Incremental Document Image Classification: A Comparative Study of Content and Layout Representations

S M Abdul Ahad[0009−0001−8029−1527], Thomas Gorges[0009−0007−0573−0992], Lukas Hüttner[0009−0001−0231−7517], Linda-Sophie Schneider[0009−0002−6195−9859], Mathias Seuret[0000−0001−9153−1031], Fei Wu[0000−0003−4196−0289], and Vincent Christlein[0000−0003−0455−3799]

Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
{abdul.ahad,thomas.gorges,lukas.huettner,linda-sophie.schneider,
mathias.seuret,river.wu,vincent.christlein}@fau.de

Abstract. Document classification has moved from convolutional neural networks to multimodal architectures that jointly encode text, images, and layout. As document streams change over time, retraining models on the full historical data is expensive and often not possible because of storage limits and privacy constraints. Incremental Learning (IL) addresses this by updating models with only the newly available data. In this work, we present a comparative study of IL strategies for document image classification rather than proposing a new method. We benchmark several representative incremental strategies, including standard IL, Extreme Value Machine (EVM)-based variants such as iEVM, and two further configurations evaluated within the same setup: RegEVM, which regularizes the Weibull scale parameters, and EVM+Out-Of-Distribution (OOD), which couples EVM with virtual-logit matching for out-of-distribution rejection. The study covers two settings, Class-Incremental Learning (CIL) and a cross-collection class-incremental setting that we also refer to as Domain-Incremental Learning (DIL), on the RVL-CDIP and Tobacco-3482 datasets. Two representative backbones are evaluated under resource-constrained, although not fully symmetric, experimental conditions. The results show that content-centric features retain prior knowledge more reliably across incremental steps, while the layout-aware backbone shows substantially larger drops under the same protocol. Among the configurations we evaluated, EVM+OOD is the most stable in CIL and behaves comparably to the others in DIL. RegEVM performs comparably to standard EVM; we therefore report it as a negative finding of the comparison rather than as an improvement over the baseline. Overall, the study is intended as a practical reference point for choosing between IL strategies for document classification under realistic compute constraints. At the same time, the results should be interpreted as a resource-constrained benchmark, not as a definitive ranking of document backbones or as a claim of a new state-of-the-art method.

Keywords: Incremental Learning · Extreme Value Machine · Document Image Classification

1 Introduction

A core challenge in modern machine learning is enabling models to adapt to evolving data distributions without degrading previously acquired knowledge. Standard neural networks are typically trained in a static setting and can exhibit Catastrophic Forgetting (CF) when updated on new data. To mitigate this limitation, incremental (or continual) learning has emerged as a paradigm that allows models to integrate new information without requiring retraining on the full dataset.

In document image analysis, deep learning has led to substantial performance gains. For example, Ensemble Self-Attention Based Mutual Learning Network for Document Image Classification (EAML) [3] achieves 97.67 % accuracy on the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) benchmark [7], while LayoutLMv3 [8] reports 95.44 % on the same dataset. More broadly, document classification has progressed from convolutional architectures to multimodal models that jointly encode text, visual content, and layout structure [11,2]. Most of this work assumes a fixed training distribution, but real document collections evolve and privacy constraints often restrict access to historical data, making continual adaptation a practical necessity.

This paper presents a comparative study of IL strategies for document image classification, rather than a new method. Several existing strategies have been explored separately in the literature, including standard regularization-based methods such as Elastic Weight Consolidation (EWC) [10], replay-based methods, Knowledge Distillation (KD) [1], Bias Correction (BC) [23], and open-set approaches based on Extreme Value Theory, EVM [16]. What is currently missing is a side-by-side evaluation of these strategies under a single framework, on representative document backbones, and under realistic compute constraints. The goal of this work is to provide that comparison. We build a modular pipeline covering both CIL and a cross-collection class-incremental setting that we also refer to as DIL. Within this pipeline, we evaluate five configurations: standard IL, EVM [16], Incremental Extreme Value Machine (iEVM) [12], and two further variants integrated in the same setup, namely Regularized Extreme Value Machine (RegEVM) (an L_2 regularized form of the EVM Weibull scale parameters) and EVM+OOD (EVM combined with ViM-based OOD). Two backbones are used to represent different feature paradigms: EAML [3], which is content-centric, and LayoutLMv3 [8], which is layout-aware. To keep methods comparable, we deliberately avoid per-method hyperparameter sweeps and run each configuration in its standard setting. The main contributions of this work are the following:

(1) A modular IL benchmark covering CIL and a cross-collection DIL setting, into which a range of IL strategies can be plugged and compared on the same backbones and datasets.
(2) A systematic side-by-side comparison of standard IL, EVM, iEVM, RegEVM, and EVM+OOD on two representative multimodal backbones (EAML and LayoutLMv3), evaluated under a shared resource-constrained protocol, with explicitly stated asymmetries for LayoutLMv3.

(3) Empirical observations from this comparison: EVM+OOD is the most stable configuration tested in CIL, RegEVM does not improve over standard EVM, and the content-centric backbone retains prior knowledge more reliably under the evaluated resource-constrained Standard Incremental Learning (standard IL) protocol.

2 Related Work

2.1 Document Analysis

Document classification is a critical task in digitization, information retrieval, and automated processing. Recent advances in deep learning and Natural Language Processing (NLP) have enabled strong performance across a wide range of document analysis tasks. Cui et al. [6] provide a comprehensive survey of Document Artificial Intelligence (AI) and its broader landscape while Kirsten et al. [11] contrast traditional methods with deep learning-based approaches and highlight the field's transition from standard Convolutional Neural Networks (CNNs) to advanced multimodal architectures. This study also evaluates the performance of various deep CNNs for document image classification on standard benchmarks, specifically the RVL-CDIP and Tobacco-3482 datasets. Their results suggest that transfer learning benefits even with small datasets and that document-specific pretraining yields crucial features.

DocFormer [2] is a multimodal transformer model designed for Visual Document Understanding (VDU). It combines text, vision, and spatial features using a novel multimodal self-attention mechanism that enables better cross-modal feature fusion. Unlike approaches that rely on pretrained object detection models, DocFormer uses ResNet50 for feature extraction and trains the entire pipeline end-to-end. Wang et al. [20] propose a language-independent transformer model LiLT for Structured Document Understanding (SDU). LiLT can be pretrained on a single language (English) and transferred to other languages by decoupling text and layout information, optimizing them separately, and then reintegrating them during fine-tuning.

Beyond DocFormer and LiLT, more recent multimodal document models have continued in this direction. DiT [14] applies image-only self-supervised pretraining on large document corpora and reaches strong results on classification and layout tasks. UDOP [18] unifies text, image, and layout in a single sequence-to-sequence model for both understanding and generation. Donut [9] takes an OCR-free approach and predicts structured output directly from the document image. These models all confirm the trend toward joint encoding of text, image, and layout, but they are largely studied in static training settings, which leaves their behavior under incremental adaptation open.

2.2 Incremental Learning

The challenges of adapting to new classes or different domain data without forgetting previously learned knowledge align with IL paradigms. iCaRL [15] enables

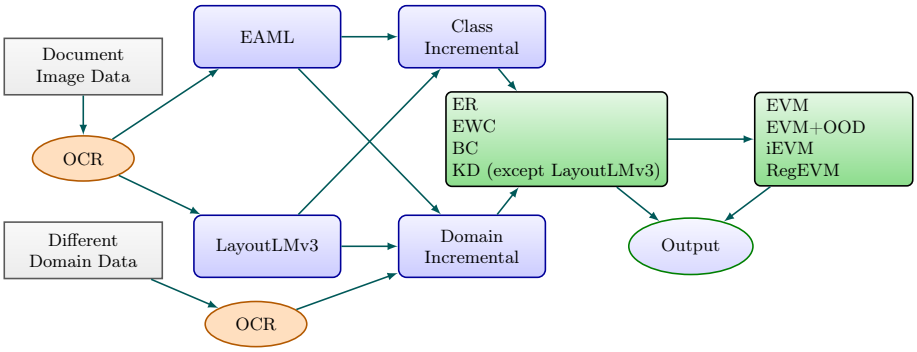

Fig. 1: Overview of the experimental pipeline and evaluated configurations.

learning new object categories over time by combining three methods: smart memory or Exemplar Replay (ER), adaptive classification, and anti-forgetting training. It prevents CF by strategically managing memory (using nearest-mean-of-exemplars) and striking a balance between new and old knowledge.

End-to-End Incremental Learning (EEIL) [5] combines ER with a cross-distillation loss to retain prior knowledge under end-to-end training. Both iCaRL [15] and EEIL [5] perform well on smaller scales but degrade under class imbalance and visually similar classes. Large Scale Incremental Learning (LSIL) [23] addresses data imbalance and visually similar classes by introducing the BC to address the scale issues. They find that the last fully connected layer has a strong bias towards the new classes and used BC to solve that. However, existing IL frameworks have not been tested on multimodal data heterogeneity, where textual content (e.g., Optical Character Recognition (OCR) tokens) and spatial layout (e.g., bounding boxes) features evolve in distinct, non-uniform ways. In this context, EVM [16] shows promise by modeling open-set recognition through statistical Extreme Value Theory (EVT), but its integration with document-specific features remains unexplored. Koch et al. explore open-world recognition, which refers to the challenge of models to distinguish known from unknown data samples, in an incremental setup by enhancing the EVM [12]. These methods target settings in which models must incrementally learn new classes over time in dynamic environments while effectively detecting known from unknown data samples and mitigating the original EVMs limitations, including inefficient model updates and unbounded model complexity during incremental learning.

More recent continual learning methods are also relevant to this study. DER and its variant DER++ [4] combine rehearsal with logit distillation and currently sit among the strongest general-purpose continual learning methods. Prompt-based methods such as L2P [22], DualPrompt [21], and CODA-Prompt [17] freeze the pretrained backbone and learn only small prompt parameters per task, which makes them a natural fit for large transformer models such as LayoutLMv3. These methods are not included in the present comparison and are listed as the main direction for follow-up work.

3 Methodology

We built an Incremental Learning (IL) framework for document image classification that integrates EVM-based decision boundaries. The framework is designed to allow a base model to adapt to novel document categories as they emerge in dynamic environments. As illustrated in Fig. 1, our pipeline employs two representative base architectures as backbones: EAML and LayoutLMv3. Both backbones are embedded into class- and domain-incremental pipelines that employ regularization, distillation, replay, and bias correction to mitigate catastrophic forgetting. We evaluate five configurations within these pipelines: standard IL, EVM, a fusion of EVM+OOD configuration, iEVM, and RegEVM.

3.1 Backbone Architectures

To evaluate the robustness of our IL strategies across different feature modalities, we utilize two representative base architectures as the foundational feature extractors.

EAML Backbone The EAML model serves as our content-centric backbone, integrating visual and textual information through a self-attention-based fusion and a mutual learning strategy [3]. We use an ensemble architecture utilizing an Inception-ResNet-V2 backbone for visual features and a BERT-based transformer for semantic text encoding. The total loss function is formulated as a weighted combination of classification and regularization terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \left(\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{text}} \right) + \lambda_{\text{kld}} \mathcal{L}_{\text{Tr-KLDReg}}, \tag{1}$$

where $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{text}}$ are supervised cross-entropy losses, and $\mathcal{L}_{\text{Tr-KLDReg}}$ represents the truncated Kullback–Leibler (KL) divergence regularization that controls mutual knowledge transfer between modalities. For training, we utilize the AdamW optimizer for stable optimization and adapt the model specifically to the requirements of the CIL and DIL pipelines.

LayoutLMv3 Backbone As a layout-aware counterpart, LayoutLMv3 jointly models textual, layout, and visual information [8]. The architecture processes document images into 16×16 patch-based visual embeddings using a Vision Transformer (ViT) and encodes textual features from OCR tokens using a BERT-base model. Spatial awareness is achieved by normalizing 2D coordinates from OCR-derived bounding boxes and adding them to corresponding text embeddings. A multi-layer fusion transformer integrates these textual, layout, and visual sequences into a unified representational space. The final classification is passed via the [CLS] token to a fully connected layer. To ensure stable convergence before incremental adaptation, the model is optimized using AdamW with a cosine annealing learning rate scheduler.

Because of the higher memory footprint of LayoutLMv3 compared to EAML, all LayoutLMv3 experiments in this paper are run under a more restricted setup

than EAML. The base model is trained on a 50 % partition of the RVL-CDIP data, and the incremental phases use only the standard IL strategy. The Knowledge Distillation-Based Incremental Learning (distillation-based IL) strategy was not feasible on our 32 GB V100 since it requires keeping a teacher copy of the model alive during distillation. Comparisons between EAML and LayoutLMv3 in this paper should therefore be read as trend-level under matched standard IL conditions, not as a fully equal comparison.

3.2 Incremental Learning Settings

Class-Incremental Learning (CIL) The CIL pipeline adapts a base model step-wise to new categories, using ER, EWC, KD, and BC to limit CF.

The framework operates in sequential learning phases in which a model initially trained on a base set adapts to unseen classes. At each incremental step t, the set of total seen classes C_t is updated to incorporate the new class c_t. The training set combines samples of the new class with a subset of exemplar data from previous classes to preserve stability.

This pipeline was evaluated using both standard IL and distillation-based IL setups for EAML, whereas LayoutLMv3 was restricted to standard IL due to resource and time constraints.

Cross-Collection Class-Incremental Learning (DIL) To extend the evaluation beyond a single dataset, we also include a cross-collection class-incremental setting. We refer to it as DIL for consistency with prior work, but the two collections used here are not fully disjoint domains. RVL-CDIP and Tobacco-3482 both originate from the IIT-CDIP, and eight of the ten Tobacco-3482 classes overlap with RVL-CDIP. The setting therefore measures how a model trained on one document collection adapts to a related but smaller collection that introduces a small number of new classes while reusing most of the old label space. The base model is trained on the 16 RVL-CDIP classes and incrementally adapted to the 10 Tobacco-3482 classes, eight of which overlap and two of which (Note, Report) are new. Because of this class structure, aggregate accuracy can be misleading: methods can score high overall by performing well on overlapping classes while quietly losing all knowledge of non-overlapping source classes. We therefore evaluate both the global metric and a class-wise breakdown that separates overlapping from non-overlapping classes, and we treat the latter as the more diagnostic view. The pipeline mirrors CIL in its modularity. EAML is evaluated with both standard IL and distillation-based IL, while LayoutLMv3 is restricted to standard IL for the same reasons as in the CIL setup.

3.3 Optimization Strategies for Incremental Adaptation

Multiple learning strategies have been implemented to mitigate catastrophic forgetting and maintain knowledge across all classification tasks. Both the CIL and DIL pipelines employ these strategies to balance plasticity, the ability to learn new document types, and stability, the ability to remember prior classes.

Incremental Learning (IL) The IL framework adapts the model to new classes by expanding the classification layer and minimizing a combined loss function consisting of cross-entropy ($\mathcal{L}_{\text{CE}}$) and Elastic Weight Consolidation ($\mathcal{L}_{\text{EWC}}$) [10]. The latter regularizes critical weights θ toward their previous optima θ^*, reducing catastrophic forgetting. The resulting base objective function is defined as

$$\mathcal{L}_{\text{IL}} = \mathcal{L}_{\text{CE}}(z, y) + \lambda_{\text{EWC}}\ \mathcal{L}_{\text{EWC}}(\theta, \theta^*), \tag{2}$$

where z represents predicted logits and y denotes ground-truth labels. When integrating EVM variants, this approach is augmented with a membership policy component, $\mathcal{L}_{\text{EVM}}$, which sharpens decision boundaries around known classes and reduces the misclassification of unfamiliar samples through the formulation

$$\mathcal{L}_{\text{IL-EVM}} = \mathcal{L}_{\text{IL}} + \lambda_{\text{EVM}} \cdot \mathcal{L}_{\text{EVM}}. \tag{3}$$

The EVM+OOD configuration further extends this objective by incorporating the out-of-distribution penalty $\mathcal{L}_{\text{OOD}}$ from Virtual-logit Matching (ViM) [19]. ViM estimates out-of-distribution uncertainty from the residual norm obtained by projecting an input feature representation against the principal subspace of the training data. This residual is scaled to form a virtual logit. This virtual logit is appended to the network's original classification logits, allowing the final OOD score to be extracted as the resulting softmax probability of this additional virtual class.

$$\mathcal{L}_{\text{IL-EVM+OOD}} = \mathcal{L}_{\text{IL-EVM}} + \lambda_{\text{OOD}} \cdot \mathcal{L}_{\text{OOD}}. \tag{4}$$

To distinguish between the two primary learning strategies, the standard IL approach is referred to as standard IL throughout this paper.

Knowledge Distillation-Based Incremental Learning The distillation-based IL strategy utilizes a teacher-student mechanism where the current model (student) is supervised by both hard labels of new data and the soft predictions z'_{old} of a frozen predecessor (teacher). The distillation loss preserves inter-class relationships beyond hard labels. The primary optimization is performed through

$$\mathcal{L}_{\text{KDIL}} = \mathcal{L}_{\text{CE}}(z, y) + \lambda_{\text{distill}}\ \mathcal{L}_{\text{distill}}\big(z, z'_{\text{old}}\big) + \lambda_{\text{EWC}}\ \mathcal{L}_{\text{EWC}}(\theta, \theta^*). \tag{5}$$

Analogous to the IL procedure, we integrate EVM-based boundaries directly into the distillation process, resulting in the expanded formulation

$$\mathcal{L}_{\text{KDIL-EVM}} = \mathcal{L}_{\text{KDIL}} + \lambda_{\text{EVM}} \cdot \mathcal{L}_{\text{EVM}}. \tag{6}$$

The final iteration combines this dual supervision with explicit outlier detection to enhance robustness in dynamic environments using the objective

$$\mathcal{L}_{\text{KDIL-EVM+OOD}} = \mathcal{L}_{\text{KDIL-EVM}} + \lambda_{\text{OOD}} \cdot \mathcal{L}_{\text{OOD}}. \tag{7}$$

In all formulations, the λ coefficients serve as hyperparameters to control the relative influence of each learning objective during the optimization phase.

Bias Correction Bias correction is applied after each incremental step to mitigate the inherent class imbalance that arises when new classes are introduced to the existing model [23]. Without correction, classifier weights tend to favor newly added classes, degrading performance on earlier categories. We address this by centering the bias vector of the classifier, a process that ensures a uniform average bias across all seen categories. This normalization improves fairness in multi-class predictions and serves as a stabilizing mechanism for the recognition of earlier document types.

Incremental Extreme Value Machine (iEVM) For open-world recognition, we employ iEVM [12], which extends traditional EVM theory to accommodate incremental data streams and shifting distributions [16]. A core feature of this implementation is a partial update mechanism designed for computational efficiency. Specifically, a new sample, x_{new}, triggers a model update only if it falls within the maximum tail distance, $d_{\tau,i}$, of an existing Extreme Vector (EV). This selective update rule is defined as

$$\theta_i^{(t+1)} = \begin{cases} \text{update}(\theta_i^{(t)}), & \text{if } d(x_i, x_{\text{new}}) < d_{\tau,i} \\ \theta_i^{(t)}, & \text{otherwise} \end{cases}, \tag{8}$$

where $\theta_i^{(t)} = \{\kappa_i, \lambda_i, d_{\tau,i}\}$ represents the Weibull shape, scale, and distance parameters at time step t. To maintain model scalability, we perform model reduction using a weighted K-set cover to retain only a subset of EVs that maximizes spatial coverage under a fixed budget. For novel samples, the model computes a class-wise inclusion probability; if the maximum inclusion score fails to meet a predefined threshold, γ, the sample is rejected as an unknown document type.

Regularized Extreme Value Machine (RegEVM) RegEVM is evaluated as a simple regularized variant of the standard EVM [16]. It applies an L_2 regularization term to the Weibull scale parameters during model fitting. We include this variant to test whether scale-parameter regularization improves robustness in the incremental setting, rather than treating it as a separate methodological contribution. The motivation is that large scale values can produce overly permissive class tails. Penalizing these values is therefore expected to yield tighter and more stable decision boundaries under noisy or limited data. This comparison tests empirically whether such regularization leads to measurable gains in an incremental setting.

Concretely, for each class c with training samples X_c, we fit a set of Weibull models

$$W_c = \{(x_i, \kappa_i, \lambda_i) \mid x_i \in X_c\}, \tag{9}$$

where κ_i and λ_i denote the Weibull shape and scale parameters associated with Extreme Vector x_i, respectively. Given the Weibull-based inclusion probability

$\Psi_i(\cdot)$, we optimize the regularized objective

$$\mathcal{L}_{\text{RegEVM}} = -\sum_{i=1}^{N_c} \log \Psi_i(x_i; \kappa_i, \lambda_i) \; + \; \alpha \sum_{i=1}^{N_c} \lambda_i^2, \tag{10}$$

where $N_c = |X_c|$ and α controls the strength of the scale regularization. This formulation encourages faithful fitting of the class-conditional tail behavior while discouraging excessively large Weibull scale parameters, thereby improving decision boundary stability and generalization for open-set document classification.

A note on design choices. Since this is a comparative study, all methods are evaluated in their standard configurations. We avoid per-method tuning and ablations to keep the comparison consistent; these remain useful directions for follow-up work.

4 Experiments and Results

4.1 Implementation Details

To ensure optimal feature extraction, preprocessing is tailored to the specific requirements of each backbone. For EAML, document images are resized to 229×229 pixels with standard normalization and augmentation. For LayoutLMv3, we utilize 224×224 patches and normalize OCR-derived bounding boxes to a $[0, 1000]$ scale to maintain spatial consistency across varying resolutions. Textual content is extracted via the Tesseract OCR engine and encoded using a *bert-base-uncased* transformer.

Experimental consistency is maintained across all pipelines using fixed hyperparameters: temperature $T = 2$, $\lambda_{\text{EWC}} = 5000$, and $\lambda_{\text{distill}} = 1.0$. For the EVM and OOD (ViM) components, scaling coefficients are set to $\lambda = 0.1$ to balance boundary regularization with classification performance. All models are optimized using the AdamW optimizer. All experiments were conducted on resource-constrained GPU environments, utilizing NVIDIA V100 with 32 GB of VRAM to ensure stable execution during the incremental training and evaluation phases.

4.2 Datasets

The RVL-CDIP [7] dataset is one of the largest publicly available document image datasets. It is mainly derived from the IIT-CDIP Test Collection 1.0 [13], a digital archive of scanned document images that was originally released for information retrieval research. This dataset consists of 400K grayscale images divided into 16 different classes.

The Tobacco-3482 dataset is a particularly small dataset with only 3482 images derived from the Tobacco Litigation Library. All the images are scanned document images of 10 categories or classes. The dataset was mainly created to

Table 1: Base Model Accuracy [%]

#of Classes	Validation	Test
11-CIL	93.21	93.02
16-DIL	92.32	91.20

(a) EAML Model Accuracy

#of Classes	Validation	Test
11-CIL	96.33	89.28
16-DIL	93.72	88.92

(b) LayoutLMv3 Model Accuracy

facilitate research in document image classification and retrieval, especially for methods involving limited training data.

The choice of RVL-CDIP [7] and Tobacco-3482 follows the EAML reference protocol and avoids starting from a weak baseline. Larger benchmarks such as DocLayNet remain important follow-up targets for testing portability beyond the IIT-CDIP family.

Performance is quantified using accuracy and the Incremental Learning Gap (G_{IL}), which measures the deviation from the base model. The G_{IL} is defined as

$$G_{\mathrm{IL}} = \frac{Acc_{\mathrm{inc}} - Acc_{\mathrm{base}}}{1 - Acc_{\mathrm{base}}}, \tag{11}$$

Here, Acc_{inc} and Acc_{base} are the incremental and base accuracies. The metric normalizes the change in accuracy by the headroom above the base model, so G = 0 means no change, G < 0 indicates forgetting, and G > 0 indicates positive transfer. As a worked example, for a base test accuracy of 93.02 % and an incremental accuracy of 83.43 %, the gap is G = (0.8343 − 0.9302) / (1 − 0.9302) ≈ −1.37, indicating a substantial drop relative to the headroom of the base model.

4.3 Results and Analysis

Base Model Performances: The foundation of our incremental learning evaluation rests upon the strong baseline results of the initial base models. For the starting CIL task, we established an initial knowledge base by training the models on 11 selected classes from the RVL-CDIP dataset: letter, form, email, handwritten, advertisement, scientific report, invoice, presentation, questionnaire, resume, and memo. This specific subset provides a diverse range of document structures, from the text-heavy nature of letters and emails to the highly structured layout of forms and questionnaires.

As demonstrated in Table 1, the EAML model achieved strong baseline metrics, with a test accuracy of 93.02 % for the 11-class subset and 91.20 % when scaled to the full 16-class RVL-CDIP dataset. These results indicate that the multimodal fusion of visual and textual features in EAML provides features that separate classes clearly for document classification.

LayoutLMv3, trained under the resource-constrained setup described in Section 3.1, reaches 89.28 % on the 11-class subset and 88.92 % on the full 16 classes,

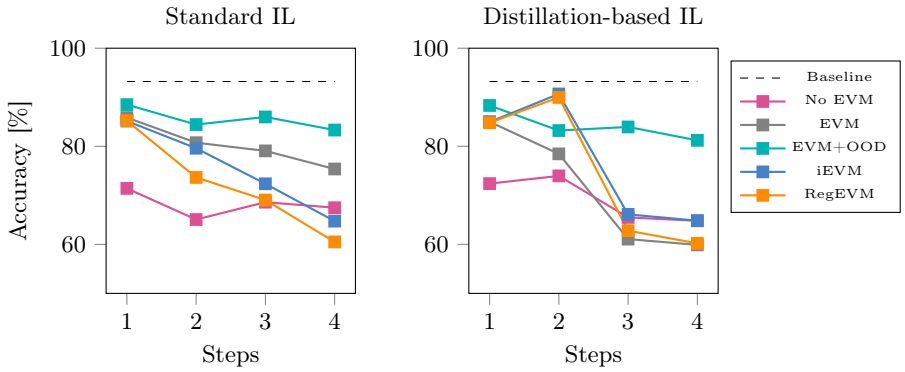

Fig. 2: EAML based CIL validation and accuracy for each incremental step. The baseline denotes the accuracy before adapting to the additional 4 classes introduced incrementally in steps 1 to 4.

which is consistent with its reported behavior on RVL-CDIP under reduced training data and serves as a stable starting point for the incremental evaluation.

Class-Incremental Learning (CIL) Performance Analysis The CIL evaluation involves fine-tuning base models, which were initially trained on 11 classes, to adapt to four novel document categories in discrete steps. As illustrated in the flow diagram in Fig. 1, the process for the content-centric EAML backbone involves implementing both standard IL and distillation-based IL strategies across five methodological configurations. Conversely, the LayoutLMv3 backbone is evaluated under the standard IL strategy using four specific methods: standard-CIL (No EVM), EVM, iEVM, and RegEVM.

EAML-based CIL The evaluation of the content-centric EAML backbone reveals significant variations in knowledge retention across the implemented configurations. As illustrated in Fig. 2, the EVM+OOD fusion emerges as the most resilient strategy, maintaining stability throughout all four incremental stages.

During the initial increment, introducing the "scientific_publication" class, this hybrid approach achieved a test accuracy of 88.04 % in the standard IL setup and 88.08 % in the distillation-based IL setup. The corresponding G_{IL} scores of -0.695 and -0.7201, respectively, indicate that the model operated near its base performance level, successfully mitigating the initial onset of Catastrophic Forgetting (CF). This stability is particularly notable when compared to the standard CIL (No EVM) baseline, which exhibited an immediate drop to 70.96 % test accuracy and a G_{IL} score of -3.2106.

A comparative analysis of the learning strategies, as detailed in Table 2, reveals that the performance decline in the standard IL configuration followed a linear trajectory, whereas the distillation-based IL strategy exhibited a sharp nonlinear collapse after the second step. Interestingly, during the second increment,

Table 2: Incremental Performance Comparison (EAML Backbone). G_B quantifies total performance deviation from the 11-class initial model, while G_P measures the relative stability between consecutive incrementally added classes (Step 1–3). Scores near zero indicate successful knowledge retention; "No EVM" represents the standard IL baseline utilizing ER, EWC, BC, and KD.

		standard IL Strategy				**distillation-based IL Strategy**			
Step	**Method**	**Val**	**Test**	G_B	G_P	**Val**	**Test**	G_B	G_P
	No EVM	71.41	70.96	-3.21	-3.21	72.39	73.77	-3.07	-3.07
1	EVM	85.85	85.58	-1.08	-1.08	85.00	84.62	-1.21	-1.21
	EVM+OOD	**88.49**	**88.04**	**-0.70**	**-0.70**	**88.32**	**88.08**	**-0.72**	**-0.72**
	No EVM	65.06	65.57	-4.15	-0.22	73.96	76.38	-2.84	+0.06
2	EVM	80.73	80.45	-1.84	-0.36	78.45	78.71	-2.17	-0.44
	EVM+OOD	**84.43**	**84.21**	**-1.29**	**-0.35**	**83.21**	**83.31**	**-1.47**	**-0.44**
	No EVM	67.47	67.61	-3.79	-0.04	64.80	64.85	-4.18	-0.02
3	EVM	75.37	75.49	-2.63	-0.28	59.89	59.78	-4.91	-0.03
	EVM+OOD	**83.31**	**83.43**	**-1.46**	**-0.19**	**81.21**	**81.37**	**-1.77**	**-0.17**

methods such as iEVM and RegEVM within the distillation-based IL framework showed temporary performance gains, reaching validation scores of 90.76 % and 89.96 %. However, this proved to be a "stability-plasticity" trade-off; by the third increment, these gains were erased by a sharp performance deficit.

The class-wise analysis further shows why EVM+OOD is more stable. At the fourth step, it maintained 81.21 % test accuracy, whereas other methods erased much of the base knowledge: five to six base classes, including *form, email, advertisement, presentation,* and *questionnaire*, reached zero accuracy. This suggests that explicit OOD uncertainty management helps prevent new class distributions from overwriting established representations.

RegEVM did not show a meaningful improvement over standard EVM across standard IL, distillation-based IL, CIL, and DIL results (Table 2 and Table 4). The L_2 penalty on Weibull scale parameters did not change the forgetting trajectory, suggesting that tail-scale instability is not the main limiting factor in this setup. We report this as a useful negative finding for similar regularization approaches.

LayoutLMv3-based CIL As detailed in Table 3, LayoutLMv3 shows a substantial performance drop from the first incremental step under the evaluated resource-constrained standard IL protocol. This suggests stronger representation drift during sequential updates, but should be interpreted as a protocol-specific observation rather than a general limitation of layout-aware models.

Specifically, upon the introduction of the "scientific_publication" class, validation accuracies decreased from the 96.33 % base performance to approximately 64 % across all configurations. The resulting G_B score of approx. -8.7 indicates

Table 3: Incremental Performance Comparison (LayoutLMv3 Backbone) using the standard IL strategy. G_B quantifies total deviation from the 11-class model, while G_P measures stability between consecutive added classes (Step 1–3). "No EVM" denotes the standard baseline (ER, EWC, BC).

	Step 1			**Step 2**				**Step 3**			
Method	Val	Test	G_B	Val	Test	G_B	G_P	Val	Test	G_B	G_P
No EVM	64.32	60.87	-8.72	49.27	60.87	-12.82	-0.42	55.45	56.53	-11.14	-0.09
EVM	64.41	64.36	-8.70	49.28	60.87	-12.82	-0.43	55.46	56.52	-11.14	+0.12
iEVM	64.43	63.37	-8.69	49.27	60.87	-12.82	-0.42	55.55	56.53	-11.11	+0.12
RegEVM	64.41	64.06	-8.70	49.28	60.87	-12.82	-0.43	55.46	56.53	-11.14	+0.12

a fundamental shift in the latent space that substantially degraded retention of base classes. This trend intensified during the second increment ("specification"), where G_B scores dropped below -12.

Although a marginal recovery in G_P ($+0.12$) was observed in the third step ("file_folder"), accuracy remained stagnant at approximately 55 %, suggesting random generalization rather than effective knowledge retention. These findings indicate that the complex spatial-textual dependencies in transformer-based layout models are highly susceptible to representation drift, leading to a breakdown in stability that standard IL strategies, including EVM-based boundary estimation, fail to mitigate. Due to these persistent performance bottlenecks and associated computational overhead, further incremental steps were discontinued for this architecture.

Domain-Incremental Learning (DIL) Performance Analysis In DIL, EAML is evaluated on top of its 16-class base model (92.32 % base accuracy), and LayoutLMv3 under standard IL only per the constraints in Section 3.1.

The EAML backbone remains stable in the cross-collection setting. As shown in Table 4, the G_{IL} scores across all configurations remain near zero, indicating a successful global adaptation. In the standard IL setup, the EVM+OOD fusion achieved a peak validation accuracy of 88.31 % ($G = -0.522$). Notably, the "No EVM" baseline also maintained high stability, suggesting that the primary challenge in DIL is not global forgetting, but rather a domain-specific representation bias.

A per-class analysis of domain-specific accuracies reveals a significant performance gap. While target domain accuracy (Acc_{Tob}) consistently reached $\approx$ 90 %, source domain accuracy (Acc_{RVL}) decreased to $\approx$ 46 %. Class-wise analysis confirms that non-overlapping source categories (e.g., *Budget, Invoice*) exhibited total CF with 0.0 % accuracy across all methods, despite the inclusion of ER. Conversely, overlapping classes (e.g., *Letter, Email*) exhibited near-perfect retention, and sometimes outperformed the base model due to positive backward transfer from the new domain samples. These findings suggest that the high

Table 4: Domain-Incremental Performance (EAML Backbone): RVL-CDIP → Tobacco-3482. Acc_{RVL} and Acc_{Tob} denote source and target domain test accuracies, respectively. G represents G_{IL} relative to the 16-class base model ($Acc = 92.32\,\%$). "No EVM" is the standard baseline (ER, EWC, BC, KD).

	standard IL Strategy				**distillation-based IL Strategy**			
Method	**Val**	Acc_{RVL}	Acc_{Tob}	G	**Val**	Acc_{RVL}	Acc_{Tob}	G
No EVM	87.55	46.15	90.44	-0.621	87.36	46.19	89.67	-0.645
EVM	87.74	46.11	90.25	-0.596	87.93	46.11	86.52	-0.572
EVM+OOD	**88.31**	**46.11**	**90.44**	**-0.522**	87.16	46.16	89.67	-0.672
iEVM	84.87	46.17	87.57	-0.970	86.97	46.18	85.09	-0.696
RegEVM	87.55	46.14	90.25	-0.621	87.36	46.13	90.06	-0.645

Table 5: Domain-Incremental Performance (LayoutLMv3 Backbone): RVL-CDIP (Source) → Tobacco-3482 (Target) using standard IL. Accuracy is reported for source (Acc_{RVL}) and target (Acc_{Tob}) domains. "No EVM" denotes the standard baseline (ER, EWC, BC).

Method	**Val**	Acc_{RVL}	Acc_{Tob}	G_B
No EVM	60.83	56.27	52.87	-5.237
EVM	58.01	60.73	55.30	-5.686
iEVM	54.43	56.27	52.59	-6.256
RegEVM	54.60	56.34	52.87	-6.229

global G_{IL} scores are partially masked by the success of overlapping features, while unique source-domain knowledge remains vulnerable to domain shift.

LayoutLMv3-based DIL: Similar to the trends observed in the CIL phase, the LayoutLMv3 backbone performs poorly in the DIL setup compared to the content-centric EAML model. As detailed in Table 5, the model's validation accuracy decreased from an initial base score of 93.72 % to a range between 54 % and 60 % across all methods. This drastic decline, reflected in the significantly low G_{IL} scores (e.g., -5.237 to -6.256), indicates that the multimodal integration of layout and content features is highly susceptible to severe Catastrophic Forgetting (CF) when exposed to new domain data.

A comparative analysis of the methods within the LayoutLMv3 framework reveals that while the "No EVM" baseline achieved slightly higher overall validation accuracy, the EVM-based method demonstrated a superior ability to balance source and target knowledge, achieving the highest source domain retention ($Acc_{RVL} = 60.73\,\%$). Despite this, the consistent performance gap between the two backbones suggests that models relying solely on content features, like EAML, provide a more efficient and stable approach for incremental setups in document image classification.

4.4 Limitations

The results of this study should be interpreted in light of the resource-constrained setting of the experiments. The main design trade-off was between breadth and repetition. We evaluated two backbones, two incremental settings, several learning strategies, and multiple EVM-based configurations, rather than repeating each configuration across several random seeds. This broader comparison prevents us from reporting confidence intervals or variance estimates. Large performance gaps, such as the difference between EVM+OOD and the standard IL baseline in CIL, indicate clear trends, but repeated runs would be required to quantify their stability.

A second limitation concerns the comparison between EAML and LayoutLMv3. LayoutLMv3 was evaluated under tighter constraints than EAML, using only 50 % of the data and the standard IL strategy, because keeping a teacher model in memory during distillation was not feasible on our 32 GB V100. As a result, the backbone comparison should not be read as a direct benchmark of absolute performance, but as a trend-level observation that the layout-aware model showed stronger representation drift under the evaluated standard IL protocol.

The cross-collection experiment should also be interpreted cautiously. Although referred to as DIL, RVL-CDIP and Tobacco-3482 are related IIT-CDIP collections with partial class overlap. The setting therefore measures adaptation across related document collections rather than robustness under a fully disjoint domain shift. Stronger domain-generalization claims would require additional datasets outside the IIT-CDIP family.

Finally, we did not perform per-method hyperparameter sweeps. Parameters such as the EVM inclusion threshold, OOD weighting, and regularization strength were kept fixed to preserve a controlled comparison, so the results reflect standard configuration performance rather than the best achievable performance of each method. The benchmark is also limited to classical IL strategies and EVM-family variants. Recent continual learning methods such as L2P [22], DualPrompt [21], CODA-Prompt [17], and DER++ [4], as well as the scalability of the EV pool to hundreds of classes, remain outside its scope. Extending the framework with stronger baselines, larger document datasets, and harder domain shifts is therefore the most direct next step.

5 Conclusion

Incremental learning addresses a practical constraint: document classification systems rarely operate on fixed collections. New document types appear, storage and privacy constraints limit access to past data, and full retraining is often too expensive. This paper examined this setting comparatively and showed that feature representation and uncertainty handling strongly affect knowledge retention.

The main finding is that incremental document classification should not be treated as a direct extension of static document classification. Backbones

that perform well in fixed training setup may still be fragile under sequential updates. In our experiments, content-centric representations were more stable, while layout-aware models showed stronger drift under the evaluated protocol. This does not reject layout-aware models generally, but suggests that they need additional stabilization in incremental settings.

The comparison also highlights the role of open-set uncertainty. EVM+OOD was the most stable configuration in the CIL setting, suggesting that explicit uncertainty handling can reduce the risk of new classes overwriting prior knowledge. In contrast, RegEVM did not improve over standard EVM, suggesting that regularizing Weibull scale parameters alone is insufficient to address forgetting in this setup.

Overall, this study should be viewed as a resource-constrained benchmark and practical reference point rather than a definitive ranking of backbones or a state-of-the-art claim. Future work should evaluate stronger continual learning baselines such as DER++, L2P, DualPrompt, and CODA-Prompt; test broader domain shifts beyond the IIT-CDIP family; and develop stabilization strategies for layout-aware document models. These directions would better reflect deployment settings where data, labels, and document distributions change continuously.

References

1. Alkhulaifi, A., Alsahli, F., Ahmad, I.: Knowledge distillation in deep learning and its applications. PeerJ Computer Science **7**, e474 (2021) 2
2. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 993–1003 (2021) 2, 3
3. Bakkali, S., Ming, Z., Coustaty, M., Rusinol, M.: Eaml: ensemble self-attention-based mutual learning network for document image classification. International Journal on Document Analysis and Recognition (IJDAR) **24**(3), 251–268 (2021) 2, 5
4. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems **33**, 15920–15930 (2020) 4, 15
5. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 233–248 (2018) 4
6. Cui, L., Xu, Y., Lv, T., Wei, F.: Document ai: Benchmarks, models and applications. arXiv preprint arXiv:2111.08609 (2021) 3
7. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 991–995. IEEE (2015) 2, 9, 10
8. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4083–4091 (2022) 2, 5
9. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 498–517. Springer Nature Switzerland, Cham (2022) 3

10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017) 2, 7
11. Kirsten, L.N., Piccoli, R., Ribani, R.: Evaluating deep neural networks for image document enhancement. In: Proceedings of the 21st ACM Symposium on Document Engineering. pp. 1–4 (2021) 2, 3
12. Koch, T., Liebezeit, F., Riess, C., Christlein, V., Köhler, T.: Exploring the open world using incremental extreme value machines. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 2792–2799. IEEE (2022) 2, 4, 8
13. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 665–666 (2006) 9
14. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: Dit: Self-supervised pre-training for document image transformer. In: Proceedings of the 30th ACM international conference on multimedia. pp. 3530–3539 (2022) 3
15. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017) 3, 4
16. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. IEEE transactions on pattern analysis and machine intelligence **40**(3), 762–768 (2017) 2, 4, 8
17. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11909–11919 (June 2023) 4, 15
18. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. arXiv preprint arXiv:2212.02623 (2022) 3
19. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022) 7
20. Wang, J., Jin, L., Ding, K.: Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7747–7757 (2022) 3
21. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European conference on computer vision. pp. 631–648. Springer (2022) 4, 15
22. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 139–149 (2022) 4, 15
23. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 374–382 (2019) 2, 4, 8