluators). Best results are in **bold**.

	LLM-as-a-Judge				Human evaluation			
Method	Concept	Evolution	Transitional	Overall	Concept	Evolution	Transitional	Overall
CitRank+HAC	4.21±0.12	3.57±0.14	2.82±0.06	4.08±0.04	4.77±0.07	4.38±0.05	4.90±0.09	4.21±0.09
Hu-CiteTaxo	4.55±0.11	4.73±0.07	3.57±0.13	4.64±0.14	6.65±0.15	5.87±0.10	5.82±0.01	6.19±0.05
TaxoAlign	5.44±0.05	4.89±0.11	4.22±0.05	4.67±0.06	5.94±0.05	5.61±0.08	5.26±0.05	6.44±0.04
Context-Aware	5.60±0.14	5.20±0.04	4.40±0.07	5.10±0.12	6.81±0.04	6.32±0.14	5.97±0.01	6.67±0.09
TaxoGen-style	**7.14±0.05**	5.66±0.11	4.97±0.04	6.27±0.12	**8.64±0.05**	6.18±0.03	5.93±0.13	7.40±0.12
EvoTree	6.03±0.13	**7.51±0.04**	**6.85±0.10**	**6.82±0.05**	7.16±0.08	**8.65±0.15**	**6.47±0.11**	**8.05±0.03**

graduate students distinct from the FS annotators, scoring independently. Both apply the same four criteria—conceptual organization, scientific evolution, transitional-paper placement, and overall quality—to anonymized structures in randomized order. Protocol details are in Appendix N.

Both protocols rank EvoTree first on overall quality and produce the same overall ordering of all six methods (Table 5). EvoTree leads on scientific evolution and transitional-paper placement, the two dimensions taxonomy induction does not target, while TaxoGen-style leads on conceptual organization. The latter follows from temporal admissibility: when a method family develops over a long span, enforcing Eq. 2 separates papers that are semantically adjacent but temporally distant, at some cost to leaf-level compactness—the same trade-off behind EvoTree's lower LP and PP. The rubric is generic across hierarchy-generation methods and does not reward EvoTree-specific mechanisms such as temporal constraints or internal-node re-attachment.

6 Conclusion

We introduced EvoTree, the first framework for automatically inducing scientific evolution trees from citation graphs, and formalized the task with an annotated benchmark covering 11 AI subfields. By decoupling conceptual backbone learning from temporal refinement, EvoTree first constructs a stable taxonomy via graph-aware encoding and distribution-based hierarchical clustering, then reattaches marginal papers to internal nodes under monotonic-path constraints, and finally invokes an LLM solely for concept verbalization without altering topology. Against taxonomy-induction, citation-only, and LLM-based baselines, EvoTree attains the best NMI, citation-direction accuracy, and annotated concept purity, is preferred by both LLM and human judges on overall quality, and is the only method that recovers transitional papers with non-trivial accuracy; ablations confirm that both the citation graph and the staged training framework are jointly necessary. We view evolution-tree induction as a step toward scalable, temporally faithful organization of scientific literature.

Limitations

Scope of domains. Our 411 ego-graphs and 11 reference trees are drawn from AI surveys. We have not verified transfer to disciplines whose citation conventions and paradigm cadences differ substantially from AI.

Scale of the few-shot reference set. The FS dataset comprises 11 curated reference trees and 352 papers, sufficient to demonstrate the few-shot regime but limiting fine-grained per-domain analyses; a larger curated set would strengthen alignment evaluation.

Reliance on survey section structure. Stage I assumes that section organization reflects a coherent conceptual taxonomy. Surveys organized by application area or chronology may provide weaker supervision than those organized by methodology.

Temporal signal noise. Publication years conflate arXiv preprint, conference, and journal dates, introducing noise into the temporal-admissibility constraint and PMR. This contributes to the modest absolute PMR values observed in our experiments.

External LLM dependence. Concept labels are generated by GPT-4.1-mini. Although the LLM is used only for verbalization and does not alter topology, this introduces a dependency on a closed-source model; substituting an open-source LLM is straightforward but may change labeling style.

References

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8).

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18370–18380.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, and 1 others. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Mohit Bansal, David Burkett, Gerard De Melo, and Dan Klein. 2014. Structured learning for taxonomy induction with belief propagation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1051.

Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. 2016. Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web*, page 381–390.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, page 113–120.

Ricardo JGB Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuiliere, and Jakob Verbeek. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2174–2183.

Santo Fortunato, Carl Bergstrom, Katy Borner, James Evans, Dirk Helbing, Stasa Milojevic, Alexander Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-Laszlo Barabasi. 2018. Science of science. *Science*, 359:eaao0185.

Yuntong Hu, Zhihan Lei, Zhongjie Dai, Allen Zhang, Abhinav Angirekula, Zheng Zhang, and Liang Zhao. 2025. Cg-rag: Research question answering by citation graph retrieval-augmented llms. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*, pages 678–687.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024. Taxonomy tree generation from citation graph.

Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. 2022. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *Proceedings of the ACM web conference 2022*, pages 925–934.

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. 2025.

Taxoadapt: Aligning llm-based multidimensional taxonomy construction to evolving research corpora. pages 29834–29850.

Avishek Lahiri, Yufang Hou, and Debarshi Kumar Sanyal. 2025. Taxoalign: Scholarly taxonomy generation using language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30191–30211.

Xiaoying Le, Pengfei Qian, Yuanzhao Zhai, Xu Zhang, Qian Liu, Feng Dawei, and Bo Ding. 2026. Evonarrator: Modeling scientific evolution for feasible hypothesis generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11846–11865.

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, pages 2819–2829.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Simin Niu, Shichao Song, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. Surveyx: Academic survey automation via large language models.

Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. 2024. Self-supervised topic taxonomy discovery in the box embedding space. *Transactions of the Association for Computational Linguistics*, 12:1401–1416.

Yuning Mao, Xiang Ren, Jiaming Shen, Xiaotao Gu, and Jiawei Han. 2018. End-to-end reinforcement learning for automatic taxonomy induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2472.

Sahil Mishra, Kumar Arjun, and Tanmoy Chakraborty. 2025. Rank, chunk and expand: Lineage-oriented reasoning for taxonomy expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12935–12953.

Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of the web conference 2020*, pages 486–497.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2180–2189.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Rion Snow, Dan Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 801–808.

Chong Wang, David Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, page 579–586.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 424–433.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: large language models can automatically write surveys. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*.

Stephen Woods. 2024. A programmatic approach to journal use and citation analysis. *portal: Libraries and the Academy*, 24:177 – 200.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. 2025. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8496–8504.

Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. 2024. Guardt2i: Defending text-to-image models from adversarial prompts. *Advances in neural information processing systems*, 37:76380–76403.

Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. 2020. Steam: Self-supervised taxonomy expansion with mini-paths. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1026–1035.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2701–2709.

Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. 2025. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 114:102701.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. 2024. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776.

Lingxiao Zhao and Leman Akoglu. 2020. Pairnorm: Tackling oversmoothing in {gnn}s. In *International Conference on Learning Representations*.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via LLM-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15616–15634.

A Case Study: Evolution Tree on Adversarial Attacks on Text-to-Image Diffusion Models

A.1 Setup

We instantiate EvoTree on the reference corpus of *Adversarial Attacks and Defenses on Text-to-Image Diffusion Models: A Survey* (Zhang et al., 2025)[1], which serves as both the ego node and the topical anchor of the resulting tree. The survey was selected because (i) it provides a recent, curated, and topically coherent snapshot of an actively evolving subfield, (ii) it spans both attack-side and defense-side work, allowing the tree to surface convergent research lines, and (iii) the ego node itself appears as paper [20] in the corpus, providing a natural reference point for the green branch. Unlike the experiments setting, we retain the survey node in this case-study visualization to anchor the visual layout and to illustrate how the ego work would be positioned by EvoTree in its own subfield.

After deduplication and minor filtering of off-topic citations, the corpus contains 61 papers published between 2021 and 2024. EvoTree organizes them into 12 leaf clusters grouped under 4 top-level branches, as visualized in Figure 2. The bracketed indices [1]–[61] refer exclusively to the corpus entries and are independent of the bibliography numbering in the main paper.

A.2 Branch and Cluster Overview

The four top-level branches partition the corpus along complementary dimensions of the research landscape:

- **Adversarial Threats and Defenses in Generative Models** (19 papers, blue) — work targeting T2I diffusion models with adversarial prompts, red-teaming protocols, and safety filter evaluation.
- **Adversarial Robustness and Defense Methods** (3 papers, green) — robustness evaluation studies and the ego survey itself.
- **Text-Conditioned Image Diffusion Models** (19 papers, red) — the foundation-model lineage on which downstream attack/defense work depends, from unconditional DDPM-class models through latent and spatially conditioned T2I systems.
- **Concept Manipulation and Optimization** (20 papers, purple) — concept erasure, model editing, and pruning techniques that intersect heavily with the safety agenda.

A.3 Analysis and Discussion

EvoTree recovers three structural properties of the subfield. (i) A clear *foundation–application* stratification: the red branch (DiT [24] (Peebles and Xie, 2023), latent diffusion [31](Rombach et al., 2022), GLIDE [32](Nichol et al., 2021), SDXL [27](Podell et al., 2024), and spatially conditioned variants [38]–[41](Couairon et al., 2023; Chefer et al., 2023; Avrahami et al., 2023; Balaji et al., 2022)) temporally and conceptually precedes the attack and editing work in the blue and purple branches. (ii) The concept-erasure cluster (entries [42]–[55](Wu et al., 2025; Zhang et al., 2024)) is the densest single thematic block, marking it as the dominant defense paradigm in this timeframe—ahead of input-side filtering or output-side detection. (iii) The blue branch is essentially confined to 2023 and later, placing adversarial-attack research on T2I (Yang et al., 2024) roughly two years behind the foundation work.

B Details of Datasets

B.1 Main Dataset

We construct 411 ego-graphs from arXiv survey papers. Each graph is centered on one survey paper: its references form the graph nodes, and citations among these references form the directed edges. The survey node itself is excluded. Each paper node carries a title, an abstract, a publication year, and a 768-dimensional SPECTER2 (Singh et al., 2022) embedding. Weak hierarchical labels are inherited from the survey's section structure: a referenced paper takes on the section path in which it appears, e.g., *Chapter 2* → *2.2* → *2.2.3*. We split at the graph level, rather than the node level, into train/validation/test sets with a 70/15/15 ratio.

Table 6: Statistics of the survey-reference graph dataset.

Statistic	Value
Graphs	411
Paper nodes	44,812
Citation edges	205,725
Avg. nodes / graph	109.0
Avg. edges / graph	500.5
Node embedding dim.	768

[1] arXiv:2407.15861.

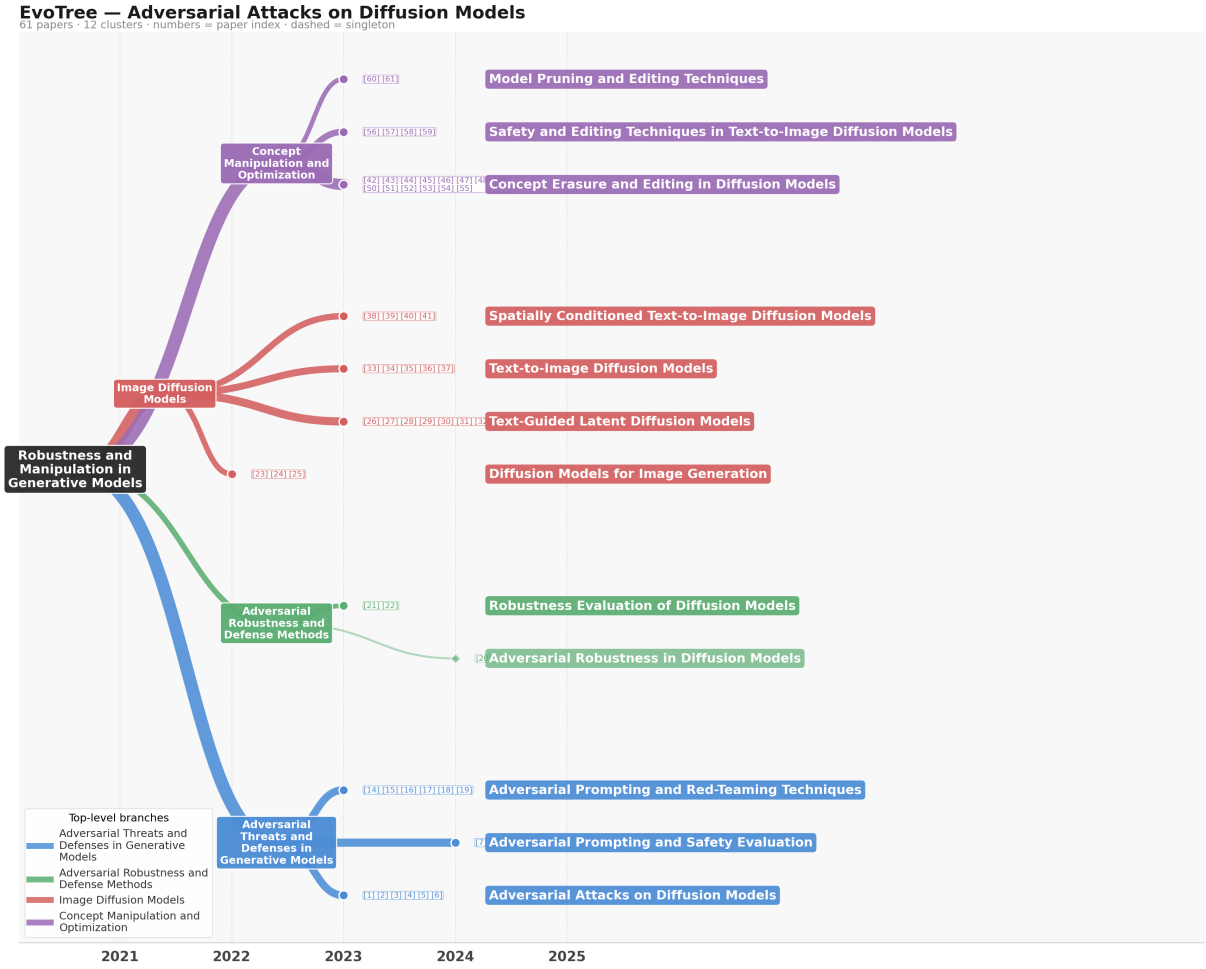


Figure 2: Evolution tree generated by EvoTree for the topic *Adversarial Attacks on Text-to-Image Diffusion Models*, instantiated on the reference corpus of the survey by Zhang et al. (2025). Cluster labels are produced by an LLM; the layout is rendered from the induced tree.

B.2 Few-shot Dataset

Table 7: Statistics of the FS dataset.

Statistic	Value
Domains	11
Paper nodes	352
FS evolution edges	90
FS marginal papers	24
Avg. nodes / domain	32.0

Annotation schema. Each domain graph contains four types of annotations: (i) `FS_concept`, which assigns each paper to a defined method concept; (ii) `FS_is_marginal`, which marks papers that lie on topic boundaries or cannot be confidently assigned to a fine-grained concept; (iii) `FS_tree_edges`, which records directed method-evolution relations, each typed as `extends`, `improves`, or (cross-domain) `adapts`, where the source paper is regarded as an extension or improvement of the target paper; and (iv) `FS_depth`, which indicates the depth of a paper in the evolution structure.

Domain coverage. The FS dataset contains 11 high-quality annotated AI subfields. We retain domains with sufficiently informative annotations for concept membership, evolution relations, and marginal-paper evaluation, and exclude domains whose evolution-edge annotations are too sparse for reliable FS-based comparison. Table 8 reports the per-domain statistics.

FS evaluation protocol. The FS benchmark participates in Stage III calibration under an 11-fold leave-one-domain-out protocol. In each fold, calibration uses the 10 remaining domains and evaluation is performed on the held-out domain; the held-out domain's concept labels, marginal-paper labels, evolution edges, and depth annotations are used neither for calibration nor for validation, threshold

Table 8: Per-domain statistics of the annotated FS dataset. N denotes the number of papers in each domain. "Concepts" denotes the number of annotated method concepts. "FS edges" denotes the number of annotated method-evolution edges. "Marginal" denotes the number of papers annotated as marginal or boundary papers.

Domain	N	**Concepts**	**FS edges**	**Marginal**
3D Vision	25	5	11	4
CV ConvNets	40	7	4	1
Vision Transformers	27	4	2	3
GAN Image Synthesis	23	6	10	3
LLM Alignment	55	5	12	1
Meta / Few-shot	27	5	13	2
NLP Transformers	35	6	3	2
Object Detection	38	6	3	1
Self-supervised Learning	29	5	11	2
Semantic Segmentation	37	5	14	2
Video Understanding	16	6	7	3
Total	**352**	**60**	**90**	**24**
Average	**32.0**	**5.5**	**8.2**	**2.2**

selection, or hyperparameter tuning. Each of the 11 domains serves once as the held-out domain; FS-aligned metrics are averaged over the 11 held-out evaluations, and we report the mean and standard deviation of this average over 5 runs.

B.3 FS Cross-domain Overlap

The leave-one-domain-out protocol assumes that the held-out domain is genuinely unseen during calibration. Because the 11 FS domains are drawn from surveys within a single broad field, this assumption could be weakened if the same papers recur across domains: a paper appearing in both the calibration and held-out sets would leak its annotation indirectly. We therefore quantify paper-level overlap across all domain pairs.

Table 9 reports the number of papers shared by each pair of domains, with domain sizes on the diagonal. Overlap is sparse throughout: the largest shared count is 3 papers, or 8.6% when normalized by the smaller domain in the pair, and most pairs share at most one or two papers. The non-negligible overlaps concentrate between semantically adjacent domains (D2/D6/D8: CV ConvNets, Object Detection, and Semantic Segmentation; D10/D11: NLP Transformers and LLM Alignment), which is expected since neighboring subfields cite a common set of foundational works. Under this level of overlap, a held-out domain's evaluation set is composed almost entirely of papers absent from the corresponding calibration folds, so the FS-aligned results are unlikely to reflect within-set memorization.

We note that overlap counts alone do not exclude all forms of information sharing—semantically similar papers may appear under different identifiers across surveys. The reported figures bound direct duplication, not conceptual proximity between domains.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
D1	**25**	1	1	0	0	1	0	1	0	0	0
D2	1	**40**	2	1	2	3	2	3	2	0	0
D3	1	2	**27**	1	1	1	2	1	1	2	1
D4	0	1	1	**23**	0	0	1	0	0	0	0
D5	0	2	1	0	**27**	1	0	0	1	0	0
D6	1	3	1	0	1	**38**	1	3	2	0	0
D7	0	2	2	1	0	1	**29**	1	0	0	0
D8	1	3	1	0	0	3	1	**37**	1	0	0
D9	0	2	1	0	1	2	0	1	**16**	0	0
D10	0	0	2	0	0	0	0	0	0	**35**	3
D11	0	0	1	0	0	0	0	0	0	3	**55**

Table 9: Pairwise paper-level overlap across the 11 FS domains. Diagonal entries (bold) give domain sizes; off-diagonal entries give the number of papers shared by the two domains.

B.4 FS Benchmark Construction and Annotation

The 11 reference evolution trees were curated from survey papers that explicitly present evolution diagrams or method-lineage organizations, rather than topic-only taxonomies. This requirement restricts the candidate pool but ensures each reference tree reflects an author-endorsed account of how the field developed, which we standardize into a unified annotation format.

Three graduate students annotated four aspects independently, following a shared guideline: concept membership (which concept node each paper belongs to), evolution edges (directed development relations between concepts), marginal-paper labels

(whether a paper is better explained at a coarser abstraction level), and paper depth. Annotators worked from paper titles, abstracts, publication years, and the citation graph, and had access to the source survey.

C Concept-Labeling Prompt Templates

This appendix provides the full prompt templates used in the bottom-up concept generation pass described in §4.6. Both templates are instantiated per node and consumed by GPT-4.1-mini at temperature 0.1.

C.1 Leaf Prompt

A leaf node represents a specific research micro-topic. The prompt supplies all paper titles with publication years and instructs the model to produce a precise 3–6 word noun phrase naming the shared method or technique, explicitly discouraging generic openers (e.g., *Advanced*, *Large*).

C.2 Internal-Node Prompt

An internal node must generalize over its children. The prompt provides the already-assigned child concept names (bottom-up guarantee) and a sample of representative papers. The model is explicitly required to produce a label that (i) is more abstract than any single child, and (ii) does not repeat any child label verbatim.

Leaf-node prompt

You are assigning a precise concept label to a leaf cluster in an academic evolutionary taxonomy. The cluster groups closely related papers that share a specific research method, task, or technique.

Papers in this cluster: `[list of titles with years]`

Note: these papers span `YYYY–YYYY`. *Reflect the temporal focus if it is distinctive.*

Requirements:

1. *Use 3–6 words; noun phrase, no verbs.*
2. *Be specific — prefer method/technique names over broad area names.*
3. *Do NOT start with 'Advanced', 'Large', or 'Modern'.*

Reply with ONLY the concept name, nothing else.

Internal-node prompt

You are assigning a concept label to an internal node in an academic evolutionary taxonomy. This node is the common parent of several sub-topics and must be named at a higher level of abstraction than its children.

Child sub-topics: `[· child_label_1, · child_label_2, . . . ]`

This subtree covers `YYYY–YYYY`.

Representative papers from this subtree (sample): `[up to 5 titles with years]`

Requirements:

1. *Use 3–6 words; noun phrase, no verbs.*
2. *The name must generalize over ALL child sub-topics listed above.*
3. *It must be more abstract than any single child concept.*
4. *Do NOT repeat a child concept name verbatim.*

Reply with ONLY the concept name, nothing else.

D Evaluation Metrics

This appendix specifies the eight evaluation metrics used in §5. We organize them into two parts: *indirect metrics* computed on citation graphs without annotation, and *few-shot metrics* computed on the 11 annotated graphs. Table 10 summarizes the role of each metric.

Throughout this appendix, $\mathcal{T}$ denotes a predicted evolutionary tree with leaf set $\mathrm{Leaves}(\mathcal{T}) = \{L_1, \ldots, L_K\}$. For each leaf L_k, $\mathcal{M}_k$ denotes the set of papers attached to L_k (i.e., $\mathcal{V}(L_k)$ restricted to direct attachment). Each paper v_i carries a publication year $\tau(v_i)$, a section-label set $\mathbf{y}_i \subseteq \mathcal{Y}$ (multi-hot weak labels from the survey vocabulary), and a section-path set $\mathbf{h}_i \subseteq \mathcal{H}$ from the pretrained section taxonomy. For papers in the annotated set $\mathcal{V}^{\mathrm{fs}}$, additional labels include a concept $g_i \in \mathcal{G}^{\mathrm{fs}}$ (single-label, deterministic), a marginal flag $y_i^{\mathrm{m,fs}} \in \{0, 1\}$, and an evolution edge set $\mathcal{E}^{\mathrm{evo}}$ over papers. We write $\delta(v_i)$ for the depth of the concept node to which v_i is attached — whether that node is a leaf or an internal node (root depth $= 0$). For methods that produce only leaf-level partitions (all baselines in this work), $\delta(v_i)$ is trivially the depth of the assigned leaf.

D.1 Indirect metrics on held-out test graphs

LP — Leaf Label Purity. The label purity of a leaf L_k is the fraction of its labeled papers covered

Table 10: Summary of evaluation metrics. "FS" indicates whether the metric requires annotation.

Metric	What it measures	Signal source	Eq.	Stage	FS?
LP	Leaf-cluster topic homogeneity	Section labels	Eq. 15	1	—
PP	Leaf-cluster path consistency	Section paths	Eq. 16	1	—
NMI	Cluster–class agreement at depth 2	Section labels	Eq. 17	1	—
CDA	Citation direction vs. tree depth	Citation graph	Eq. 18	2	—
PMR	Whole-path temporal monotonicity	Publication years	Eq. 20	2	—
FS-CP	Leaf vs. concept	Concepts	Eq. 21	3	✓
FS-EDA	Evolution direction vs. tree depth	Evolution edges	Eq. 22	3	✓
FS-Mar	Marginal-paper detection (AUROC)	Marginal flags	Eq. 23	3	✓

by its single most frequent section label:

$$\mathrm{Purity}(L_k) = \frac{\max_{\ell \in \mathcal{Y}} \left|\{v_i \in \mathcal{M}_k : \ell \in \mathbf{y}_i\}\right|}{\left|\{v_i \in \mathcal{M}_k : \mathbf{y}_i \neq \emptyset\}\right|}. \quad (14)$$

LP averages this across leaves:

$$\mathrm{LP} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{Purity}(L_k). \quad (15)$$

LP measures topic homogeneity within each leaf cluster.

PP — Leaf Path Purity. PP applies the same definition but operates over section-path labels $\mathbf{h}_i$ instead of label sets $\mathbf{y}_i$:

$$\mathrm{PP} = \frac{1}{K} \sum_{k=1}^{K} \frac{\max_{h \in \mathcal{H}} \left|\{v_i \in \mathcal{M}_k : h \in \mathbf{h}_i\}\right|}{\left|\{v_i \in \mathcal{M}_k : \mathbf{h}_i \neq \emptyset\}\right|}. \quad (16)$$

PP is stricter than LP: papers grouped in the same leaf must agree not only on labels but on the full hierarchical path leading to them in the survey taxonomy.

NMI — Normalized Mutual Information. To compare predicted trees against flat-clustering baselines under equal granularity, we truncate $\mathcal{T}$ at depth 2, obtaining a per-paper cluster assignment $\hat{c}_i \in \{1, \ldots, K'\}$. The reference class for paper v_i is its lexicographically smallest label $y_i = \min(\mathbf{y}_i)$. Restricting to papers with $\mathbf{y}_i \neq \emptyset$ and $\hat{c}_i$ defined,

$$\mathrm{NMI} = \frac{2\, I(\hat{C}; Y)}{H(\hat{C}) + H(Y)}, \quad (17)$$

where $I(\cdot;\cdot)$ is mutual information and $H(\cdot)$ is entropy, following the arithmetic-mean normalization in scikit-learn.

CDA — Citation Direction Accuracy. For a citation edge $(v_s, v_d) \in \mathcal{E}_\mathcal{G}$ in which v_s cites v_d, v_s is typically the more recent work building on v_d. We expect v_s to attach at a depth no shallower than v_d. To suppress citations within the same year (which carry weak directional information), we restrict to edges with year gap $\tau(v_s) - \tau(v_d) \geq 1$:

$$\mathrm{CDA} = \frac{\left|\{(v_s, v_d) \in \mathcal{E}_\mathcal{G} : \delta(v_s) \geq \delta(v_d),\ \tau(v_s) - \tau(v_d) \geq 1\}\right|}{\left|\{(v_s, v_d) \in \mathcal{E}_\mathcal{G} : \tau(v_s) - \tau(v_d) \geq 1\}\right|}. \quad (18)$$

CDA assesses whether the predicted depth structure aligns with the directionality of citations.

PMR — Path Monotonicity Rate. PMR evaluates whether root-to-leaf paths reflect a temporally forward-progressing trajectory at the level of the entire concept body, including descendants. For each concept node c, we use the *subtree-mean time* $\bar{\tau}_\mathcal{T}(c) = \frac{1}{|\mathcal{V}(c)|} \sum_{v_i \in \mathcal{V}(c)} \tau(v_i)$ as the path statistic (Eq. 1). Let Π be the set of root-to-leaf paths in $\mathcal{T}$ containing at least two nodes with valid $\bar{\tau}$. A path $\pi = (c_0, c_1, \ldots, c_L) \in \Pi$ is temporally monotone if

$$\bar{\tau}_\mathcal{T}(c_0) \leq \bar{\tau}_\mathcal{T}(c_1) \leq \cdots \leq \bar{\tau}_\mathcal{T}(c_L). \quad (19)$$

PMR is the fraction of monotone paths:

$$\mathrm{PMR} = \frac{\left|\{\pi \in \Pi : \pi \text{ is temporally monotone}\}\right|}{|\Pi|}. \quad (20)$$

D.2 Few-shot metrics on annotated graphs

FS-CP — FS Concept Purity. FS-CP replicates LP using single-label concepts g_i in place of multi-label section labels:

$$\mathrm{FS\text{-}CP} = \frac{1}{K} \sum_{k=1}^{K} \frac{\max_{g \in \mathcal{G}^{\mathrm{fs}}} \left|\{v_i \in \mathcal{M}_k : g_i = g\}\right|}{|\mathcal{M}_k|}. \quad (21)$$

FS-CP directly measures how well the predicted leaf clusters align with research concepts and is the primary structural-alignment metric in our evaluation.

FS-EDA — FS Evolutionary Direction Accuracy. FS-EDA mirrors CDA but uses annotated evolution edges $\mathcal{E}^{\text{evo}}$, in which (v_s, v_d) indicates that v_s is an evolutionary descendant (typically more recent) of v_d:

$$\text{FS-EDA} = \frac{\left|\{(v_s, v_d) \in \mathcal{E}^{\text{evo}} : \delta(v_s) \geq \delta(v_d)\}\right|}{|\mathcal{E}^{\text{evo}}|}. \tag{22}$$

Compared with CDA, FS-EDA replaces the noisy citation graph with a clean signal: an edge in $\mathcal{E}^{\text{evo}}$ reflects "v_s extends/improves v_d", whereas a citation in $\mathcal{E}_{\mathcal{G}}$ may indicate any of comparison, background, or methodological reuse.

FS-Mar — FS Marginal AUROC. Our model produces a per-paper marginality score $h_i \in [0, 1]$ (the marginal-head output from Eq. 53). With $y_i^{\text{m,fs}} \in \{0, 1\}$ as ground truth,

$$\text{FS-Mar} = \frac{\left|\{(v_i, v_j) : y_i^{\text{m,fs}} = 1,\ y_j^{\text{m,fs}} = 0,\ h_i > h_j\}\right|}{\left|\{(v_i, v_j) : y_i^{\text{m,fs}} = 1,\ y_j^{\text{m,fs}} = 0\}\right|}, \tag{23}$$

which equals the standard AUROC of h against $y^{\text{m,fs}}$. Marginal papers are those occupying domain boundaries: relevant by topic, but whose research focus deviates from the core methodological lineage. FS-Mar evaluates a capability unique to our framework—neither standard clustering baselines nor LLM baselines produce a marginality score (reported as "–" in Table 2; a constant score would give the trivial AUROC of 0.500).

E Notation Summary

Table 11 summarizes the notation used throughout the paper. Symbols are grouped by category for ease of reference.

F Graph-Aware Paper Encoder

This appendix expands the abstract encoder f_θ introduced in §4.2.1 into concrete forward computations and justifies the three design choices discussed in the main text.

F.1 Forward computation

A pretrained scientific encoder (SPECTER2 (Singh et al., 2022)) produces a fixed semantic representation

$$\mathbf{x}_i = f_{\text{LM}}(s_i) \in \mathbb{R}^{d_{\text{LM}}}, \tag{24}$$

which is then passed through two parallel branches.

Semantic branch. A lightweight two-layer MLP produces the semantic stream

$$\mathbf{z}_i^{\text{sem}} = \mathbf{W}_2\, \text{GELU}(\mathbf{W}_1 \mathbf{x}_i), \tag{25}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{\text{LM}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times d}$.

Graph branch. A stack of L controlled propagation blocks computes the graph stream. With $\mathbf{h}_i^{(0)} = \text{MLP}_{\text{graph}}(\mathbf{x}_i)$, each layer $l = 0, \ldots, L-1$ updates

$$\mathbf{m}_i^{(l)} = \frac{1}{|\mathcal{N}_{\mathcal{G}}(v_i)|} \sum_{j \in \mathcal{N}_{\mathcal{G}}(v_i)} \mathbf{W}_{\text{msg}}\, \mathbf{h}_j^{(l)}, \tag{26}$$

$$\mathbf{u}_i^{(l)} = \beta\, \mathbf{h}_i^{(l)} + (1 - \beta)\, \mathbf{m}_i^{(l)}, \tag{27}$$

$$\mathbf{h}_i^{(l+1)} = \text{PairNorm}\Big(\alpha\, \mathbf{x}_i + (1 - \alpha)\, \mathbf{u}_i^{(l)}\Big). \tag{28}$$

For isolated nodes $v_i \in \mathcal{V}_o$ where $|\mathcal{N}_{\mathcal{G}}(v_i)| = 0$, neighborhood aggregation is undefined; we set $\mathbf{m}_i^{(l)} = \mathbf{0}$ so that the propagation reduces to $\mathbf{h}_i^{(l+1)} = \text{PairNorm}(\alpha \mathbf{x}_i + (1 - \alpha)\beta \mathbf{h}_i^{(l)})$, i.e., a pure semantic residual. The final graph-stream output is $\mathbf{z}_i^{\text{graph}} = \mathbf{h}_i^{(L)}$.

Adaptive fusion gate. The two streams are combined through a learnable connectivity-biased gate:

$$\lambda_i = \sigma\Big(\mathbf{W}_g[\mathbf{z}_i^{\text{graph}} \,\|\, \mathbf{z}_i^{\text{sem}}] + \delta\, \mathbb{1}[\text{iso}(v_i)]\Big), \tag{29}$$

$$\mathbf{z}_i^{\text{fused}} = \lambda_i\, \mathbf{z}_i^{\text{graph}} + (1 - \lambda_i)\, \mathbf{z}_i^{\text{sem}}. \tag{30}$$

The clustering embedding is $\mathbf{e}_i = \text{L}_2\text{Norm}(\mathbf{z}_i^{\text{fused}})$. An independent two-layer MLP produces the projection representation $\mathbf{p}_i = \text{MLP}_{\text{proj}}(\mathbf{z}_i^{\text{fused}})$, used exclusively by the contrastive objective in §4.3 and never as input to clustering or downstream attachment.

The three design choices flagged in the main text correspond to specific equations above:

Table 11: Notation summary. Symbols introduced in the problem formulation (§3) are used consistently across the Methodology and Appendix.

Symbol	Meaning	Introduced
Input citation graph		
$\mathcal{G} = (\mathcal{V}, \mathcal{E}_{\mathcal{G}})$	Citation graph from survey reference set	§3
$v_i \in \mathcal{V}$	A paper node	§3
s_i	Title and abstract text of paper v_i	§3
$\tau(v_i)$	Publication year of paper v_i	§3
$\mathcal{N}_{\mathcal{G}}(v_i)$	Citation neighborhood of v_i	App. F
Trees and concepts		
$\mathcal{T}^{\text{tax}}$	Time-agnostic taxonomy backbone	§3
$\mathcal{T}$	Target evolutionary tree	§3
$\mathcal{T}^{(t)}, \mathcal{C}^{(t)}$	Tree and concept set at EM iteration t	§4.4.1
$c \in \mathcal{C}$	Concept node (also indexes a cluster)	§3
r	Root concept node	§4.4.2
$\mathcal{E}_C, \mathcal{E}_P$	Concept edges, paper-attachment edges	§3
$\mathcal{V}(c)$	Papers attached to c or its descendants	§3
$\text{children}(c)$	Direct children of c in $\mathcal{T}$	§4.6
$\text{Leaves}(\mathcal{T})$	Leaf concept nodes of $\mathcal{T}$	§4.4.3
$\phi(c)$	Concept label of c	§3
$\tau^{\text{att}}(c), \bar{\tau}(c)$	Attached-mean / subtree-mean publication year of c	§3
ϵ	Temporal tolerance margin	§3
Representations		
$\mathbf{x}_i$	Pretrained LM embedding of s_i	§4.2.1
$\mathbf{z}_i^{\text{sem}}, \mathbf{z}_i^{\text{graph}} (= \mathbf{h}_i^{(L)})$	Semantic / graph stream output	§4.2.1
$\mathbf{e}_i$	Fused clustering embedding	§4.2.1
$\mathbf{p}_i$	Projection representation (contrastive only)	§4.2.1
$\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2$	Diagonal Gaussian parameters of concept c	§4.2.2
w_{ic}	Soft membership $p(v_i \in C_c)$	App. G
Marginal mechanism		
$\mathcal{M}$	Marginal paper set	§3
$m(v_i, c)$	Paper-to-concept compatibility score	§3
A_i^{leaf}	Maximum leaf compatibility for v_i	§4.4.3
η	Marginal threshold	§3
$s_{\text{sem}}, s_{\text{graph}}, s_{\text{time}}$	Compatibility-score components	App. K
$\gamma_1, \gamma_2, \gamma_3$	Compatibility-score weights	App. K
Constraints and constants		
ρ	Major-branch minimum size ratio	§4.4.2
$\text{MBC}(\mathcal{T})$	Major Branch Count of $\mathcal{T}$	§4.4.2
$[d_{\min}, d_{\max}]$	Drift-window bounds	App. I

- **Anchored residual:** the $\alpha \mathbf{x}_i$ term in Eq. 28 preserves the original semantic vector at every layer, preventing the propagation from collapsing distinct papers to similar representations in shallow citation graphs.
- **Density-aware normalization:** PairNorm in Eq. 28 centers features and rescales pairwise distances, accommodating the highly uneven cluster densities induced by hub papers.
- **Connectivity-biased fusion gate:** the term $\delta \mathbb{1}[\text{iso}(v_i)]$ in Eq. 29, with δ initialized to -2.0, drives λ_i toward 0 for isolated nodes so that $\mathbf{z}_i^{\text{fused}} \approx \mathbf{z}_i^{\text{sem}}$ at initialization, while remaining learnable.

G Distributional Tree Builder

G.1 Base clustering with soft membership

HDBSCAN is applied to the L2-normalized embeddings $\{\mathbf{e}_i\}$ to identify dense regions. To obtain soft membership, we use the algorithm's native extension based on the mutual reachability tree (Campello et al., 2013), which yields

$$w_{ic} = p(v_i \in C_c) \in [0, 1], \quad c = 1, \ldots, K, \quad (31)$$

where K is the number of discovered base clusters. Each cluster's diagonal Gaussian is fitted by

weighted maximum likelihood:

$$\boldsymbol{\mu}_c = \frac{\sum_i w_{ic}\,\mathbf{e}_i}{\sum_i w_{ic}}, \tag{32}$$

$$\sigma^2_{c,d} = \max\left(\frac{\sum_i w_{ic}\,(e_{i,d}-\mu_{c,d})^2}{\sum_i w_{ic}},\ \sigma^2_{\min}\right), \tag{33}$$

where d indexes embedding dimensions and $\sigma^2_{\min}$ is a numerical floor preventing degenerate Gaussians.

G.2 2-Wasserstein distance between diagonal Gaussians

For two diagonal Gaussians $C_p \sim \mathcal{N}(\boldsymbol{\mu}_p, \mathrm{diag}(\boldsymbol{\sigma}_p^2))$ and $C_q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathrm{diag}(\boldsymbol{\sigma}_q^2))$, the squared 2-Wasserstein distance has a closed form:

$$\mathcal{W}_2^2(C_p, C_q) = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\|_2^2 + \sum_d (\sigma_{p,d} - \sigma_{q,d})^2. \tag{34}$$

G.3 Bottom-up agglomeration

Starting from the K base clusters, we greedily merge the pair $(C_p^\star, C_q^\star) = \arg\min_{p\neq q} \mathcal{W}_2(C_p, C_q)$ at each step. The merged parent's Gaussian is re-estimated from the union of member papers $M = \mathrm{members}(C_p^\star) \cup \mathrm{members}(C_q^\star)$ via unweighted maximum likelihood:

$$\boldsymbol{\mu}_{\mathrm{parent}} = \frac{1}{|M|}\sum_{i\in M} \mathbf{e}_i,$$

$$\sigma^2_{\mathrm{parent},d} = \max\left(\frac{1}{|M|}\sum_{i\in M}(e_{i,d} - \mu_{\mathrm{parent},d})^2,\ \sigma^2_{\min}\right). \tag{35}$$

Re-estimation relies only on running sums $\sum_{i\in M}\mathbf{e}_i$ and sum of squares $\sum_{i\in M}\mathbf{e}_i^{\odot 2}$, which can be incrementally maintained as sufficient statistics during agglomeration. The procedure terminates when a single root cluster remains. Algorithm 1 summarizes the full builder.

H Stage I Loss Details

This appendix expands the three pre-training loss groups summarized in §4.3:

$$\mathcal{L}^{\mathrm{tax}} = \mathcal{L}_{\mathrm{sec}} + \mathcal{L}_{\mathrm{cite}} + \mathcal{L}_{\mathrm{geo}}.$$

Algorithm 1 Distributional tree builder.

Require: Embeddings $\{\mathbf{e}_i\}$; HDBSCAN parameters; (optional) temporal admissibility filter $\mathcal{F}_\epsilon$.
1: $\{w_{ic}\}, K \leftarrow$ HDBSCAN_SoftMembership($\{\mathbf{e}_i\}$)
2: **for** $c = 1, \ldots, K$ **do**
3: Compute $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2$ via Eq. 32–33.
4: **end for**
5: $\mathcal{C} \leftarrow \{C_1, \ldots, C_K\}$; record member papers per cluster.
6: **while** $|\mathcal{C}| > 1$ **do**
7: Compute $\mathcal{W}_2(C_p, C_q)$ for all pairs (Eq. 34).
8: Select $(C_p^\star, C_q^\star) \leftarrow \arg\min \mathcal{W}_2$ **satisfying** $\mathcal{F}_\epsilon$ **if provided**.
9: Re-estimate parent via Eq. 35; remove children, add parent.
10: **end while**
11: **return** Hierarchical tree with $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$ at every node.

H.1 Section supervision $\mathcal{L}_{\mathrm{sec}}$

Survey section structure supplies weak labels: each paper v_i inherits a multi-hot section vector $\mathbf{y}_i^{\mathrm{lbl}} \in \{0,1\}^{|\mathcal{S}|}$ over the union of section labels $\mathcal{S}$, plus a hierarchical level $y_i^{\mathrm{lvl}} \in \{1, \ldots, L_{\max}\}$. Let $\hat{\mathbf{y}}_i^{\mathrm{lbl}} = \mathrm{sigmoid}(\mathbf{W}_{\mathrm{lbl}}\mathbf{e}_i)$ and $\hat{\mathbf{y}}_i^{\mathrm{lvl}} = \mathrm{softmax}(\mathbf{W}_{\mathrm{lvl}}\mathbf{e}_i)$. We use masked binary cross-entropy and categorical cross-entropy over labeled nodes $\mathcal{V}_{\mathrm{lbl}} \subseteq \mathcal{V}$:

$$\mathcal{L}_{\mathrm{label}} = -\frac{1}{|\mathcal{V}_{\mathrm{lbl}}|}\sum_{i\in\mathcal{V}_{\mathrm{lbl}}} \mathrm{BCE}\left(\hat{\mathbf{y}}_i^{\mathrm{lbl}}, \mathbf{y}_i^{\mathrm{lbl}}\right), \tag{36}$$

$$\mathcal{L}_{\mathrm{level}} = -\frac{1}{|\mathcal{V}_{\mathrm{lbl}}|}\sum_{i\in\mathcal{V}_{\mathrm{lbl}}} \log \hat{y}^{\mathrm{lvl}}_{i,y_i^{\mathrm{lvl}}}, \tag{37}$$

$$\mathcal{L}_{\mathrm{sec}} = w_{\mathrm{lbl}}\mathcal{L}_{\mathrm{label}} + w_{\mathrm{lvl}}\mathcal{L}_{\mathrm{level}}. \tag{38}$$

H.2 Citation consistency $\mathcal{L}_{\mathrm{cite}}$

Two terms preserve citation structure: link prediction on $\mathbf{e}_i$ and contrastive ranking on $\mathbf{p}_i$.

Link prediction. For each observed edge $(v_i, v_j) \in \mathcal{E}_{\mathcal{G}}$ (positive) and randomly sampled non-edge (v_i, v_{j^-}) (negative), define the logit $\hat{y}_{ij} = \tau_{\mathrm{lnk}}\,\mathbf{e}_i^\top \mathbf{e}_j$ with temperature τ_{lnk} (we use 10 to scale cosine similarities into a useful logit range):

$$\mathcal{L}_{\mathrm{link}} = \mathrm{BCE}(\mathrm{sigmoid}(\hat{y}_{ij}),\, y_{ij}), \tag{39}$$

where $y_{ij} \in \{0,1\}$ indicates edge existence. We use one negative per positive.

Contrastive ranking on $\mathbf{p}_i$. For two papers v_i, v_j sharing a section label, they form a positive pair. Hard negatives v_{k^-} are mined as the most similar paper to v_i not sharing any label of v_i. With cosine similarity $\mathrm{sim}(a,b) = \mathbf{p}_a^\top \mathbf{p}_b / (\|\mathbf{p}_a\|\|\mathbf{p}_b\|)$

and temperature τ_{ctr}:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(v_i, v_j)/\tau_{\text{ctr}})}{\sum_{k\in\{j\}\cup\mathcal{K}_i^-} \exp(\text{sim}(v_i, v_k)/\tau_{\text{ctr}})}, \tag{40}$$

with $\mathcal{K}_i^-$ the hard-negative set ($|\mathcal{K}_i^-| = 8$). Note that $\mathcal{L}_{\text{contrast}}$ acts on $\mathbf{p}_i$ and never on $\mathbf{e}_i$, keeping the clustering geometry untouched.

Combination.

$$\mathcal{L}_{\text{cite}} = w_{\text{lnk}}\mathcal{L}_{\text{link}} + w_{\text{ctr}}\mathcal{L}_{\text{contrast}}. \tag{41}$$

H.3 Cluster geometry $\mathcal{L}_{\text{geo}}$

A prototype-anchoring term pulls each labeled node toward its class centroid. For each label $s \in \mathcal{S}$, let $\mathcal{V}_s = \{v_i : s \in \mathbf{y}_i^{\text{lbl}}\}$ and define the stop-gradient centroid

$$\bar{\mathbf{e}}_s = \text{sg}\left(\frac{1}{|\mathcal{V}_s|}\sum_{i\in\mathcal{V}_s}\mathbf{e}_i\right). \tag{42}$$

The anchor loss is:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{1}{|\mathcal{V}_s|}\sum_{i\in\mathcal{V}_s}\Big(1 - \mathbf{e}_i^\top \bar{\mathbf{e}}_s\Big). \tag{43}$$

$\mathcal{L}_{\text{geo}} = w_{\text{anc}}\mathcal{L}_{\text{anchor}}$.

I Stage II Loss Details

This appendix expands the three evolution loss groups summarized in §4.4.1:

$$\mathcal{L}^{\text{evo}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{marg}}.$$

All terms operate on the current tree $\mathcal{T}^{(t)}$ with concept Gaussians $\{(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)\}_{c\in\mathcal{C}^{(t)}}$. Let $p(c)$ denote the parent of concept c and $p(i)$ the parent concept of paper v_i under attachment a.

I.1 Semantic evolution consistency $\mathcal{L}_{\text{sem}}$

Three sub-terms enforce parent–child semantic coherence.

Node-to-parent pull. Each paper is pulled toward the centroid of its parent concept:

$$\mathcal{L}_{\text{sem-n}} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\Big(1 - \cos(\mathbf{e}_i, \boldsymbol{\mu}_{p(i)})\Big). \tag{44}$$

Cluster-to-parent pull. Each concept is pulled toward the centroid of its parent concept:

$$\mathcal{L}_{\text{sem-c}} = \frac{1}{|\mathcal{C}^{(t)} \setminus \{r\}|}\sum_{c\in\mathcal{C}^{(t)}\setminus\{r\}}\Big(1 - \cos(\boldsymbol{\mu}_c, \boldsymbol{\mu}_{p(c)})\Big). \tag{45}$$

Drift window. Let $\delta_c = \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{p(c)}\|_2$. A two-sided hinge confines δ_c to $[d_{\min}, d_{\max}]$:

$$\mathcal{L}_{\text{drift}} = \sum_{c\in\mathcal{C}^{(t)}\setminus\{r\}}\big[\max(0, d_{\min} - \delta_c) + \max(0, \delta_c - d_{\max})\big]. \tag{46}$$

Combination.

$$\mathcal{L}_{\text{sem}} = w_{\text{semn}}\mathcal{L}_{\text{sem-n}} + w_{\text{semc}}\mathcal{L}_{\text{sem-c}} + w_{\text{drift}}\mathcal{L}_{\text{drift}}. \tag{47}$$

I.2 Temporal coherence $\mathcal{L}_{\text{temp}}$

Three sub-terms enforce forward-progressing evolution.

Path smoothness. For a root-to-leaf concept path $P = (c_0, c_1, \ldots, c_L)$, define velocity $\Delta_l = \boldsymbol{\mu}_{c_l} - \boldsymbol{\mu}_{c_{l-1}}$. We penalize second-order changes:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}}\frac{1}{L-1}\sum_{l=2}^{L}\|\Delta_l - \Delta_{l-1}\|_2^2, \tag{48}$$

where $\mathcal{P}$ is the set of root-to-leaf paths in $\mathcal{T}^{(t)}$.

Path monotonicity. Let $\tau_T(c) = \min_{v\in\mathcal{V}(c)}\tau(v)$ denote the earliest publication year in the subtree of c (also used in Appendix J). A hinge encourages child concepts to be no earlier than parents:

$$\mathcal{L}_{\text{path-mono}} = \sum_{(c_p,c_c)\in\mathcal{E}_C}\max(0,\ \tau_T(c_p) - \tau_T(c_c) - \epsilon). \tag{49}$$

Citation directionality. For each citation $(v_i, v_j) \in \mathcal{E}_\mathcal{G}$ with v_i citing v_j, the citing paper should attach to a deeper node. Let $q_{i,c}$ be the soft attachment probability of v_i to concept c:

$$q_{i,c} = \frac{\exp\big(-\|\mathbf{e}_i - \boldsymbol{\mu}_c\|_2^2/\tau_{\text{att}}\big)}{\sum_{c'\in\mathcal{C}^{(t)}}\exp\big(-\|\mathbf{e}_i - \boldsymbol{\mu}_{c'}\|_2^2/\tau_{\text{att}}\big)}. \tag{50}$$

Define the expected depth $\bar{d}(v_i) = \sum_c q_{i,c}\,\text{depth}(c)$. The citation directionality loss is:

$$\mathcal{L}_{\text{cite-dir}} = \sum_{(v_i,v_j)\in\mathcal{E}_\mathcal{G}}\max(0,\ \bar{d}(v_j) - \bar{d}(v_i) + \xi), \tag{51}$$

penalizing cases where the cited paper v_j ends up at least ξ levels deeper than the citing v_i.

Combination.

$$\mathcal{L}_{\text{temp}} = \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{path-mono}} + \mathcal{L}_{\text{cite-dir}}. \tag{52}$$

I.3 Marginal identification and re-attachment $\mathcal{L}_{\text{marg}}$

Let $h_i = \text{MLP}_{\text{mar}}(\mathbf{e}_i) \in [0,1]$ be the marginal-head output. Pseudo-labels are $\tilde{y}_i^{\text{m}} = \mathbb{1}[A_i^{\text{leaf}} < \eta]$.

Marginal classification.

$$\mathcal{L}_{\text{mar}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\big[\tilde{y}_i^{\text{m}}\log h_i + (1-\tilde{y}_i^{\text{m}})\log(1-h_i)\big]. \quad (53)$$

Attachment ranking. For each flagged marginal paper $v_m \in \mathcal{M}^{(t)}$, let $c_m^+ = a(v_m)$ (the optimal attachment from Eq. 11) and c^- a candidate concept that fails the temporal admissibility check. With score

$$s_{\text{att}}(c,m) = \log\mathcal{N}\big(\mathbf{e}_m;\, \boldsymbol{\mu}_c, \text{diag}(\boldsymbol{\sigma}_c^2)\big)\,,$$

$$\mathcal{L}_{\text{att}}^{\text{rank}} = \sum_{v_m\in\mathcal{M}^{(t)}}\sum_{c^-}\max\big(0,\, s_{\text{att}}(c^-,m) - s_{\text{att}}(c_m^+,m) + \zeta\big), \quad (54)$$

with margin ζ.

Centroid repulsion. To prevent internal concept centroids from collapsing, sibling concepts are encouraged to maintain pairwise separation. For each internal concept c with children $\text{children}(c)$:

$$\mathcal{L}_{\text{repulsion}} = \sum_{c}\sum_{\substack{c_a,c_b\in\text{children}(c)\\ a\neq b}}\max\big(0,\, \zeta_r - \|\boldsymbol{\mu}_{c_a} - \boldsymbol{\mu}_{c_b}\|_2\big), \quad (55)$$

with separation margin ζ_r.

Combination.

$$\mathcal{L}_{\text{marg}} = \mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{att}}^{\text{rank}} + \mathcal{L}_{\text{repulsion}}. \quad (56)$$

J Structural Constraints

The main text describes the three most consequential constraints (temporal consistency, backbone preservation, tree legality). Here we formalize the two remaining constraints deferred from §4.4.2.

J.1 Path temporal consistency

For every root-to-leaf concept path $(c_0, c_1, \ldots, c_L)$ in $\mathcal{T}$,

$$\tau_T(c_0) \le \tau_T(c_1) + \epsilon \le \cdots \le \tau_T(c_L) + L\epsilon, \quad (57)$$

where $\tau_T(c) = \min_{v\in\mathcal{V}(c)}\tau(v)$ and ϵ is the per-edge tolerance margin. This rules out global temporal inversions in which local pairs satisfy the temporal order but the path as a whole accumulates a violation.

J.2 Path membership coherence

For every concept $c \in \mathcal{C}$, its evolutionary ancestor chain $\text{anc}^{\text{evo}}(c) = (c, p(c), p(p(c)), \ldots, r)$ must retain non-trivial distributional overlap with its taxonomy-ancestor chain $\text{anc}^{\text{tax}}(c)$. Formally, defining overlap as the maximum normalized intersection between any ancestor in the two chains,

$$\Omega(c) = \max_{c_e\in\text{anc}^{\text{evo}}(c),\, c_t\in\text{anc}^{\text{tax}}(c)}\frac{|\mathcal{V}(c_e)\cap\mathcal{V}(c_t)|}{\min(|\mathcal{V}(c_e)|, |\mathcal{V}(c_t)|)}, \quad (58)$$

and requiring $\Omega(c) \ge \theta_{\text{ov}}$. We use $\theta_{\text{ov}} = 0.3$. This constraint prevents the temporal reorientation in Stage II from rewiring a concept into a semantically unrelated branch.

K Marginal Compatibility Score

This appendix specifies the compatibility score $m(v_i, c)$ in §4.4.3.

For paper v_i and concept c with Gaussian $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$ and member set $\mathcal{V}(c)$:

Semantic similarity.

$$s_{\text{sem}}(i,c) = \frac{1}{2}(1 + \cos(\mathbf{e}_i, \boldsymbol{\mu}_c)) \in [0,1]. \quad (59)$$

Citation-neighborhood overlap.

$$s_{\text{graph}}(i,c) = \frac{|\mathcal{N}_{\mathcal{G}}(v_i)\cap\mathcal{V}(c)|}{\max(|\mathcal{N}_{\mathcal{G}}(v_i)|, 1)} \in [0,1]. \quad (60)$$

Temporal compatibility. Let $\bar{\tau}_c = |\mathcal{V}(c)|^{-1}\sum_{v\in\mathcal{V}(c)}\tau(v)$ be the nominal time of c. With time-scale σ_t:

$$s_{\text{time}}(i,c) = \exp\left(-\frac{|\tau(v_i) - \bar{\tau}_c|}{\sigma_t}\right) \in (0,1]. \quad (61)$$

Combined compatibility.

$$m(v_i,c) = \gamma_1\, s_{\text{sem}}(i,c) + \gamma_2\, s_{\text{graph}}(i,c) + \gamma_3\, s_{\text{time}}(i,c), \quad (62)$$

with $\gamma_1 + \gamma_2 + \gamma_3 = 1$ and $m(v_i,c) \in [0,1]$.

L Stage III Calibration Loss Details

This appendix expands the three calibration loss terms summarized in §4.5:

$$\mathcal{L}^{\text{cal}} = \mathcal{L}_{\text{marg}}^{\text{cal}} + \mathcal{L}_{\text{concept}}^{\text{cal}} + \mathcal{L}_{\text{edge}}^{\text{cal}}.$$

The FS dataset (§5.1) provides per-paper labels $y_i^{\text{m,fs}} \in \{0,1\}$ (marginal), $g_i \in \mathcal{G}^{\text{fs}}$ (concept), and the evolution edge set $\mathcal{E}^{\text{evo}}$ over papers (Appendix D).

L.1 FS marginal classification

Class-weighted BCE to handle the imbalance between core and marginal papers:

$$\mathcal{L}^{\text{cal}}_{\text{marg}} = -\frac{1}{|\mathcal{V}^{\text{fs}}|}\sum_{i\in\mathcal{V}^{\text{fs}}}\big[w^{+}y_i^{\text{m,fs}}\log h_i + w^{-}(1-y_i^{\text{m,fs}})\log(1-h_i)\big], \tag{63}$$

with $w^{+} = (1-\bar{y}^{\text{m}})/\bar{y}^{\text{m}}$, $w^{-} = 1$, and $\bar{y}^{\text{m}}$ the marginal fraction in the FS set.

L.2 FS concept alignment

For each FS concept $g \in \mathcal{G}^{\text{fs}}$, define its empirical centroid

$$\bar{\mathbf{e}}_g = \text{sg}\left(\frac{1}{|\mathcal{V}^{\text{fs}}_g|}\sum_{i\in\mathcal{V}^{\text{fs}}_g}\mathbf{e}_i\right), \tag{64}$$

with $\mathcal{V}^{\text{fs}}_g = \{v_i : g_i = g\}$. An InfoNCE objective aligns each paper with its FS centroid:

$$\mathcal{L}^{\text{cal}}_{\text{concept}} = -\frac{1}{|\mathcal{V}^{\text{fs}}|}\sum_{i\in\mathcal{V}^{\text{fs}}}\log\frac{\exp(\mathbf{e}_i^{\top}\bar{\mathbf{e}}_{g_i}/\tau_{\text{ic}})}{\sum_{g'\in\mathcal{G}^{\text{fs}}}\exp(\mathbf{e}_i^{\top}\bar{\mathbf{e}}_{g'}/\tau_{\text{ic}})}. \tag{65}$$

L.3 FS edge directionality

For each annotated edge $(v_s, v_d) \in \mathcal{E}^{\text{evo}}$, v_s is the evolutionary descendant of v_d. A triplet-margin loss requires the predicted depth of v_s to exceed that of v_d, with cross-domain `adapts` relations receiving a smaller margin:

$$\mathcal{L}^{\text{cal}}_{\text{edge}} = \sum_{(v_s,v_d)\in\mathcal{E}^{\text{evo}}}\max\big(0,\ \bar{d}(v_d) - \bar{d}(v_s) + \mu(v_s,v_d)\big), \tag{66}$$

where $\bar{d}(\cdot)$ is the expected depth from Eq. 50, and

$$\mu(v_s,v_d) = \begin{cases}\mu_{\text{strict}} & \texttt{extends / improves},\\ \mu_{\text{relax}} & \texttt{adapts}.\end{cases}$$

M EM-style Optimization

This appendix details the E-step / M-step alternation summarized in §4.4.1.

E-step frequency. In practice we run the E-step once per epoch. Running it more frequently (e.g., every 100 steps) did not improve performance and increased compute by $\sim 3\times$.

Temporal admissibility filter $\mathcal{F}_\epsilon$. A candidate merge of clusters C_p, C_q is admissible if for each subsequent merging step the resulting parent does not violate parental temporal ordering by more than $\epsilon = 1$ year, where the nominal time of a cluster is τ^{att} (Eq. 1), the mean publication year of its directly attached papers, as in Eq. 8.

Algorithm 2 Stage II EM-style optimization.

Require: Pre-trained encoder f_θ with adapters θ_a; taxonomy backbone $\mathcal{T}^{\text{tax}}$; epochs T.
Initialize $\mathcal{T}^{(0)} \leftarrow \mathcal{T}^{\text{tax}}$.
for $t = 1, \dots, T$ **do**
  E-step:
    Freeze θ_a; compute $\{\mathbf{e}_i\}$ via current encoder.
    Run Algorithm 1 with temporal admissibility filter $\mathcal{F}_\epsilon$.
    Obtain $\mathcal{T}^{(t)}, \mathcal{C}^{(t)}$ and identify marginal candidates $\mathcal{M}^{(t)}$.
    Generate pseudo-labels $\{\tilde{y}_i^{\text{m}}\}$ via Eq. 3.
  M-step:
    Freeze $\mathcal{T}^{(t)}$; update θ_a by one epoch of SGD on $\mathcal{L}^{\text{evo}}$.
end for
return Adapters θ_a, final tree $\mathcal{T} = \mathcal{T}^{(T)}$.

N Generation-quality Evaluation Protocol

N.1 Evaluation Setup

Both protocols receive identical supporting information: paper titles, publication years, citation relations, and the anonymized generated hierarchy. Method names, model descriptions, and metric results are withheld. To reduce presentation bias, structures are assigned randomized anonymous identifiers and presented in randomized order. The LLM judge is the same model used for concept labeling, evaluating each structure over multiple independent runs under a fixed prompt and decoding configuration. Human evaluation is conducted by three CS graduate students who did not participate in constructing the FS annotations; they evaluate each structure independently and do not discuss individual cases before submitting their judgments.

N.2 Criteria and Aggregation

Both protocols score four dimensions—conceptual organization, scientific evolution, transitional-paper placement, and overall explanatory quality—on a 1–10 scale under the same rubric, each accompanied by a brief evidence-based justification. Scores are averaged over evaluation runs (LLM) or evaluators (human), then over the FS domains, and we report the mean and standard deviation per method and dimension in Table 5.

The following prompt is given to the LLM judge; human evaluators receive the same criteria in written form.

Generation-quality evaluation prompt

You are an expert researcher. Given a citation graph and a generated hierarchy, evaluate the hierarchy on the following aspects.

Citation graph: `[titles, publication years, citation relations]`

Generated hierarchy: `[anonymized structure]`

Aspects:

1. ***Conceptual Organization*** *— Are papers grouped into semantically coherent concepts? Are concept names accurate?*
2. ***Scientific Evolution*** *— Does the hierarchy reflect the historical development of the field? Are parent–child relationships reasonable?*
3. ***Transitional Papers*** *— Are transitional / bridge papers attached at an appropriate position?*
4. ***Overall Quality*** *— How well does the hierarchy explain the organization and evolution of this research field?*

Give a score from 1–10 for each aspect and provide a brief explanation.