

HybriDLA: Hybrid Generation for Document Layout Analysis

Yufan Chen¹, Omar Moured¹, Ruiping Liu¹, Junwei Zheng¹, Kunyu Peng¹,
Jiaming Zhang^{2,*}, Rainer Stiefelhagen¹

¹ Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology

² School of Artificial Intelligence and Robotics, Hunan University

{yufan.chen, omar.moured, ruiping.liu, junwei.zheng, kunyu.peng, rainer.stiefelhagen}@kit.edu, jiamingzhang@hnu.edu.cn

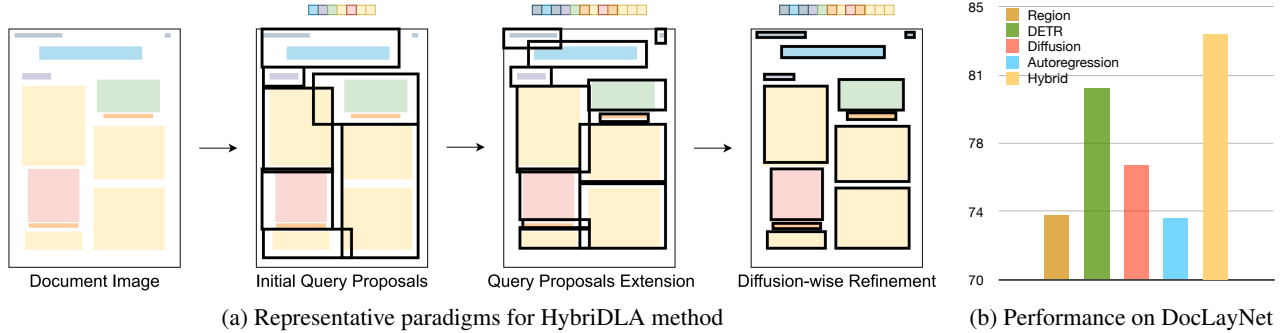


Figure 1: (a) The progressive prediction behaviors of representative paradigms in our proposed HybriDLA ensure successive intermediate outputs obtained during the forward pass from left to right. (b) On the DocLayNet (Pfitzmann et al. 2022) dataset, we compare the performance (mAP in %) of the best models from five different DLA methods, *i.e.*, traditional region-based method, DETR-based method, diffusion-based method, autoregressive method, and our proposed HybriDLA method.

Abstract

Conventional document layout analysis (DLA) traditionally depends on empirical priors or a fixed set of learnable queries executed in a single forward pass. While sufficient for early-generation documents with a small, predetermined number of regions, this paradigm struggles with contemporary documents, which exhibit diverse element counts and increasingly complex layouts. To address challenges posed by modern documents, we present **HybriDLA**, a novel generative framework that unifies diffusion and autoregressive decoding within a single layer. The diffusion component iteratively refines bounding-box hypotheses, whereas the autoregressive component injects semantic and contextual awareness, enabling precise region prediction even in highly varied layouts. To further enhance detection quality, we design a multi-scale feature-fusion encoder that captures both fine-grained and high-level visual cues. This architecture elevates performance to **83.5%** mean Average Precision (mAP). Extensive experiments on the DocLayNet and M⁶Doc benchmarks demonstrate that **HybriDLA** sets a state-of-the-art performance, outperforming previous approaches. All data and models will be made publicly available at <https://yufanchen96.github.io/projects/HybriDLA>.

Introduction

Understanding the layout of a document image is a fundamental task in document analysis. Accurate document layout

analysis (DLA) is crucial for downstream applications, *e.g.*, document understanding and information extraction. However, a persistent challenge in DLA is the extreme variability in the number and arrangement of layouts across diverse documents. A simple form might contain only a handful of text fields, while a dense scientific article page could contain dozens of paragraphs, figures, and tables. This variability makes it strenuous for approaches that assume a predefined number of object queries to perform well across all cases.

Early layout analysis methods as cascade R-CNN (Cai and Vasconcelos 2018), often treat the problem similarly to generic object detection, using a limited set of proposals or queries to find layout elements on a single document page. Transformer-based detectors, *e.g.*, DETR, adopt a predetermined number of learnable queries to predict all layout elements. While effective on scenes with relatively consistent object counts, fixed-query methods struggle when real layout elements, which could range from 2 to 200, greatly differ from the predefined number. If fewer queries are used, the model misses layouts that need to be detected; if a larger preset number is used, it must handle useless “no object” queries, leading to inefficiency and potential disturbance. This rigidity contrasts with how humans approach reading a document, *i.e.*, we do not decide in advance exactly how many items we will find. Instead, we first scan the page to identify major regions, then gradually delve into each region to find sub-components, adjusting our expectations.

Motivated by this human reading strategy (Furnas 1986; Liu 2005), we present the hybrid generation method

*Corresponding author.

for document layout analysis, *i.e.*, **HybriDLA**, a unified transformer-based model that, for the first time, explicitly simulates the human-like coarse-to-fine reading strategy within a single end-to-end trainable architecture. Our model dynamically adjusts the number of object queries hierarchically, allowing it to handle documents with widely varying numbers of elements efficiently. As illustrated in Figure 1, HybriDLA initially uses only a small set of object queries to obtain a rough sketch of the page layout. Each of these initial queries attends to a broad region of the page, potentially spanning multiple actual elements. Then, in subsequent decoding layers, the model progressively increases its resolution: it refines the localization of these coarse regions and spawns new queries to delve into detailed regions. During this continually iterative process, each decoder layer focuses on a finer level of granularity than the previous one. By the final layer, the model has transitioned from a coarse overview to a detailed, precise identification of every individual layout element on the page of the document image.

To enable this hierarchical expansion and refinement, HybriDLA employs the hybrid generative decoding mechanism that combines the strengths of diffusion-based and autoregressive modeling. These two components work in tandem at each stage of the decoder. The diffusion-based part continuously refines the spatial coordinates of all current queries, analogous to fine-tuning the attention on a region for precise boundaries. Meanwhile, the autoregressive part handles the semantic and structural aspect, *i.e.*, it determines if a query represents multiple elements that should be further split. If further detail is needed, the autoregressive decoder generates new queries for the next layer, targeting the sub-elements within a coarse region.

To summarize, our contributions are as follows:

- We propose **HybriDLA**, a novel document layout analysis model that, for the first time, integrates a human-inspired coarse-to-fine querying strategy.
- HybriDLA couples the feature fusion encoder with the hybrid generative decoder that unifies autoregressive query expansion and diffusion-based refinement, enabling precise layout analysis.
- Extensive experiments on DocLayNet and M⁶Doc demonstrate the state-of-the-art performance of our proposed HybriDLA method.

Related Work

Document Layout Analysis

Document Layout Analysis (DLA) plays a crucial role in comprehending the structure and content organization within documents. Both traditional machine learning techniques (Diem, Kleber, and Sablatnig 2011; Garz, Diem, and Sablatnig 2010) and modern deep learning approaches (Long et al. 2022; Gemelli et al. 2022; Peng et al. 2022; Zhu et al. 2022; Coquenot, Chatelain, and Paquet 2023; Yang and Hsu 2022) have witnessed substantial advancements with the availability of diverse datasets and benchmarks (Li et al. 2020; Shihab et al. 2023; Zhong, Tang, and Yepes 2019; Shen, Zhang, and Dell 2020; Moured et al.

2023). DLA has been explored using single-modal framework Faster R-CNN (Ren et al. 2017), Mask R-CNN (He et al. 2017), and DocSegTr (Biswas et al. 2022), as well as multi-modal models including LayoutLMv3 (Huang et al. 2022) and DiT (Li et al. 2022). Furthermore, text grid-based techniques (Zhang et al. 2021) demonstrate the effective fusion of textual layout with visual cues. More recently, transformer-based architectures (Coquenot, Chatelain, and Paquet 2023; Yang and Hsu 2022; Cheng et al. 2023; Tang et al. 2023; Li et al. 2022; Arroyo, Postels, and Tombari 2021; Wang, Jin, and Ding 2022) have gained prominence in this field. Self-supervised pretraining methods (Xu et al. 2020, 2021; Li et al. 2021; Appalaraju et al. 2021; Luo et al. 2022; Huang et al. 2022; Luo et al. 2023), such as DocFormer (Appalaraju et al. 2021) and LayoutLMv3 (Huang et al. 2022), have also attracted significant interest. Zhang et al. (Zhang et al. 2025b) propose a unified feature-conductive end-to-end document image translation framework. Chen et al. (Chen et al. 2025) propose graph-based document layout analysis focusing on cross-page long-range document analysis. Zhang et al. (Zhang et al. 2025a) propose SAIL, a finetuning-free method for Document Information Extraction that leverages textual and layout similarity to guide LLMs, achieving strong performance without full training. Shen et al. (Shen et al. 2025) introduces ProcTag, a process-based tagging method for evaluating document instruction data, and DocLayPrompt, a layout-aware prompting strategy. Constum et al. (Constum, Tranouez, and Paquet 2025) presents DANIEL, combining layout analysis, handwriting recognition, and named entity recognition in handwritten documents using a language-model-based decoder.

Generic Detection Models

DETR-Based Detection Model. DETR (Carion et al. 2020) reframed object detection as a bipartite set prediction task, *i.e.*, a fixed collection of N learnable queries is fed to a transformer decoder with Hungarian matching that aligns the N predictions with the ground-truth objects and supervises them with a permutation-invariant loss. While conceptually elegant, vanilla DETR converges slowly and struggles to localize small objects because each query must attend the full feature map at every decoder layer. Deformable DETR (Zhu et al. 2021) tackles this by sampling a sparse set of keypoints around each reference point, reducing quadratic attention to linear cost and accelerating convergence. DINO (Zhang et al. 2022), which unifies the denoising, contrastive query selection, and look-forward twice refinement, achieves state-of-the-art accuracy with fewer epochs. RoDLA (Chen et al. 2024) suppresses local noise responses by inserting channel attention and average pooling into the self-attention mechanism, allowing it to remain stable under distortions.

Autoregressive Detection Model. Pix2Seq (Chen et al. 2021) pioneered a language-model view of detection: the model serializes each object as a sequence token $\langle x y w h c \rangle$ and generates tokens autoregressively with teacher forcing. Pix2Seq-v2 (Chen et al. 2022b) extends this formulation to a unified multitask framework, improves box precision with relative coordinate encoding, and decodes up to 300 objects per image on COCO. While autoregressive decoding grace-

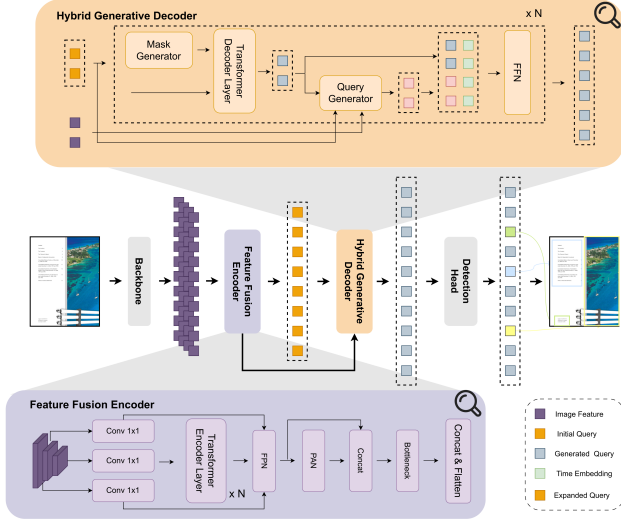


Figure 2: **Overview of the HybriDLA architecture.** The framework consists of a feature fusion encoder and a hybrid generative decoder. The encoder aggregates multi-scale visual features via convolutional and transformer layers, producing a layout-aware representation. The decoder operates in two mechanisms: it performs autoregressive query expansion to propose hierarchical layout regions, then applies a diffusion-style refinement with residual correction to denoise and adjust spatial predictions. Auxiliary queries and intermediate supervision facilitate convergence. This coarse-to-fine pipeline enables precise and adaptive generation of layout detection results.

fully accommodates sequences of arbitrary length and captures rich inter-object dependencies, its computational cost grows linearly with the number of targets.

Diffusion-based Detection Model. Inspired by score-based generative modeling, DiffusionDet (Chen et al. 2022a) samples N Gaussian noise boxes at time step T and trains the network to denoise them towards ground-truth boxes through T steps. During inference, only four to seven predictor steps are executed with a learned variational sampler, achieving strong detection performance and ensemble-like robustness. Despite iterative quality gains, diffusion detectors still rely on a prefixed pool of initial proposals.

HybriDLA Framework

Pipeline Overview

The HybriDLA framework introduces a human-inspired coarse-to-fine layout generation pipeline, integrating a top-down coarse prediction with a bottom-up refinement. As shown in Figure 2, the HybriDLA framework follows a two-stage, hierarchical generation pipeline composed of a multi-scale feature fusion encoder and a hybrid generative decoder. In the first stage, the encoder processes the input document image features, which are extracted through the backbone, to produce coarse layout region predictions that capture the approximate locations and extents of major content blocks.

These coarse predictions serve as high-level layout priors rather than final outputs. In the second stage, the hybrid decoder utilizes these coarse region priors, along with the rich feature representations of encoders, to refine them into a final set of layout elements with precise boundaries and semantic labels. This two-tier decoding strategy mirrors how a human would parse a page, first sketching high-level regions, then focusing on details. By combining top-down and bottom-up approaches, HybriDLA achieves high performance on accuracy in analysis. In the following, we will provide an in-depth description of these modules.

Feature Fusion Encoder

The feature fusion encoder is designed as a unified hierarchical representation extraction module that transforms a set of multi-scale feature maps $F_{l=1}^L$ into a single spatial-aware representation G . In essence, this encoder combines information from different scales in an organized, multi-stage manner so that the resulting representation G captures the informative layout context and element content of documents.

Local Feature Encoding. For each input feature map F_l (at scale l), the encoder first performs a local encoding to produce an enriched feature map $H_l = \phi(F_l)$. The local feature encoding function ϕ typically includes a combination of self-attention and convolutional operations applied within the module. The self-attention component enables the model to capture long-range dependencies and relationships within the feature map, complementing the convolution, which focuses on local patterns. In practice, this means ϕ can learn intra-scale patterns, *e.g.*, global information such as textures and part-whole relationships that might span across the spatial extent of F_l . The outcome is that each H_l is a refined version of F_l at scale l . It preserves important fine-grained details of that scale while also embedding rich contextual information. This step ensures that, before any cross-scale fusion, features in each scale are locally enhanced.

Cross-Scale Fusion. After obtaining the set of enhanced features $H_{l=1}^L$, the encoder then applies an attention-based fusion function Ψ to merge information across these scales. The cross-scale fusion $\Psi(H_l)$ is designed to perform adaptive information exchange between different feature scales. In practice, this involves a cross-scale attention where features at one focused scale integrate information from features at another scale. This allows the fine-detailed feature map to incorporate global context from a low-scale feature map, and vice versa, a coarse feature to be informed about fine details from a low-scale map. Additionally, lateral convolutional layers are used to align and combine feature maps from different scales. These lateral convolutions ensure that features from various scales are projected into a common space and integrated smoothly, allowing the network to learn a coherent, multi-scale representation.

Through local feature encoding and the cross-scale fusion module, the resulting representation $G = \Psi(H_{l=1}^L)$ is a hierarchical feature embedding which provides a global context understanding of the document layout while still retaining the necessary detail for precise prediction tasks. The representation G could function as the initial coarse proposals of the important layout elements, which guide subsequent de-

coder processing. Besides, G could also act as a source of multi-scale information for the hybrid generative decoder. Concretely, the decoder’s generated and expanded queries are guided by the learned feature information encoded in G .

Hybrid Generative Decoder

The decoder of HybriDLA is a hybrid generative module that combines an autoregressive query expansion mechanism with a diffusion-based refinement process. This design enables the adaptive generation of layout elements, transitioning from coarse proposals to fine-grained results within these two mutually complementary mechanisms.

Autoregressive Query Expansion (AQE). We formulate the query expansion mechanism as an autoregressive generative process operating over spatial layout regions. Given an extracted document image feature denoted X , the model defines a probability distribution over a variable-length sequence of region queries $Q = (q_1, q_2, \dots, q_N)$. Each query q_t represents a candidate region with associated spatial and semantic parameters, and is generated conditionally on the image feature and all previously generated queries. Formally, the joint distribution over the query sequence factorizes as:

$$\begin{aligned} P(Q | X) &= P(q_1, q_2, \dots, q_N, \text{EOS} | X) \\ &= \prod_{t=1}^N P(q_t | X, q_{1:t-1}) \cdot P(\text{EOS} | X, q_{1:N}) \end{aligned} \quad (1)$$

where EOS is a special end-of-sequence token that terminates the process. This formulation implies that at each step t , a learned proposal distribution $P(q_t | X, q_{<t})$, which is parameterized by the decoder layer, proposes the next query given the current context, and eventually emits EOS to adaptively halt expansion. The decoder layer can be viewed as a parametric function $D_{\theta,t}$ at iteration t that operates over the growing query set. At each iteration step t , $D_{\theta,t}$ conditions on the input X and the generated queries $q_1, \dots, q_{t-1}$ from the previous step, and further conditions the next conditional distribution for the new query q_t . In this way, the model autoregressively “queries the queries,” *i.e.*, it conditions on its own past outputs to decide future ones. During this generative process, each q_t is generated in contextual content. The cardinality N of the query set is not fixed a priori but is controlled in a data-dependent manner by the learned stopping criterion $P(\text{EOS} | X, q_{1:N})$. Sequentially, after every new set of queries has been proposed, the decoder layer assesses whether the current set of queries sufficiently explores the whole layout of the input document image X . If there remain unexplained regions or high residual uncertain regions, the model will continue to propose new queries. Conversely, when the layout has been fully accounted for, the EOS token is emitted, terminating the expansion.

Diffusion-based Refinement (DR). We model the layout refinement process as an implicit denoising operation, similar to a diffusion model but without any forward noise injection. Each decoder layer treats its input prediction as a noisy estimation and produces a residual correction to progressively remove errors. Formally, $\hat{y}^{(t)}$ denote the layout

prediction after t refinement steps, the update rule could be formulated as $\hat{y}^{(t+1)} = \hat{y}^{(t)} + \Delta^{(t)}$, where $\Delta^{(t)}$ is the predicted residual at step t . This residual formulation ensures that the model focuses on correcting deviations from the previous prediction, effectively denoising $\hat{y}^{(t)}$ to approach the accurate layout. During the iteratively refinement by decoder layers, the self-attention mechanism allows decoder queries to share contextual information, while cross-attention integrates relevant visual features, enabling precise localized updates. The feed-forward network then adjusts each query embedding, applying the learned residual $\Delta^{(t)}$ to yield the refined prediction $\hat{y}^{(t+1)}$. Through these attention and feed-forward operations, successive decoder layers systematically reduce spatial errors and converge toward a high-fidelity layout estimation. To train this refinement mechanism, we incorporate denoising queries and intermediate supervision. During training, a subset of decoder queries is initialized with perturbed ground-truth layouts, challenging the network to reconstruct the correct layout from a degraded version. The model is supervised with auxiliary loss at each decoder layer, which forces intermediate predictions $\hat{y}^{(t)}$ to stay close to the ground truth, and with auxiliary heads for fast convergence. These auxiliary training strategy guides the decoder to robustly remove noise and refine predictions at every step, ensuring theoretical accuracy and stable convergence.

By combining autoregressive query expansion with diffusion-based refinement, HybriDLA can flexibly accommodate varying document complexities. If a document with many small elements, more queries will be generated in the expansion stage, and the diffusion refinement will meticulously adjust each one. If the layout is simple, the model generates fewer queries and still refines them for precision. This progressive refinement approach allows the model to iteratively approach the truth, focusing on coarse structure first, then detail, like a human would (Liu 2005; Furnas 1986).

Experiments

Experiment Setup

Dataset. We conduct experiments on the DocLayNet (Pfaffmann et al. 2022) and M6Doc (Cheng et al. 2023) datasets. DocLayNet is a large human-annotated document layout dataset containing 80,863 page images with labeled bounding-box annotations for 11 distinct layout classes, *i.e.*, Caption, Footnote, Formula, List-item, Page-footer, Page-header, Picture, Section-header, Table, Text, Title. M6Doc is a recent multi-format, multi-type, multi-layout dataset comprising 9,080 document pages with fine-grained bounding-box annotations for 74 categories, totaling 237,116 annotated instances. All annotations in both datasets are provided as axis-aligned bounding boxes for each layout element.

Evaluation Protocol. We adopt mean Average Precision (mAP) as the primary evaluation metric, following the standard definition. Specifically, mAP is computed by averaging the Average Precision over a range of Intersection over Union (IoU) thresholds from 0.50 to 0.95 in 0.05 increments. This mAP summarizes detection accuracy across multiple overlap criteria, and we report it as an overall indicator of

performance, *i.e.*, higher mAP indicates better layout detection. All results are reported on the respective official setting of each dataset.

Implementation Details. We benchmark HybriDLA against a comprehensive list of analyzers reproduced under the same training setting as in Table 1, which include

- **Traditional region-based method:** Classical two-stage pipelines remain the standard for layout analysis. We reproduce the strongest public variants with both convolutional and transformer backbones. These methods establish baselines and are still widely used in industry.
- **DETR-based method:** We cover the evolution from vanilla DETR to domain-specialized DLAFormer with both convolutional and Transformer backbones. Set-prediction makes DETR variants the most competitive contemporary baselines.
- **Diffusion-based method:** DiffusionDet (Chen et al. 2022a) represents the emerging class of generative detectors. We evaluate on both convolutional and transformer backbones to examine whether coarse-to-fine refinement benefits complex page structures.
- **Autoregressive method:** We include Pix2Seq (Chen et al. 2021) with transformer backbone. These methods sequentially predict objects as token strings, offering an alternative that unifies detection and language modeling.
- **Hybrid method (ours):** We test HybriDLA method with 5 backbones to systematically validate the generality.

All backbones are initialized from the official weights, while analyzers are trained from scratch. Every model is optimized with a batch size of 40, except for DLAFormer, retrained in the same setting to ensure comparability.

Experiment Results on DocLayNet

Table 1 summarizes the detection performance of various methods on the DocLayNet (Pfitzmann et al. 2022) benchmark. Traditional region-based methods with only vision achieve less than 73.5% mAP. ResNet-101 (He et al. 2015) with Mask R-CNN (He et al. 2017) reaches 73.5% mAP, followed by Faster R-CNN (Ren et al. 2017) at 73.4%. While self-attention integrated DocSegTr (Biswas et al. 2022) yields 69.3%, a specialized document-transformer backbone DiT (Li et al. 2022) with Cascade R-CNN (Cai and Vasconcelos 2018) attains only 62.1%. Incorporating textual input provides a modest boost to LayoutLMv3 (Huang et al. 2022), improving to 75.1% mAP, which indicates the benefit of multimodal cues over purely visual baselines.

In contrast, DETR-based methods achieve higher accuracy on DocLayNet (Pfitzmann et al. 2022), with mAP more than 73%. Vanilla DETR (Carion et al. 2020) with InternImage (Wang et al. 2023) backbone obtains 74.3% mAP, while Deformable DETR (Zhu et al. 2021) and DINO (Zhang et al. 2022) reach 76.1% and 75.7% mAP, respectively. Co-DINO (Zong, Song, and Liu 2023) further improves to 77.2% mAP, and a Swin-based SwinDocSegmenter (Banerjee et al. 2023) achieves 76.9% mAP. The strongest vision-only DETR-based method is RoDLA (Chen et al. 2024) at 80.5% mAP, representing a substantial improvement. Besides, DiffusionDet (Chen et al. 2022a) yields up to 76.3%

Table 1: Experiment results on DocLayNet dataset. V, L, and T denote the Visual, Layout, and Textual modalities.

Backbone	Detector	#Params	Modality			mAP
			V	L	T	
Region-based Method						
ResNet-101	Mask R-CNN	63M	✓	✗	✗	73.5
ResNet-101	Faster R-CNN	61M	✓	✗	✗	73.4
ResNet-101	DocSegTr	103M	✓	✗	✗	69.3
DiT	Cascade R-CNN	304M	✓	✗	✗	62.1
LayoutLMv3	Cascade R-CNN	368M	✓	✓	✓	75.1
DETR-based Method						
InternImage	DETR	352M	✓	✗	✗	74.3
InternImage	Deformable DETR	353M	✓	✗	✗	76.1
InternImage	DINO	358M	✓	✗	✗	75.7
InternImage	Co-DINO	363M	✓	✗	✗	77.2
Swin-L	SwinDocSegmenter	223M	✓	✗	✗	76.9
InternImage	RoDLA	323M	✓	✗	✗	80.5
ResNet-50	DLAFormer	—	✓	✓	✓	83.8
Diffusion-based Method						
ResNet-50	DiffusionDet	111M	✓	✗	✗	73.7
Swin-L	DiffusionDet	283M	✓	✗	✗	76.3
Autoregressive Method						
ViT-L	Pix2Seq	341M	✓	✗	✗	72.5
FIT-L	Pix2Seq	370M	✓	✗	✗	73.4
Hybrid Method						
ResNet-50	Ours	95M	✓	✗	✗	74.4
ResNet-101	Ours	114M	✓	✗	✗	76.9
ViT-L	Ours	385M	✓	✗	✗	78.8
Swin-L	Ours	270M	✓	✗	✗	80.4
InternImage	Ours	392M	✓	✗	✗	83.5

mAP with Swin Transformer (Liu et al. 2021) backbone, while autoregressive model Pix2Seq (Chen et al. 2021) with ViT (Dosovitskiy et al. 2021) backbone reach 72.5% mAP.

Compared to these baseline models, our HybriDLA model obtains a state-of-the-art performance among vision-only document layout analysis models. As shown in the bottom section of Table 1, the hybrid approach outperforms prior methods across diverse backbones. With the InternImage (Wang et al. 2023) backbone, HybriDLA achieves 83.5% mAP, which is the best result for vision-only layout analysis model. This nearly matches the performance of 83.8% mAP by DLAFormer (Wang, Hu, and Huo 2024), indicating that HybriDLA narrows the performance gap to multi-modal systems using only visual features. HybriDLA also yields an average 3% mAP gains with other backbones. Even with a smaller ResNet-50, HybriDLA achieves 74.4%, essentially matching the DiffusionDet baseline. These consistent gains across different backbone models demonstrate the effectiveness and generality of our HybriDLA model.

Experiment Results on M⁶Doc

Table 2 presents the detection accuracy of various approaches on the challenging M⁶Doc (Cheng et al. 2023) dataset. For traditional region-based methods, ResNet-101 (He et al. 2015) with Mask R-CNN (He et al. 2017) achieves 61.9% mAP, and Faster R-CNN (Ren et al.

Table 2: Experiment results on M⁶Doc dataset.

Backbone	Detector	#Params	Modality			mAP
			V	L	T	
<i>Region-based Method</i>						
ResNet-101	Mask R-CNN	63M	✓	✗	✗	61.9
ResNet-101	Faster R-CNN	61M	✓	✗	✗	62.0
ResNet-101	DocSegTr	103M	✓	✗	✗	60.3
DiT	Cascade R-CNN	304M	✓	✗	✗	70.2
LayoutLMv3	Cascade R-CNN	368M	✓	✓	✓	64.3
<i>DETR-based Method</i>						
InternImage	DETR	352M	✓	✗	✗	54.2
InternImage	Deformable DETR	353M	✓	✗	✗	61.2
InternImage	DINO	358M	✓	✗	✗	66.8
Swin-L	SwinDocSegmenter	223M	✓	✗	✗	47.1
InternImage	RoDLA	323M	✓	✗	✗	70.0
<i>Diffusion-based Method</i>						
Swin-L	DiffusionDet	283M	✓	✗	✗	62.7
<i>Autoregressive Method</i>						
ViT-L	Pix2Seq	341M	✓	✗	✗	54.9
FIT-L	Pix2Seg	370M	✓	✗	✗	54.9
<i>Hybrid Method</i>						
ResNet-50	Ours	95M	✓	✗	✗	62.1
ResNet-101	Ours	114M	✓	✗	✗	64.7
ViT-L	Ours	385M	✓	✗	✗	68.6
Swin-L	Ours	270M	✓	✗	✗	68.1
InternImage	Ours	392M	✓	✗	✗	71.4

2017) yields a similar 62.0%. The self-attention-based DocSegTr (Biswas et al. 2022) attains 60.3% mAP, slightly lower than the R-CNN models. Notably, the document-specific transformer backbone DiT (Li et al. 2022) combined with Cascade R-CNN (Cai and Vasconcelos 2018) stands out with a significantly higher 70.2% mAP, indicating the benefit of pretraining on document layouts even without text input. We also observe that incorporating textual modality provides only a modest boost on M⁶Doc (Cheng et al. 2023) dataset, the multimodal LayoutLMv3 (Huang et al. 2022) with multimodal features only reaches 64.3% mAP.

Transformer-based DETR-style methods exhibit a wide range of results. The vanilla DETR (Carion et al. 2020) with an InternImage (Wang et al. 2023) backbone reaching only 54.2% mAP, likely due to the difficulty with a large number of classes and diverse layouts. Introducing multi-scale deformable attention significantly improves performance. Deformable DETR (Zhu et al. 2021) rises to 61.2% mAP, and DINO (Zhang et al. 2022) achieves 66.8% mAP. Interestingly, the SwinDocSegmenter (Banerjee et al. 2023) underperforms on this dataset, with only 47.1% mAP. Among DETR-based approaches, the recent RoDLA (Chen et al. 2024) obtains the highest result at 70.0% mAP, nearly matching the DiT-based region method.

Beyond DETR-style models, DiffusionDet (Chen et al. 2022a) achieves 62.7% mAP, which is better than the initial DETR and R-CNN baselines but still below the

Table 3: Ablation study of autoregressive query expansion (AQE) mechanism. Deformable stands for deformable attention mechanism, and DINO stands for improved denoising anchor boxes mechanism.

Backbone	DETR	Deformable	DINO	AQE	mAP
ResNet-50	✓	✗	✗	✗	74.2
ResNet-50	✓	✗	✗	✓	74.4
ResNet-50	✓	✓	✗	✗	75.1
ResNet-50	✓	✓	✗	✓	76.3
ResNet-50	✓	✓	✓	✗	75.3
ResNet-50	✓	✓	✓	✓	76.8
Swin-L	✓	✓	✓	✓	77.2

SOTA-performing transformer methods. Autoregressive sequence prediction approaches perform less competitively, both Pix2Seq (Chen et al. 2021) and its improved variant with FIT (Chen and Li 2023) achieve 54.9% mAP.

In contrast to the above baselines, our proposed HybriDLA approach establishes a state-of-the-art performance for vision-only document layout analysis on M⁶Doc (Cheng et al. 2023) dataset. As shown in the bottom section of Table 2, HybriDLA consistently outperforms prior methods across all backbone architectures. With the powerful InternImage backbone, HybriDLA achieves 71.4% mAP and even outperforms the multimodal LayoutLMv3 (Huang et al. 2022) by over 7% mAP. With ResNet-101 (He et al. 2015) backbone, HybriDLA reaches 64.7%, which is a 2% obvious improvement over traditional R-CNN methods on the same backbone. Similarly, with the ViT (Dosovitskiy et al. 2021) and Swin Transformer (Liu et al. 2021) backbones, HybriDLA yields 68.6% mAP and 68.1% mAP, respectively, considerably higher than the corresponding DETR-based and DiffusionDet (Chen et al. 2022a) results. These consistent gains across diverse backbone types demonstrate the effectiveness and generality of our hybrid approach.

Ablation Study

To comprehensively evaluate our proposed HybriDLA method, we conducted ablation studies from five perspectives on the DocLayNet (Pfaffmann et al. 2022) dataset.

The effect of autoregressive query expansion (AQE) mechanism. As shown in Table 3, enabling the AQE mechanism consistently improves detection performance across different model variants. Compared to a marginal mAP increase of the ResNet-50 DETR baseline, stronger DINO DETR (Zhang et al. 2022) raises mAP from 75.3% to 76.8%, indicating a marginal effect. This trend suggests that the benefits of AQE become more pronounced as the layout analyzer is more advanced, which shows the complementary nature of the proposed AQE mechanism.

The effect of feature fusion encoder (FFE). Table 4 contrasts the proposed FFE with the standard deformable encoder (DE). On high-capacity, multi-scale backbones, the FFE delivers clear benefits. With Swin-L, mAP rises from 77.2% to 79.1% even before any diffusion refinement mech-

Table 4: Ablation study on feature fusion encoder, diffusion-based refinement mechanism and backbone selection. FFE stands for feature fusion encoder, DE stands for normal deformable attention encoder, DR stands for diffusion-based refinement mechanism, and AQE stands for autoregressive query expansion mechanism.

Backbone	Encoder	DR	AQE	mAP
ResNet-50	DE	✓	✓	76.2
ResNet-50	FFE	✗	✓	74.4
ResNet-50	FFE	✓	✓	74.4
Swin-L	DE	✓	✓	78.1
Swin-L	FFE	✗	✓	79.1
Swin-L	FFE	✓	✓	80.4
ResNet-101	DE	✓	✓	76.6
ResNet-101	FFE	✗	✓	76.1
ResNet-101	FFE	✓	✓	76.9
ViT-L	DE	✓	✓	79.0
ViT-L	FFE	✗	✓	77.2
ViT-L	FFE	✓	✓	78.8
InternImage	DE	✓	✓	82.4
InternImage	FFE	✗	✓	81.3
InternImage	FFE	✓	✓	83.5

anism, and reaches 80.4% mAP once FFE is paired with DR, which contains 2.3% mAP gains compared to DE with DR. FFE with InternImage (Wang et al. 2023) shows a similar trend, climbing from 82.4% mAP to 83.5% mAP. ResNet-101 (He et al. 2015) also benefits slightly. In contrast, smaller models see limited impact, *e.g.*, ResNet-50 (He et al. 2015) drops from 76.2% mAP to 74.4% mAP without DR. These results indicate that the FFE is advantageous when the backbone provides context-rich features.

The effect of diffusion-based refinement (DR) mechanism. As shown in Table 4, integrating the DR mechanism yields a consistent performance uplift across all backbones with the sole exception of ResNet-50 (He et al. 2015), whose performance remains unchanged irrespective of whether DR is employed. The ResNet-101 (He et al. 2015) gains 0.8% mAP improvement, and similarly, the mAP result of Swin Transformer (Liu et al. 2021) increases from 79.1% to 80.4%. The magnitude of improvement varies with model capacity. Smaller backbones tend to gain relatively more, while larger models slightly improve, implying that a larger backbone may require more steps to eliminate residual noise while more detailed features are extracted from images.

The generalization of backbone selection. Table 4 also demonstrates monotonic scaling with model capacity, as the backbone network capacity increases, performance increases continuously without exception. The mAP results rise from 74.4% with ResNet-50 (He et al. 2015), to 83.5% with InternImage (Wang et al. 2023), showing a clear upward trend. This monotonic improvement indicates that our approach leverages the representational strength of larger backbones, achieving higher accuracy at each step up in

Table 5: Ablation study on the impact of the number of initial proposal queries and AQE queries.

Backbone	#Init. Q.	#AQE Q.	mAP
ResNet-50	300	300	74.4
ResNet-101	300	300	76.9
ViT-L	300	300	78.8
Swin-L	300	300	80.4
InternImage	300	300	83.5
ResNet-50	300	30	77.4
ResNet-101	300	30	76.9
ViT-L	300	30	78.2
Swin-L	300	30	78.2
InternImage	300	30	80.7
ResNet-50	900	30	75.5
ResNet-101	900	30	76.6
ViT-L	900	30	79.0
Swin-L	900	30	75.7
InternImage	900	30	81.4

model size. Moreover, the HybriDLA framework exhibits remarkable stability across diverse types of backbones, maintaining stable results regardless of backbone types.

The analysis for hyper-parameters. As shown in Table 5, the number of initial queries and expanded autoregressive queries has a significant impact on detection performance. For smaller backbones, reducing the expanded queries from 300 to 30 does not degrade performance. In contrast, larger backbones benefit from more queries, cutting the expanded queries to 30 leads to notable drops in mAP. Moreover, increasing the initial queries from 300 to 900 produces inconsistent results. The highest-capacity models show slight improvements, while others suffer a decline. These results suggest that each model has an optimal query budget. Smaller models saturate with fewer queries, while larger models can leverage more queries for better performance.

Conclusion

In this work, we introduced HybriDLA, a hybrid approach to document layout analysis that addresses complex and diverse document layouts. HybriDLA employs a coarse-to-fine generation strategy via diffusion-based layout refinement and autoregressive query expansion mechanisms, effectively capturing both global layout context and fine-grained details. Empirical evaluations on datasets demonstrate that HybriDLA achieves state-of-the-art performance in document layout analysis. Moreover, the architecture-agnostic design of HybriDLA enables seamless integration with various backbone structures, illustrating the generality and broad applicability. However, the current framework relies solely on visual inputs, which constitutes a notable limitation of analysis accuracy. As future work, we intend to address this issue by incorporating multimodal features to guide the layout generation process for better performance.

Acknowledgments

This work was supported in part by Helmholtz Association of German Research Centers, in part by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, and in part by National Natural Science Foundation of China under Grant No. 62503166. This work was partially performed on the HoreKa supercomputer funded by the MWK and by the Federal Ministry of Education and Research, partially on the HAICORE@KIT partition supported by the Helmholtz Association Initiative and Networking Fund, and partially on bwForCluster Helix supported by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

- Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. DocFormer: End-to-End Transformer for Document Understanding. In *ICCV*.
- Arroyo, D. M.; Postels, J.; and Tombari, F. 2021. Variational transformer networks for layout generation. In *CVPR*.
- Banerjee, A.; Biswas, S.; Lladós, J.; and Pal, U. 2023. Swin-DocSegmenter: An End-to-End Unified Domain Adaptive Transformer for Document Instance Segmentation. *ICDAR*.
- Biswas, S.; Banerjee, A.; Lladós, J.; and Pal, U. 2022. DocSegTr: An Instance-Level End-to-End Document Image Segmentation Transformer. *arXiv preprint arXiv:2201.11438*.
- Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *CVPR*.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 213–229. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58451-1.
- Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2022a. Diffusion-Det: Diffusion Model for Object Detection. *arXiv preprint arXiv:2211.09788*.
- Chen, T.; and Li, L. 2023. FIT: Far-reaching Interleaved Transformers. *arXiv preprint arXiv:2305.12689*.
- Chen, T.; Saxena, S.; Li, L.; Fleet, D. J.; and Hinton, G. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*.
- Chen, T.; Saxena, S.; Li, L.; Lin, T.-Y.; Fleet, D. J.; and Hinton, G. 2022b. A Unified Sequence Interface for Vision Tasks. *arXiv preprint arXiv:2206.07669*.
- Chen, Y.; Liu, R.; Zheng, J.; Wen, D.; Peng, K.; Zhang, J.; and Stiefelhagen, R. 2025. Graph-based Document Structure Analysis. *arXiv preprint arXiv:2502.02501*.
- Chen, Y.; Zhang, J.; Peng, K.; Zheng, J.; Liu, R.; Torr, P.; and Stiefelhagen, R. 2024. RoDLA: Benchmarking the Robustness of Document Layout Analysis Models. In *CVPR*.
- Cheng, H.; Zhang, P.; Wu, S.; Zhang, J.; Zhu, Q.; Xie, Z.; Li, J.; Ding, K.; and Jin, L. 2023. M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis. In *CVPR*.
- Constum, T.; Tranouez, P.; and Paquet, T. 2025. DANIEL: A fast Document Attention Network for Information Extraction and Labelling of handwritten documents. *IJDAR*.
- Coquenot, D.; Chatelain, C.; and Paquet, T. 2023. DAN: a segmentation-free document attention network for handwritten document recognition. *TPAMI*.
- Diem, M.; Kleber, F.; and Sablatnig, R. 2011. Text classification and document layout analysis of paper fragments. In *ICDAR*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.
- Furnas, G. W. 1986. Generalized fisheye views. *SIGCHI Bull.*, 17(4): 16–23.
- Garz, A.; Diem, M.; and Sablatnig, R. 2010. Detecting text areas and decorative elements in ancient manuscripts. In *ICFHR*.
- Gemelli, A.; Biswas, S.; Civitelli, E.; Lladós, J.; and Marinai, S. 2022. Doc2graph: a task agnostic document understanding framework based on graph neural networks. In *ECCV*.
- He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.
- Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACMMM*.
- Li, J.; Xu, Y.; Lv, T.; Cui, L.; Zhang, C.; and Wei, F. 2022. DiT: Self-supervised Pre-training for Document Image Transformer. *arXiv preprint arXiv:2203.02378*.
- Li, M.; Xu, Y.; Cui, L.; Huang, S.; Wei, F.; Li, Z.; and Zhou, M. 2020. DocBank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*.
- Li, Y.; Qian, Y.; Yu, Y.; Qin, X.; Zhang, C.; Liu, Y.; Yao, K.; Han, J.; Liu, J.; and Ding, E. 2021. StrucTexT: Structured Text Understanding with Multi-Modal Transformers. In *ACMMM*.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Liu, Z. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation*, 61(6): 700–712.

- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. Los Alamitos, CA, USA: IEEE Computer Society.
- Long, S.; Qin, S.; Panteleev, D.; Bissacco, A.; Fujii, Y.; and Raptis, M. 2022. Towards end-to-end unified scene text detection and layout analysis. In *CVPR*.
- Luo, C.; Cheng, C.; Zheng, Q.; and Yao, C. 2023. Geo-LayoutLM: Geometric Pre-training for Visual Information Extraction. In *CVPR*.
- Luo, C.; Tang, G.; Zheng, Q.; Yao, C.; Jin, L.; Li, C.; Xue, Y.; and Si, L. 2022. Bi-VLDoc: Bidirectional Vision-Language Modeling for Visually-Rich Document Understanding. *ArXiv*.
- Moured, O.; Zhang, J.; Roitberg, A.; Schwarz, T.; and Stiefelhagen, R. 2023. Line Graphics Digitization: A Step Towards Full Automation. In *ICDAR*, 438–453. Springer.
- Peng, Q.; Pan, Y.; Wang, W.; Luo, B.; Zhang, Z.; Huang, Z.; Hu, T.; Yin, W.; Chen, Y.; Zhang, Y.; et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.
- Pfitzmann, B.; Auer, C.; Dolfi, M.; Nassar, A. S.; and Staar, P. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *SIGKDD*.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *TPAMI*.
- Shen, Y.; Luo, C.; Zhu, Z.; Chen, Y.; Zheng, Q.; Yu, Z.; Bu, J.; and Yao, C. 2025. Protag: Process tagging for assessing the efficacy of document instruction data. In *AAAI*.
- Shen, Z.; Zhang, K.; and Dell, M. 2020. A large dataset of historical Japanese documents with complex layouts. In *CVPR*.
- Shihab, M. I. H.; Hasan, M. R.; Emon, M. R.; Hossen, S. M.; Ansary, M. N.; Ahmed, I.; Rakib, F. R.; Dhruvo, S. E.; Dip, S. S.; Pavel, A. H.; et al. 2023. Badlad: A large multi-domain bengali document layout analysis dataset. In *ICDAR*.
- Tang, Z.; Yang, Z.; Wang, G.; Fang, Y.; Liu, Y.; Zhu, C.; Zeng, M.; Zhang, C.; and Bansal, M. 2023. Unifying vision, text, and layout for universal document processing. In *CVPR*.
- Wang, J.; Hu, K.; and Huo, Q. 2024. DLAFormer: An End-to-End Transformer For Document Layout Analysis. *arXiv:2405.11757*.
- Wang, J.; Jin, L.; and Ding, K. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *ACL*.
- Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. 2023. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*.
- Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*.
- Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; Zhang, M.; and Zhou, L. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *ACL*.
- Yang, H.; and Hsu, W. 2022. Transformer-based approach for document layout understanding. In *ICIP*.
- Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.
- Zhang, J.; You, Z.; Wang, J.; and Le, X. 2025a. Sail: Sample-centric in-context learning for document information extraction. In *AAAI*.
- Zhang, P.; Li, C.; Qiao, L.; Cheng, Z.; Pu, S.; Niu, Y.; and Wu, F. 2021. VSR: a unified framework for document layout analysis combining vision, semantics and relations. In *ICDAR*.
- Zhang, Z.; Zhang, Y.; Liang, Y.; Ma, C.; Xiang, L.; Zhao, Y.; Zhou, Y.; and Zong, C. 2025b. Understand Layout and Translate Text: Unified Feature-Conductive End-to-End Document Image Translation. *TPAMI*.
- Zhong, X.; Tang, J.; and Yepes, A. J. 2019. PubLayNet: largest dataset ever for document layout analysis. In *ICDAR*.
- Zhu, F.; Lei, W.; Feng, F.; Wang, C.; Zhang, H.; and Chua, T.-S. 2022. Towards complex document understanding by discrete reasoning. In *ACMMM*.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Zong, Z.; Song, G.; and Liu, Y. 2023. DETRs with Collaborative Hybrid Assignments Training. In *ICCV*.

Social Impact and Limitations

Social Impact. We demonstrate that the advancement of HybriDLA in document layout analysis can directly benefit several real-world domains. In digital archiving, our method helps preserve and digitize complex historical documents by accurately recognizing diverse layouts, which aids libraries and museums in making cultural heritage accessible online. It also enhances screen-reader accessibility by converting visually rich layouts into structured content flows, empowering visually impaired users to navigate documents that were previously inaccessible due to complex formatting. Similarly, in education, HybriDLA can automatically organize content into logical sections, figures, and captions, facilitating interactive and adaptive learning materials. Additionally, by accurately extracting and grouping fields in forms and administrative documents, our system can streamline data entry and information retrieval in business and e-governance applications. These use cases highlight HybriDLA’s potential for broad positive social impact through improved access to information and automation of document workflows.

Limitations Despite its strengths, our approach has several limitations. First, HybriDLA relies on document images without leveraging textual context or OCR results, which means it might miss semantic cues that could disambiguate layout elements, *e.g.*, distinguishing figure captions from body text purely by appearance. Second, the hybrid generation mechanism of our model comes with a high inference cost, which combines multiple analysis stages and generative components, making it computationally intensive and less practical for real-time or large-scale processing. This could limit deployment on edge devices or slow down analysis for massive archives. We are exploring several future directions to address these issues. For instance, incorporating multimodal inputs or metadata could enrich the model’s understanding beyond pure visuals, and optimizing the architecture or using model distillation can reduce inference overhead. These steps should make HybriDLA more practical and reliable for widespread adoption.

Experiment Details

Evaluation Metric Details.

We report **mean Average Precision (mAP)** following the COCO object-detection evaluation protocol (Lin et al. 2014). For each class $c \in \mathcal{C}$ and IoU threshold $t \in \mathcal{T}$, we compute the Average Precision (AP) as the area under the precision–recall curve:

$$AP_{c,t} = \int_0^1 \text{Prec}_{c,t}(r) dr, \quad (2)$$

where $\text{Prec}_{c,t}(r)$ denotes the maximum precision obtained for recall r after 101-point interpolation used in the COCO API. The mAP we report averages AP over all classes and over the 10 IoU thresholds

$$\mathcal{T} = \{0.50, 0.55, \dots, 0.95\}, \quad (3)$$

yielding

$$\text{mAP} = \frac{1}{|\mathcal{C}| |\mathcal{T}|} \sum_{c \in \mathcal{C}} \sum_{t \in \mathcal{T}} AP_{c,t}. \quad (4)$$

This metric rewards detectors that are simultaneously accurate in classification and precise in localization across a spectrum of overlap criteria, and it is the standard for fair comparison in modern document layout analysis benchmarks. All results in the main paper and supplementary material are computed with the official COCO (Lin et al. 2014) evaluation script to ensure reproducibility.

Implementation Details.

To comprehensively benchmark both classic and state-of-the-art object-detection paradigms for document layout analysis (DLA), we trained 15 distinct backbone–detector combinations, which span region-based, DETR-based, diffusion-based, and autoregressive methods on the DocLayNet and M6Doc datasets. In addition, we evaluated five backbone variants of our proposed HybriDLA detector, enabling a direct comparison against existing SOTA methods. The architectural characteristics, strengths of all evaluated models are summarized below:

Traditional region-based method:

- **ResNet-101** (He et al. 2015) with **Faster R-CNN** (Ren et al. 2017). Faster R-CNN (Ren et al. 2017) is a classic two-stage detector that employs a ResNet-101 (He et al. 2015) backbone with a Region Proposal Network to generate candidate object regions. It prioritizes accuracy and provides a strong baseline for layout detection.
- **ResNet-101** (He et al. 2015) with **Mask R-CNN** (He et al. 2017). Mask R-CNN (He et al. 2017) is a two-stage, region-based detector that extends the Faster R-CNN (Ren et al. 2017) architecture by adding a parallel mask segmentation branch for each detected object.
- **ResNet-101** (He et al. 2015) with **DocSegTr** (Biswas et al. 2022). DocSegTr (Biswas et al. 2022) is a self-attention integrated model for documents that combines visual features with a twin-attention module. As one of the first attention-driven approaches for document layout analysis, it emphasizes efficient semantic reasoning and achieves competitive accuracy on complex layouts.
- **DiT** (Li et al. 2022) with **Cascade R-CNN** (Cai and Vasconcelos 2018). This combination integrates the Document Image Transformer (DiT) (Li et al. 2022) as the backbone in a Cascade R-CNN (Cai and Vasconcelos 2018) detection framework. The DiT (Li et al. 2022) backbone, pre-trained on large-scale unlabeled document images, provides rich document-specific features, while the cascade of detectors, which is trained with increasing IoU thresholds, refines layout element predictions at multiple stages for improved quality.
- **LayoutLMv3** (Huang et al. 2022) with **Cascade R-CNN** (Cai and Vasconcelos 2018). LayoutLMv3 (Huang et al. 2022) is a pre-trained multimodal Transformer that jointly encodes text, layout coordinates, and image signals with unified masking strategies. We employ LayoutLMv3 as the backbone in a Cascade R-CNN – leveraging its cross-modal document representations – and

refine predictions through multiple stages, which yields state-of-the-art accuracy in document layout analysis.

DETR-based method:

- **InternImage** (Wang et al. 2023) with **DETR** (Carion et al. 2020). This combination pairs the large CNN-based InternImage (Wang et al. 2023) backbone with the DETR (Carion et al. 2020) set-prediction framework for object detection. The backbone uses deformable convolutions to attain a large effective receptive field and adaptive spatial aggregation, complementing the end-to-end Transformer decoder of DETR (Carion et al. 2020) to produce strong layout detection results without the need for non-maximum suppression.
- **InternImage** (Wang et al. 2023) with **Deformable DETR** (Zhu et al. 2021). This model combines with Deformable DETR (Zhu et al. 2021), an improved DETR variant that employs multi-scale deformable attention for faster convergence and better small-object detection. The image features together with efficient deformable attention allow the detector to learn document layouts more effectively, improving both training speed and detection accuracy over the DETR (Carion et al. 2020) approach.
- **InternImage** (Wang et al. 2023) with **DINO** (Zhang et al. 2022). In this combination, we use the DINO (Zhang et al. 2022) framework as a detector, which is a DETR (Carion et al. 2020) with improved denoising and query initialization techniques. By using contrastive denoising training and hybrid anchor box queries, DINO (Zhang et al. 2022) enhances detection performance and, coupled with strong visual features, achieves high mAP on the document layout analysis task.
- **InternImage** (Wang et al. 2023) with **Co-DINO** (Zong, Song, and Liu 2023). This model builds upon Collaborative DETR (Zong, Song, and Liu 2023) training schemes applied to the DINO (Zhang et al. 2022) detector. It utilizes a collaborative hybrid label assignment strategy, *i.e.*, training multiple auxiliary heads with one-to-many assignments to boost the learning of the model. The Co-DINO (Zong, Song, and Liu 2023) approach strengthens the encoder’s feature extraction and improves query decoding, resulting in improved detection precision compared to the standard DINO (Zhang et al. 2022).
- **Swin Transformer** (Liu et al. 2021) with **Swin-DocSegmenter** (Banerjee et al. 2023). SwinDocSegmenter (Banerjee et al. 2023) is an end-to-end transformer for document layout analysis that introduces a domain-adaptive training approach with contrastive learning and mixed query initialization in the transformer decoder, enabling robust analysis of complex document layouts and achieving superior accuracy.
- **InternImage** (Wang et al. 2023) with **RoDLA** (Chen et al. 2024). RoDLA is a DINO-inspired (Zhang et al. 2022) detector focused on robustness under image perturbations. It incorporates channel-attention and average-pooling blocks into the DINO (Zhang et al. 2022) transformer encoder and uses the InternImage (Wang et al. 2023) backbone pre-trained on ImageNet-22K. This design enhances multi-scale feature stability, yielding a

highly robust layout detector that achieves state-of-the-art performance on document layout analysis tasks even under challenging conditions.

- **ResNet-50** (He et al. 2015) with **DLAFormer** (Wang, Hu, and Huo 2024). DLAFormer (Wang, Hu, and Huo 2024) is an end-to-end transformer model for Document Layout Analysis that unifies multiple sub-tasks in a single DETR-like (Carion et al. 2020) framework. Using a ResNet-50 (He et al. 2015) backbone for feature extraction, it treats detection, classification, and even reading order as joint relation predictions in a unified label space. This approach, along with type-specific query embeddings with text and layout modalities integration, allows DLAFormer to concurrently detect layout elements and infer their roles, achieving top performance. However, there is no detailed description of the model-specific internal module structure, and the code was not open-sourced after publication. We are unable to reproduce its structural design and can only compare it with the results provided in the paper for reference.

Diffusion-based method:

- **ResNet-50** (He et al. 2015) with **DiffusionDet** (Chen et al. 2022a). DiffusionDet (Chen et al. 2022a) formulates object detection as a denoising diffusion process that transforms noisy bounding boxes into accurate object boxes. Implemented with a ResNet-50 (He et al. 2015) backbone, the model generates a set of random boxes and iteratively refines them through the diffusion model, allowing a flexible number of proposals and improved results with more refinement steps. This diffusion-based detector provides a novel paradigm for layout analysis by treating detection as progressive generation rather than direct regression.
- **Swin Transformer** (Liu et al. 2021) with **Diffusion-Det** (Chen et al. 2022a). This combination applies the same diffusion-based detection approach using a Swin Transformer (Liu et al. 2021) backbone for feature extraction. The stronger backbone captures richer visual context from document images, which improves the quality of the denoising process and yields higher detection accuracy. By combining the features with iterative diffusion refinement, this model achieves the best results among diffusion-based methods in our comparison.

Autoregressive method:

- **ViT** (Dosovitskiy et al. 2021) with **Pix2Seq** (Chen et al. 2021). Pix2Seq (Chen et al. 2021) is an autoregressive object detection method that casts detection as a language modeling task, *i.e.*, the model outputs a sequence of tokens representing bounding box coordinates and class labels. This combination uses a ViT (Dosovitskiy et al. 2021) backbone to encode the input image into a sequence of tokens, then a Transformer decoder generates the sequence of layout element descriptions. This approach allows the model to learn to read out object locations and classes sequentially, yielding competitive detection results with a unified generative framework.
- **FIT** (Chen and Li 2023) with **Pix2Seq** (Chen et al. 2021). This combination employs the Far-reaching Interleaved

Transformer (FIT) (Chen and Li 2023) architecture as the backbone for the Pix2Seq (Chen et al. 2021) framework. FIT (Chen and Li 2023) is a large Transformer that interleaves local and global self-attention layers, enabling efficient modeling of very long token sequences. By using a powerful encoder, this Pix2Seq (Chen et al. 2021) model can handle high-resolution document images and long output sequences more effectively.

Our Proposed hybrid method:

- **ResNet-50** (He et al. 2015) with **HybriDLA**. Using ResNet-50 (He et al. 2015) as a lightweight visual backbone, our HybriDLA detector retains the feature-fusion encoder and hybrid generative decoder pipeline, allowing coarse-to-fine query expansion even on limited compute.
- **ResNet-101** (He et al. 2015) with **HybriDLA**. Replacing the backbone with a deeper ResNet-101 (He et al. 2015) provides richer convolutional features that are fed into the same hybrid decoder, improving spatial detail during diffusion-style refinement.
- **ViT** (Dosovitskiy et al. 2021) with **HybriDLA**. This variant utilizes a ViT (Dosovitskiy et al. 2021) encoder, whose global self-attention is paired with query decoding from HybriDLA, allowing the model to capture long-range layout dependencies with minimal inductive bias.
- **Swin Transformer** (Liu et al. 2021) with **HybriDLA**. Integrating a Swin Transformer (Liu et al. 2021) backbone supplies multi-scale window self-attention features that align with our feature-fusion encoder, yielding stronger localization for small layout elements.
- **InternImage** (Wang et al. 2023) with **HybriDLA**. Coupling HybriDLA with the large deformable-convolution InternImage (Wang et al. 2023) backbone maximizes receptive-field coverage and adaptive feature aggregation before hybrid decoding.

All models were trained under comparable settings for a fair evaluation with the Nvidia A100 GPUs. We trained all models for 36 epochs on both DocLayNet and M⁶Doc. Training uses the AdamW optimizer with an initial learning rate of 2×10^{-4} for a base batch size of 40 images and weight decay of 1×10^{-4} . The learning rate follows a multi-step schedule, dropping by $10\times$ at epoch 30. We employed gradient clipping with max norm 0.5 to stabilize training, and maintained an exponential moving average (EMA) of model weights with decay rate 0.9999 for more robust convergence.

Additional Detailed Analysis of Ablation Study

To further validate our design choices, we conduct two additional sets of ablation experiments, focusing on the influence of architectural components and hyperparameters on detection performance. In particular, we examine the effect of the Transformer decoder depth together with our diffusion-based refinement (DR) mechanism and feature fusion encoder (FFE), and we investigate the interplay between the encoder type, the presence of the DR module, and the number of queries used in the autoregressive query expansion (AQE) pipeline. All experiments are performed on the DocLayNet (Pfzmann et al. 2022) dataset using two back-

bone architectures, *i.e.*, ResNet-50 (He et al. 2015) and Swin Transformer (Liu et al. 2021), to observe these effects.

Table 6: Impact of number of decoder layers under different combinations of the diffusion-based refinement (DR) mechanism with the feature fusion encoder (FFE). All models use 300 initial queries and 300 AQE queries on the DocLayNet (Pfzmann et al. 2022) dataset.

Backbone	Detector	Decoder Layers	Encoder	DR	AQE	mAP
ResNet-50	HybriDLA	1	FFE	✗	✓	74.20
ResNet-50	HybriDLA	1	FFE	✓	✓	72.80
ResNet-50	HybriDLA	3	FFE	✗	✓	74.40
ResNet-50	HybriDLA	3	FFE	✓	✓	73.10
ResNet-50	HybriDLA	6	FFE	✗	✓	74.40
ResNet-50	HybriDLA	6	FFE	✓	✓	74.40
Swin-L	HybriDLA	1	FFE	✗	✓	76.80
Swin-L	HybriDLA	1	FFE	✓	✓	76.60
Swin-L	HybriDLA	3	FFE	✗	✓	78.90
Swin-L	HybriDLA	3	FFE	✓	✓	78.20
Swin-L	HybriDLA	6	FFE	✗	✓	79.10
Swin-L	HybriDLA	6	FFE	✓	✓	80.40

The ablation study of the decoder depth. Table 6 examines how the decoder depths influence mAP under different settings of our diffusion-based refinement (DR) mechanism with the feature fusion encoder fixed. We report results on DocLayNet for a lightweight ResNet-50 (He et al. 2015) backbone with a large Swin Transformer (Liu et al. 2021) backbone, each evaluated with 1, 3, and 6 decoder layers.

For ResNet-50 (He et al. 2015), increasing the decoder depth from 1 to 3 layers yields a marginal gain, *i.e.*, mAP from 74.2% to 74.4% without DR, and adding more layers beyond that shows no further improvement from 3 to 6 layers, remaining at 74.4% mAP. Interestingly, enabling the DR mechanism with an insufficient decoder depth can hurt performance. For example, with only 1 decoder layer, the model with DR actually performs worse than the one without DR. At 3 layers, the DR variant still lags behind. Only when the decoder is deep enough, *i.e.*, 6 layers, does the ResNet-50 model fully recover the performance with DR. This suggests that for a smaller backbone, the diffusion refinement introduces extra complexity that a shallow decoder cannot effectively capitalize on. For Swin Transformer (Liu et al. 2021), the backbone consistently benefits from deeper decoders, and it can more effectively leverage the DR mechanism when enough decoder layers are present. Without DR, increasing decoder layers from 1 to 6 yields steady improvements. With DR, a single decoder layer is not yet advantageous, and 3 layers with DR reach 78.2% mAP. Crucially, at 6 decoder layers, the Swin Transformer (Liu et al. 2021) with DR beats the model without DR, achieving 80.4% mAP. This indicates that when paired with a powerful backbone, the diffusion-based refinement mechanism can significantly improve performance. Overall, the ablation highlights that more transformer decoder layers generally improve detection performance, and the effectiveness of the DR mechanism is contingent on model capacity. It can slightly degrade results if the decoder is too shallow, espe-

cially on a weaker backbone, but becomes highly beneficial with a stronger backbone and a deeper decoder.

Table 7: Performance under different architectural components and query numbers on DocLayNet (Pfitzmann et al. 2022). Each row specifies the encoder type (feature fusion encoder (FFE) or deformable encoder (DE)), diffusion-based refinement (DR) mechanism, and autoregressive query expansion (AQE) mechanism, along with the number of initial and AQE queries.

Backbone	Encoder	DR	AQE	#Init. Q.	#AQE Q.	mAP
ResNet-50	DE	✓	✓	300	300	76.2
ResNet-50	FFE	✗	✓	300	300	74.4
ResNet-50	FFE	✓	✓	300	300	74.4
ResNet-50	DE	✓	✓	300	30	77.8
ResNet-50	FFE	✗	✓	300	30	76.3
ResNet-50	FFE	✓	✓	300	30	77.4
ResNet-50	DE	✓	✓	900	30	74.2
ResNet-50	FFE	✗	✓	900	30	73.5
ResNet-50	FFE	✓	✓	900	30	75.5
Swin-L	DE	✓	✓	300	300	78.1
Swin-L	FFE	✗	✓	300	300	79.1
Swin-L	FFE	✓	✓	300	300	80.4
Swin-L	DE	✓	✓	300	30	75.3
Swin-L	FFE	✗	✓	300	30	77.6
Swin-L	FFE	✓	✓	300	30	78.2
Swin-L	DE	✓	✓	900	30	76.7
Swin-L	FFE	✗	✓	900	30	75.7
Swin-L	FFE	✓	✓	900	30	77.3

The ablation study of the interplay between the architectural components and hyperparameters. As shown in Table 7, we evaluated detection mAP on DocLayNet under various encoder configurations and query counts, comparing the deformable encoder (DE) baseline to our feature fusion encoder (FFE) with and without the diffusion-based refinement (DR) and autoregressive query expansion (AQE). On the ResNet-50 (He et al. 2015) backbone, the deformable encoder that integrates DR yields the highest accuracy unless the FFE is also coupled with DR. When DR is enabled, FFE attains parity and in several settings overtakes the deformable encoder, demonstrating that iterative refinement can compensate for weaker geometric modeling on a light backbone. On the Swin Transformer (Liu et al. 2021) backbone, the situation is reversed. FFE alone already beats the deformable encoder, and when DR is counted, the FFE becomes the clear optimum. The richer multi-scale representations of Swin Transformer (Liu et al. 2021) allow FFE to exploit DR more effectively than the deformable counterpart. Besides, a large pool of initial queries, *e.g.*, 900, consistently lowers mean average precision for both backbones. Insufficient capacity in the decoder to process this overload causes optimization difficulties that DR can only partially help with. Empirically, assigning approximately 300 initial queries and 300 autoregressively expanded queries gives the most stable and accurate results, balancing coverage with computational tractability. Last, the benefit of DR is strongly correlated with overall model size. Deeper decoders and stronger backbones extract more value from the iterative refinement of queries, while shallow models show limited and

negative returns when DR is introduced preemptively.

Quantitative Result Analysis and Failure Case

Quantitative Result Analysis.

As shown in Figure 3, the HybriDLA method accurately identifies diverse document layout elements such as Page-header, Caption, Picture, Section-header, Table, and Text in complex layouts. It demonstrates robust performance across different document styles, including scientific articles, scanned newspaper pages, and black-and-white documents, highlighting its high detection precision and adaptability to varied layouts.

Failure Case.

Figure 4 gathers 8 representative failure pages and reveals three frequent patterns:

- i) **Correct detections that are not present in the ground-truth.** In several pages, our detector identifies legitimate layout elements, *e.g.*, small Page-header banners or thin rules that separate a figure and its caption that were omitted by the human annotators in DocLayNet. These show up as false positives in quantitative scores, but are correct predictions, highlighting an upper-bound limitation imposed by imperfect annotations.
- ii) **Over-refinement with too fragmented boxes.** The hybrid diffusion stage occasionally produces multiple highly overlapping or over-divided boxes for one logical element, especially around dense text blocks and when a picture contains textual graphics. This is mainly from the iterative diffusion-based refinement steps.
- iii) **Dominant background and atypical layouts.**
 - **Full-bleed photographs with embedded text.** Sparse textual regions on top of large images confuse the class priors and lead to caption and section-header swaps or misses.
 - **Code snippets and dense marginalia.** Extremely small mono-spaced text blocks are mistaken for tables or captions.
 - **Overlapping translucent layers or cross-page elements.** Transparent overlays invalidate positional cues, yielding duplicate detections and missed list-items.

These cases indicate that a portion of the apparent errors are due to incomplete annotations, the diffusion-based refinement can over-divide in visually cluttered areas, and extremely unbalanced foreground-background ratios or non-Manhattan layouts remain challenging.

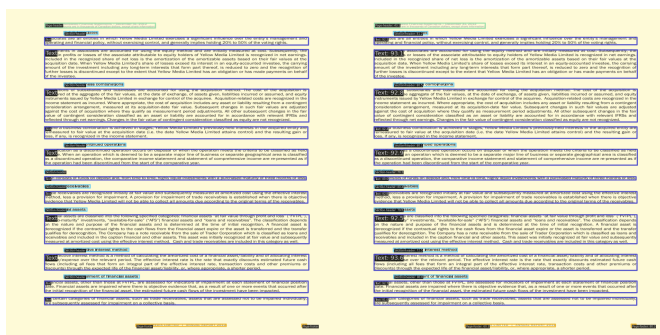
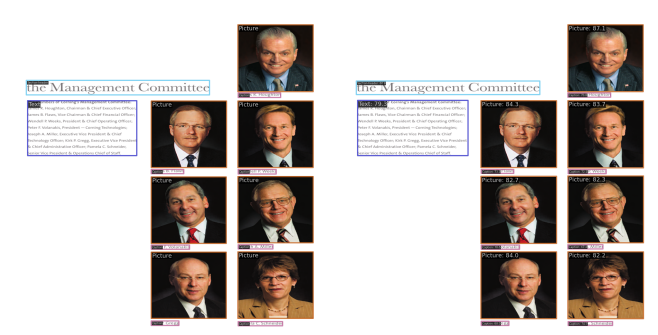
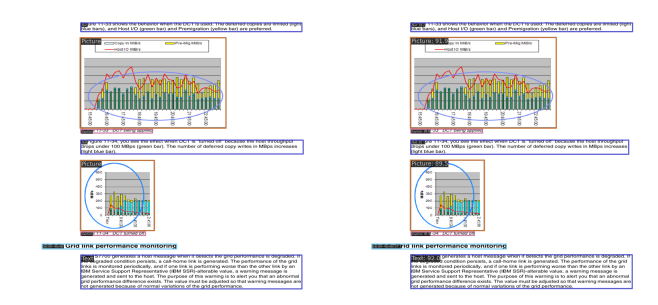



Figure 3: HybriDLA with InternImage (Wang et al. 2023) Results on DocLayNet (Pfitzmann et al. 2022). Each subfigure shows a document page with ground-truth annotations on the left and the results of the model on the right.

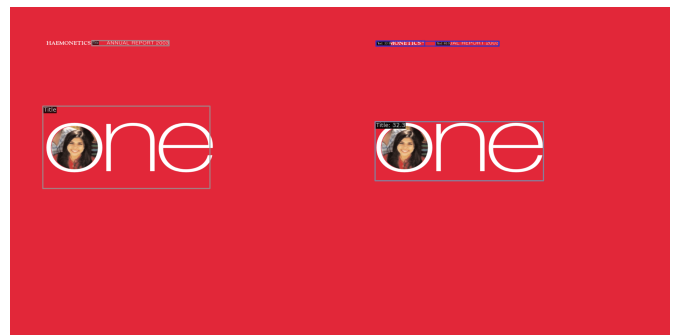
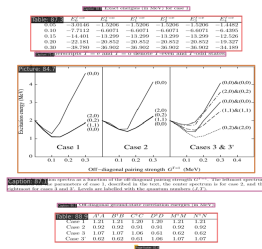
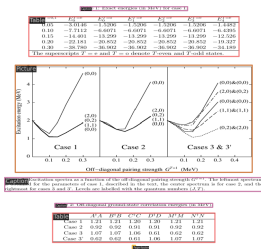
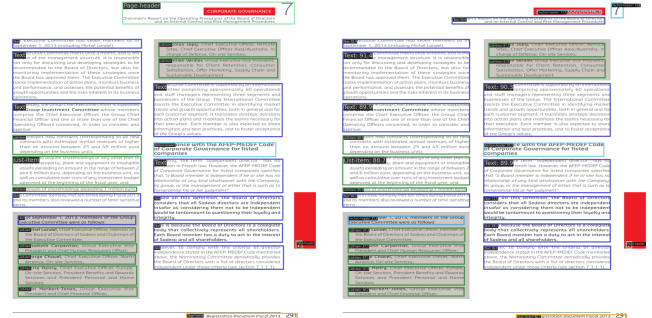
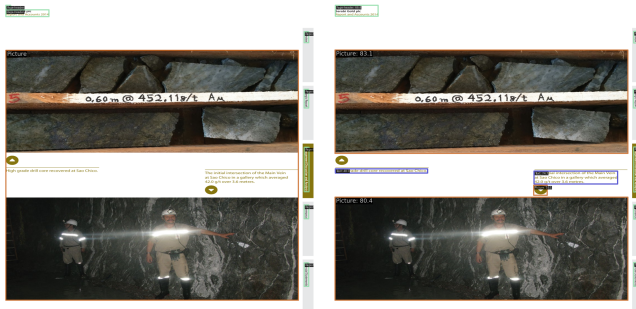
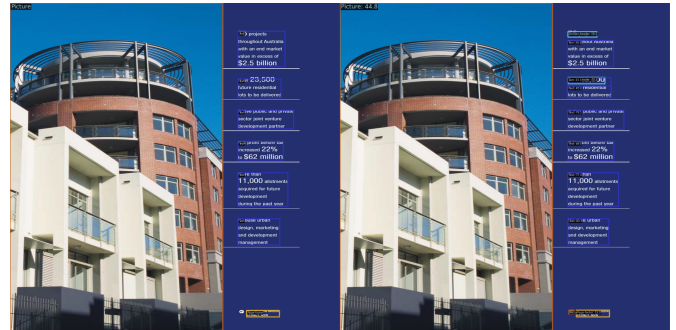
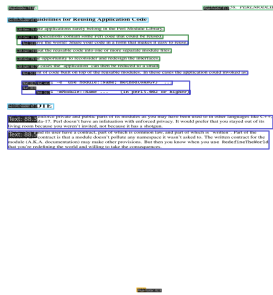
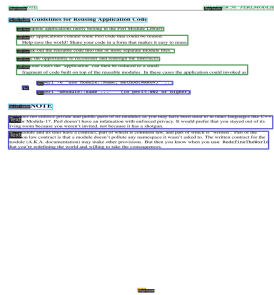
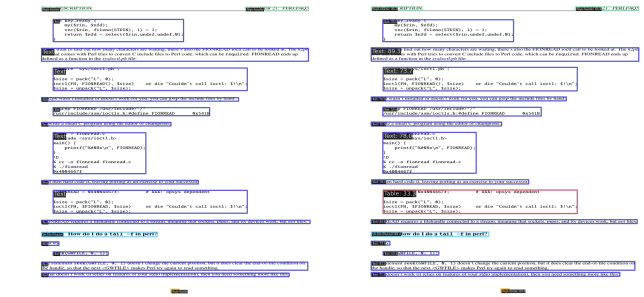


Figure 4: Typical failure cases of HybriDLA with InternImage (Wang et al. 2023) on DocLayNet (Pfützmann et al. 2022). For each page, the left half shows ground-truth annotations, and the right half shows prediction results.