ADAPTIVE FEATURE EVOLUTION LEARNING WITH MULTI-LABEL EVIDENCE TRANSFER: ALGORITHM AND THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-label classification (MLC) conventionally assumes a stationary feature space, whereas real-world data streams often undergo simultaneous feature disappearance and emergence, disrupting label-discriminative evidence and rendering historical label dependencies increasingly mismatch. To address this chanllenges, we propose AFET, a unified framework that recasts dynamic-feature MLC as preserving and transferring label-relevant evidence across evolving feature spaces. Specifically, AFET exploits survived features as a cross-stage semantic anchor, recovering information associated with vanished features while preserving historical discriminative semantics through label-conditioned latent alignment. It further integrates recovered, survived and augmented evidence through instance-adaptive transformations, while adapting evolving label dependencies via structure-aware distillation and dynamically calibrating label-wise decision thresholds online. Theoretical analysis characterizes prediction stability across feature recovery, discriminative alignment, structural transfer and threshold adaptation. Extensive experiments on seven benchmarks under both one-shot and multi-shot settings demonstrate consistent improvements over competitive baselines.

1 INTRODUCTION

Multi-label classification (MLC) aims to assign multiple correlated labels to each instance, requiring models to capture both feature-to-label associations and inter-label dependencies (Tsoumakas & Katakis, 2007; Yu et al., 2014; Li et al., 2026; 2025a;b; Luo et al., 2025). Most existing MLC methods, however, are developed under the assumption that the feature space remains fixed throughout training and prediction (Liu et al., 2023; Si et al., 2024). In many real-world streaming environments, this assumption is difficult to maintain (Hou et al., 2021). New sensing sources (Zhu et al., 2018; Chow, 2024), modalities (Zhang et al., 2022; Astruc et al., 2025), or attributes may become available over time, while previously observed features may disappear because of sensor replacement, acquisition constraints (Fivush & Slackman, 2026), or privacy considerations (Liu et al., 2022; Pasham, 2023). The coexistence of feature emergence and disappearance gives rise to MLC with incremental and decremental features, where models must maintain reliable predictions despite continual changes in the observable feature space (Mi, 2025; Qiu et al., 2025).

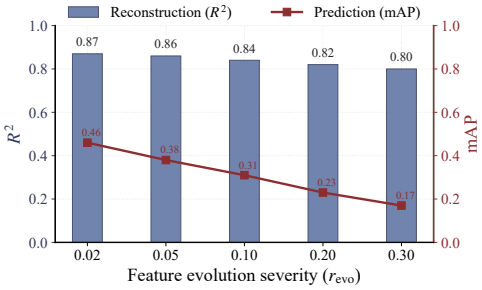


Figure 1: Feature reconstruction without preserving label-discriminative information may attenuate critical label-specific regions or even induce erroneous feature mixing.

Existing studies mainly cope with such feature evolution by establishing correspondence across successive feature spaces, including linear mappings (Tu et al., 2024), matrix factorization (Hou et al., 2017), feature transformation, and feature reconstruction (Gu et al., 2022). More recent approaches further employ deep representations to capture nonlinear cross-space relationships (Lian et al., 2023), while multi-label learning has begun to consider incremental and decremental features

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

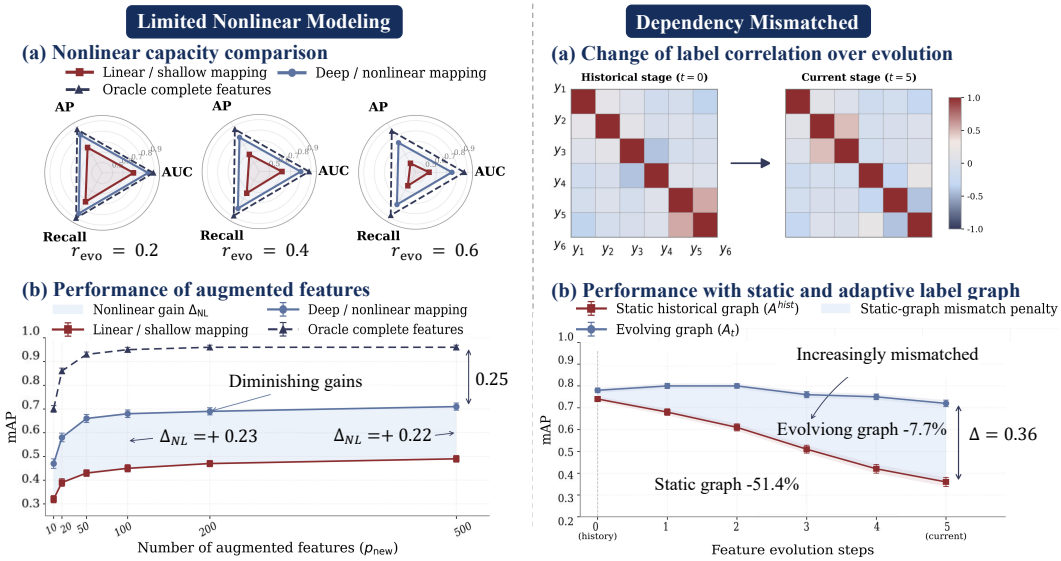


Figure 2: Key challenges of MLC with feature evolution. Left: limited nonlinear modeling capacity on Enron under increasing feature evolution. Right: mismatch between historical and current label dependencies on Scene.

jointly (Jiang et al., 2026). Nevertheless, these methods largely treat adaptation or reconstruction of the feature space as the primary objective. For MLC, however, preserving feature values and preserving predictive evidence are fundamentally different goals. A reconstructed feature representation can remain close to the original in feature space while losing information that is critical for individual label decisions.

More specifically, MLC under feature evolution faces three intertwined challenges. First, shallow mappings and explicit feature transformations may lack sufficient expressive capacity to capture nonlinear interactions between evolving feature representations and multiple labels. Second, feature recovery does not imply label-evidence preservation. Even accurate reconstruction of vanished features does not necessarily retain the discriminative information required for downstream label prediction; as illustrated in Fig. 1, label-specific evidence may be attenuated, blurred, or incorrectly mixed during reconstruction. Third, feature evolution can change the predictive regime beyond the feature space itself. Historical inter-label dependencies may become less informative as available evidence changes, causing static relational structures to become increasingly mismatched to current predictions, as exemplified in Fig. 2. Meanwhile, shifts in label-specific score distributions can make fixed decision thresholds poorly calibrated. These challenges call for expressive cross-stage representation learning, explicit preservation of historical label evidence, adaptive modeling of predictive label dependencies, and decision-level calibration under distributional drift.

To address these challenges, we propose **AFET**, a unified framework for adaptive feature-evolution learning that preserves and transfers multi-label predictive evidence across evolving feature spaces. AFET exploits survived features as the cross-stage interface between historical and newly available information. Before feature evolution, a teacher learns a generative survived-to-vanished recovery pathway and performs label-conditioned latent-space alignment to preserve discriminative information from the complete historical feature space, while historical label co-occurrence is encoded as a structural prior. After feature evolution, a student inherits the recovery mapping, combines surrogate vanished features with survived and augmented features, and learns instance-adaptive representations through transformations of different depths. The student further constructs a prediction-induced label graph and fuses it with the historical prior, while structure-aware distillation transfers both label-wise predictive responses and inter-label dependency structure from the teacher. Online label-wise threshold calibration additionally adapts decision boundaries to shifts in label-specific score distributions. Together, these mechanisms preserve historical predictive evidence while incorporating information introduced by augmented features. We further provide a theoretical characterization of prediction stability under feature evolution and conduct extensive experiments on seven

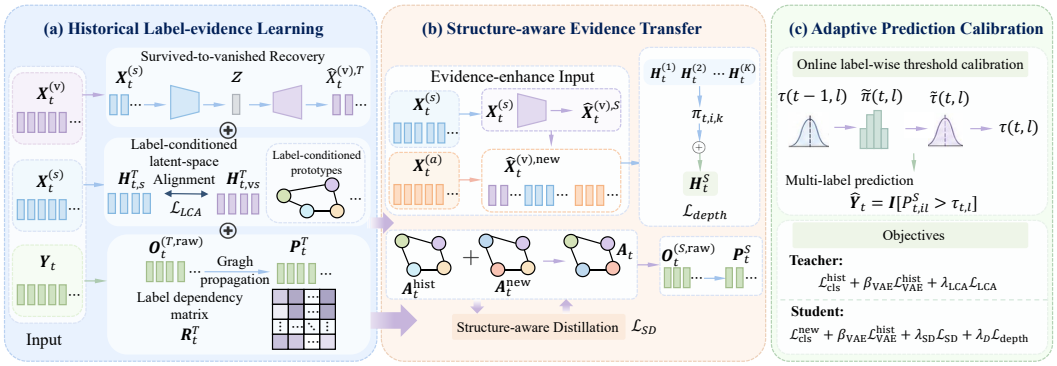


Figure 3: Overall architecture of AFET. (a) Historical label-evidence learning preserves discriminative information from historical features. (b) Structure-aware evidence transfer propagates predictive and relational knowledge across evolutionary stages. (c) Adaptive prediction calibration dynamically adjusts label-wise decisions under feature evolution.

benchmark datasets under both one-shot and multi-shot settings, demonstrating consistent improvements over competitive baselines across multiple multi-label metrics.

Our main contributions are summarized as follows:

- We introduce a label-guided learning strategy for multi-label classification with incremental and decremental features, emphasizing the preservation of label-discriminative information throughout feature evolution. By incorporating label supervision into cross-stage representation learning, this strategy preserves the predictive value of historical features beyond feature-space recovery alone.
- We develop **AFET**, a unified framework for preserving historical predictive evidence while incorporating newly emerged information under simultaneous feature disappearance and emergence. By exploiting survived features as a cross-stage semantic anchor, AFET integrates label-conditioned evidence preservation, instance-adaptive representation learning, evolving label-dependency modeling, structure-aware knowledge transfer, and online label-wise calibration into a coherent learning framework.
- We establish a theoretical characterization of prediction stability under feature evolution, connecting feature recovery, label-conditioned alignment, relational knowledge transfer, and online calibration to complementary aspects of predictive stability. The analysis elucidates how these mechanisms respectively control information loss, prediction inconsistency, relational drift, and threshold deviation as the feature space evolves.

2 Methodology

2.1 Problem Setup

Let $\mathcal{Y} = \{1, \ldots, L\}$ denote a fixed label space. At time step t, the learner receives a mini-batch $\mathcal{D}_t = (\mathbf{X}_t, \mathbf{Y}_t)$, where $\mathbf{Y}_t \in \{0,1\}^{n_t \times L}$. We consider multi-label learning under simultaneous incremental and decremental feature evolution. The overall feature space is partitioned into three disjoint subsets, $\mathcal{F} = V \cup S \cup A$, corresponding to *vanished*, *survived*, and *augmented* features, respectively. The historical and new-stage inputs are $\mathbf{X}_t^{\text{hist}} = [\mathbf{X}_t^{(v)}, \mathbf{X}_t^{(s)}]$, $\mathbf{X}_t^{\text{new}} = [\mathbf{X}_t^{(s)}, \mathbf{X}_t^{(a)}]$. Hence, S constitutes the only explicitly shared feature interface across stages. We follow a one-pass streaming protocol. Each batch is accessed once and discarded after updating, so historical instances cannot be revisited after feature evolution. Our goal is to learn a time-dependent predictor $f_t : \mathbf{X}_t^{\text{new}} \mapsto [0,1]^L$ while preserving predictive evidence acquired from the inaccessible features.

2.2 Label-Conditioned Latent-Space Alignment

Before feature evolution, AFET employs a teacher model to consolidate both label-specific and relational evidence from the complete historical feature space. The teacher first learns a generative cor-

respondence from survived features S to vanished features V. Given $\mathbf{X}_t^{(s)}$, a latent variable is sampled from a variational encoder and decoded to recover the vanished features, $z_{t,s}^T \sim q_S^T\left(z \mid \mathbf{X}_t^{(s)}\right)$, $\widehat{\mathbf{X}}_t^{(v),T} = D_v^T\left(z_{t,s}^T\right)$. This pathway establishes the correspondence $S \to V$ before V becomes unavailable. The complete variational parameterization and reconstruction objective are deferred to the supplementary material. Feature reconstruction alone does not ensure preservation of label-discriminative information. We therefore align the survived-feature representation $\mathbf{H}_{t,s}^T$ with the complete historical representation $\mathbf{H}_{t,vs}^T = E_{vs}^T(\mathbf{X}_t^{\text{hist}})$ within label-conditioned regions. For a representation matrix $\mathbf{H}$, the prototype of label l is

$$c_l(\mathbf{H}; \mathbf{Y}_t) = \frac{\sum_{i=1}^{n_t} Y_{t,il}\mathbf{H}_{i,:}}{\sum_{i=1}^{n_t} Y_{t,il} + \epsilon}, \qquad \mathcal{L}_{\text{LCA}}^t = \frac{1}{L}\sum_{l=1}^{L} \left\| c_l(\mathbf{H}_{t,s}^T; \mathbf{Y}_t) - c_l(\mathbf{H}_{t,vs}^T; \mathbf{Y}_t) \right\|_2^2. \tag{1}$$

Unlike global alignment, $\mathcal{L}_{\text{LCA}}^t$ explicitly preserves information associated with individual labels that may otherwise disappear with V.

To additionally capture relational evidence, we construct a batch-wise historical label graph from label co-occurrence:

$$\mathbf{A}_t^{\text{hist}} = \text{Norm}\big(\mathbf{Y}_t^\top \mathbf{Y}_t - \text{Diag}\left(\text{diag}(\mathbf{Y}_t^\top \mathbf{Y}_t)\right)\big), \qquad \text{Norm}(\mathbf{M}) = \frac{\mathbf{M}}{\|\mathbf{M}\|_2 + \epsilon}. \tag{2}$$

The graph directly propagates information in the teacher prediction space, $\mathbf{O}_t^T = \mathbf{O}_t^{T,\text{raw}} + \alpha \mathbf{O}_t^{T,\text{raw}} \mathbf{A}_t^{\text{hist}}$, with $\mathbf{P}_t^T = \sigma(\mathbf{O}_t^T)$. The final historical graph is retained as the structural prior $\mathbf{A}^{\text{hist}}$ for the new stage. The teacher is optimized by

$$\mathcal{L}_{\text{hist}}^t = \mathcal{L}_{\text{cls}}^{\text{hist},t} + \beta_{\text{VAE}} \mathcal{L}_{\text{VAE}}^{\text{hist},t} + \lambda_{\text{LCA}} \mathcal{L}_{\text{LCA}}^t, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ is the standard multi-label binary cross-entropy loss and $\mathcal{L}_{\text{VAE}}$ denotes the variational recovery objective.

2.3 Structure-Aware Evidence Transfer

After feature evolution, V becomes inaccessible and only $[S, A]$ is observed. AFET introduces a student model that uses S as the cross-stage bridge to recover historical evidence while incorporating newly available information. The student inherits the survived-to-vanished recovery mapping learned by the teacher and obtains surrogate vanished features $\widehat{\mathbf{X}}_t^{(v),S}$ from $\mathbf{X}_t^{(s)}$. These are combined with the observable features to form $\widetilde{\mathbf{X}}_t^{\text{new}} = \left[\widehat{\mathbf{X}}_t^{(v),S}, \mathbf{X}_t^{(s)}, \mathbf{X}_t^{(a)}\right]$, which integrates recovered historical, survived, and newly augmented evidence. Historical label relations need not remain equally informative after feature evolution. We therefore infer a new-stage graph from preliminary student predictions $\widetilde{\mathbf{P}}_t^S = \sigma(\mathbf{O}_t^{S,\text{raw}})$ and combine it with the historical prior:

$$\mathbf{A}_t^{\text{new}} = \text{Norm}\left((\widetilde{\mathbf{P}}_t^S)^\top \widetilde{\mathbf{P}}_t^S\right), \qquad \mathbf{A}_t = (1-\eta_t)\mathbf{A}^{\text{hist}} + \eta_t \mathbf{A}_t^{\text{new}}. \tag{4}$$

The fused graph balances historical structural preservation and new-stage adaptation. It produces the structure-aware student prediction $\mathbf{O}_t^S = \mathbf{O}_t^{S,\text{raw}} + \alpha \mathbf{O}_t^{S,\text{raw}} \mathbf{A}_t$, with $\mathbf{P}_t^S = \sigma(\mathbf{O}_t^S)$.

Feature recovery alone cannot preserve the predictive behavior accumulated by the historical model. We therefore distill both marginal label responses and inter-label dependencies from the frozen teacher. Define $\mathbf{R}_t^\bullet = \text{Norm}\left((\mathbf{P}_t^\bullet)^\top \mathbf{P}_t^\bullet\right)$, $\bullet \in \{T, S\}$. The structure-aware distillation objective contains response transfer and dependency transfer:

$$\mathcal{L}_{\text{SD}}^t = \frac{1}{n_t L}\left\|\mathbf{P}_t^S - \text{sg}(\mathbf{P}_t^T)\right\|_F^2 + \omega \left\|\mathbf{R}_t^S - \text{sg}(\mathbf{R}_t^T)\right\|_F^2, \tag{5}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The student therefore inherits historical evidence beyond individual prediction probabilities.

2.4 Adaptive Prediction Calibration

To capture nonlinear feature–label interactions with instance-dependent complexity, $\widetilde{\mathbf{X}}_t^{\text{new}}$ is processed through transformations of different depths. Let $\mathbf{H}_t^{(k)}$ denote the representation matrix after

the k-th transformation block. A routing network predicts instance-specific weights, yielding

$$\boldsymbol{\pi}_{t,i} = \mathrm{Softmax}\left(\frac{g(\mathbf{H}_t^{(0)}[i,:])}{T_d}\right), \qquad \mathbf{H}_t^S[i,:] = \sum_{k=1}^{K} \pi_{t,ik}\mathbf{H}_t^{(k)}[i,:]. \tag{6}$$

A lightweight depth regularizer $\mathcal{L}_{\mathrm{depth}}^t$ discourages systematically selecting unnecessarily deep transformations; its explicit form is provided in the supplementary material. The overall new-stage objective is

$$\mathcal{L}_{\mathrm{new}}^t = \mathcal{L}_{\mathrm{cls}}^{\mathrm{new},t} + \beta_{\mathrm{VAE}}\mathcal{L}_{\mathrm{VAE}}^{\mathrm{new},t} + \lambda_{\mathrm{SD}}\mathcal{L}_{\mathrm{SD}}^t + \lambda_D\mathcal{L}_{\mathrm{depth}}^t. \tag{7}$$

Feature evolution may shift label-specific score distributions, making fixed decision thresholds suboptimal. For label l, AFET estimates the current pseudo-positive rate and adapts its threshold as

$$\widehat{\pi}_{t,l} = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathbb{I}\left[P_{t,il}^S > \tau_{t-1,l}\right], \qquad r_{t,l} = \mathrm{clip}\left((1+\rho)\widehat{\pi}_{t,l} + m, r_{\min}, r_{\max}\right), \tag{8}$$

followed by $\widetilde{\tau}_{t,l} = \mathrm{Quantile}\left(\{P_{t,il}^S\}_{i=1}^{n_t} 1 - r_{t,l}\right)$, $\tau_{t,l} = \gamma\tau_{t-1,l} + (1-\gamma)\widetilde{\tau}_{t,l}$.

The final prediction is $\widehat{Y}_{t,il} = \mathbb{I}[P_{t,il}^S > \tau_{t,l}]$. This calibration operates entirely at the decision level and introduces no additional trainable parameters. Detailed variational objectives, routing regularization, and optimization procedures are provided in the supplementary material.

3 Theoretical Analysis

We analyze AFET from the perspective of label-evidence preservation under feature evolution. Our analysis characterizes the stability induced by vanished-feature recovery, label-conditioned alignment, structure-aware knowledge transfer, label-graph evolution, and online threshold calibration. All proofs and additional derivations are deferred to the supplementary material.

Prediction stability under vanished-feature recovery. Define the oracle complete input as $\mathbf{X}_t^\star = \left[\mathbf{X}_t^{(v)}, \mathbf{X}_t^{(s)}, \mathbf{X}_t^{(a)}\right]$, and the evidence-enhanced input used by AFET as $\widetilde{\mathbf{X}}_t^{\mathrm{new}} = \left[\widehat{\mathbf{X}}_t^{(v),S}, \mathbf{X}_t^{(s)}, \mathbf{X}_t^{(a)}\right]$. Let F_t^S denote the complete student mapping to the structure-aware logits, and let $\mathcal{L}_{\mathrm{BCE}}$ denote the average binary cross-entropy loss in logit form.

Theorem 3.1 (Recovery-Induced Prediction Stability)**.** *If F_t^S is locally K_t-Lipschitz in a region containing $\widetilde{\mathbf{X}}_t^{\mathrm{new}}$ and $\mathbf{X}_t^\star$, then*

$$\left|\mathcal{L}_{\mathrm{BCE}}\left(F_t^S(\widetilde{\mathbf{X}}_t^{\mathrm{new}}), \mathbf{Y}_t\right) - \mathcal{L}_{\mathrm{BCE}}\left(F_t^S(\mathbf{X}_t^\star), \mathbf{Y}_t\right)\right| \le \frac{K_t}{\sqrt{n_t L}}\left\|\widehat{\mathbf{X}}_t^{(v),S} - \mathbf{X}_t^{(v)}\right\|_F. \tag{9}$$

Label-conditioned discriminative stability. For label l with $\mathcal{I}_{t,l} = \{i : Y_{t,il} = 1\} \neq \emptyset$, define $\mathbf{c}_{t,l}^s = c_l(\mathbf{H}_{t,s}^T; \mathbf{Y}_t)$, $\mathbf{c}_{t,l}^{vs} = c_l(\mathbf{H}_{t,vs}^T; \mathbf{Y}_t)$. For $q \in \{s, vs\}$, let $s_{t,l}^q = \frac{1}{|\mathcal{I}_{t,l}|}\sum_{i\in\mathcal{I}_{t,l}}\left\|\mathbf{H}_{t,q}^T[i,:] - \mathbf{c}_{t,l}^q\right\|_2$.

Proposition 3.2 (Label-Conditioned Prediction Stability)**.** *If the label-specific classifier f_l is K_l-Lipschitz, then*

$$\frac{1}{|\mathcal{I}_{t,l}|}\sum_{i\in\mathcal{I}_{t,l}}\left|f_l\left(\mathbf{H}_{t,s}^T[i,:]\right) - f_l\left(\mathbf{H}_{t,vs}^T[i,:]\right)\right| \le K_l\left(s_{t,l}^s + s_{t,l}^{vs} + \left\|\mathbf{c}_{t,l}^s - \mathbf{c}_{t,l}^{vs}\right\|_2\right). \tag{10}$$

Structural transfer and graph adaptation. Let $\mathbf{G}_t^\bullet = (\mathbf{P}_t^\bullet)^\top\mathbf{P}_t^\bullet$, $\mathbf{R}_t^\bullet = \mathrm{Norm}(\mathbf{G}_t^\bullet)$, $\bullet \in \{T, S\}$, and define $\mathcal{L}_{\mathrm{KD}}^t = \frac{\|\mathbf{P}_t^S - \mathbf{P}_t^T\|_F^2}{n_t L}$.

Proposition 3.3 (Response-to-Structure Consistency)**.** *The dependency discrepancies satisfy*

$$\left\|\mathbf{G}_t^S - \mathbf{G}_t^T\right\|_F \le 2n_t L\sqrt{\mathcal{L}_{\mathrm{KD}}^t}, \qquad \left\|\mathbf{R}_t^S - \mathbf{R}_t^T\right\|_F^2 \le 4K_N^2 n_t^2 L^2 \mathcal{L}_{\mathrm{KD}}^t, \tag{11}$$

where the second inequality holds when $\mathrm{Norm}(\cdot)$ is K_N-Lipschitz over the relevant bounded region.

For graph adaptation, let $\mathbf{Q}_t = \mathbf{O}_t^{S,\mathrm{raw}}$, $\Phi(\mathbf{Q}, \mathbf{A}) = \mathbf{Q} + \alpha\mathbf{Q}\mathbf{A}$, and let $\mathbf{A}_t^\star$ denote an oracle current-stage label graph. Define $\epsilon_{\mathrm{hist},t} = \left\|\mathbf{A}^{\mathrm{hist}} - \mathbf{A}_t^\star\right\|_F$, $\epsilon_{\mathrm{new},t} = \|\mathbf{A}_t^{\mathrm{new}} - \mathbf{A}_t^\star\|_F$.

Proposition 3.4 (Stability under Label-Graph Evolution). *For $\mathbf{A}_t = (1-\eta_t)\mathbf{A}^{\mathrm{hist}} + \eta_t\mathbf{A}_t^{\mathrm{new}}$, we have*

$$\|\Phi(\mathbf{Q}_t, \mathbf{A}_t) - \Phi(\mathbf{Q}_t, \mathbf{A}_t^\star)\|_F \le \alpha\|\mathbf{Q}_t\|_2 \left[(1-\eta_t)\epsilon_{\mathrm{hist},t} + \eta_t\epsilon_{\mathrm{new},t}\right], \tag{12}$$

$$\|\sigma(\Phi(\mathbf{Q}_t, \mathbf{A}_t)) - \sigma(\Phi(\mathbf{Q}_t, \mathbf{A}_t^\star))\|_F \le \frac{\alpha}{4}\|\mathbf{Q}_t\|_2 \left[(1-\eta_t)\epsilon_{\mathrm{hist},t} + \eta_t\epsilon_{\mathrm{new},t}\right]. \tag{13}$$

Dynamic tracking of label-wise decision boundaries. For label l, define $\pi_{t,l}^\star = \Pr\left(P_{t,l}^S > \tau_{t-1,l}\right)$, $r_{t,l}^\star = \mathrm{clip}\left((1+\rho)\pi_{t,l}^\star + m, r_{\min}, r_{\max}\right)$, and $\tau_{t,l}^\star = F_{t,l}^{-1}\left(1 - r_{t,l}^\star\right)$, where $F_{t,l}$ denotes the population CDF of the student score for label l.

Theorem 3.5 (Dynamic Tracking of Label-Wise Thresholds). *Conditioned on the preceding history, assume that the current label-l scores are independently sampled from $F_{t,l}$, $\left|\tau_{t,l}^\star - \tau_{t-1,l}^\star\right| \le \Delta_{t,l}$, and that the score density is lower bounded by $m_l > 0$ in a neighborhood of the relevant target quantiles. Then, with probability at least $1-\delta_\tau$,*

$$\left|\widetilde{\tau}_{t,l} - \tau_{t,l}^\star\right| \le \xi_{t,l}, \qquad e_{t,l} \le \gamma e_{t-1,l} + \gamma\Delta_{t,l} + (1-\gamma)\xi_{t,l}, \tag{14}$$

where $\xi_{t,l} = \frac{2+\rho}{m_l}\sqrt{\frac{\log(2/\delta_\tau)}{2n_t}}$, $e_{t,l} = \left|\tau_{t,l} - \tau_{t,l}^\star\right|$, for the update $\tau_{t,l} = \gamma\tau_{t-1,l} + (1-\gamma)\widetilde{\tau}_{t,l}$.

Remark 3.6. Taken together, the above results provide a unified stability characterization of label-evidence preservation throughout feature evolution. Theorem 3.1 establishes that the predictive loss incurred by replacing inaccessible vanished features is explicitly controlled by the corresponding recovery error, thereby connecting feature-level recovery to downstream predictive stability. Proposition 3.2 further refines this guarantee at the label-discriminative level: by reducing the discrepancy between label-conditioned prototypes, the proposed latent alignment constrains the prediction variation between survived-feature and complete historical representations within concentrated label regions. Beyond marginal label responses, Proposition 3.3 shows how response-level consistency propagates to the induced inter-label dependency structure, while Proposition 3.4 characterizes the effect of graph evolution as an explicit preservation–adaptation trade-off between historical structural mismatch and current graph-estimation error. Finally, Theorem 3.5 extends the stability analysis to the decision level, showing that the online label-wise calibration tracks time-varying target thresholds with an error jointly determined by finite-sample uncertainty, distributional drift, and the smoothing factor γ. Collectively, these results characterize AFET as preserving predictive evidence at multiple complementary levels, from recovered features and label-conditioned discriminative representations to inter-label dependencies and final decision boundaries, rather than relying on feature reconstruction alone.

4 Experiment

4.1 Configurations

Datasets and Evaluation metrics. To evaluate the effectiveness of our approach, we conduct experiments on seven widely used benchmark datasets spanning diverse application domains: Languagelog (Read, 2010), Enron (Read et al., 2008), Scene (Boutell et al., 2004) , Bibtex (Katakis et al., 2008), Rcv1v2(subset1) (Lewis et al., 2004), NUS-WIDE (Chua et al., 2009) and Tmc2007 (Srivastava & Zane-Ulman, 2005). Performance is assessed using six standard MLC metrics: Ranking Loss (RL), Recall, Average Precision (AP), Normalized Discounted Cumulative Gain (NDCG), Hamming Loss (HL), and the adapted Area Under the Curve (AUC). For clarity, we report 1-RL and 1-HL. So that higher values consistently indicate better performance.

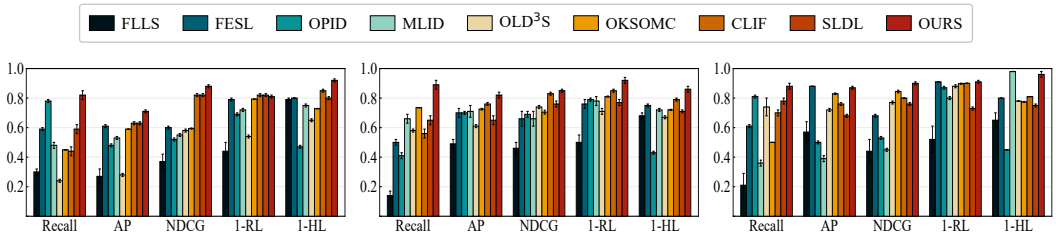
Implementation Details. For both training and testing, we assume a uniform class distribution across all batches. Training is conducted over R stages with a fixed total number of samples, while the number of samples per batch may vary. In multi-shot case, the data are evenly partitioned across stages. The full feature set is likewise divided into equally sized blocks according to the original feature ordering. At each prediction step, the available features consist of those retained from the previous step together with newly introduced features.

Comparison Methods. We compare AFET with eight methods, i.e., FLLS (Gu et al., 2022), FESL (Hou et al., 2017), OPID (C Hou, 2017), OLD^{3}S (Lian et al., 2023), MLID (Jiang et al., 2026), OKS-OMC (Zhai & Liu, 2025), CLIF (Hang & Zhang, 2022) and SLDL (Zhao et al., 2025). We note that the first three baselines are representative methods designed for streaming scenarios with dynamic feature evolution. Among them, MLID is a linear model specifically developed for multi-label learning with both incremental and decremental features. The remaining three baselines are originally designed for conventional MLC settings under a closed-world assumption and are evaluated only at the post-incremental stage. Among these methods, OLD^{3}S, CLIF, and SLDL are deep learning-based approaches. To ensure a fair comparison, all hyperparameters are configured according to the settings recommended in their original papers or official implementations.

4.2 Comparative Study

Data set	n_i	FLLS	FESL	OPID	OLD^{3}S	MLID	OKS-OMC	CLIF	SLDL	OURS
Languagelog	15	.4383 (.0595)	.7867 (.0080)	.6831 (.0093)	.4366 (.0070)	.7230 (.1244)	.7921 (.0013)	.7833 (.0122)	.7077 (.0220)	**.8118 (.0111)**
	30	.4610 (.0830)	.7918 (.0142)	.6875 (.0146)	.4366 (.0070)	.7216 (.0128)	.7921 (.0013)	.7889 (.0107)	.7078 (.0220)	**.8138 (.0112)**
	60	.4678 (.0893)	.7931 (.0116)	.6888 (.0083)	.4366 (.0070)	.7177 (.0173)	.7921 (.0013)	.7825 (.0110)	.7078 (.0220)	**.8122 (.0110)**
	120	.4553 (.0747)	.8091 (.0040)	.6838 (.0107)	.4366 (.0070)	.7215 (.0171)	.7921 (.0013)	.7711 (.0126)	.7078 (.0220)	**.8161 (.0102)**
Enron	20	.5001 (.0494)	.7626 (.0304)	.7800 (.0150)	.4893 (.0153)	.7760 (.0252)	.8059 (.0042)	.7554 (.0562)	.7052 (.0200)	**.8174 (.0127)**
	40	.5121 (.0490)	.7536 (.0026)	.7896 (.0079)	.4893 (.0153)	.7749 (.0231)	.8059 (.0042)	.7484 (.0443)	.7052 (.0200)	**.8283 (.0145)**
	80	.5592 (.0389)	.7577 (.0030)	.7861 (.0190)	.4893 (.0153)	.7724 (.0301)	.8059 (.0042)	.7554 (.0152)	.7052 (.0200)	**.8308 (.0101)**
	160	.5366 (.0522)	.7877 (.0195)	.7835 (.0160)	.4893 (.0153)	.7949 (.0196)	.8059 (.0042)	.7302 (.0181)	.7052 (.0200)	**.8215 (.0142)**
Scene	30	.5324 (.0290)	.8301 (.0405)	.9072 (.0072)	.6003 (.0241)	.8795 (.0106)	.8990 (.0035)	.9260 (.0066)	.8791 (.0146)	**.9321 (.0075)**
	60	.5423 (.0203)	.8566 (.0230)	.9035 (.0246)	.6543 (.0219)	.8862 (.0172)	.8990 (.0035)	.9201 (.0070)	.8791 (.0146)	**.9204 (.0061)**
	120	.5191 (.0203)	.8552 (.0170)	.9038 (.0323)	.7386 (.0201)	.8850 (.0139)	.8990 (.0035)	.9075 (.0054)	.8791 (.0146)	**.9267 (.0059)**
	240	.4996 (.0165)	.8755 (.0141)	.9079 (.0282)	.7995 (.0150)	.8849 (.0037)	.8990 (.0035)	.8805 (.0094)	.8791 (.0146)	**.9136 (.0096)**
Rcv1v2(subset1)	60	.5359 (.0225)	.5537 (.0080)	.9406 (.0041)	.4859 (.0021)	.7145 (.0051)	.8927 (.0006)	.7315 (.0108)	.7330 (.0181)	**.9581 (.0131)**
	120	.5305 (.0186)	.5517 (.0107)	.9405 (.0041)	.4865 (.0099)	.7036 (.0053)	.8925 (.0004)	.7234 (.0112)	.7330 (.0181)	**.9508 (.0195)**
	240	.5305 (.0186)	.5598 (.0073)	.9405 (.0041)	.4865 (.0098)	.7050 (.0021)	.8930 (.0002)	.6543 (.0173)	.7330 (.0181)	**.9507 (.0149)**
	480	.5305 (.0186)	.5581 (.0080)	.9399 (.0119)	.4865 (.0051)	.7056 (.0041)	.8918 (.0003)	.6010 (.0151)	.7330 (.0181)	**.9570 (.0186)**
Bibtex	85	.4949 (.0162)	.8241 (.0101)	.8654 (.0074)	.6251 (.0097)	.5023 (.0046)	.8091 (.0025)	.7676 (.0120)	.8221 (.0069)	**.9229 (.0124)**
	175	.4896 (.0193)	.8410 (.0075)	.8651 (.0077)	.6251 (.0097)	.5089 (.0012)	.8096 (.0037)	.6826 (.0096)	.8221 (.0069)	**.9235 (.0111)**
	350	.4896 (.0193)	.8489 (.0055)	.8616 (.0064)	.6251 (.0097)	.5021 (.0051)	.8121 (.0028)	.5755 (.0067)	.8221 (.0069)	**.9115 (.0180)**
	700	.4856 (.0190)	.8763 (.0144)	.8611 (.0027)	.6251 (.0097)	.5008 (.0062)	.8089 (.0032)	.7942 (.0061)	.8221 (.0069)	**.9224 (.0337)**
Tmc2007	50	.5238 (.0928)	.4845 (.0168)	.5788 (.0072)	.7044 (.0149)	.5184 (.0061)	.7030 (.0007)	.8453 (.0025)	.7547 (.0108)	**.8564 (.0047)**
	100	.5238 (.0928)	.4800 (.0034)	.5855 (.0317)	.6884 (.0130)	.5048 (.0032)	.7174 (.0004)	.8460 (.0025)	.7547 (.0108)	**.8555 (.0037)**
	200	.5238 (.0928)	.4794 (.0042)	.5848 (.0069)	.7338 (.0144)	.4879 (.0067)	.7005 (.0031)	.8457 (.0032)	.7547 (.0108)	**.8562 (.0058)**
	400	.5265 (.0873)	.4799 (.0032)	.6008 (.0515)	.7369 (.0112)	.5032 (.0032)	.7023 (.0006)	.8518 (.0022)	.7547 (.0108)	**.8543 (.0041)**
NUS-WIDE	500	.5156 (.0211)	.8555 (.0042)	.9016 (.0026)	.6807 (.0112)	.6146 (.0045)	.7564 (.0017)	.7942 (.0061)	.5277 (.0043)	**.9187 (.0046)**
	1000	.5156 (.0211)	.8643 (.0483)	.9016 (.0026)	.6874 (.0111)	.6178 (.0022)	.7542 (.0023)	.7444 (.0078)	.5277 (.0043)	**.9207 (.0051)**
	2000	.5286 (.0312)	.8518 (.0040)	.9003 (.0025)	.6870 (.0018)	.6193 (.0056)	.7550 (.0030)	.6670 (.0073)	.5277 (.0043)	**.9043 (.0135)**
	4000	.5286 (.0313)	.8645 (.0038)	.9005 (.0025)	.6851 (.0043)	.6187 (.0072)	.7582 (.0035)	.6088 (.0075)	.5277 (.0043)	**.9125 (.0084)**

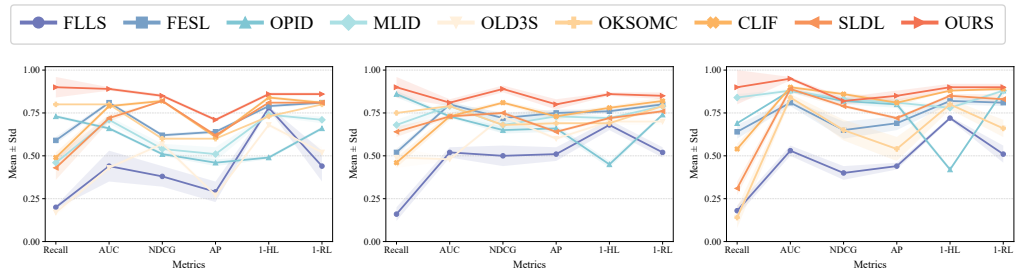
Table 1: The one-shot testing AUC (mean std) between AFET and the compared methods on 7 data sets with different number of training and testing examples.



(a) Languaglog (b) Enron (c) Tmc2007

Figure 4: Performance of various metrics across methods and datasets with n_i fixed as the minimum batch number in one-shot case.

To evaluate the effectiveness of our method, we conduct experiments on seven datasets against eight state-of-the-art baselines. Table 1 summarizes the AUC results across methods and datasets, while Fig.(7) reports performance under the remaining evaluation metrics. For completeness, additional experimental results are provided in the supplementary materials. Based on these results, we highlight several key observations below. (1) AFET consistently outperforms the competing baselines across a broad range of evaluation metrics and experimental conditions. In particular, the substantial gains in AUC and AP highlight the importance of effectively preserving and transferring multi-label information as the feature space evolves. (2) Under the same experimental settings, AFET consistently surpasses all competing MLC methods. The results suggest that the latent con-



(a) Task1: Languagelog. (b) Task1: Enron. (c) Task1: Scene.

(d) Task2: Languagelog. (e) Task2: Enron. (f) Task2: Scene.

Figure 5: Performance of various methods in predicting $\hat{\mathbf{Y}}_{T+2}$ and $\hat{\mathbf{Y}}_{T_i+t+2}$ on 3 datasets with n_i fixed as the minimum batch number in two tasks during multi-shot case.

ditional alignment mechanism effectively mitigates label-information loss and semantic mismatch across stages induced by feature addition and removal. Compared with MLID, AFET also exhibits a stronger ability to capture complex nonlinear dependencies under the same evolving-feature scenarios. (3) AFET further demonstrates strong performance in streaming-data settings, outperforming competitive baselines such as OPID and OLD^{3}S. Its adaptive thresholding strategy, together with label-correlation graph modeling, enables the model to better exploit latent semantic dependencies among labels, which are often insufficiently captured by conventional classification approaches. (4) AFET remains consistently superior to all baselines across a wide range of n_i settings and retains strong robustness under the multi-shot case. As shown in Fig. (5), AFET maintains the best overall performance across evaluation metrics despite repeated feature addition and removal. Notably, it incurs only a 4% performance degradation, substantially lower than FLLS (18%), FESL (32%), OPID (40%), MLID (24%), and OKSOMC (45%). This markedly smaller degradation indicates that AFET is considerably less sensitive to repeated changes in feature spaces, underscoring its robustness in highly dynamic streaming environments.

Table 2: Scenario ablation study. $S_1 - S_6$ are combinations to quantify the contributions of different predictive evidence across evolutionary stages. ✔ and ✗ denote whether the corresponding information source is enabled or disabled, respectively.

Scenario	History		New			Enron		NUS-WIDE		Tmc2007	
	$\mathcal{V}$	$\mathcal{S}$	$\hat{\mathcal{V}}$	$\mathcal{S}$	$\mathcal{A}$	AP	Recall	AP	Recall	AP	Recall
S_1	✗	✔	✗	✔	✗	0.81	0.70	0.63	0.65	0.82	0.78
S_2	✗	✗	✗	✔	✔	0.82	0.71	0.63	0.70	0.83	0.79
S_3	✗	✗	✗	✗	✔	0.78	0.67	0.57	0.62	0.79	0.76
S_4	✗	✔	✗	✔	✔	0.83	0.72	0.64	0.60	0.84	0.79
S_5	✔	✔	✔	✔	✗	0.82	0.77	0.61	0.72	0.83	0.84
S_6	✔	✔	✔	✔	✔	**0.85**	**0.79**	**0.72**	**0.74**	**0.85**	**0.85**

4.3 Further Analysis

4.3.1 Ablation Study

To assess the feature-evolution setting and quantify the contribution of each component in AFET, we conduct both scenario-level and module-level ablation studies on Enron, NUS-WIDE and Tmc2007. The corresponding results are reported in Tables 2 and 3.

At the scenario level, the full configuration (S_6) consistently achieves the best performance in terms of AP and recall, indicating that historical features and newly available features provide complementary predictive evidence. In particular, excluding historical information that disappears in later stages leads to a pronounced performance degradation, highlighting the importance of preserving label-

discriminative evidence accumulated from earlier stages. Meanwhile, the additional gains brought by surviving and newly emerging features further demonstrate their respective roles in bridging knowledge across evolutionary stages and incorporating novel predictive evidence.

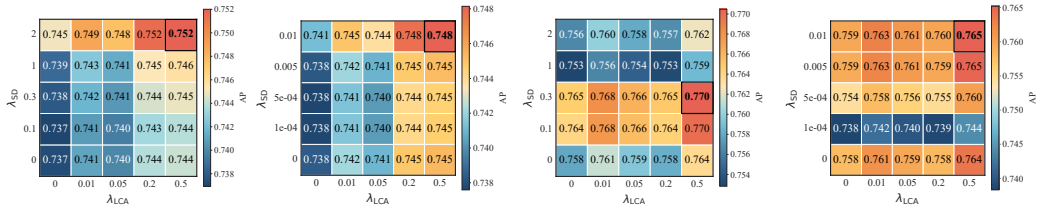
At the module level, removing any of the three components results in performance degradation, confirming that they contribute complementary functionality to the overall framework. In particular, eliminating label-conditioned latent-space alignment (M_1) causes a substantial performance drop, supporting the necessity of preserving label-specific discriminative information beyond feature-level reconstruction. Structure-aware evidence transfer (M_2) further improves performance by transferring both predictive responses and relational knowledge across stages. Adaptive prediction calibration (M_3) is especially critical for recall, removing this component reduces recall from 0.74 to 0.40, underscoring the importance of adaptively adjusting learned representations and decision boundaries under feature evolution. Overall, these results validate the complementary roles of evidence preservation, structure-aware transfer, and adaptive prediction in AFET.

4.3.2 Parameter Sensitivity

To evaluate the sensitivity of AFET to its hyperparameters, we investigate the interplay between λ_{LCA} and λ_{SD}, as well as between λ_{LCA} and λ_{D}. Specifically, λ_{LCA} is varied over $\{0, 0.01, 0.05, 0.2, 0.5\}$, while λ_{SD} is explored within $[0, 2]$ and λ_{D} is restricted to values below 0.01. As illustrated in the Fig. (6), AFET maintains relatively stable performance over a broad range of λ_{SD} and λ_{D}, demonstrating its robustness and low sensitivity to these hyperparameter choices. In contrast, the performance is comparatively more sensitive to λ_{LCA}, suggesting that its magnitude needs to be carefully controlled. This observation further highlights the importance of latent alignment in the label space. One possible explanation is that overly strong label-space alignment may excessively constrain the learned representations, thereby reducing their flexibility in preserving task-relevant and instance-specific information. Consequently, a moderate alignment strength appears to provide a better trade-off between enforcing structural consistency in the label space and retaining sufficient representational freedom.

Table 3: Module ablation study. M_1, M_2 and M_3 evaluate the contribution by selectively disabling each component. ✔ and ✗ denote the inclusion and exclusion of the corresponding module, respectively.

M_1	M_2	M_3	Enron		NUS-WIDE		Tmc2007	
			AP	Recall	AP	Recall	AP	Recall
✔	✔	✗	0.81	0.66	0.64	0.40	0.72	0.52
✗	✔	✔	0.72	0.70	0.53	0.51	0.73	0.78
✔	✗	✔	0.71	0.72	0.68	0.62	0.78	0.80
✔	✔	✔	**0.85**	**0.79**	**0.72**	**0.74**	**0.85**	**0.85**



(a) Enron. (b) Tmc2007

Figure 6: Parameter sensitivity analysis among $\lambda_{LCA}, \lambda_{SD}$ and λ_D on 2 datasets in one-shot setting.

5 Conclusion

The paper presented AFET, a unified framework for preserving and transferring label evidence under simultaneous incremental and decremental feature evolution. AFET combines label-conditioned latent alignment to retain historical label-discriminative information, structure-aware distillation and an evolving label graph to transfer and adapt inter-label knowledge, and instance-adaptive transformations with online label-wise calibration to accommodate evolving predictive regimes. We further provided a theoretical characterization of prediction stability and extended the evaluation to multi-shot feature evolution. Extensive experiments across diverse benchmark datasets demonstrate the consistent effectiveness and robustness of AFET. In the future, our work will investigate more challenging open-world settings in which augmented features may arise from heterogeneous modalities.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing of the manuscript, including improving grammar, wording, clarity, and readability. We did not use generative AI tools for any other tasks requiring disclosure, and the remaining required disclosure tasks are not applicable to this work. We did not use generative AI tools for other research-related tasks, such as developing research ideas or methods, formulating scientific claims, designing experiments, writing or modifying code, analyzing data, interpreting results, or conducting literature searches. All AI-assisted edits were carefully reviewed by the authors to ensure that they preserved the intended scientific meaning and did not introduce inaccurate or unsupported content. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19530–19540. IEEE, 2025.

Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

ZH Zhou C Hou. One-pass learning with incremental and decremental features. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2017.

Chi-Wai Chow. Recent advances and future perspectives in optical wireless communication, free space optical communication and sensing for 6g. *Journal of Lightwave Technology*, 42(11):3972–3980, 2024.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–9, 2009.

Jiahua Dong, Yang Cong, Gan Sun, Tao Zhang, Xu Tang, and Xiaowei Xu. Evolving metric learning for incremental and decremental features. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2290–2302, 2021.

Robyn Fivush and Elizabeth Slackman. The acquisition and development of scripts. In *Event Knowledge*, pp. 71–96. Routledge, 2026.

Shilin Gu, Yuhua Qian, and Chenping Hou. Incremental feature spaces learning with label scarcity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–26, 2022.

Jun-Yi Hang and Min-Ling Zhang. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis Machine Intelligence*, pp. 9860–9871, 2022.

Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. Learning with feature evolvable streams. *Advances in Neural Information Processing Systems*, 30, 2017.

Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. Prediction with unpredictable feature evolution. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5706–5715, 2021.

Mingdie Jiang, Quanjiang Li, Tingjin Luo, Yiping Song, and Chenping Hou. Multi-label classification with incremental and decremental features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22345–22353, 2026.

Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, 75:2008, 2008.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Quanjiang Li, Tingjin Luo, Mingdie Jiang, Zhangqi Jiang, Chenping Hou, and Feijiang Li. Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18430–18438, 2025a.

Quanjiang Li, Tingjin Luo, and Jiahui Liao. Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20706–20715, 2025b.

Quanjiang Li, Tianxiang Xu, Tingjin Luo, Yan Zhong, Yang Li, Yiyun Zhou, and Chenping Hou. Theory-driven label-specific representation for incomplete multi-view multi-label learning. *Advances in Neural Information Processing Systems*, 38:121996–122037, 2026.

Heng Lian, Di Wu, Bo-Jian Hou, Jian Wu, and Yi He. Online learning from evolving feature spaces with deep variational models. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4144–4162, 2023.

Yanfang Liu, Xiaocong Fan, Wenbin Li, and Yang Gao. Online passive-aggressive active learning for trapezoidal data streams. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (10):6725–6739, 2022.

Zhifeng Liu, Chuanjing Tang, Stanley Ebhohimhen Abhadiomhen, Xiang-Jun Shen, and Yangyang Li. Robust label and feature space co-learning for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11846–11859, 2023.

Tingjin Luo, Quanjiang Li, Mingdie Jiang, and Chenping Hou. Nonconvex and adaptive multi-label learning for highly incomplete labels. *Knowledge-Based Systems*, pp. 114784, 2025.

Yun-Long Mi. Concept neural network based on time-delay regret for dynamic stream learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3796–3814, 2025.

Sai Dikshit Pasham. Privacy-preserving data sharing in big data analytics: A distributed computing approach. *The Metascience*, 1(1):149–184, 2023.

Jinjie Qiu, Shengda Zhuo, Philip S. Yu, Changdong Wang, and Shuqiang Huang. Online learning for noisy labeled streams. *ACM Transactions on Knowledge Discovery from Data*, pp. 1–29, 2025.

Jesse Read. *Scalable Multi-Label Classification*. PhD thesis, University of Waikato, 2010.

Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Multi-label classification using ensembles of pruned sets. In *2008 eighth IEEE international conference on data mining*, pp. 995–1000. IEEE, 2008.

Chongjie Si, Yuheng Jia, Ran Wang, Min-Ling Zhang, Yanghe Feng, and Chongxiao Qu. Multi-label classification with high-rank and high-order label correlations. *IEEE Transactions on Knowledge and Data Engineering*, pp. 4076–4088, 2024.

Ashok N Srivastava and Brett Zane-Ulman. Discovering recurring anomalies in text reports regarding complex space systems. In *2005 IEEE aerospace conference*, pp. 3853–3862. IEEE, 2005.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Jiahang Tu, Shilin Gu, and Chenping Hou. Online imbalance learning with unpredictable feature evolution and label scarcity. *Neurocomputing*, 610:128476, 2024.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, pp. 593–601. PMLR, 2014.

Tingting Zhai and Wei Liu. Online kernel selection for online multi-label classification. *Neural Networks*, pp. 107958, 2025.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022.

Peng Zhang, Hongpeng Yin, Xuanhong Deng, and Sheng-Qing Lv. One-pass online learning under feature evolution data streams with a fast rate. *IEEE Transactions on Knowledge and Data Engineering*, pp. 6075–6090, 2025.

Xingyu Zhao, Yuexuan An, Lei Qi, and Xin Geng. Scalable label distribution learning for multi-label classification. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2025.

Lingli Zhu, Juha Suomalainen, Jingbin Liu, Juha Hyyppä, Harri Kaartinen, Henrik Haggren, et al. A review: Remote sensing sensors. *Multi-Purposeful Application of Geospatial Data*, 19:19–42, 2018.

A Appendix

This supplementary material presents additional details of our proposed method, along with extended experimental results. The content is organized into the following four parts: (1) Related Work; (2) Detailed Method Formulation; (3) Theoretical Analysis; (4) Experimental Results.

A.1 Related Work

A.1.1 Multi-Label Classification

Multi-label classification (MLC) aims to assign multiple relevant labels to a single instance. Unlike conventional single-label classification, MLC must not only identify multiple semantic categories associated with each instance but also model the underlying co-occurrence patterns, dependencies and mutual exclusions among labels. Consequently, effectively exploiting label correlations has long been a central problem in multi-label learning.

Traditional approaches to label-correlation modeling primarily operate from two perspectives, the classifier and the label matrix. From the classifier perspective, features associated with correlated labels are assumed to admit a shared low-rank representation. Based on this assumption, methods such as LEML (Yu et al., 2014) and RLFSCL (Liu et al., 2023) impose various low-rank constraints, including trace-norm regularization, to capture shared structures across labels. From the label-matrix perspective, HOMI (Si et al., 2024) observes that an approximately full-rank label matrix may indicate the presence of label-specific information that cannot be reconstructed from other labels. It therefore explicitly models label self-representation and incorporates local graph-structure embedding to preserve the high-rank characteristics of the label space.

With the increasing prevalence of heterogeneous, incomplete, and structurally complex data sources, MLC has gradually evolved toward more challenging learning scenarios. For large-scale nonlinear classification tasks, OKSOMC (Zhai & Liu, 2025) introduces a novel online kernel-selection strategy for multi-label learning. Compared with conventional methods, neural networks offer greater flexibility in modeling nonlinear and high-order label relationships. However, many existing approaches treat semantic relationships among labels as fixed and immutable priors. To address this limitation, CLIF (Hang & Zhang, 2022) jointly learns label semantics and deep label-specific features for multi-label classification. The learned label semantics guide the extraction of label-specific features, while the discriminative information encoded in these features is propagated back to improve the learning of label semantics. To address asymmetric label relationships and the scalability bottleneck caused by a large number of labels, SLDL (Zhao et al., 2025) represents each label as a distribution in a latent space. This formulation allows label correlations to be asymmetric while making the latent dimensionality independent of the number of labels. Moreover, MLC becomes substantially more challenging in open environments, where the feature space and high-order label relationships may continuously evolve over time. MLID (Jiang et al., 2026) addresses this setting by compressing information contained in disappearing features into the surviving feature space, thereby facilitating the propagation of label dependencies. It further employs low-rank regularization over the classifiers and constructs an adaptive classification mechanism based on augmented features.

As data representations become increasingly semantically expressive, multi-label learning methods have demonstrated considerable potential for handling low-quality data and improving predictive accuracy across complex learning scenarios.

A.1.2 Dynamic Feature Space Learning

Dynamic feature space learning investigates learning problems in which the feature space evolves across time, environments or instances. Conventional supervised learning typically assumes that the training and test data lie in the same fixed feature space, such that all instances are represented using an identical set of feature dimensions. The assumption, however, is frequently violated in real-world applications. Changes in data acquisition devices, feature extraction pipelines, or available modalities may introduce new features, cause existing features to disappear, or result in inconsistent feature views across instances.

Existing studies have primarily investigated three representative scenarios. (1) Trapezoidal incremental data streams, as exemplified by FLLS (Gu et al., 2022), PAA$_{TS}$ (Liu et al., 2022) and OIFL (Tu et al., 2024). Methods in this category typically accommodate newly emerging features by integrating them with classifiers learned from the previous feature space, thereby reusing historical predictive knowledge through ensemble learning. (2) Data streams with overlapping instances, including methods such as FESL (Hou et al., 2017), OLD3S (Liu et al., 2022) and PUFE (Hou et al., 2021). A common strategy is to exploit the overlapping period to learn the correspondence between the old and new feature spaces, which can subsequently be used to reconstruct or compensate for disappearing features. For high-dimensional and nonlinear multimedia data, OLD3S identifies a shared latent subspace that aggregates information from both the old and new feature spaces. By establishing an intermediate feature mapping through this shared representation, it achieves a trade-off between computational efficiency and representational capacity. (3) Data streams with partially overlapping feature spaces, as studied by OPID (C Hou, 2017) and MLID. Under a single-pass learning protocol with sequentially arriving feature blocks, these methods typically reuse features observed during the overlapping period to transfer predictive knowledge and compress information associated with disappearing features. This setting has also been extended to metric learning for sensor data in EML (Dong et al., 2021).

Beyond structural changes in the feature space, real-world online data streams often exhibit additional forms of non-stationarity, including distribution shifts, label noise, and mismatches between the data arrival rate and the model update rate. Consequently, dynamic feature space learning is evolving from isolated feature representation and knowledge-transfer mechanisms toward the joint modeling of complex non-stationary streams, with increasing emphasis on online adaptability, predictive robustness, and update efficiency. ConceptNN (Mi, 2025) addresses the limitations of conventional neural networks when processing high-speed data streams, particularly under settings in which the underlying concepts may evolve over time. To handle distributional discrepancies across instances and noisy labels in streaming data, OLNLS (Qiu et al., 2025) develops an online confidence-learning framework for features associated with noisy supervision. It combines online confidence inference with geometric structure learning to improve robustness across arbitrary data streams and noisy annotations. When data arrive faster than the model can be updated, OLFE-FR (Zhang et al., 2025) introduces the notion of relative rate to adaptively determine whether the model should operate in prediction or update mode at each point in the stream. It further proposes an adaptive learning-rate strategy derived from an upper bound on dynamic regret.

A.2 Additional Methodological Details

This section provides additional formulations and implementation details of the three methodological components introduced in Sec. 2 of the main paper. To avoid redundancy, we only elaborate on details omitted from the main text.

A.2.1 Label-Conditioned Latent-Space Alignment

The first component of AFET learns historical label evidence before feature evolution. The main paper introduces the survived-to-vanished recovery pathway and label-conditioned latent-space alignment. Here we provide the corresponding variational parameterization and recovery objective.

Given the survived features $\mathbf{X}_t^{(s)}$, the teacher employs a variational encoder to parameterize $q_S^T(z \mid \mathbf{X}_t^{(s)}) = \mathcal{N}(\boldsymbol{\mu}_{t,s}^T, \mathrm{diag}((\boldsymbol{\sigma}_{t,s}^T)^2))$. During training, the latent variable is sampled using the reparameterization trick, $z_{t,s}^T = \boldsymbol{\mu}_{t,s}^T + \boldsymbol{\sigma}_{t,s}^T \odot \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For deterministic inference, the posterior mean is used instead. The latent variable is decoded along two complementary pathways. The survived-feature decoder reconstructs $\widehat{\mathbf{X}}_t^{(s),T} = D_s^T(z_{t,s}^T)$, while the vanished-feature decoder produces $\widehat{\mathbf{X}}_t^{(v),T} = D_v^T(z_{t,s}^T)$. The same latent variable is projected to the hidden representation $\mathbf{H}_{t,s}^T$ used in the label-conditioned alignment objective defined in Eq. (1) of the main paper.

The complete historical recovery objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}^{\mathrm{hist},t} = \frac{1}{n_t d_s}\left\|\widehat{\mathbf{X}}_t^{(s),T} - \mathbf{X}_t^{(s)}\right\|_F^2 + \frac{1}{n_t d_v}\left\|\widehat{\mathbf{X}}_t^{(v),T} - \mathbf{X}_t^{(v)}\right\|_F^2 \\ + \beta_{\mathrm{KLD}} D_{\mathrm{KL}}\left(q_S^T(z \mid \mathbf{X}_t^{(s)}) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right),\end{aligned} \tag{15}$$

where d_s and d_v denote the dimensions of the survived and vanished feature spaces, respectively. The first reconstruction term regularizes the survived-conditioned latent representation, whereas the second explicitly learns the $S \to V$ correspondence required after feature evolution.

The complete historical input $[\mathbf{X}_t^{(v)}, \mathbf{X}_t^{(s)}]$ is not processed by another generative model. Instead, AFET uses a discriminative evidence encoder E_{vs}^T to produce $\mathbf{H}_{t,vs}^T = E_{vs}^T([\mathbf{X}_t^{(v)}, \mathbf{X}_t^{(s)}])$. This representation serves as the complete historical reference in the label-conditioned alignment objective and is also used for historical multi-label prediction. No reconstruction objective is imposed on this complete-input branch. In our implementation, E_{vs}^T is a two-layer MLP with BatchNorm, ReLU, and dropout. The survived-feature variational pathway consists of an encoder, two posterior parameter heads for the mean and log-variance, a reconstruction decoder, and a latent-to-hidden projection. The log-variance is clipped to a bounded numerical range before reparameterization for stable optimization. The label-wise predictors introduced in the main paper share the common historical representation but maintain separate decision functions. Each label-specific head is implemented by Dropout–Linear–ReLU–Linear and outputs one scalar logit. The resulting logits are concatenated before applying the historical label-graph propagation defined in the main paper.

A.2.2 Structure-Aware Evidence Transfer

The second component transfers historical predictive evidence after the feature space changes. In addition to the structure-aware distillation and evolving label graph defined in the main paper, AFET explicitly transfers the learned recovery pathway from the historical teacher to the new-stage student.

At the feature-evolution boundary, the student inherits both the complete survived-feature variational pathway and the survived-to-vanished decoder from the teacher:

$$\Theta_{\mathrm{VAE}_S}^S \leftarrow \Theta_{\mathrm{VAE}_S}^T, \qquad \Theta_{D_v}^S \leftarrow \Theta_{D_v}^T. \tag{16}$$

Here, Θ_{VAE_S} includes the survived-feature encoder, posterior parameterization, survived-feature reconstruction decoder, and latent-to-hidden mapping. The transferred parameters are used as initialization rather than a frozen recovery module. They remain trainable during the new stage and are jointly optimized together with the student representation encoder and multi-label classifier. Thus, the student starts from the historical S to V correspondence while retaining the ability to adapt the recovery pathway to the new-stage data distribution.

Given the currently observable survived features, the student employs $q_S^S(z \mid \mathbf{X}_t^{(s)}) = \mathcal{N}(\boldsymbol{\mu}_{t,s}^S, \mathrm{diag}((\boldsymbol{\sigma}_{t,s}^S)^2))$ and generates both the survived-feature reconstruction $\widehat{\mathbf{X}}_t^{(s),S}$ and the vanished-feature surrogate $\widehat{\mathbf{X}}_t^{(v),S}$. Since the true vanished features are inaccessible after feature evolution, they cannot be used as new-stage reconstruction targets. Accordingly, the new-stage variational objective is

$$\mathcal{L}_{\mathrm{VAE}}^{\mathrm{new},t} = \frac{1}{n_t d_s}\left\|\widehat{\mathbf{X}}_t^{(s),S} - \mathbf{X}_t^{(s)}\right\|_F^2 + \beta_{\mathrm{KLD}} D_{\mathrm{KL}}\left(q_S^S(z \mid \mathbf{X}_t^{(s)}) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right). \tag{17}$$

Therefore, no loss term directly supervises $\widehat{\mathbf{X}}_t^{(v),S}$ with inaccessible $\mathbf{X}_t^{(v)}$ after feature evolution. Historical vanished-feature evidence is retained through initialization from the teacher and is further constrained indirectly by supervised prediction and structure-aware distillation.

The historical teacher is frozen throughout the new stage. However, because the original vanished features can no longer be observed, directly evaluating the teacher on its original historical input is impossible. AFET therefore constructs a teacher-side surrogate input using the teacher's own historical recovery pathway.

Given $\mathbf{X}_t^{(s)}$, the frozen teacher first recovers $\widehat{\mathbf{X}}_t^{(v),T}$. Since the teacher operates in evaluation mode, the posterior mean is used instead of stochastic sampling. The teacher-side inference pathway is

$$\widehat{\mathbf{X}}_t^{(v),T} = D_v^T(\boldsymbol{\mu}_{t,s}^T), \qquad \widetilde{\mathbf{X}}_t^{\mathrm{hist},T} = [\widehat{\mathbf{X}}_t^{(v),T}, \mathbf{X}_t^{(s)}], \qquad \mathbf{P}_t^T = F^T\left(\widetilde{\mathbf{X}}_t^{\mathrm{hist},T}; \mathbf{A}^{\mathrm{hist}}\right). \tag{18}$$

Here, $F^T(\cdot)$ denotes the complete frozen historical predictor, including the historical evidence encoder, label-specific prediction heads, historical graph propagation, and sigmoid activation. The resulting $\mathbf{P}_t^T$ provides the teacher-side predictive and relational targets used in Eq. (5) of the main paper. Importantly, this procedure requires neither the inaccessible ground-truth vanished features nor mean-value imputation.

The prediction-induced new-stage graph is constructed from the student's raw predictions after detaching them from the computational graph. Hence, the graph acts as an adaptive structural context for prediction, while gradients do not propagate through the graph-construction operation itself. Gradients are instead propagated through the graph-refined student predictions and the structure-aware distillation objective.

For each historical mini-batch, the teacher is optimized for a fixed number of local updates using Eq. (3) of the main paper, after which the batch is discarded. At the beginning of the new stage, the teacher is frozen and the student is initialized according to Eq. 16. For each incoming new-stage mini-batch, the frozen teacher first generates its surrogate historical prediction according to Eq. 18. This teacher target is computed once and detached from the computation graph. The student then performs its local gradient updates using Eq. (7) of the main paper. All student components, including the inherited recovery pathway, are optimized jointly. We use AdamW for neural parameter optimization and clip the global gradient norm to 5 for numerical stability. Consistent with the one-pass protocol, each streaming mini-batch is discarded after completing its local updates.

A.2.3 Adaptive Prediction Calibration

The third component of AFET combines instance-adaptive representation learning with online label-wise decision calibration. The main paper provides the routing formulation and the threshold update rule. We complement them with the detailed transformation architecture, the explicit routing regularizer, and the update order used in the streaming implementation.

The evidence-enhanced input $\widetilde{\mathbf{X}}_t^{\mathrm{new}}$ is first projected to the shared hidden space using a Linear–BatchNorm–ReLU–Dropout layer. The resulting representation is subsequently processed by K residual MLP blocks. Each block consists of Linear–BatchNorm–ReLU–Dropout–Linear, followed by residual addition, LayerNorm, and ReLU.

Let $\mathbf{H}_t^{(k)}$ denote the representation produced after the k-th block. A scalar routing logit is generated for every available depth. During training, the depth representations are combined through the soft routing distribution defined in Eq. (6) of the main paper. Transformation depths below the predefined minimum depth are masked before the routing softmax, ensuring that each instance receives at least the prescribed minimum amount of nonlinear transformation.

Let $\pi_{t,ik}$ denote the routing probability assigned to depth k for instance i. The expected depth of instance i is $\sum_{k=1}^{K} k\pi_{t,ik}$. AFET normalizes the average expected depth by the maximum available depth and defines

$$\mathcal{L}_{\mathrm{depth}}^t = \frac{1}{n_t K} \sum_{i=1}^{n_t} \sum_{k=1}^{K} k\, \pi_{t,ik}. \tag{19}$$

Minimizing this term discourages the routing network from systematically assigning unnecessary depth. It does not impose a fixed shallow architecture: instances requiring more complex nonlinear transformations can still receive larger routing mass at deeper blocks. In implementation, the depth temperature controls the sharpness of the routing distribution. The same soft depth mixture is used

by default during inference, while the implementation also supports an optional hard early-exit mode based on the learned depth gates.

The label-wise threshold calibration defined in Eq. (8) of the main paper is performed entirely outside gradient-based model optimization. For each new-stage streaming mini-batch, AFET first completes all local student parameter updates. The updated student then performs one additional forward pass on the current mini-batch, from which the label-wise thresholds are updated. Thus, threshold calibration is performed once per streaming mini-batch, rather than once per local optimization epoch. The resulting thresholds are subsequently used for binary multi-label decisions until they are updated by the next incoming batch. No ground-truth test labels are involved in this calibration process. This ordering decouples representation adaptation from decision-level calibration. The former is learned through gradient descent, whereas the latter tracks label-specific changes in predictive score distributions without introducing additional trainable parameters.

A.3 Theoretical Analysis

This section provides detailed theoretical analyses of AFET. We first restate the main theoretical results for completeness and then present their complete proofs. Additional analyses on conditional generalization, sample complexity, and computational complexity are provided in later sections. The theoretical analysis is developed from the perspective of preserving label-related evidence under feature evolution. Specifically, we characterize how vanished-feature recovery, label-conditioned alignment, structure-aware distillation, evolving label dependency modeling, and adaptive threshold calibration contribute to stable adaptation from the historical stage to the new stage.

A.3.1 Preliminary Lemmas

We first introduce several technical lemmas used throughout the theoretical analysis.

Lipschitz Continuity of Multi-Label Classification Loss For a logit matrix $\mathbf{O} \in \mathbb{R}^{n_t \times L}$, the average binary cross-entropy loss is defined as

$$\mathcal{L}_{\mathrm{BCE}}(\mathbf{O}, \mathbf{Y}_t) = \frac{1}{n_t L} \sum_{i=1}^{n_t} \sum_{l=1}^{L} \left[\log(1 + \exp(O_{il})) - Y_{t,il} O_{il} \right].$$

Lemma A.1 (Lipschitz Continuity of Multi-Label BCE Loss)**.** *For any two logit matrices* $\mathbf{O}, \mathbf{O}' \in \mathbb{R}^{n_t \times L}$*, the BCE loss satisfies*

$$|\mathcal{L}_{\mathrm{BCE}}(\mathbf{O}, \mathbf{Y}_t) - \mathcal{L}_{\mathrm{BCE}}(\mathbf{O}', \mathbf{Y}_t)| \leq \frac{1}{\sqrt{n_t L}} \|\mathbf{O} - \mathbf{O}'\|_F . \tag{20}$$

Proof. For an individual prediction-label pair, define $\phi(o, y) = \log(1+\exp(o)) - yo$. The derivative with respect to the logit is $\frac{\partial \phi(o,y)}{\partial o} = \sigma(o) - y$, where $\sigma(\cdot)$ denotes the sigmoid function. Since $\sigma(o) \in [0, 1]$ and $y \in \{0, 1\}$, $\left| \frac{\partial \phi(o,y)}{\partial o} \right| \leq 1$. Therefore, the gradient of the averaged BCE loss satisfies

$$\|\nabla_{\mathbf{O}} \mathcal{L}_{\mathrm{BCE}}(\mathbf{O}, \mathbf{Y}_t)\|_F^2 = \frac{1}{(n_t L)^2} \sum_{i=1}^{n_t} \sum_{l=1}^{L} |\sigma(O_{il}) - Y_{t,il}|^2 \leq \frac{1}{n_t L}. \tag{21}$$

Applying the mean-value inequality gives

$$|\mathcal{L}_{\mathrm{BCE}}(\mathbf{O}, \mathbf{Y}_t) - \mathcal{L}_{\mathrm{BCE}}(\mathbf{O}', \mathbf{Y}_t)| \leq \frac{1}{\sqrt{n_t L}} \|\mathbf{O} - \mathbf{O}'\|_F. \tag{22}$$

□

Lipschitz Continuity of Graph Normalization AFET normalizes label dependency matrices using $\mathrm{Norm}(\mathbf{M}) = \frac{\mathbf{M}}{\|\mathbf{M}\|_2 + \epsilon}$, $\epsilon > 0$, where $\|\cdot\|_2$ denotes the matrix spectral norm.

Lemma A.2 (Lipschitz Continuity of Matrix Normalization)**.** *Assume* $\|\mathbf{M}_1\|_F, \|\mathbf{M}_2\|_F \le G_{\max}$. *Then*

$$\|\mathrm{Norm}(\mathbf{M}_1) - \mathrm{Norm}(\mathbf{M}_2)\|_F \le K_{\mathcal{N}} \|\mathbf{M}_1 - \mathbf{M}_2\|_F, \tag{23}$$

where $K_{\mathcal{N}} = \frac{1}{\epsilon} + \frac{G_{\max}}{\epsilon^2}$.

Proof. Let $a = \|\mathbf{M}_1\|_2$, $b = \|\mathbf{M}_2\|_2$. By adding and subtracting $\mathbf{M}_2/(a+\epsilon)$,

$$\left\| \frac{\mathbf{M}_1}{a+\epsilon} - \frac{\mathbf{M}_2}{b+\epsilon} \right\|_F \le \frac{\|\mathbf{M}_1 - \mathbf{M}_2\|_F}{a+\epsilon} + \|\mathbf{M}_2\|_F \left| \frac{1}{a+\epsilon} - \frac{1}{b+\epsilon} \right|. \tag{24}$$

The first term satisfies

$$\frac{\|\mathbf{M}_1 - \mathbf{M}_2\|_F}{a+\epsilon} \le \frac{1}{\epsilon} \|\mathbf{M}_1 - \mathbf{M}_2\|_F. \tag{25}$$

For the second term,

$$\|\mathbf{M}_2\|_F \left| \frac{1}{a+\epsilon} - \frac{1}{b+\epsilon} \right| = \|\mathbf{M}_2\|_F \frac{|a-b|}{(a+\epsilon)(b+\epsilon)}.$$

Using the reverse triangle inequality, $|a-b| \le \|\mathbf{M}_1 - \mathbf{M}_2\|_F$, and $\|\mathbf{M}_2\|_F \le G_{\max}$, we obtain

$$\|\mathbf{M}_2\|_F \left| \frac{1}{a+\epsilon} - \frac{1}{b+\epsilon} \right| \le \frac{G_{\max}}{\epsilon^2} \|\mathbf{M}_1 - \mathbf{M}_2\|_F. \tag{26}$$

Combining the two terms completes the proof. □

A.3.2 Recovery Stability

We first analyze the influence of vanished-feature recovery on new-stage prediction. Let the oracle complete input be $\mathbf{X}_t^\star = [\mathbf{X}_t^{(v)}, \mathbf{X}_t^{(s)}, \mathbf{X}_t^{(a)}]$, which is unavailable during the new stage. The recovered input constructed by AFET is $\widetilde{\mathbf{X}}_t^{\mathrm{new}} = [\widehat{\mathbf{X}}_t^{(v),\mathrm{S}}, \mathbf{X}_t^{(s)}, \mathbf{X}_t^{(a)}]$. Let $\mathcal{F}_t^{\mathrm{S}}$ denote the complete student mapping from input features to logits.

Theorem A.3 (Recovery-Induced Prediction Stability)**.** *Assume that* $\mathcal{F}_t^{\mathrm{S}}$ *is locally* K_t*-Lipschitz around* $\widetilde{\mathbf{X}}_t^{\mathrm{new}}$ *and* $\mathbf{X}_t^\star$*. Then,*

$$\left| \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}_t^{\mathrm{S}}(\widetilde{\mathbf{X}}_t^{\mathrm{new}}), \mathbf{Y}_t) - \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}_t^{\mathrm{S}}(\mathbf{X}_t^\star), \mathbf{Y}_t) \right| \le \frac{K_t}{\sqrt{n_t L}} \|\widehat{\mathbf{X}}_t^{(v),\mathrm{S}} - \mathbf{X}_t^{(v)}\|_F. \tag{27}$$

Proof. According to Lemma A.1,

$$\begin{aligned} &\left| \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}_t^{\mathrm{S}}(\widetilde{\mathbf{X}}_t^{\mathrm{new}}), \mathbf{Y}_t) - \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}_t^{\mathrm{S}}(\mathbf{X}_t^\star), \mathbf{Y}_t) \right| \\ &\le \frac{1}{\sqrt{n_t L}} \|\mathcal{F}_t^{\mathrm{S}}(\widetilde{\mathbf{X}}_t^{\mathrm{new}}) - \mathcal{F}_t^{\mathrm{S}}(\mathbf{X}_t^\star)\|_F. \end{aligned} \tag{28}$$

Using the local Lipschitz property,

$$\|\mathcal{F}_t^{\mathrm{S}}(\widetilde{\mathbf{X}}_t^{\mathrm{new}}) - \mathcal{F}_t^{\mathrm{S}}(\mathbf{X}_t^\star)\|_F \le K_t \|\widetilde{\mathbf{X}}_t^{\mathrm{new}} - \mathbf{X}_t^\star\|_F. \tag{29}$$

Because the two inputs differ only in the vanished-feature component, $\|\widetilde{\mathbf{X}}_t^{\mathrm{new}} - \mathbf{X}_t^\star\|_F = \|\widehat{\mathbf{X}}_t^{(v),\mathrm{S}} - \mathbf{X}_t^{(v)}\|_F$. Combining the above inequalities proves the theorem. □

A.3.3 Label Evidence Preservation

Feature recovery alone does not necessarily preserve the discriminative information required for multi-label prediction. AFET therefore introduces label-conditioned alignment to explicitly maintain label-related evidence during feature evolution.

For label l, define the positive instance set $\mathcal{I}_{t,l} = \{i : Y_{t,il} = 1\}$. Given teacher representations from the survived-feature pathway and the survived-plus-recovered pathway, define the corresponding label-conditioned prototypes as $\mathbf{c}_{t,l}^{s} = \frac{\sum_{i\in\mathcal{I}_{t,l}} \mathbf{H}_{t,s}^{\mathrm{T}}[i,:]}{|\mathcal{I}_{t,l}|}$ and $\mathbf{c}_{t,l}^{vs} = \frac{\sum_{i\in\mathcal{I}_{t,l}} \mathbf{H}_{t,vs}^{\mathrm{T}}[i,:]}{|\mathcal{I}_{t,l}|}$. The label-conditioned alignment objective is

$$\mathcal{L}_{\mathrm{LCA}}^{t} = \frac{1}{L}\sum_{l=1}^{L} \left\|\mathbf{c}_{t,l}^{s} - \mathbf{c}_{t,l}^{vs}\right\|_2^2. \tag{30}$$

To characterize the influence of LCA, we define the within-label dispersion:

$$s_{t,l}^{s} = \frac{1}{|\mathcal{I}_{t,l}|}\sum_{i\in\mathcal{I}_{t,l}} \left\|\mathbf{H}_{t,s}^{\mathrm{T}}[i,:] - \mathbf{c}_{t,l}^{s}\right\|_2, \tag{31}$$

$$s_{t,l}^{vs} = \frac{1}{|\mathcal{I}_{t,l}|}\sum_{i\in\mathcal{I}_{t,l}} \left\|\mathbf{H}_{t,vs}^{\mathrm{T}}[i,:] - \mathbf{c}_{t,l}^{vs}\right\|_2. \tag{32}$$

Theorem A.4 (Label-Conditioned Prediction Stability)**.** *Assume that the label-specific classifier $f_l(\cdot)$ is K_l-Lipschitz. Then the prediction discrepancy within the positive region of label l satisfies*

$$\frac{1}{|\mathcal{I}_{t,l}|}\sum_{i\in\mathcal{I}_{t,l}} \left|f_l(\mathbf{H}_{t,s}^{\mathrm{T}}[i,:]) - f_l(\mathbf{H}_{t,vs}^{\mathrm{T}}[i,:])\right| \le K_l\left(s_{t,l}^{s} + s_{t,l}^{vs} + \|\mathbf{c}_{t,l}^{s} - \mathbf{c}_{t,l}^{vs}\|_2\right). \tag{33}$$

Proof. For any $i \in \mathcal{I}_{t,l}$, the Lipschitz continuity of f_l gives

$$\left|f_l(\mathbf{H}_{t,s}^{\mathrm{T}}[i,:]) - f_l(\mathbf{H}_{t,vs}^{\mathrm{T}}[i,:])\right| \le K_l\|\mathbf{H}_{t,s}^{\mathrm{T}}[i,:] - \mathbf{H}_{t,vs}^{\mathrm{T}}[i,:]\|_2. \tag{34}$$

The representation discrepancy can be decomposed through the two label prototypes:

$$\|\mathbf{H}_{t,s}^{\mathrm{T}}[i,:] - \mathbf{H}_{t,vs}^{\mathrm{T}}[i,:]\|_2 \le \|\mathbf{H}_{t,s}^{\mathrm{T}}[i,:] - \mathbf{c}_{t,l}^{s}\|_2 + \|\mathbf{c}_{t,l}^{s} - \mathbf{c}_{t,l}^{vs}\|_2 + \|\mathbf{H}_{t,vs}^{\mathrm{T}}[i,:] - \mathbf{c}_{t,l}^{vs}\|_2.$$

Averaging over all positive instances of label l yields

$$\frac{1}{|\mathcal{I}_{t,l}|}\sum_{i\in\mathcal{I}_{t,l}} \left|f_l(\mathbf{H}_{t,s}^{\mathrm{T}}[i,:]) - f_l(\mathbf{H}_{t,vs}^{\mathrm{T}}[i,:])\right| \le K_l(s_{t,l}^{s} + s_{t,l}^{vs} + \|\mathbf{c}_{t,l}^{s} - \mathbf{c}_{t,l}^{vs}\|_2), \tag{35}$$

which completes the proof. □

The proposition reveals that LCA controls the prediction deviation caused by feature disappearance through three complementary factors. The compactness of label-specific regions, the alignment of their centers, and the stability of the recovered representations. Among these terms, only the prototype displacement is explicitly optimized by the LCA objective, while the dispersion terms characterize the intrinsic difficulty of the corresponding label regions.

Corollary A.5 (Prototype-Level Probability Stability)**.** *Let $K_{\max} = \max_l K_l$. Then,*

$$\frac{1}{L}\sum_{l=1}^{L} \left|f_l(\mathbf{c}_{t,l}^{s}) - f_l(\mathbf{c}_{t,l}^{vs})\right|^2 \le K_{\max}^2\mathcal{L}_{\mathrm{LCA}}^{t}. \tag{36}$$

Moreover, after sigmoid transformation,

$$\frac{1}{L}\sum_{l=1}^{L} \left|\sigma(f_l(\mathbf{c}_{t,l}^{s})) - \sigma(f_l(\mathbf{c}_{t,l}^{vs}))\right|^2 \le \frac{K_{\max}^2}{16}\mathcal{L}_{\mathrm{LCA}}^{t}. \tag{37}$$

Proof. By the Lipschitz property of f_l,

$$\left|f_l(\mathbf{c}_{t,l}^{s}) - f_l(\mathbf{c}_{t,l}^{vs})\right| \le K_l\|\mathbf{c}_{t,l}^{s} - \mathbf{c}_{t,l}^{vs}\|_2. \tag{38}$$

Squaring both sides and averaging over labels gives

$$\frac{1}{L}\sum_{l} \left|f_l(\mathbf{c}_{t,l}^{s}) - f_l(\mathbf{c}_{t,l}^{vs})\right|^2 \le \frac{1}{L}\sum_{l} K_l^2\|\mathbf{c}_{t,l}^{s} - \mathbf{c}_{t,l}^{vs}\|_2^2 \le K_{\max}^2\mathcal{L}_{\mathrm{LCA}}^{t}. \tag{39}$$

Since the sigmoid function is $1/4$-Lipschitz, $|\sigma(a) - \sigma(b)| \le \frac{1}{4}|a - b|$, which completes the proof. □

The proposition also explains the mechanism behind the stability guarantee. The prototype alignment term in LCA preserves label-related evidence by reducing the prediction discrepancy.

A.3.4 STRUCTURE-AWARE DISTILLATION CONSISTENCY

AFET transfers historical label evidence through two complementary signals: instance-level label responses and label-level dependency structures. Let $\mathbf{P}_t^{\mathrm{T}}, \mathbf{P}_t^{\mathrm{S}} \in [0,1]^{n_t \times L}$ denote teacher and student prediction matrices. The response-level distillation loss is

$$\mathcal{L}_{\mathrm{KD}}^t = \frac{1}{n_t L} \|\mathbf{P}_t^{\mathrm{S}} - \mathbf{P}_t^{\mathrm{T}}\|_F^2. \tag{40}$$

The corresponding dependency matrices are $\mathbf{G}_t^{\mathrm{T}} = (\mathbf{P}_t^{\mathrm{T}})^\top \mathbf{P}_t^{\mathrm{T}}$, and $\mathbf{G}_t^{\mathrm{S}} = (\mathbf{P}_t^{\mathrm{S}})^\top \mathbf{P}_t^{\mathrm{S}}$.

Theorem A.6 (Response-to-Structure Consistency)**.** *The discrepancy between teacher and student dependency matrices satisfies*

$$\|\mathbf{G}_t^{\mathrm{S}} - \mathbf{G}_t^{\mathrm{T}}\|_F \le 2 n_t L \sqrt{\mathcal{L}_{\mathrm{KD}}^t}. \tag{41}$$

Furthermore, for normalized dependency matrices $\mathbf{R}_t^{\mathrm{S}} = \mathrm{Norm}(\mathbf{G}_t^{\mathrm{S}})$ *and* $\mathbf{R}_t^{\mathrm{T}} = \mathrm{Norm}(\mathbf{G}_t^{\mathrm{T}})$*, we have*

$$\mathcal{L}_{\mathrm{dep}}^t \le 4 K_{\mathcal{N}}^2 n_t^2 L^2 \mathcal{L}_{\mathrm{KD}}^t. \tag{42}$$

Proof. The dependency discrepancy can be decomposed as

$$\mathbf{G}_t^{\mathrm{S}} - \mathbf{G}_t^{\mathrm{T}} = (\mathbf{P}_t^{\mathrm{S}} - \mathbf{P}_t^{\mathrm{T}})^\top \mathbf{P}_t^{\mathrm{S}} + (\mathbf{P}_t^{\mathrm{T}})^\top (\mathbf{P}_t^{\mathrm{S}} - \mathbf{P}_t^{\mathrm{T}}). \tag{43}$$

Using matrix norm submultiplicativity,

$$\|\mathbf{G}_t^{\mathrm{S}} - \mathbf{G}_t^{\mathrm{T}}\|_F \le (\|\mathbf{P}_t^{\mathrm{S}}\|_2 + \|\mathbf{P}_t^{\mathrm{T}}\|_2) \|\mathbf{P}_t^{\mathrm{S}} - \mathbf{P}_t^{\mathrm{T}}\|_F. \tag{44}$$

Because every prediction lies in $[0,1]$,$\|\mathbf{P}_t^{\mathrm{S}}\|_2$, $\|\mathbf{P}_t^{\mathrm{T}}\|_2 \le \sqrt{n_t L}$. Therefore,

$$\|\mathbf{G}_t^{\mathrm{S}} - \mathbf{G}_t^{\mathrm{T}}\|_F \le 2\sqrt{n_t L} \|\mathbf{P}_t^{\mathrm{S}} - \mathbf{P}_t^{\mathrm{T}}\|_F. \tag{45}$$

According to the definition of $\mathcal{L}_{\mathrm{KD}}^t$,

$$\|\mathbf{P}_t^{\mathrm{S}} - \mathbf{P}_t^{\mathrm{T}}\|_F = \sqrt{n_t L \mathcal{L}_{\mathrm{KD}}^t}. \tag{46}$$

Substituting gives

$$\|\mathbf{G}_t^{\mathrm{S}} - \mathbf{G}_t^{\mathrm{T}}\|_F \le 2 n_t L \sqrt{\mathcal{L}_{\mathrm{KD}}^t}. \tag{47}$$

Finally, applying Lemma A.2,

$$\|\mathbf{R}_t^{\mathrm{S}} - \mathbf{R}_t^{\mathrm{T}}\|_F \le K_{\mathcal{N}} \|\mathbf{G}_t^{\mathrm{S}} - \mathbf{G}_t^{\mathrm{T}}\|_F, \tag{48}$$

and squaring both sides proves the second result. □

A.3.5 EVOLVING LABEL-GRAPH STABILITY

Beyond transferring label responses, AFET further preserves historical label dependencies through an evolving graph mechanism. This section analyzes the stability of graph-based prediction under the transition from historical-stage dependencies to new-stage dependencies. Let $\mathbf{Q}_t = \mathbf{O}_t^{\mathrm{S,raw}}$ denote the raw student logits before graph propagation. The graph propagation operator is defined as $\Phi(\mathbf{Q}_t, \mathbf{A}) = \mathbf{Q}_t + \alpha \mathbf{Q}_t \mathbf{A}$, where α controls the contribution of label dependency propagation. AFET constructs the evolving label graph as $\mathbf{A}_t = (1-\eta_t)\mathbf{A}^{\mathrm{hist}} + \eta_t \mathbf{A}_t^{\mathrm{new}}$. Let $\mathbf{A}_t^\star$ denote the oracle label dependency structure under the current new-stage distribution. Define the historical and new-stage graph estimation errors $\epsilon_{\mathrm{hist},t} = \|\mathbf{A}^{\mathrm{hist}} - \mathbf{A}_t^\star\|_F$ and $\epsilon_{\mathrm{new},t} = \|\mathbf{A}_t^{\mathrm{new}} - \mathbf{A}_t^\star\|_F$.

Theorem A.7 (Stability under Label-Graph Evolution)**.** *The perturbation of graph-propagated logits satisfies*

$$\|\Phi(\mathbf{Q}_t, \mathbf{A}_t) - \Phi(\mathbf{Q}_t, \mathbf{A}_t^\star)\|_F \le \alpha \|\mathbf{Q}_t\|_2 \left[(1-\eta_t)\epsilon_{\mathrm{hist},t} + \eta_t \epsilon_{\mathrm{new},t}\right]. \tag{49}$$

Moreover, after sigmoid transformation,

$$\|\sigma(\Phi(\mathbf{Q}_t, \mathbf{A}_t)) - \sigma(\Phi(\mathbf{Q}_t, \mathbf{A}_t^\star))\|_F \le \frac{\alpha}{4} \|\mathbf{Q}_t\|_2 \left[(1-\eta_t)\epsilon_{\mathrm{hist},t} + \eta_t \epsilon_{\mathrm{new},t}\right]. \tag{50}$$

Proof. By the definition of the propagation operator, $\Phi(\mathbf{Q}_t, \mathbf{A}_t) - \Phi(\mathbf{Q}_t, \mathbf{A}_t^\star) = \alpha \mathbf{Q}_t(\mathbf{A}_t - \mathbf{A}_t^\star)$. Using matrix norm submultiplicativity,

$$\|\Phi(\mathbf{Q}_t, \mathbf{A}_t) - \Phi(\mathbf{Q}_t, \mathbf{A}_t^\star)\|_F \leq \alpha \|\mathbf{Q}_t\|_2 \|\mathbf{A}_t - \mathbf{A}_t^\star\|_F. \tag{51}$$

According to the graph fusion rule,

$$\|\mathbf{A}_t - \mathbf{A}_t^\star\|_F = \|(1-\eta_t)(\mathbf{A}^{\text{hist}} - \mathbf{A}_t^\star) + \eta_t(\mathbf{A}_t^{\text{new}} - \mathbf{A}_t^\star)\|_F \leq (1-\eta_t)\epsilon_{\text{hist},t} + \eta_t \epsilon_{\text{new},t}. \tag{52}$$

Combining the above inequalities proves the logit perturbation bound. For the probability-level bound, the derivative of sigmoid is bounded by $\max_z \sigma(z)(1-\sigma(z)) \leq \frac{1}{4}$. Therefore,

$$\|\sigma(\mathbf{U}) - \sigma(\mathbf{V})\|_F \leq \frac{1}{4}\|\mathbf{U} - \mathbf{V}\|_F, \tag{53}$$

which completes the proof. □

The proposition characterizes the preservation and adaptation trade-off introduced by graph evolution. A larger η_t enables faster adaptation toward the new-stage dependency structure, while a smaller η_t retains more historical structural knowledge. Therefore, graph fusion provides a controlled transition mechanism rather than a heuristic interpolation between two independent graphs.

Corollary A.8 (Variance-Optimal Graph Fusion under Unbiased Perturbations)**.** *Assume the graph estimation errors $\mathbf{E}_t^{\text{hist}} = \mathbf{A}^{\text{hist}} - \mathbf{A}_t^\star$ and $\mathbf{E}_t^{\text{new}} = \mathbf{A}_t^{\text{new}} - \mathbf{A}_t^\star$ are zero-mean and mutually uncorrelated. Define $v_{\text{hist},t} = \mathbb{E}[\|\mathbf{E}_t^{\text{hist}}\|_F^2]$, and $v_{\text{new},t} = \mathbb{E}[\|\mathbf{E}_t^{\text{new}}\|_F^2]$. Then the expected graph estimation error $\mathbb{E}[\|\mathbf{A}_t - \mathbf{A}_t^\star\|_F^2]$ is minimized by*

$$\eta_t^\star = \frac{v_{\text{hist},t}}{v_{\text{hist},t} + v_{\text{new},t}}. \tag{54}$$

Proof. Under the assumptions,

$$\mathbb{E}[\|\mathbf{A}_t - \mathbf{A}_t^\star\|_F^2] = \mathbb{E}[\|(1-\eta_t)\mathbf{E}_t^{\text{hist}} + \eta_t \mathbf{E}_t^{\text{new}}\|_F^2] = (1-\eta_t)^2 v_{\text{hist},t} + \eta_t^2 v_{\text{new},t}. \tag{55}$$

Taking the derivative with respect to η_t, $-2(1-\eta_t)v_{\text{hist},t} + 2\eta_t v_{\text{new},t} = 0$. Solving this equation gives

$$\eta_t^\star = \frac{v_{\text{hist},t}}{v_{\text{hist},t} + v_{\text{new},t}}, \tag{56}$$

which completes the proof. □

A.3.6 Dynamic Threshold Tracking

AFET dynamically calibrates label-wise thresholds to accommodate evolving prediction distributions. We next provide a formal analysis of the tracking behavior. For label l, define the population pseudo-positive rate $\pi_{t,l}^\star = \Pr(P_{t,l}^{\text{S}} > \tau_{t-1,l})$. The empirical estimate from the current mini-batch is $\widehat{\pi}_{t,l} = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathbb{I}[P_{t,il}^{\text{S}} > \tau_{t-1,l}]$. The desired pseudo-positive ratio is $r_{t,l}^\star = \text{clip}((1+\rho)\pi_{t,l}^\star + m, r_{\min}, r_{\max})$. The corresponding ideal threshold is $\tau_{t,l}^\star = F_{t,l}^{-1}(1 - r_{t,l}^\star)$, where $F_{t,l}$ denotes the population CDF of prediction scores.

Theorem A.9 (Dynamic Tracking of Label-Wise Thresholds)**.** *Assume that the prediction-score density around the target quantile is lower bounded by $m_l > 0$. Furthermore, assume the threshold drift satisfies $|\tau_{t,l}^\star - \tau_{t-1,l}^\star| \leq \Delta_{t,l}$. Then, with probability at least $1 - \delta_\tau$,*

$$|\widetilde{\tau}_{t,l} - \tau_{t,l}^\star| \leq \xi_{t,l}, \tag{57}$$

where $\xi_{t,l} = \frac{2+\rho}{m_l}\sqrt{\frac{\log(2/\delta_\tau)}{2n_t}}$. For the exponential update $\tau_{t,l} = \gamma\tau_{t-1,l} + (1-\gamma)\widetilde{\tau}_{t,l}$, the tracking error $e_{t,l} = |\tau_{t,l} - \tau_{t,l}^\star|$ satisfies

$$e_{t,l} \leq \gamma e_{t-1,l} + \gamma\Delta_{t,l} + (1-\gamma)\xi_{t,l}. \tag{58}$$

Proof. Conditioned on the previous stage history, $\tau_{t-1,l}$ is fixed with respect to the current mini-batch. Using the Dvoretzky–Kiefer–Wolfowitz inequality,

$$\sup_z |\widehat{F}_{t,l}(z) - F_{t,l}(z)| \leq \sqrt{\frac{\log(2/\delta_\tau)}{2n_t}} \tag{59}$$

with probability at least $1 - \delta_\tau$. Therefore,

$$|\widehat{\pi}_{t,l} - \pi^\star_{t,l}| \leq \sqrt{\frac{\log(2/\delta_\tau)}{2n_t}}. \tag{60}$$

Because the clipping operation is non-expansive,

$$|r_{t,l} - r^\star_{t,l}| \leq (1+\rho)\sqrt{\frac{\log(2/\delta_\tau)}{2n_t}}. \tag{61}$$

The inverse CDF is locally Lipschitz with constant $1/m_l$. Combining the empirical CDF error and target-ratio estimation error yields

$$|\widetilde{\tau}_{t,l} - \tau^\star_{t,l}| \leq \frac{2+\rho}{m_l}\sqrt{\frac{\log(2/\delta_\tau)}{2n_t}}. \tag{62}$$

For the EMA update,

$$\tau_{t,l} - \tau^\star_{t,l} = \gamma(\tau_{t-1,l} - \tau^\star_{t-1,l}) + \gamma(\tau^\star_{t-1,l} - \tau^\star_{t,l}) + (1-\gamma)(\widetilde{\tau}_{t,l} - \tau^\star_{t,l}). \tag{63}$$

Taking absolute values and applying the triangle inequality gives

$$e_{t,l} \leq \gamma e_{t-1,l} + \gamma\Delta_{t,l} + (1-\gamma)\xi_{t,l}. \tag{64}$$

□

Corollary A.10 (Uniform Threshold Tracking Bound)**.** *Suppose that for all stages, $\Delta_{t,l} \leq \Delta_l$, and $\xi_{t,l} \leq \xi_l$. Then,*

$$e_{t,l} \leq \gamma^t e_{0,l} + \frac{\gamma}{1-\gamma}\Delta_l + \xi_l. \tag{65}$$

Proof. Unrolling the recursion in Theorem S2 gives

$$e_{t,l} \leq \gamma^t e_{0,l} + \sum_{s=1}^{t} \gamma^{t-s}(\gamma\Delta_l + (1-\gamma)\xi_l). \tag{66}$$

The geometric series satisfies

$$\sum_{s=1}^{t} \gamma^{t-s} \leq \frac{1}{1-\gamma}. \tag{67}$$

Therefore,

$$e_{t,l} \leq \gamma^t e_{0,l} + \frac{\gamma}{1-\gamma}\Delta_l + \xi_l, \tag{68}$$

which completes the proof. □

A.3.7 Conditional Generalization Analysis

The previous analyses establish the stability of individual components in AFET. We further provide a conditional generalization characterization to quantify how finite-sample estimation, vanished-feature recovery, and label-graph approximation affect new-stage prediction.

We emphasize that this analysis does not attempt to provide an end-to-end optimization guarantee for the complete non-convex framework. Instead, conditioned on the learned representation and adaptation mechanisms, we characterize the statistical behavior of the final prediction layer.

Assumption A.11 (Conditional New-Stage Sampling)**.** Conditioned on the historical observations and the learned recovery and representation mappings, samples within the current new stage are independently drawn from the new-stage distribution $\mathcal{D}_t$.

Assumption A.12 (Bounded Representation and Classifier)**.** The recovered representation $\widehat{\mathbf{h}}$ and oracle representation $\mathbf{h}^\star$ satisfy $\|\widehat{\mathbf{h}}\|_2, \|\mathbf{h}^\star\|_2 \le \Lambda$. The classifier parameter satisfies $\|\mathbf{W}\|_F \le M$. The label dependency matrices satisfy $\|\mathbf{A}\|_2$, $\|\mathbf{A}_t^\star\|_2 \le a$.

Define the graph propagation matrix $\mathbf{B}(\mathbf{A}) = \mathbf{I}_L + \alpha\mathbf{A}$ and let $\kappa = 1 + \alpha a$. The oracle population risk is defined as

$$R_t^\star(\mathbf{W}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_t}\left[\ell(f_{\mathbf{W},\mathbf{A}_t^\star}(\mathbf{h}^\star), \mathbf{y})\right]. \tag{69}$$

The empirical risk evaluated with recovered representations is

$$\widehat{R}_t(\mathbf{W}, \mathbf{A}) = \frac{1}{n_t}\sum_{i=1}^{n_t} \ell(f_{\mathbf{W},\mathbf{A}}(\widehat{\mathbf{h}}_i), \mathbf{y}_i). \tag{70}$$

We define the representation discrepancy $\Delta_{\psi,t} = \mathbb{E}[\|\widehat{\mathbf{h}} - \mathbf{h}^\star\|_2]$, and graph discrepancy $\Delta_{A,t} = \|\mathbf{A} - \mathbf{A}_t^\star\|_2$.

Theorem A.13 (Conditional Generalization Bound)**.** *Under Assumptions A.11 and A.12, for any $\delta_g \in (0,1)$, with probability at least $1-\delta_g$,*

$$R_t^\star(\mathbf{W}) \le \widehat{R}_t(\mathbf{W}, \mathbf{A}) + \frac{2\sqrt{2}\rho_\ell\kappa\Lambda M}{\sqrt{n_t L}} + 3B\sqrt{\frac{\log(2/\delta_g)}{2n_t}} + \frac{\rho_\ell\kappa M}{\sqrt{L}}\Delta_{\psi,t} + \frac{\rho_\ell\alpha\Lambda M}{\sqrt{L}}\Delta_{A,t}. \tag{71}$$

Proof. The proof consists of statistical estimation error, representation mismatch and graph mismatch.

Step 1: Statistical estimation error. For any admissible graph, $\|\mathbf{B}(\mathbf{A})\|_2 \le 1+\alpha a = \kappa$. Therefore, $\|\mathbf{W}\mathbf{B}(\mathbf{A})\|_F \le \kappa M$. The empirical Rademacher complexity of the induced classifier class satisfies

$$\mathfrak{R}_{n_t} \le \frac{\sqrt{2}\rho_\ell\kappa\Lambda M}{\sqrt{n_t L}}. \tag{72}$$

Applying the standard Rademacher generalization inequality gives

$$R_t(\mathbf{W}, \mathbf{A}) \le \widehat{R}_t(\mathbf{W}, \mathbf{A}) + \frac{2\sqrt{2}\rho_\ell\kappa\Lambda M}{\sqrt{n_t L}} + 3B\sqrt{\frac{\log(2/\delta_g)}{2n_t}}. \tag{73}$$

Step 2: Representation mismatch. The loss function is ρ_ℓ-Lipschitz. Therefore,

$$|\ell(f_{\mathbf{W},\mathbf{A}_t^\star}(\mathbf{h}^\star), \mathbf{y}) - \ell(f_{\mathbf{W},\mathbf{A}_t^\star}(\widehat{\mathbf{h}}), \mathbf{y})| \le \frac{\rho_\ell}{\sqrt{L}}\|(\mathbf{h}^\star - \widehat{\mathbf{h}})^\top\mathbf{W}\mathbf{B}(\mathbf{A}_t^\star)\|_2. \tag{74}$$

Using matrix norm submultiplicativity,

$$\frac{\rho_\ell}{\sqrt{L}}\|(\mathbf{h}^\star - \widehat{\mathbf{h}})^\top\mathbf{W}\mathbf{B}(\mathbf{A}_t^\star)\|_2 \le \frac{\rho_\ell\kappa M}{\sqrt{L}}\|\mathbf{h}^\star - \widehat{\mathbf{h}}\|_2. \tag{75}$$

Taking expectation yields $\frac{\rho_\ell\kappa M}{\sqrt{L}}\Delta_{\psi,t}$.

Step 3: Graph mismatch. Similarly,

$$|\ell(f_{\mathbf{W},\mathbf{A}_t^\star}(\widehat{\mathbf{h}}), \mathbf{y}) - \ell(f_{\mathbf{W},\mathbf{A}}(\widehat{\mathbf{h}}), \mathbf{y})| \le \frac{\rho_\ell\alpha\Lambda M}{\sqrt{L}}\|\mathbf{A} - \mathbf{A}_t^\star\|_2. \tag{76}$$

Combining the three components completes the proof. □

Corollary A.14 (Recovery Error Induces Representation Error)**.** *Assume that the representation mapping is $K_{\psi,t}$-Lipschitz. Then*

$$\Delta_{\psi,t} \le K_{\psi,t}\Delta_{\mathrm{rec},t}, \tag{77}$$

where $\Delta_{\mathrm{rec},t} = \mathbb{E}[\|\widehat{\mathbf{X}}_t^{(v),\mathrm{S}} - \mathbf{X}_t^{(v)}\|_2]$.

Proof. The oracle and recovered inputs differ only in the vanished-feature component. Therefore,

$$\|\widehat{\mathbf{h}} - \mathbf{h}^\star\|_2 = \|\psi_t(\widehat{\mathbf{X}}_t) - \psi_t(\mathbf{X}_t^\star)\|_2 \le K_{\psi,t}\|\widehat{\mathbf{X}}_t^{(v),\mathrm{S}} - \mathbf{X}_t^{(v)}\|_2. \tag{78}$$

Taking expectation completes the proof. □

A.3.8 Threshold Calibration Sample Requirement

The conditional generalization theorem provides a sufficient condition on the number of new-stage samples required to achieve a target risk level. Define $C_t = \frac{2\sqrt{2}\rho_\ell \kappa \Lambda M}{\sqrt{L}} + 3B\sqrt{\frac{\log(2/\delta_g)}{2}}$, $B_t = \frac{\rho_\ell \kappa M}{\sqrt{L}}\Delta_{\psi,t} + \frac{\rho_\ell \alpha \Lambda M}{\sqrt{L}}\Delta_{A,t}$. Then Theorem S3 becomes $R_t^\star \le \widehat{R}_t + \frac{C_t}{\sqrt{n_t}} + B_t$.

Corollary A.15 (Sufficient New-Stage Sample Size)**.** *Suppose* $\widehat{R}_t \le \epsilon_{\text{emp}}$ *and* $\epsilon_{\text{tar}} > \epsilon_{\text{emp}} + B_t$. *Then it is sufficient that*

$$n_t \ge \left\lceil \frac{C_t^2}{(\epsilon_{\text{tar}} - \epsilon_{\text{emp}} - B_t)^2} \right\rceil \tag{79}$$

to guarantee $R_t^\star \le \epsilon_{\text{tar}}$.

Proof. The result follows by requiring

$$\frac{C_t}{\sqrt{n_t}} \le \epsilon_{\text{tar}} - \epsilon_{\text{emp}} - B_t. \tag{80}$$

Solving for n_t yields the result. □

The threshold tracking analysis can also be converted into a finite-sample requirement. Using Corollary A.10, for all labels simultaneously, the threshold error satisfies

$$\max_l |\tau_{t,l} - \tau_{t,l}^\star| \le \frac{2+\rho}{m_{\min}}\sqrt{\frac{\log(2L/\delta_\tau)}{2n_t}}. \tag{81}$$

Therefore, to guarantee a threshold precision ϵ_τ,

$$n_t \ge \left\lceil \frac{(2+\rho)^2 \log(2L/\delta_\tau)}{2m_{\min}^2 \epsilon_\tau^2} \right\rceil. \tag{82}$$

A.3.9 Computational Complexity Analysis

We finally analyze the computational complexity of AFET. Let $d_{\text{full}} = d_v + d_s + d_a$ denote the complete feature dimension. Let r denote the VAE latent dimension, h the hidden representation dimension, K the number of residual transformation blocks, and L the number of labels.

Historical Stage Complexity. The main computational components include:

- VAE encoding: $\mathcal{O}(n_t d_s r)$.
- Reconstruction of survived and vanished features $\mathcal{O}(n_t r(d_s + d_v))$.
- Feature projection: $\mathcal{O}(n_t d_{\text{hist}} h)$.
- Label-conditioned prediction: $\mathcal{O}(n_t L h)$.
- Label dependency construction: $\mathcal{O}(n_t L^2)$.

Therefore, the historical-stage complexity is $\mathcal{O}(n_t[d_s r + r(d_s + d_v) + d_{\text{hist}} h + Lh + L^2])$.

New Stage Complexity. The student model contains:

- Feature recovery: $\mathcal{O}(n_t r(d_s + d_v))$
- Evidence-enhanced projection: $\mathcal{O}(n_t d_{\text{full}} h)$.
- Residual transformation: $\mathcal{O}(K n_t h^2)$.
- Label prediction: $\mathcal{O}(n_t L h)$.
- Graph construction and propagation: $\mathcal{O}(n_t L^2)$.

The total new-stage complexity is $\mathcal{O}(n_t[r(d_s + d_v) + d_{\text{full}}h + Kh^2 + Lh + L^2])$.

Space Complexity. During new-stage training, AFET stores teacher and student parameters, current mini-batch representations, label dependency matrices and label-wise thresholds. Therefore, the memory complexity is $\mathcal{O}(P_{\text{T}} + P_{\text{S}} + n_t h + n_t L + L^2)$.

Importantly, AFET does not store historical instances. Thus, unlike replay-based approaches, the memory requirement does not grow with the number of previously observed samples. If the label graph is sparsified with at most q neighbors per label, the graph-related complexity can be reduced from $\mathcal{O}(n_t L^2)$ to $\mathcal{O}(n_t L q)$.

A.4 Experimental Results

This section provides additional experimental details and reports the remaining results. The detailed statistics of the experimental datasets are summarized in Table 4, which includes the dataset name (Data), the number of label categories (c), the number of samples in the first stage ($\sum_{i=1}^{T_1} n_i$), the number of samples in each batch (n_i) and the feature dimensionality of each part (d_v, d_s, d_a). We conducted experiments on datasets spanning seven distinct application domains. Specifically, Scene and NUS-WIDE represent visual multimedia data, while the others pertain to text classification tasks. To comprehensively evaluate the performance of the proposed approach in comparison with baseline methods, we adopted six evaluation metrics.

Data	c	$\sum_{i=1}^{T_1} n_i$	n_i	d_v	d_s	d_a
Languagelog	75	689	15, 30, 60, 120	251	502	251
Enron	53	833	20, 40, 80, 160	250	500	251
Scene	6	1179	30, 60, 120, 240	73	147	74
Rcv1v2(subset1)	101	2940	60, 120, 240, 480	25	51	25
Bibtex	159	3623	85, 175, 350, 700	296	593	297
Tmc2007	22	14012	50, 100, 200, 400	12265	24530	12265
NUS-WIDE	81	134824	500, 1000, 2000, 4000	32	64	32

Table 4: Details about experimental setting in one-shot case.

- **Recall:** Recall quantifies the proportion of relevant labels correctly predicted by the model with respect to all ground-truth relevant labels. It reflects the model's ability to comprehensively capture the full set of relevant labels. $\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i|}$, where N denotes the total number of samples. Y_i represents the ground-truth label set and $\hat{Y}_i$ denotes the predicted label set for the i-th sample.
- **The Area Under the Curve (AUC):** AUC is a ranking-based metric that quantifies a model's capability to consistently rank relevant labels ahead of irrelevant ones. For evaluation, the ROC curve can be computed individually for each label in a one-vs-rest manner, followed by calculating the corresponding AUC. The results can then be aggregated using macro or micro-averaging strategies to provide an overall performance summary. $\text{AUC} = \frac{1}{L}\sum_{j=1}^{L}\text{AUC}_j$, where L is the number of labels. AUC_j is the ROC-AUC value for the j-th label, which is typically computed based on the ranking of positive and negative sample pairs.
- **Average Precision (AP):** AP evaluates a model's capability to position the correct labels at the top of the predicted ranking. It is computed as a weighted mean of the precision scores measured at multiple cutoff thresholds along the ranking. $\text{AP}_i = \frac{1}{|Y_i|}\sum_{k=1}^{|L|} P_i(k)\cdot\text{rel}_i(k)$. $\text{MAP} = \frac{1}{N}\sum_{i=1}^{N}\text{AP}_i$, where $P_i(k)$ is the precision at the top k predicted labels for the i-th sample; $\text{rel}_i(k) \in \{0, 1\}$ indicates whether the k-th predicted label is a true label for the i-th sample. MAP is the mean value averaged over all samples.
- **Normalized Discounted Cumulative Gain (NDCG):** NDCG assesses the quality of a model's label ranking, with an emphasis on placing highly relevant labels at the top positions. It is widely adopted in multi-label ranking tasks due to its ability to account for the graded relevance of labels and the position of correct predictions within the ranking. $\text{DCG}_i = \sum_{k=1}^{|L|}\frac{2^{\text{rel}_i(k)}-1}{\log_2(k+1)}$. $\text{NDCG}_i = \frac{\text{DCG}_i}{\text{IDCG}_i}$. $\text{NDCG} = \frac{1}{N}\sum_{i=1}^{N}\text{NDCG}_i$, where $\text{rel}_i(k) \in \{0, 1\}$ indicates whether the k-th label is relevant for the i-th sample. IDCG_i is the DCG under the ideal (perfect) ranking for the i-th sample.
- **Hamming Loss (HL):** HL measures the degree of inconsistency between predicted labels and true labels. It is the average of the proportion of incorrect labels (including mul-

tiple predictions and missed predictions). A lower value indicates better performance. Hamming Loss $= \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbf{1}[y_{ij} \neq \hat{y}_{ij}]$, where $y_{ij} \in \{0, 1\}$ indicates whether sample i is associated with label j. $\hat{y}_{ij}$ denotes the corresponding prediction result.

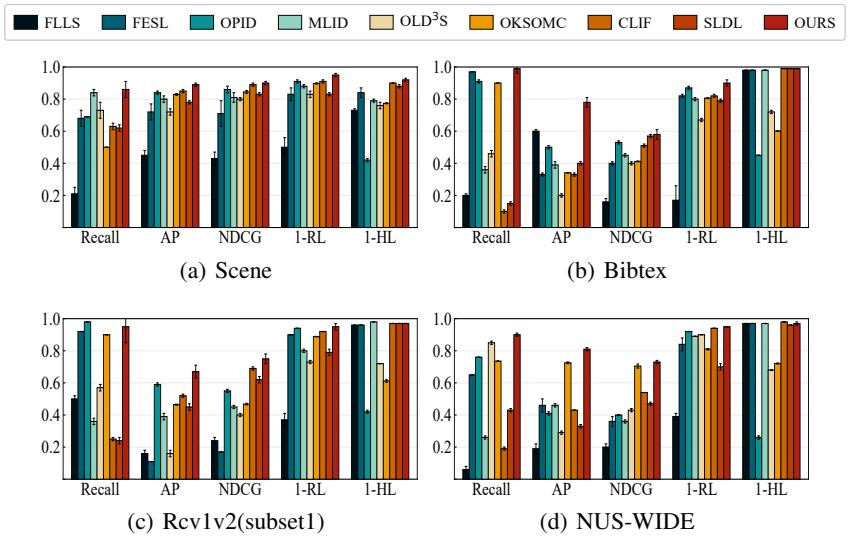
- **Ranking Loss (RL):** RL quantifies the fraction of label pairs in which a relevant label is ranked lower than an irrelevant one. Lower values indicate better performance, with an ideal score of zero representing a perfectly ordered ranking. Ranking Loss $= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|Y_i||\bar{Y}_i|} |\{(y^+, y^-) \mid f_i(y^+) \leq f_i(y^-)\}|$, where Y_i is the set of relevant labels, and $\bar{Y}_i$ is the set of irrelevant labels for sample i. $f_i(y)$ is the predicted score of label y for sample i. The more incorrectly ordered label pairs there are, the higher RL will be.

A.4.1 Comparative Study

We compare AFET with eight state-of-the-art methods.

- FLLS proposes a novel incremental feature-space sparsity learning framework, in which an objective function incorporating a truncated sparsification ratio $B \in \{0.25, 0.5, 0.75\}$ and regularization parameters $\lambda \in 10^{[-2:2]}$ is designed to effectively handle incremental data.
- FESL learns a mapping during the overlapping period to reconstruct the vanished features, and subsequently trains two separate models based on the recovered features and the current feature space, respectively. We set the step size τ_t to be $1/c\sqrt{t}$ where c is searched in the range $\{1, 10, 50, 100, 150\}$.
- OPID proposes compressing the informative content of vanished features into a functional representation over the survived features, which is subsequently extended to incorporate augmented features for addressing one-pass multi-class classification tasks. The hyperparameters are in the range of $\{1, 2, \cdots, 9\}$.
- OLD^{3}S treats model capacity as learnable semantics and dynamically adapts the model architecture to the complexity and nonlinearity of incoming data streams, jointly optimizing the model depth and parameters in an online manner. Set $\eta = 8\sqrt{1/lnT}$ and $\beta = 0.9$, while p and $\alpha^{(l)}$ are adaptively updated during training.
- MLID seeks to compress the informational content of vanished features into the domain of the survived features, facilitates the propagation of label dependencies via low-rank regularization imposed on the classifier, and further integrates augmented features to construct an adaptive classification mechanism. All hyper-parameters are set to the recommended value of 0.1 as specified in the original paper.
- OKS-OMC decomposes the joint optimization of multi-label kernel classifiers and their combination coefficients into two tractable subproblems, which can be efficiently solved in an incremental manner while admitting a meaningful regret bound. Search for the optimal η within [0.25,1] and select λ from a range of $\{0.001, 0.01, 0.1\}$.
- CLIF integrates a graph autoencoder that captures semantic dependencies in the label space with a dedicated feature decoding module that extracts label-specific representations. All hyper-parameters are set to the values recommended in the original paper.
- SLDL learns a mapping from the feature space to the latent space, rendering the computational complexity independent of the number of labels. Finally, a nearest-neighbor-based decoding strategy is employed to recover the final predictions from the learned latent representations. Tune $\hat{c}$ within the range $[64, 96]$.

One-shot Case Fig. (7) presents a comparative evaluation of AFET and competing methods across different datasets and evaluation metrics. AFET exhibits consistently strong performance across the four datasets when n_i is fixed to the minimum batch number. On Scene, Bibtex, and Rcv1v2(subset1), AFET achieves leading or near-leading results across most of the five evaluation metrics, with particularly clear gains in AP, NDCG, and 1-RL. This indicates that AFET is effective not only in predicting positive labels but also in preserving their relative ranking under feature-space evolution. On NUS-WIDE, the competing methods show larger metric-dependent variations; although AFET is not uniformly superior on every individual metric, it remains highly competitive on Recall, NDCG, 1-RL, and 1-HL, resulting in a more balanced overall performance. Taken

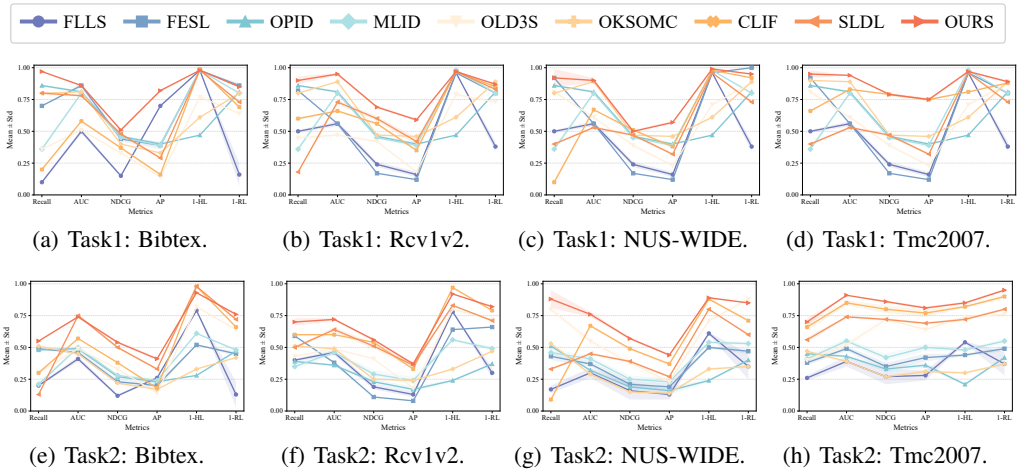


(a) Scene (b) Bibtex

(c) Rcv1v2(subset1) (d) NUS-WIDE

Figure 7: Performance of various metrics across methods and datasets with n_i fixed as the minimum batch number.

together, the results demonstrate that the proposed feature-evidence recovery and structure-aware transfer mechanisms remain effective even under the minimum-batch setting, where the amount of new-stage data available for online adaptation is relatively limited.



(a) Task1: Bibtex. (b) Task1: Rcv1v2. (c) Task1: NUS-WIDE. (d) Task1: Tmc2007.

(e) Task2: Bibtex. (f) Task2: Rcv1v2. (g) Task2: NUS-WIDE. (h) Task2: Tmc2007.

Figure 8: Performance of various methods in predicting $\hat{\mathbf{Y}}_{T+2}$ and $\hat{\mathbf{Y}}_{T_i+t+2}$ on other datasets with n_i fixed as the minimum batch number in two task.

Multi-shot Case As shown in Fig. (8), AFET remains consistently competitive under both multi-shot prediction tasks, demonstrating that its effectiveness extends beyond a single feature-evolution event. In Task 1, while several competing methods exhibit pronounced metric-dependent fluctuations, AFET maintains leading or near-leading performance across Bibtex, Rcv1v2, NUS-WIDE, and Tmc2007. This suggests that the survived-feature bridge can effectively carry historical predictive evidence across successive feature-space transitions. Rather than relying solely on feature reconstruction, AFET reconstructs vanished evidence from survived features and further preserves both label-wise responses and inter-label dependencies through structure-aware distillation. The advantage becomes more consistent in Task 2, where the effect of successive feature changes is more strongly accumulated, particularly on NUS-WIDE and Tmc2007. In this more challenging setting, the evolving label graph continuously adapts historical dependencies to the current prediction structure, while instance-adaptive transformations and online label-wise threshold calibration

Table 5: Extended scenario ablation study on the remaining four datasets. ✔and ✗denote whether the corresponding information source is enabled or disabled, respectively.

Scenario	History		New			Languagelog		Scene		Bibtex		Rcv1v2(subset1)	
	V	S	$\hat{V}$	S	A	AP	Recall	AP	Recall	AP	Recall	AP	Recall
S_1	✗	✔	✗	✔	✗	0.52	0.65	0.77	0.55	0.26	0.66	0.47	0.70
S_2	✗	✗	✗	✔	✔	0.52	0.46	0.78	0.56	0.47	0.78	0.48	0.61
S_3	✗	✗	✗	✗	✔	0.59	0.52	0.57	0.53	0.23	0.72	0.44	0.78
S_4	✗	✔	✗	✔	✔	0.63	0.67	0.78	0.57	0.28	0.79	0.49	0.82
S_5	✔	✔	✔	✔	✗	0.68	0.76	0.82	0.66	0.54	0.87	0.53	0.81
S_6	✔	✔	✔	✔	✔	**0.75**	**0.78**	**0.83**	**0.79**	**0.65**	**0.90**	**0.64**	**0.87**

mitigate representation and score-distribution shifts. These results indicate that AFET can progressively transfer and adapt label-relevant evidence across multiple evolution stages, thereby reducing the degradation caused by accumulated feature-space drift.

A.4.2 Ablation Study

Scenario Ablation The scenario-level ablation study is designed to examine a central question of feature-evolving MLC, which is the information observable at the current stage sufficient for prediction, or is it necessary to preserve and transfer predictive evidence accumulated before feature evolution? To answer this question, we construct six scenarios by systematically controlling the historical information available at the historical stage and the feature evidence used for prediction after evolution.

Let V, S, and A denote the vanished, survived, and augmented feature subsets, respectively, and let $\widehat{V}$ denote the surrogate vanished evidence recovered from S. (1) Shared survived only (S_1). This scenario retains only the survived features throughout both stages. This setting evaluates whether the explicitly shared feature subspace by itself can provide sufficient information to bridge the two stages. (2) New observable only (S_2). This scenario completely removes historical-stage learning and trains the new-stage learner directly from the currently observable features $[S, A]$. It therefore serves as the most direct current-stage baseline and tests whether simply training on all features available after evolution can replace the use of historical information. (3) Augmented only (S_3). To further isolate the contribution of newly introduced information, this scenario discards both historical learning and the survived features and uses only A for new-stage prediction. Comparing S_3 with S_2 reveals how much the survived features contribute in addition to the newly emerged features when no information is transferred from the historical stage. (4) No vanished evidence (S_4). This scenario uses S during the historical stage and $[S, A]$ during the new stage, but never accesses V or $\widehat{V}$. In contrast to S_2, the learner has already acquired historical knowledge from the shared survived features before evolution. Therefore, the comparison between S_2 and S_4 directly evaluates whether information accumulated at the previous stage remains beneficial even when vanished-feature evidence is completely removed. (5) Inherited vanished features (S_5). This scenario trains the historical model from the complete historical feature space $[V, S]$ and transfers the learned survived-to-vanished recovery pathway to the new stage. The new-stage predictor uses $[\widehat{V}, S]$, while the augmented features A are excluded. Here, the survived features are still required because they provide the cross-stage anchor from which $\widehat{V}$ is recovered. Comparing S_5 with S_1 quantifies the additional benefit of preserving information associated with the vanished features beyond the shared survived features alone. (6) Full evolution (S_6). The complete AFET configuration combines the historical evidence learned from $[V, S]$ with both the recovered vanished evidence and the newly available feature evidence, resulting in the new-stage representation $[\widehat{V}, S, A]$. This scenario therefore simultaneously exploits historical, cross-stage shared, and newly emerged information.

To ensure that the comparisons primarily reflect differences in available evidence rather than model capacity, all scenarios use the same underlying student architecture and optimization protocol. Feature blocks excluded by a particular scenario are masked with zeros before being passed to the feature encoder, instead of changing the dimensionality of the encoder. Thus, the representation net-

Table 6: Module ablation study on the remaining datasets. M_1, M_2 and M_3 denote label-conditioned latent-space alignment, structure-aware evidence transfer, and adaptive prediction calibration, respectively. ✔ and ✗ indicate whether the corresponding module is enabled or disabled.

M_1	M_2	M3	Languagelog		Scene		Bibtex		Rcv1v2(subset1)	
			AP	Recall	AP	Recall	AP	Recall	AP	Recall
✔	✔	✗	0.69	0.45	0.80	0.55	0.53	0.66	0.60	0.39
✗	✔	✔	0.56	0.54	0.71	0.66	0.42	0.62	0.51	0.70
✔	✗	✔	0.68	0.76	0.78	0.67	0.43	0.82	0.53	0.71
✔	✔	✔	**0.75**	**0.78**	**0.83**	**0.79**	**0.65**	**0.90**	**0.64**	**0.87**

work retains the same input dimensionality and parameterization across scenarios. The handling of historical information is also strictly controlled. For S_1 and S_4, historical learning is performed using S only; no vanished-feature teacher, recovered $\widehat{V}$, or teacher-side evidence transfer is used. For S_2 and S_3, the historical stage is skipped entirely, so no statistics or parameters from the historical data are introduced. For S_5 and S_6, the teacher is trained from $[V, S]$, and the learned survived-feature recovery pathway is transferred to the student before new-stage adaptation. When a configuration does not use $\widehat{V}$, teacher-side transfer signals associated with vanished-feature evidence are also disabled, preventing hidden leakage of historical vanished information into the corresponding baseline. For configurations that use historical data, the label-structural prior is estimated from the historical stage. In contrast, configurations that entirely skip the historical stage construct the corresponding structural information using new-stage training data only, ensuring that the no-history baselines do not indirectly access historical labels. All other non-ablated components and optimization hyperparameters are kept unchanged. Results are shown in Table. 5.

The six scenarios are designed to permit several complementary comparisons. Most importantly, S_2 and S_4 share the same new-stage input $[S, A]$ but differ in whether information has been acquired from the historical stage. Their comparison therefore directly tests the necessity of carrying previous-stage information forward. S_1 and S_5 both exclude augmented features, while S_5 additionally preserves evidence associated with the vanished features; their comparison evaluates the benefit of vanished-evidence inheritance beyond the shared survived features. Comparing S_5 with S_6 isolates the contribution of newly augmented evidence once historical information has already been preserved. Finally, S_4 and S_6 use the same currently observable features $[S, A]$, but S_6 additionally restores and transfers the information associated with V, providing a direct evaluation of the complete historical-evidence inheritance mechanism.

Module Ablation We further investigate the roles of the three major components of AFET by examining their metric- and dataset-specific effects. The ablation results reveal that the three modules exhibit clearly different functional characteristics rather than providing homogeneous performance improvements.

As shown in Table 6,an important observation is that no individual component dominates across all datasets and evaluation metrics. M_1 generally has a strong influence on AP by preserving label-specific historical evidence, M3 produces particularly pronounced improvements in Recall by adapting representations and label-wise decisions, while M_2 provides consistent gains by transferring and updating inter-label relational knowledge. This differentiated behavior indicates that the three components address distinct sources of performance degradation caused by feature evolution. Specifically, M_1 focuses on preserving label-specific historical evidence, M_2 maintains and adapts inter-label relational evidence, and M3 accommodates instance- and label-specific predictive shifts at the new stage. Their contributions are therefore complementary rather than redundant, which explains why the complete AFET model consistently achieves the strongest AP and Recall performance across the three datasets.

We note that the current module-level ablation removes M_3 as a whole. Therefore, the observed performance degradation reflects the combined contribution of instance-adaptive depth transformation and online label-wise threshold calibration, rather than either mechanism in isolation. The strong Recall degradation suggests that decision-level adaptation is particularly important under feature

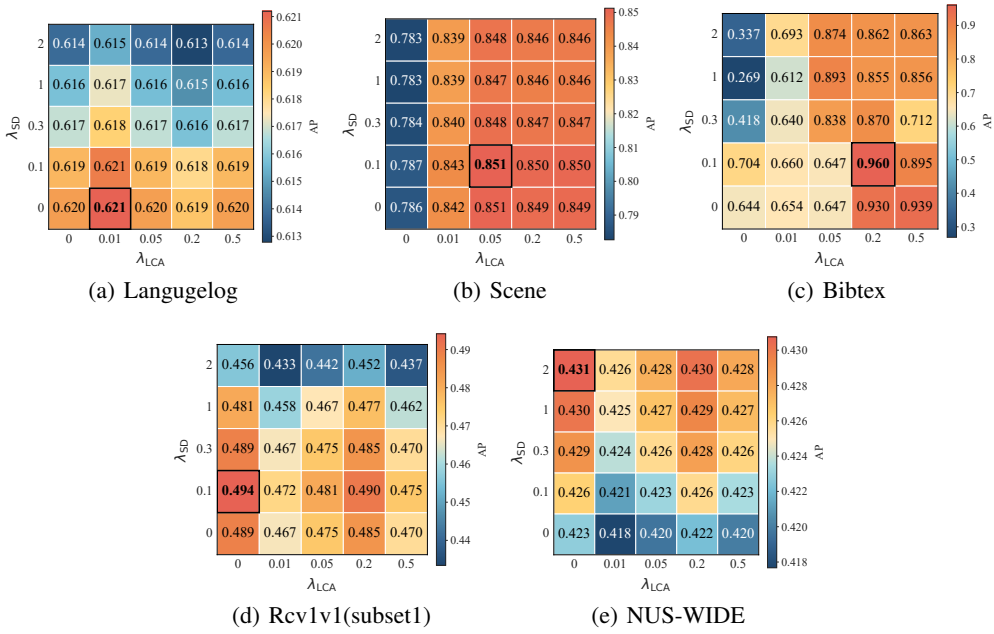


Figure 9: Parameter influence of AFET in one-shot scenario between λ_{LCA} and λ_{SD}.

evolution, while the simultaneous decrease in AP indicates that representation adaptation also contributes to maintaining discriminative predictive scores.

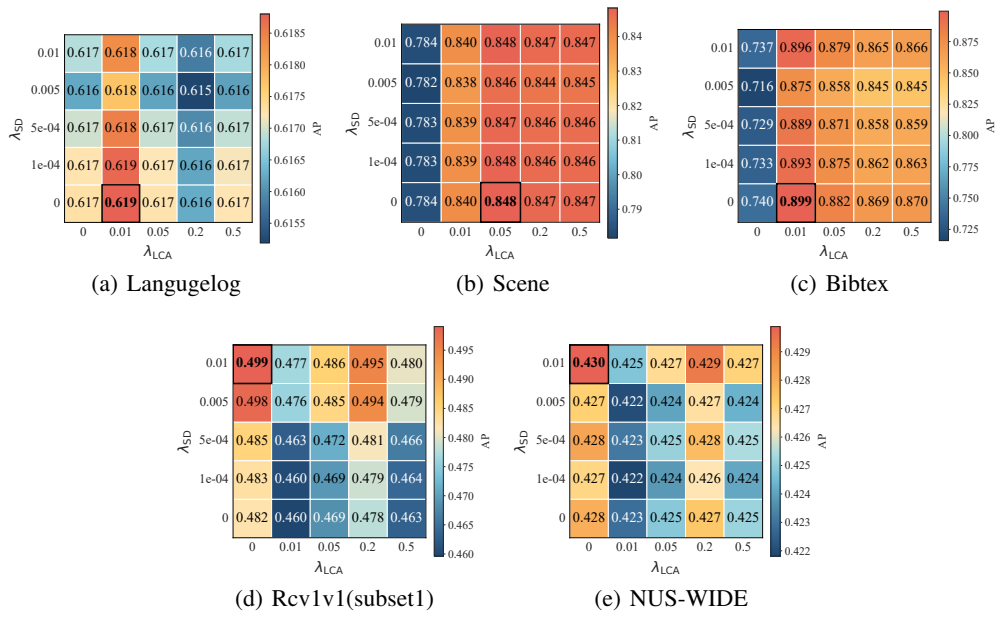
A.4.3 Parameter Sensitivity Analysis

To investigate the sensitivity of AFET to its key hyper-parameters, we conduct a grid search over λ_{LCA} together with λ_{SD} and λ_D, respectively. The corresponding AP results on five representative datasets are illustrated in Fig. (9) and Fig. (10).

Effect of λ_{LCA} and λ_{SD}. As shown in Fig. (9), the influence of λ_{LCA} and λ_{SD} is generally dataset-dependent. On Languagelog and NUS-WIDE, the performance is relatively stable over a wide range of parameter combinations. In particular, the AP on Languagelog varies only from 0.613 to 0.621, while that on NUS-WIDE remains within 0.418 and 0.431, indicating that AFET is not sensitive to two hyper-parameters on these datasets. Scene exhibits moderate variation, where introducing a non-zero λ_{LCA} consistently improves the performance compared with $\lambda_{\text{LCA}} = 0$. The best AP of 0.851 is obtained around $\lambda_{\text{LCA}} = 0.05$ with a relatively small λ_{SD}. A similar but milder trend can be observed on Rcv1v1(subset1), where the best result of 0.494 is achieved at $\lambda_{\text{LCA}} = 0$ and $\lambda_{\text{SD}} = 0.1$.

Bibtex is comparatively more sensitive to the interaction between these two parameters. Its AP varies considerably across different combinations, and the best performance is achieved at $\lambda_{\text{LCA}} = 0.2$ and $\lambda_{\text{SD}} = 0.1$. In contrast, inappropriate combinations, especially those involving a large λ_{SD} together with a weak LCA regularization, can lead to a substantial performance drop. This observation suggests that the relative strengths of the LCA and SD components need to be properly balanced on datasets such as Bibtex.

Effect of λ_{LCA} and λ_D. Fig. (10) further reports the joint influence of λ_{LCA} and λ_D. Compared with λ_{SD}, varying λ_D generally results in smoother performance changes on most datasets. For Languagelog, the AP remains between 0.615 and 0.619, and the best performance is obtained around $\lambda_{\text{LCA}} = 0.01$. NUS-WIDE shows a similarly stable pattern, with AP ranging only from approximately 0.422 to 0.430. Scene also exhibits a relatively broad high-performance region. When λ_{LCA} is set around 0.05, AP remains close to its maximum value of 0.848 under different choices of λ_D.



(a) Langugelog (b) Scene (c) Bibtex

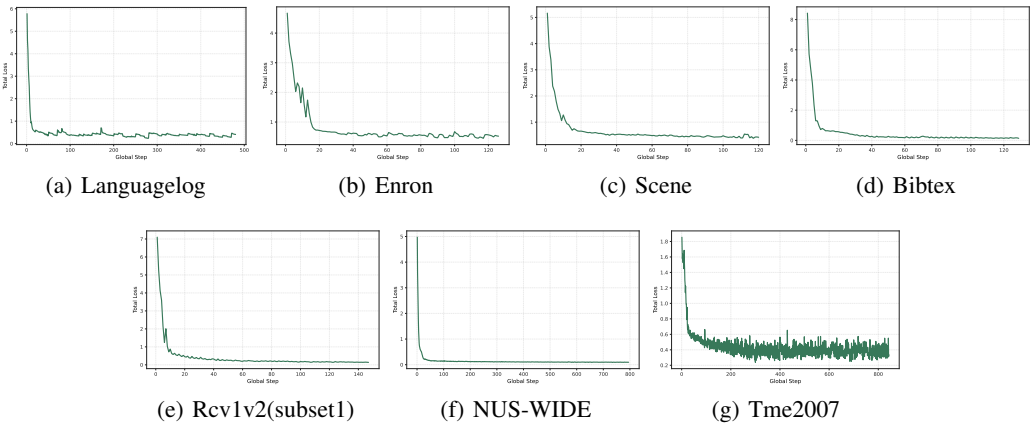
(d) Rcv1v1(subset1) (e) NUS-WIDE

Figure 10: Parameter influence of AFET in one-shot scenario between λ_{LCA} and λ_D.

For Bibtex, the choice of λ_{LCA} has a more pronounced impact than λ_D. The performance is consistently high around $\lambda_{\text{LCA}} = 0.01$, and the highest AP of 0.899 is obtained when $\lambda_D = 0$. On Rcv1v1(subset1), a larger λ_D is generally beneficial, while a small λ_{LCA} is preferred. The best AP reaches 0.499 at $\lambda_{\text{LCA}} = 0$ and $\lambda_D = 0.01$.

Overall, the above results demonstrate that AFET is relatively robust to hyperparameter variations on Languagelog, Scene, and NUS-WIDE, where broad regions of the parameter space yield comparable performance. Rcv1v1(subset1) shows moderate sensitivity, whereas Bibtex exhibits a stronger interaction among different regularization terms and therefore benefits more from dataset-specific parameter tuning. These results also indicate that the different components of AFET play complementary roles, and appropriately balancing their contributions can further improve the performance.

A.4.4 Convergence Study



(a) Languagelog (b) Enron (c) Scene (d) Bibtex

(e) Rcv1v2(subset1) (f) NUS-WIDE (g) Tme2007

Figure 11: Optimization trajectories of $\mathcal{L}_{\text{hist}}$ on seven benchmark datasets.

We further examine the optimization behavior of AFET during the historical stage. Specifically, we record the historical-stage training objective $\mathcal{L}_{\text{hist}}$, which jointly optimizes MLC, variational feature

recovery, and label-conditioned latent-space alignment. Fig. (11) reports the loss trajectories on all seven benchmark datasets.

As shown in Fig. (11), a consistent optimization pattern is observed across datasets with substantially different feature and label characteristics. The objective decreases sharply during the early optimization steps and subsequently enters a stable regime. In particular, the loss on Languagelog, Scene, Bibtex, Rcv1v2(subset1), and NUS-WIDE rapidly approaches a plateau, while Enron exhibits several transient oscillations at the beginning before stabilizing. Tmc2007 shows relatively more persistent mini-batch-level fluctuations after the initial descent; nevertheless, its loss remains confined within a stable range and exhibits no systematic upward trend. Overall, these results provide empirical evidence that the joint historical-stage objective can be optimized stably across diverse multi-label datasets, despite simultaneously involving classification, feature recovery, and label-conditioned alignment terms.