HybriDLA: Hybrid Generation for Document Layout Analysis

Yufan Chen[1], Omar Moured[1], Ruiping Liu[1], Junwei Zheng[1], Kunyu Peng[1], Jiaming Zhang[2,*], Rainer Stiefelhagen[1]

[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

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 M^{6}Doc benchmarks demonstrate that **HybriDLA** sets a state-of-the-art performance, outperforming previous approaches.

Code — https://yufanchen96.github.io/projects/HybriDLA

Introduction

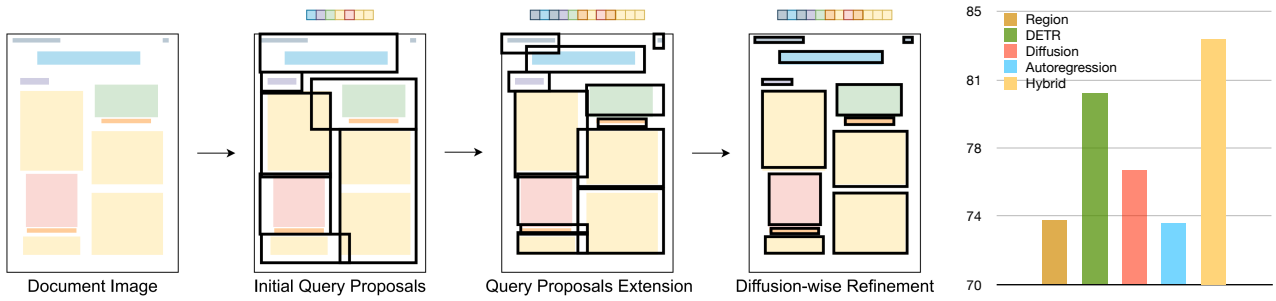
Understanding the layout of a document image is a fundamental task in document analysis. Accurate document layout analysis (DLA) is crucial for downstream applications, *e.g.*, document understanding and information extraction. However, a persistent challenge in DLA is the extreme variability in the 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 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

*Corresponding author.



(a) Representative paradigms for HybriDLA method (b) Performance on DocLayNet

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.

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 M^{6}Doc 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; Coquenet, 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 (Coquenet, 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. 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.

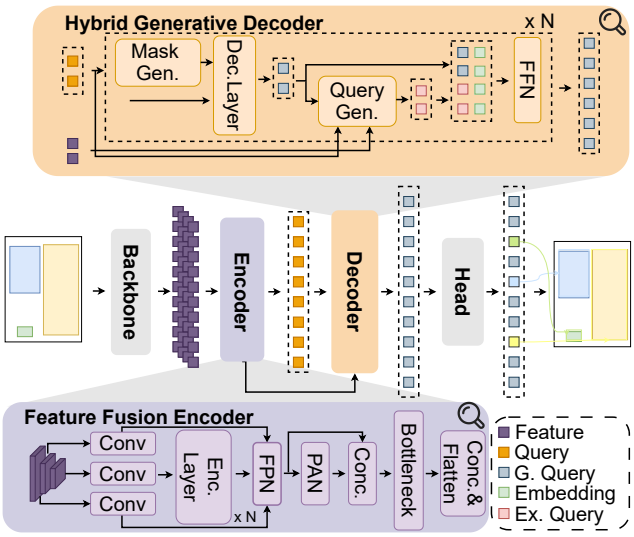


Figure 2: Overview of the HybriDLA architecture. It has a feature fusion encoder and a hybrid generative decoder. The encoder aggregates multi-scale visual features via convolutional and transformer layers, producing a layout-aware representation. The decoder operates in two mechanisms: it performs autoregressive 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.

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 gracefully 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 decoder 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 $Q = (q_1, q_2, \ldots, q_N)$. 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. The joint distribution over the query sequence factorizes as:

$$\begin{aligned} P(Q \mid X) &= P(q_1, q_2, \ldots, q_N, \texttt{EOS} \mid X) \\ &= \prod_{t=1}^{N} P(q_t \mid X, q_{1:t-1}) \cdot P(\texttt{EOS} \mid X, q_{1:N}) \end{aligned} \tag{1}$$

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 $P(q_t \mid X, q_{<t})$, 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 $q_1, \ldots, q_{t-1}$ 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 $P(\texttt{EOS} \mid X, q_{1:N})$. 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 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 (Pfitzmann 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 stan-

dard 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 as an overall indicator of performance, *i.e.*, higher mAP indicates better layout detection. All results are reported in the 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, which 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% 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.

Backbone	Detector	#Params	Modality V	Modality L	Modality T	mAP
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**

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

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 M^{6}Doc

Table 2 presents the detection accuracy of various approaches on the challenging M^{6}Doc (Cheng et al. 2023) dataset. For traditional region-based methods, ResNet-

Backbone	Detector	#Params	Modality			mAP
			V	L	T	
Region-based Method						
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
DETR-based Method						
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
Diffusion-based Method						
Swin-L	DiffusionDet	283M	✓	✗	✗	62.7
Autoregressive Method						
ViT-L	Pix2Seq	341M	✓	✗	✗	54.9
FIT-L	Pix2Seg	370M	✓	✗	✗	54.9
Hybrid Method						
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**

Table 2. Experiment results on M^{6}Doc dataset.

101 (He et al. 2015) with Mask R-CNN (He et al. 2017) achieves 61.9% mAP, and Faster R-CNN (Ren et al. 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 M^{6}Doc (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 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 M^{6}Doc (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.

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

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.

Ablation Study

To comprehensively evaluate our proposed HybriDLA method, we conducted ablation studies from five perspectives on the DocLayNet (Pfitzmann 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

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

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

FFE delivers clear benefits. With Swin-L, mAP rises from 77.2% to 79.1% even before any diffusion refinement mechanism, 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 model size. Moreover, the HybriDLA framework exhibits remarkable stability across diverse types of backbones, maintaining stable results regardless of backbone types.

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

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

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. SwinDocSegmenter: 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. DiffusionDet: 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*.

Coquenet, 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*.

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. GeoLayoutLM: 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. Proctag: 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*.