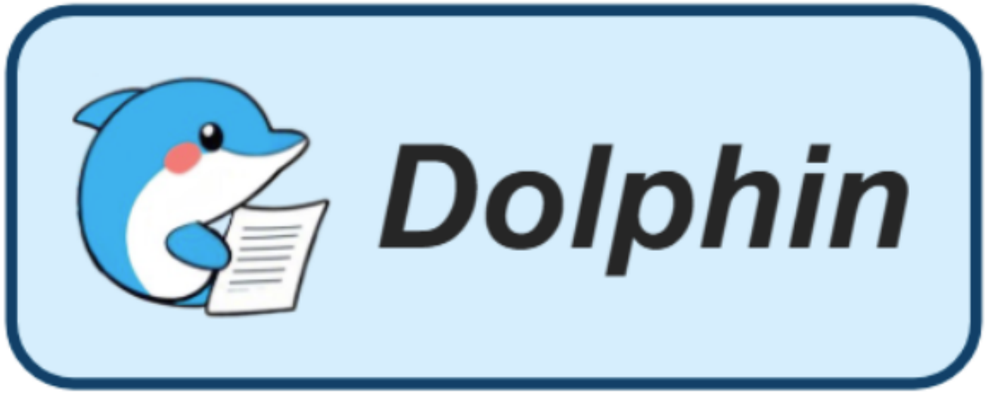
Dolphin - v2

:Universal Document Parsing via Scalable Anchor Prompting

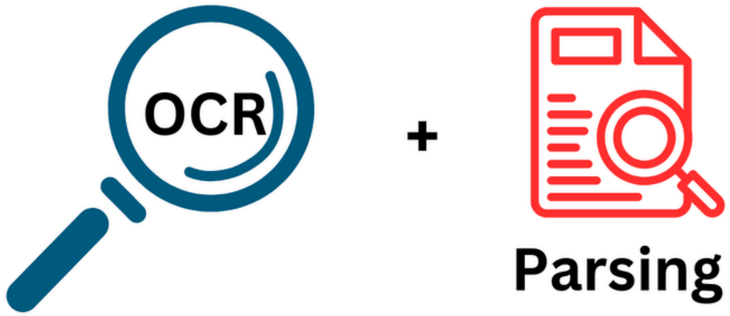


- Dolphin에서 Dolphin-v2로의 확장
- Feng, H., et al. “Dolphin: Document Image Parsing via Heterogeneous Anchor Prompting.” Findings of the Association for Computational Linguistics: ACL 2025.
- Feng, H., et al. “Dolphin-v2: Universal Document Parsing via Scalable Anchor Prompting.” arXiv preprint arXiv:2602.05384, 2026.

Contents

DeepShark Lab

연구 배경



문서 이미지 파싱

다양한 요소들(텍스트 단락, 그림, 표, 수식 등)을 포함한 이미지에서 구조화된 내용 추출
→ 시각적으로 표현된 문서 내용을 기계가 읽을 수 있는 형식으로 변환

DeepShark Lab

연구 배경 기존 문서 이미지 파싱의 딜레마

OCR 전문화 모델 통합 방식

OCR 작업에 특화

- 광학 문자 인식
- 레이아웃 검출 / 읽기 순서 예측 / 텍스트 인식 / 표 · 수식 인식
- 모델들을 순차적으로 연결하는 다단계 파이프라인

장점

- 높은 인식 성능
- 텍스트, 표, 수식 등 특정 요소의 정밀한 추출

단점

- 복잡한 모델 파이프라인 통합 과정
- 각 모델의 독립적인 학습 · 최적화 필요
- 모델 간 출력 형식 및 오류 조정 필요
- 복잡한 문서 레이아웃 구조 이해 부족

DeepShark Lab

연구 배경 기존 문서 이미지 파싱의 딜레마

VLM 기반 auto-regressive 방식

End-to-End Autoregressive Parsing

- 문서 이미지 전체를 입력
- VLM이 구조화된 문서 내용을 토큰 단위로 순차 생성
- 별도의 레이아웃 · 인식 모델 없이 하나의 모델로 처리

장점

- 단순한 시스템 구성
- 시각 정보와 언어 정보의 통합
- 페이지 전체 문맥 이해 가능

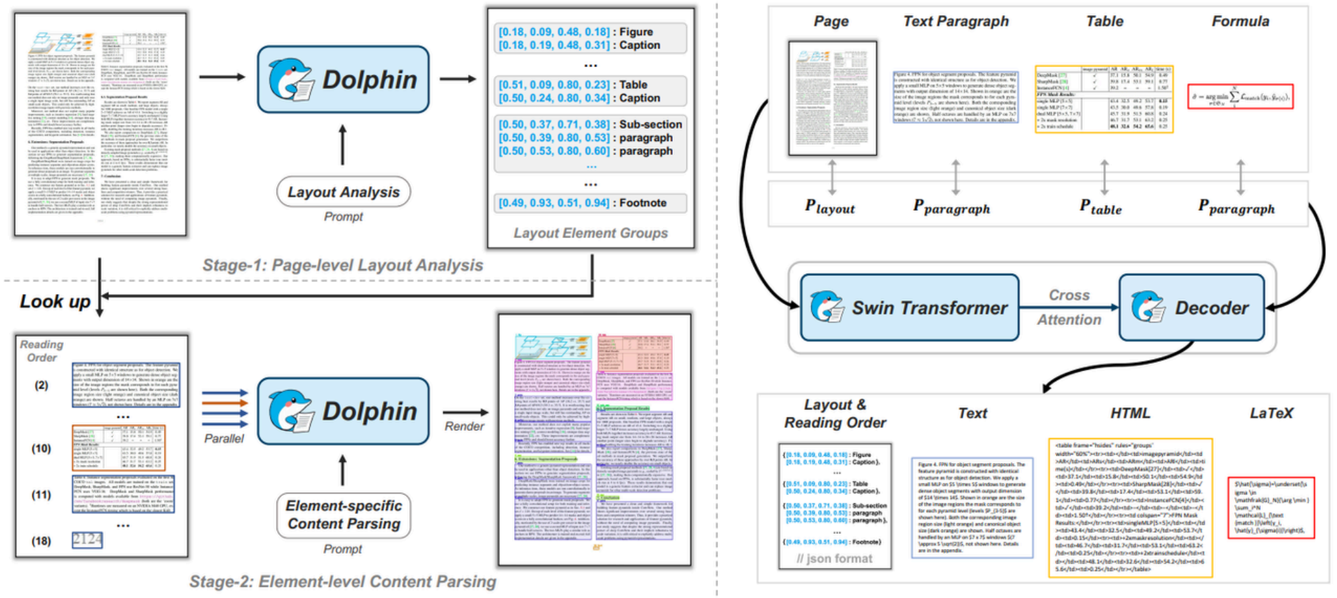
Autoregressive (자기 회귀)
이전까지의 생성된 출력 결과를 조건으로 다음 시점의 값을 하나씩 순차적으로 예측하는 생성 방식

단점

- 긴 문서 처리 시 효율성 저하
- 복잡한 레이아웃에서의 문서 구조 저하
- 긴 출력 생성 시 환각 발생 가능

DeepShark Lab

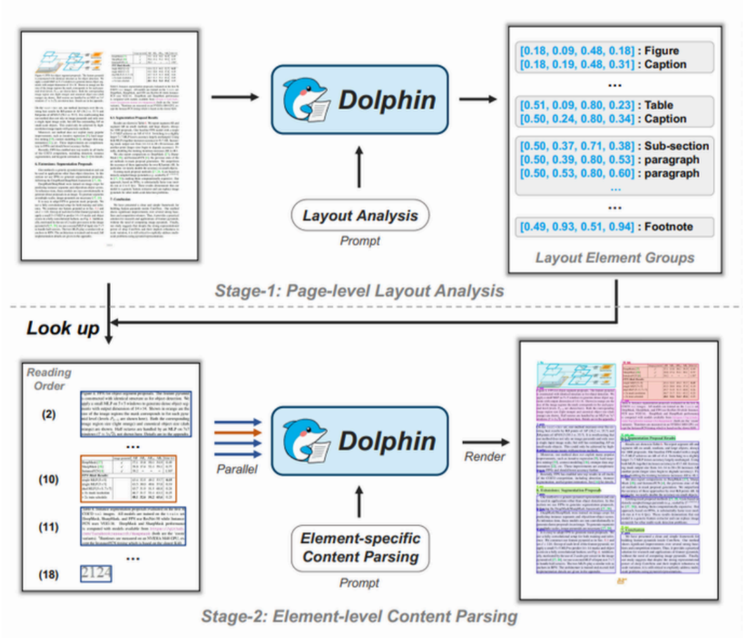
Dolphin 개요



Analyze-then-Parse

DeepShark Lab

Dolphin 개요



개요

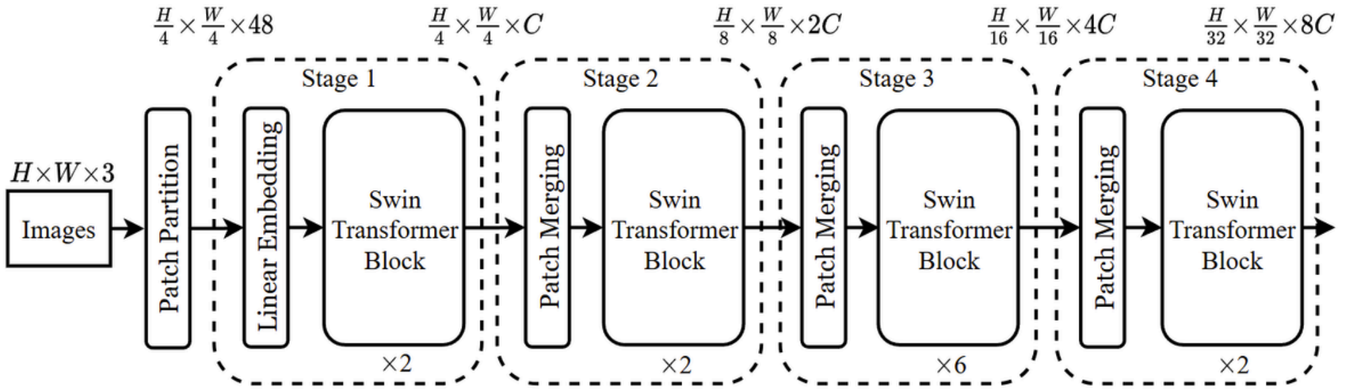
- **Analyze - then - Parse** 패러다임
- 페이지 수준 레이아웃 분석 → 요소 추출
- 추출한 요소를 앵커로 사용
- 유형별 프롬프트로 개별 요소의 내용을 병렬로 파싱

모델 구조

- 인코더 - 디코더 Transformer 기반 VLM
- 시각 인코더: **Swin Transformer**
- 언어 디코더: **mBART**
- 두 단계에서 동일한 모델 파라미터 공유

DeepShark Lab

Swin Transformer



특징

- stage가 깊어질수록 해상도 감소, 특징 차원 증가
- 4개의 stage로 계층적 시각 특징 추출

입력 이미지

- (H, W, 3)

Patch Partition

- Patch 크기 : 4 x 4
- 이미지를 겹치지 않는 patch로 분할
- 각 패치를 하나의 시각 토큰으로 처리
- 각 patch의 feature 차원은 48

Stage 1

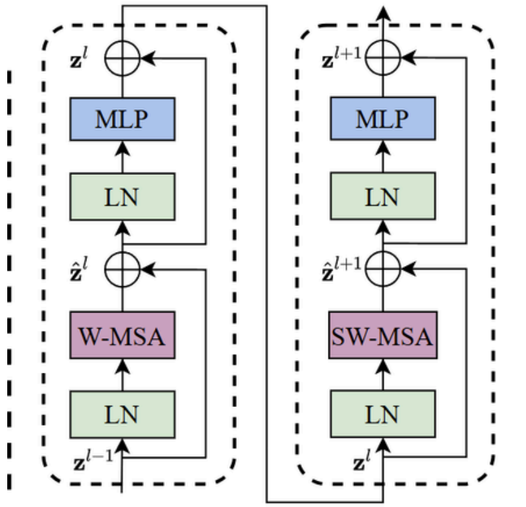
- Linear Embedding
  → 48차원 패치 특징을 C차원으로 투영
- Swin Transformer Block
- 해상도(토큰) 유지

Stage 2~4

- Patch Merging
  → 인접한 2x2 패치 특징 결합
- Swin Tranformer Block
- 토큰 $\frac{1}{4}$ 씩 감소

DeepShark Lab

Swin Tranformer Block



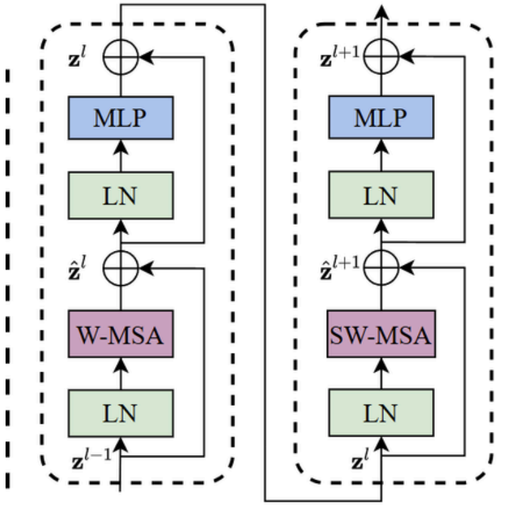
Two Successive Swin Transformer Blocks

Block 구성

- 두 개의 연속된 Swin Transformer Block으로 구성
- 첫 번째 Block: W-MSA
- 두 번째 Block: SW-MSA
- 각 MSA와 MLP 앞에 Layer Normalization 적용
- 각 모듈 뒤에 잔차 연결 적용

DeepShark Lab

Swin Tranformer Block



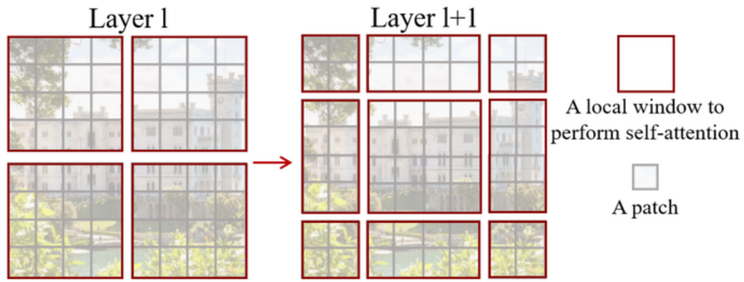
Two Successive Swin Transformer Blocks

W-MSA

- 특징 맵을 겹치지 않는 Local Window로 분할
- 각 윈도우 내부에서 Self-Attention 계산

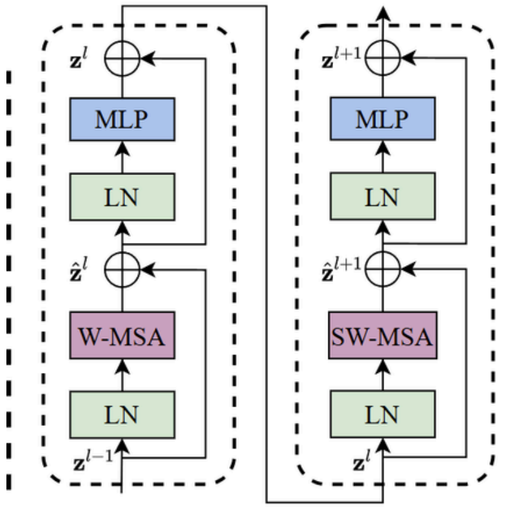
SW-MSA

- 다음 Block에서 윈도우 분할 위치 이동
- 이전 윈도우의 경계를 가로지르는 윈도우 구성
- 서로 다른 윈도우 사이의 연결 형성



DeepShark Lab

Swin Tranformer Block



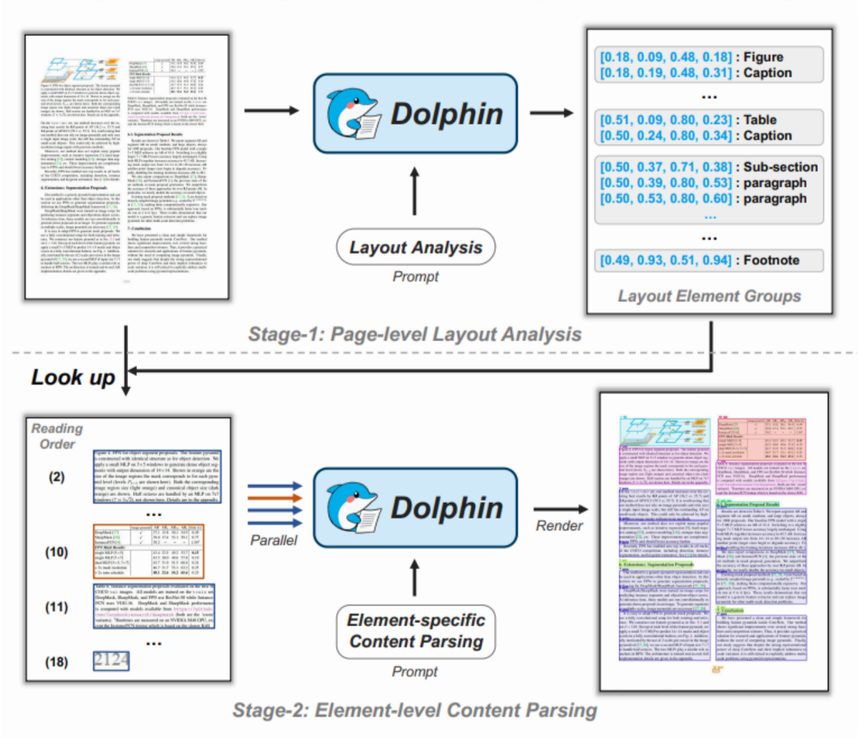
Two Successive Swin Transformer Blocks

$$\hat{\mathbf{z}}^l = \text{W-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1},$$
$$\mathbf{z}^l = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l,$$
$$\hat{\mathbf{z}}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(\mathbf{z}^l\right)\right) + \mathbf{z}^l,$$
$$\mathbf{z}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1},$$

DeepShark Lab

Dolphin 개요

Stage 1. Page-level Layout Analysis



목적

- 문서의 레이아웃 요소 및 읽기 순서 식별
- 요소별 유형 및 위치를 구조화된 시퀀스로 생성
- 생성된 요소를 Stage 2의 앵커로 사용

Page Image Encoding

- 입력 문서 이미지 I
- 비율 유지한 고정 크기 $H \times W$로 크기 조정 및 패딩
- Swin Transformer로 페이지 이미지 인코딩
- 시각 임베딩 $z \in \mathbb{R}^{d \times N}$ 생성 (d: 임베딩 차원, N: 이미지 패치 수)

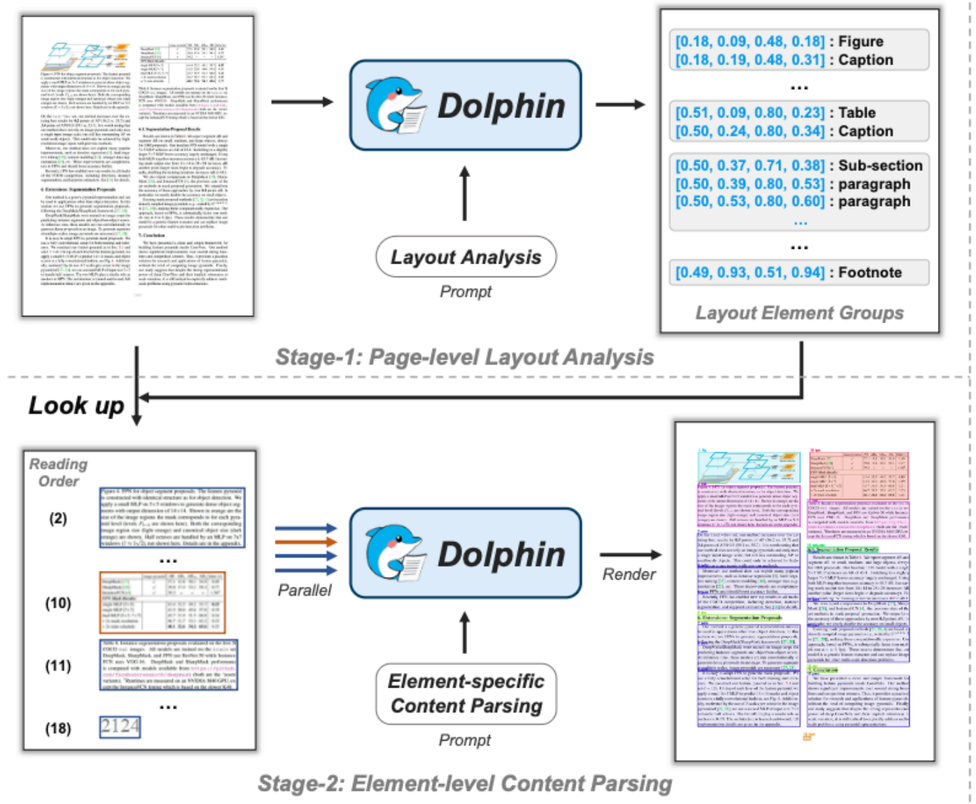
Layout Sequence Generation

- P_{layout} 입력: 레이아웃 분석 프롬프트 → 디코더 입력
- “Parse the reading order of this document.”
- mBART 디코더가 Cross-Attention으로 시각 특징 참조
- 문서 요소를 자연스러운 읽기 순서로 생성
- $L = \{l_1, l_2, \ldots, l_n\}$
- l_i : 요소 유형 + Bounding Box
- Ex) Figure, Caption, Table, Section title, Paragraph

DeepShark Lab

Dolphin 개요

Stage 2. Element-level Content Parsing



목적

- Stage 1에서 분석한 레이아웃 요소를 앵커로 사용
- 각 요소의 내용 파싱
- 요소별 처리 전문성 및 효율성 유지

Element Image Encoding

- 각 요소의 영역을 원본 이미지에서 crop
  → Local View I_i 생성
- Swin Transformer로 Local View 인코딩
- 요소별 시각 특징 생성

Parallel Content Parsing

- 요소 유형 별 프롬프트 적용
- Table: P_{table} → HTML
- Formula: $P_{paragraph}$ → LaTeX
- Paragraph: $P_{paragraph}$ → 텍스트
- 요소별 결과를 읽기 순서에 따라 결합
- 최종 Markdown 문서 생성

DeepShark Lab

Dolphin Dataset

Training Dataset

Source	Granularity	#Samples	Task Types
Mixed Documents	Page	0.12M	Layout
HTML	Page	4.37M	Parsing
LaTeX	Page	0.5M	Parsing
Markdown	Page	0.71M	Parsing
Table	Element	1.57M	Parsing
Formula	Element	23M	Parsing
Total	-	30.27M	-

Mixed Documents
- 시험지 , 교과서, 잡지 등
- 요소 경계, 읽기 순서, 주석 제공

HTML
- 중국어 + 영어 Wikipedia 문서
- Web 렌더링하여 생성

LaTeX
- arXiv 문서 렌더링
- 요소 유형, 계층 관계, 공간 위치 추출

Markdown
- GitHub에서 수집한 문서를 PDF 변환
- 문단, 줄, 단어 수준의 텍스트 주석 구성

Formula
- arXiv 소스에서 수집한 LaTeX 수식
- 수식 이미지로 렌더링

Table
- PubTabNet + PubTab1M
- 과학 출판물에서 추출된 대규모 표 데이터셋

DeepShark Lab

Dolphin Dataset

Evaluation Dataset

Page-level Evaluation

Fox-Page

- 순수 텍스트 문서
- 영어 112페이지 + 중국어 100페이지
- 단일 열 및 다중 열 구성
- 이미지 파싱 평가

Dolphin-Page

- 복합 문서 210페이지
- 순수 텍스트 111페이지
- 표, 수식, 그림 혼합
- 읽기 순서로 수동 주석 처리
- 문서 파싱 능력 평가

Element-level Evaluation

텍스트 문단

- 순수 텍스트 인식 평가
- Fox-Block: 텍스트 문단 이미지 포함
- Dolphin-Block: 복합 문서의 텍스트 문단 추출

수식

- 수식 인식 평가
- SPE(인쇄), SCE(화면 캡처), CPE(복잡한 수식)

표

- 표 인식 평가
- PubTabNet, PubTab1M

Evaluation metrics

ED

- Edit Distance
- 정답 텍스트로 바꾸기 위해 예측 텍스트를 수정한 횟수

CDM

- Character Difference Metric
- 예측 수식과 정답 사이의 문자 수준 차이를 기반으로 수식 인식 성능 평가

TEDS

- Tree-Edit-Distance-based Similarity
- 표를 HTML로 변환 후, 예측한 표와 정답 표의 트리 구조 비교

DeepShark Lab

Dolphin Experiment

모델 설정

Encoder

- Swin Transformer
- Window size: 7
- Encoder layers: [2, 2, 14, 2]
- Attention heads: [4, 8 , 16, 32]

Decoder

- 10개 Transformer layer
- Hidden dimension: 1024

학습 설정

- Optimizer: AdamW
- Learning rate: 5e-5
- Cosine decay schedule
- 2 epochs
- Batch size: GPU 장치 당 16
- Gradient accumulation 적용

입력 처리

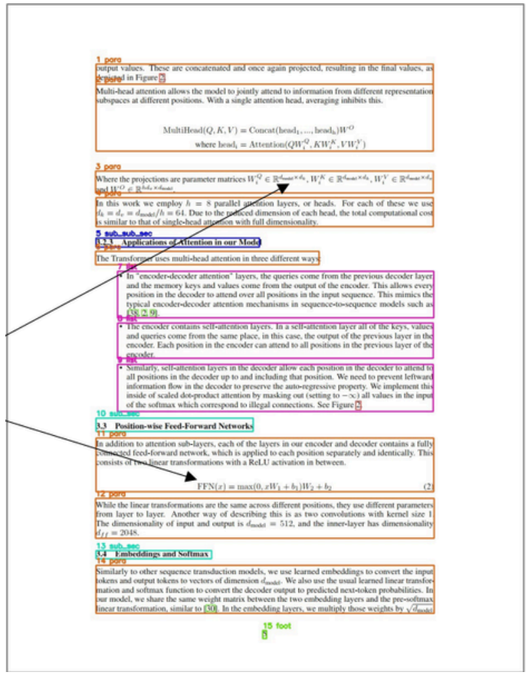
- 이미지 비율 유지
- 패딩 적용한 896 x 896 이미지
- Bounding Box 좌표 정규화

DeepShark Lab

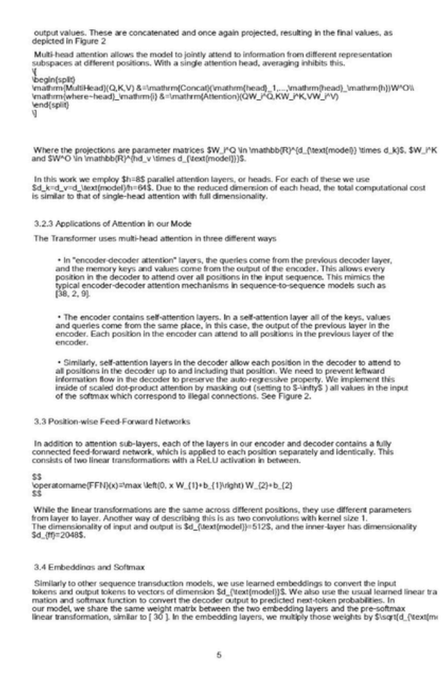
Dolphin Experiment

Page-level Parsing Process

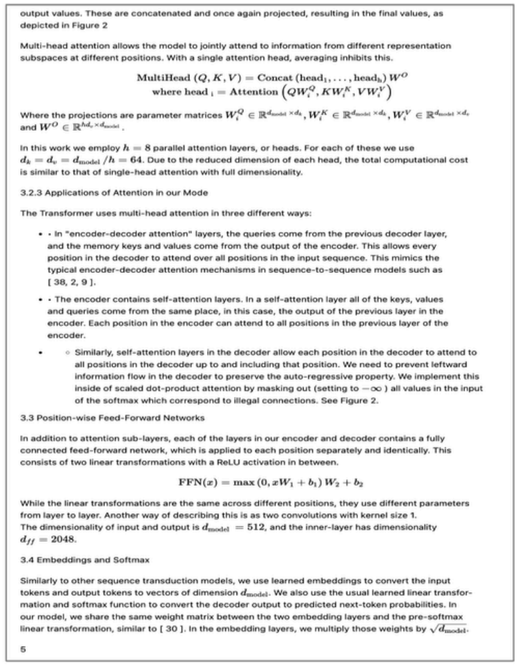
English Single-column Document



레이아웃 요소 식별 및 읽기 순서 예측



요소별 파싱 결과



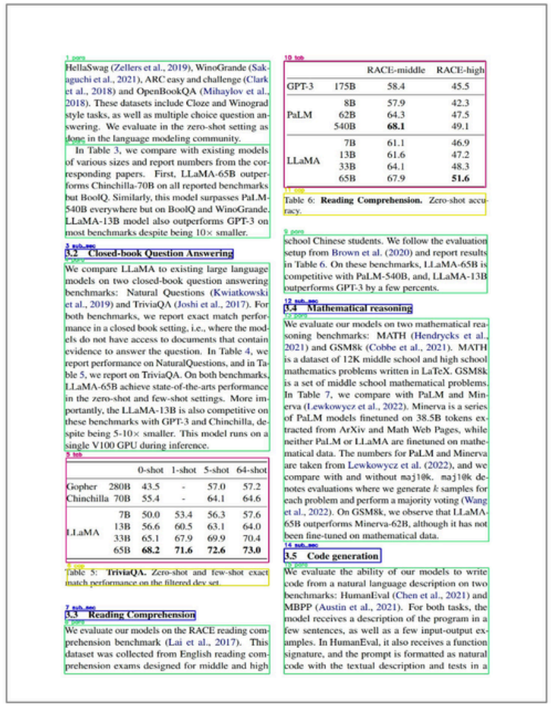
최종 마크다운 렌더링

DeepShark Lab

Dolphin Experiment

Page-level Parsing Process

English Double-column Document with Table



레이아웃 요소 식별 및 읽기 순서 예측

요소별 파싱 결과

최종 마크다운 렌더링

DeepShark Lab

Dolphin Experiment

Page-level Parsing Process

Chinese Single-column Document

레이아웃 요소 식별 및 읽기 순서 예측

요소별 파싱 결과

최종 마크다운 렌더링

DeepShark Lab

Dolphin Experiment

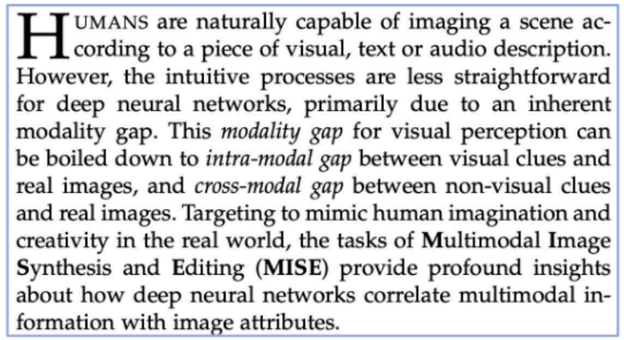
Page-level Evaluation

Category	Method	Model Size	Plain Doc (ED ↓)		Complex Doc (ED ↓)	Avg. ED	FPS ↑
			Fox-Page-EN	Fox-Page-ZH	Dolphin-Page		
Integration-based	MinerU	1.2B	0.0685	0.0702	0.2770	0.1732	0.0350
	Mathpix	-	0.0126	0.0412	0.1586	0.0924	0.0944
Expert VLMs	Nougat	250M	0.1036	0.9918	0.7037	0.6131	0.0673
	Kosmos-2.5	1.3B	0.0256	0.2932	0.3864	0.2691	0.0841
	Vary	7B	0.092*	0.113*	-	-	-
	Fox	1.8B	0.046*	0.061*	-	-	-
	GOT	580M	0.035*	0.038*	0.2459	0.1411	0.0604
	olmOCR	7B	0.0235	0.0366	0.2000	0.1148	0.0427
	SmolDocling	256M	0.0221	0.7046	0.5632	0.4636	0.0140
	Mistral-OCR	-	0.0138	0.0252	0.1283	0.0737	0.0996
General VLMs	InternVL-2.5	8B	0.3000	0.4546	0.4346	0.4037	0.0444
	InternVL-3	8B	0.1139	0.1472	0.2883	0.2089	0.0431
	MiniCPM-o 2.6	8B	0.1590	0.2983	0.3517	0.2882	0.0494
	GLM4v-plus	9B	0.0814	0.1561	0.3797	0.2481	0.0427
	Gemini-1.5 pro	-	0.0996	0.0529	0.1920	0.1348	0.0376
	Gemini-2.5 pro	-	0.0560	0.0396	0.2382	0.1432	0.0231
	Claude3.5-Sonnet	-	0.0316	0.1327	0.1923	0.1358	0.0320
	GPT-4o-202408	-	0.0585	0.3580	0.2907	0.2453	0.0368
	GPT-41-250414	-	0.0489	0.2549	0.2805	0.2133	0.0337
	Step-1v-8k	-	0.0248	0.0401	0.2134	0.1227	0.0417
	Qwen2-VL	7B	0.1236	0.1615	0.3686	0.2550	0.0315
	Qwen2.5-VL	7B	0.0135	0.0270	0.2025	0.1112	0.0343
Ours	**Dolphin**	322M	**0.0114**	**0.0131**	**0.1028**	**0.0575**	**0.1729**

DeepShark Lab

Dolphin Experiment

Element-level Parsing Process



HUMANS are naturally capable of imaging a scene according to a piece of visual, text or audio description. However, the intuitive processes are less straightforward for deep neural networks, primarily due to an inherent modality gap. This *modality gap* for visual perception can be boiled down to *intra-modal gap* between visual clues and real images, and *cross-modal gap* between non-visual clues and real images. Targeting to mimic human imagination and creativity in the real world, the tasks of **Multimodal Image Synthesis and Editing (MISE)** provide profound insights about how deep neural networks correlate multimodal information with image attributes.

HUMANS are naturally capable of imaging a scene according to a piece of visual, text or audio description. However, the intuitive processes are less straightforward for deep neural networks, primarily due to an inherent modality gap. This modality gap for visual perception can be boiled down to intra-modal gap between visual clues and real images, and cross-modal gap between non-visual clues and real images. Targeting to mimic human imagination and creativity in the real world, the tasks of Multimodal Image Synthesis and Editing (MISE) provide profound insights about how deep neural networks correlate multimodal information with image attributes.

中华医学会 中华医学会杂志社 中华医学会全科医学分会 中华医学会《中华全科医师杂志》编辑委员会 心血管系统疾病基层诊疗指南编写专家组

通信作者：孙艺红，中日友好医院心脏科，北京 100029，Email：yihongsun72@163.com；胡大一，北京大学人民医院心血管病研究所 100044，Email：dayi.hu@china-heart.org

【关键词】 指南； 胸痛

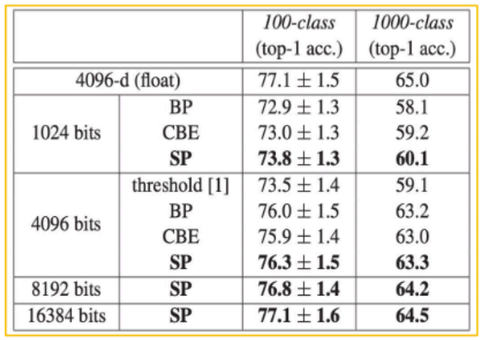
DOI:10.3760/cma.j.issn.1671-7368.2019.10.004

Guideline for primary care of chest pain(2019)

Chinese Medical Association, Chinese Medical Journals Publishing House, Chinese Society of General Practice, Editorial Board of Chinese Journal of General Practitioners of Chinese Medical Association, Expert Group of Guidelines for Primary Care of Cardiovascular Disease

Corresponding author: Sun Yihong, Department of Cardiology, China-Japan Friendship Hospital, Beijing 100029, China, Email: yihongsun72@163.com; Hu Dayi, Institute of Cardiovascular Disease, Peking University People's Hospital, Beijing 100044, China, Email: dayi.hu@china-heart.org

中华医学会 中华医学会杂志社 中华医学会全科医学分会 中华医学会《中华全科医师杂志》编辑委员会 心血管系统疾病基层诊疗指南编写专家组
通信作者: 孙艺红, 中日友好医院心脏科, 北京 100029, Email: yihongsun72@163.com;
胡大一, 北京大学人民医院心血管病研究所 100044, Email: dayi.hu@china-heart.org
【关键词】指南; 胸痛
DOI:10.3760/cma.j.issn.1671-7368.2019.10.004
Guideline for primary care of chest pain(2019)
Chinese Medical Association, Chinese Medical Journals Publishing House, Chinese Society of General Practice, Editorial Board of Chinese Journal of General Practitioners of Chinese Medical Association, Expert Group of Guidelines for Primary Care of Cardiovascular Disease
Corresponding author: Sun Yihong,Department of Cardiology,China-Japan Fiendship Hospital, Beijing 100029, China,Email:yihongsun72@163.com; Hu Dayi,Institute of Cardiovascular Disease, Peking University People's Hospital, Beijing 100044, China,Email:dayi.hu@china-heart.org



		100-class (top-1 acc.)	*1000-class* (top-1 acc.)
4096-d (float)		77.1 ± 1.5	65.0
1024 bits	BP	72.9 ± 1.3	58.1
	CBE	73.0 ± 1.3	59.2
	SP	**73.8 ± 1.3**	**60.1**
4096 bits	threshold [1]	73.5 ± 1.4	59.1
	BP	76.0 ± 1.5	63.2
	CBE	75.9 ± 1.4	63.0
	SP	**76.3 ± 1.5**	**63.3**
8192 bits	**SP**	**76.8 ± 1.4**	**64.2**
16384 bits	**SP**	**77.1 ± 1.6**	**64.5**

		100-class (top-1 acc.)	1000-class (top-1 acc.)
4096-d (float)		77.1 ± 1.5	65.0
1024 bits	BP	72.9 ± 1.3	58.1
	CBE	73.0 ± 1.3	59.2
	SP	73.8 ± 1.3	60.1
4096 bits	threshold [1]	73.5 ± 1.4	59.1
	BP	76.0 ± 1.5	63.2
	CBE	75.9 ± 1.4	63.0
	SP	76.3 ± 1.5	63.3
8192 bits	SP	76.8 ± 1.4	64.2
16384 bits	SP	77.1 ± 1.6	64.5

DeepShark Lab

Dolphin Experiment

Element-level Evaluation

Category	Method	Text Paragraph (ED ↓)			Formula (CDM ↑)			Table (TEDS ↑)	
		Fox-Block		Dolphin-Block	SPE	SCE	CPE	PubTabNet	PubTab1M
		EN	ZH						
Expert Models	UnimerNet-base	-	-	-	**0.9914***	0.94*	0.9595*	-	-
	Mathpix	-	-	-	0.9729*	0.9318*	**0.9671***	-	-
	Pix2tex	-	-	-	0.9619*	0.2453*	0.6489*	-	-
Expert VLMs	TabPedia	-	-	-	-	-	-	**0.9541**	0.9511
	GOT	0.0181	0.0452	0.0931	0.8501	0.7369	0.7197	0.3684	0.3269
General VLMs	GLM-4v-plus	0.0170	0.0400	0.1786	0.9651	0.9585	0.7055	0.5462	0.6018
	Qwen2-VL-7B	0.0910	0.1374	0.1012	0.5339	0.6797	0.1220	0.3973	0.5101
	Qwen2.5-VL-7B	0.0803	0.0301	0.0712	0.9486	0.9484	0.8309	0.6169	0.6462
	Gemini-1.5 pro	0.0108	0.0461	0.0857	0.9572	0.9469	0.7171	0.7571	0.7776
	Claude3.5-Sonnet	0.0375	0.1177	0.0746	0.8995	0.9464	0.7543	0.5431	0.7127
	GPT-4o-202408	0.0170	0.1019	0.0489	0.9570	0.9402	0.7722	0.6692	0.7243
	Step-1v-8k	0.0098	0.0175	0.0252	0.9526	0.9336	0.7519	0.6808	0.6588
Ours	**Dolphin**	**0.0029**	**0.0121**	**0.0136**	0.9850	**0.9685**	0.8739	0.9515	**0.9625**

DeepShark Lab

Dolphin 한계

세로쓰기 문서 처리 제한

- 가로쓰기 문서 중심
- 세로쓰기 문서 파싱 성능 제한

다국어 지원 제한

- 영어 + 중국어 문서 중심
- 다양한 언어 확장 필요

손글씨 인식 성능 부족

- 인쇄 문서 중심 학습
- 손글씨 인식 능력 추가 개선 필요

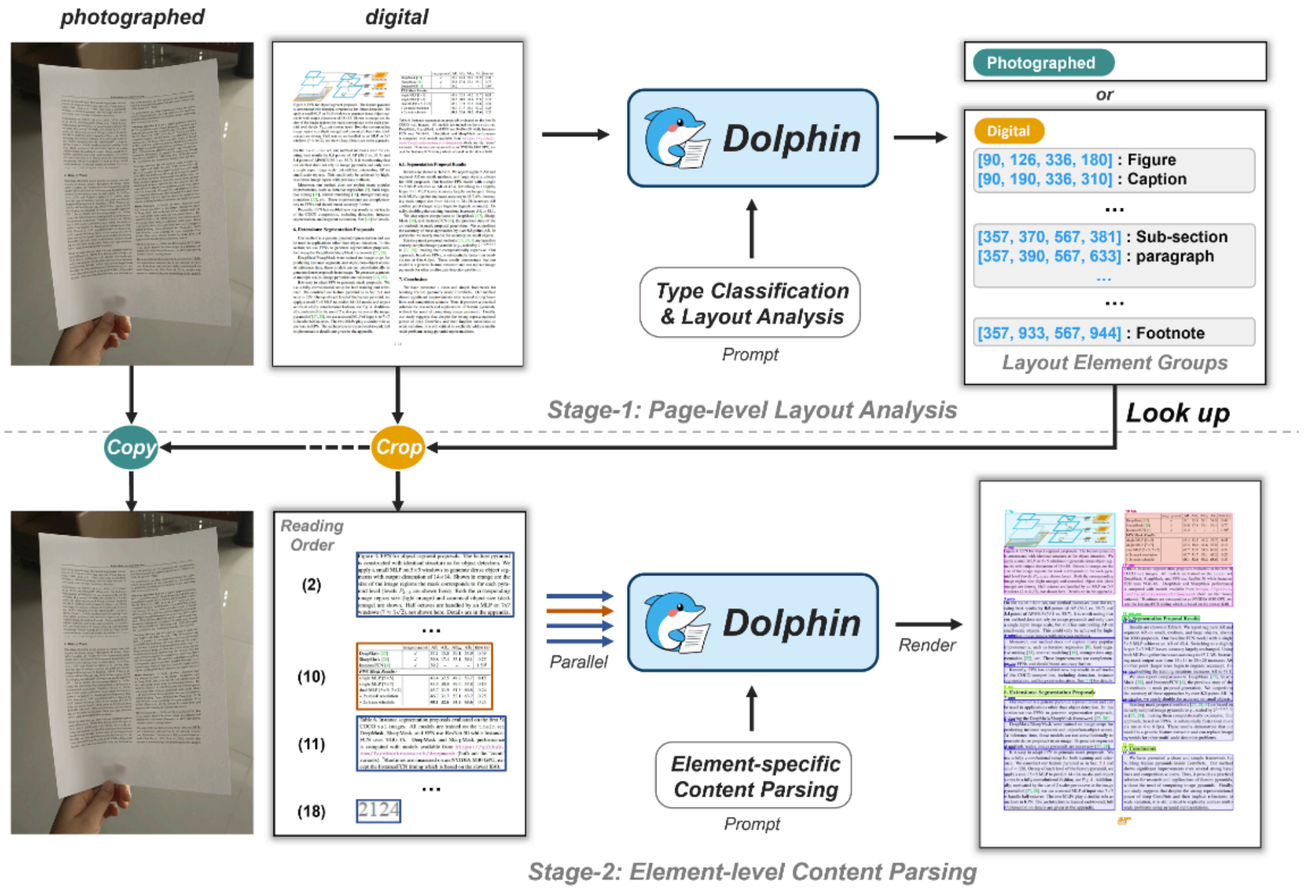
병렬 처리 최적화 필요

- 현재는 요소 단위 병렬 파싱 수행
- 텍스트 줄 + 표 셀 단위의
  추가 병렬화

DeepShark Lab

Dolphin - v2 개요

VLM 기반 2단계 구조



Stage 1

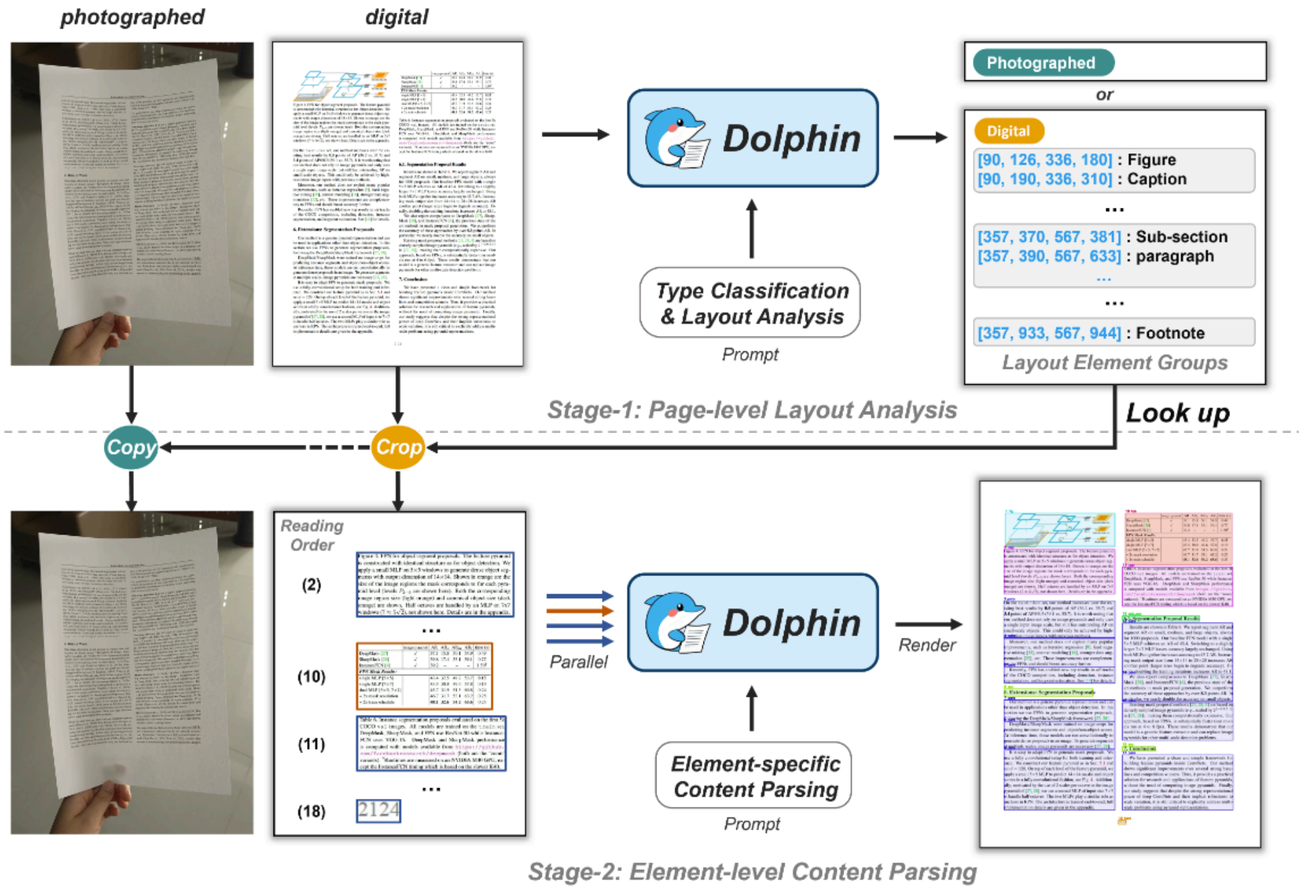
- 문서 유형 분류
- 페이지 수준 레이아웃 분석
- 읽기 순서 예측

Stage2

- 촬영 문서: 전체 페이지 파싱
- 디지털 문서: 요소 단위 병렬 파싱

DeepShark Lab

Dolphin - v2 개요

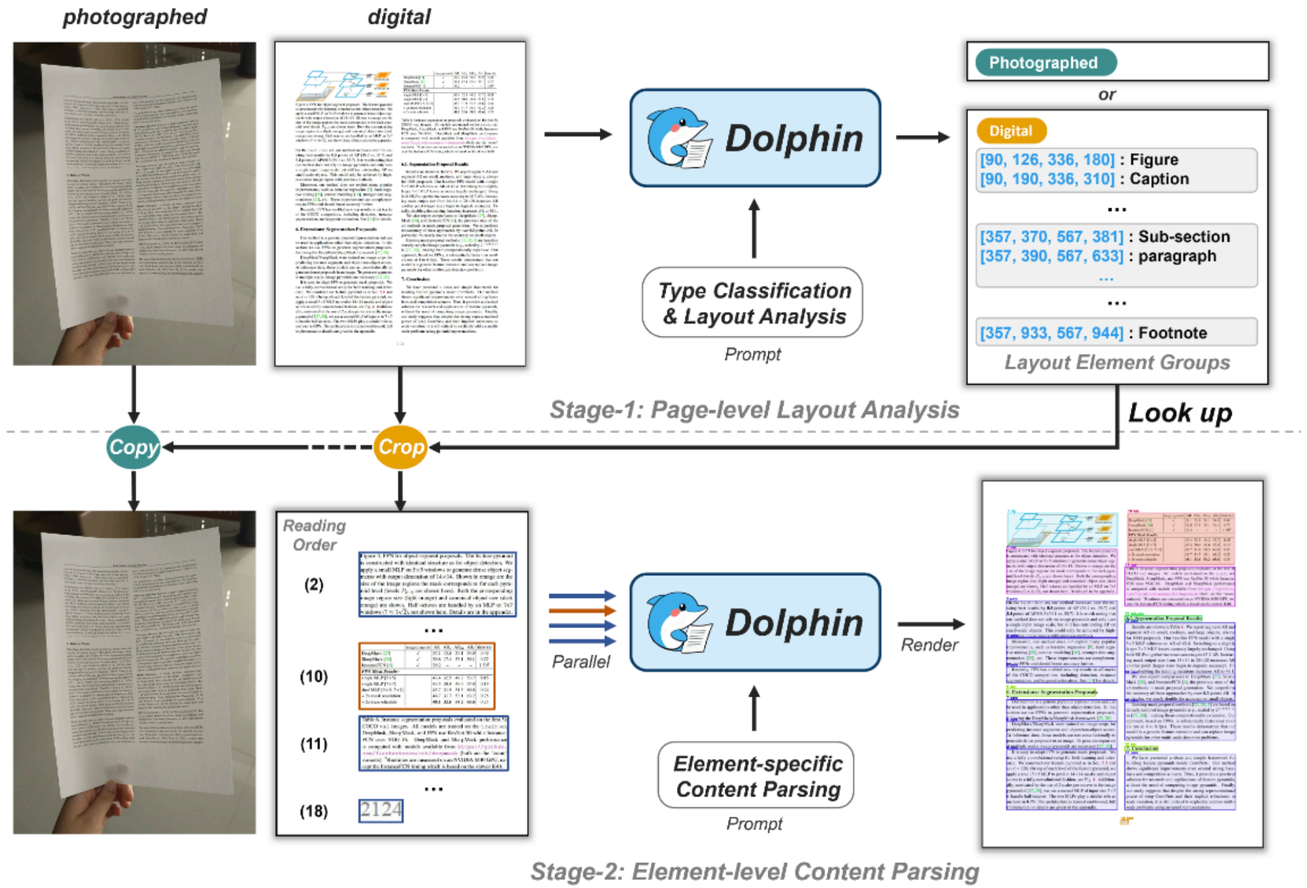


Page Image Encoding

- 시각 인코더: Qwen2.5-VL 기반 NaViT
- 원본 해상도 이미지 직접 처리
- 크기 조정 및 패딩 불필요
- 시각 임베딩 $z \in \mathbb{R}^{d \times N}$ 생성

DeepShark Lab

Dolphin - v2 개요



문서 분류 및 레이아웃 생성

1. 문서 유형 우선 예측

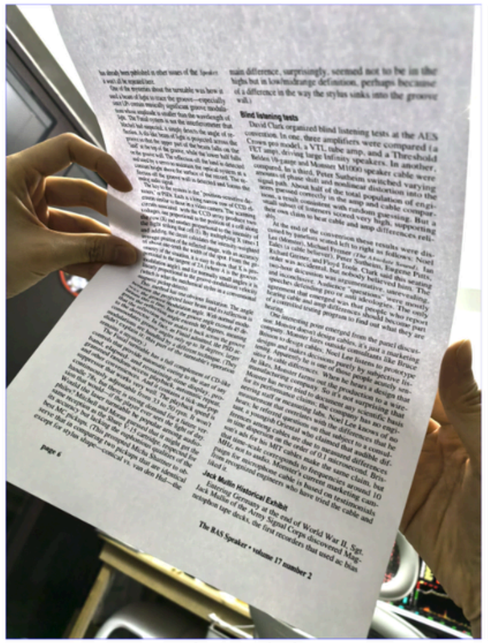
2. 촬영 문서 → 레이아웃 생성 종료

3. 디지털 문서
   → 요소 유형 및 Bounding Box 생성
   → 읽기 순서에 따라 순차 출력
   → 의미 속성 추출

4. 출력
- 요소 유형 $L = \{l_1, l_2, \ldots, l_n\}$
- Bounding Box
- 읽기 순서
- 의미 속성

DeepShark Lab

Dolphin - v2 개요



Input Photographed Image

has already been published in other issues of the Spopter it won't all be repeated here.

One of the mysteries about the turntable was how it used a beam of light to trace the groove—especially since LPs contain musically significant groove modulations whose amplitude is smaller than the wavelength of light. The Final system is not the interferometer that Mitchell had suspected; it simply detects the angle of reflection. A full-like beam-of light is projected across the groove so that the upper part of the beam falls on the "land" at the top of the groove, while the lower half falls on the groove wall. The reflection off the land is detected and used by a servo to maintain the optical system at a constant height above the surface of the record. The reflection off the groove wall is detected and forms the analog audio signal.

The key to the system is the "position-sensitive detectors," or PSDs. Each is a long narrow row of CCD elements similar to those in a video camera. The scanning circuits associated with the CCD array produce two voltages, one proportional to the position of a cell along the row (X) and another proportional to the intensity of the light striking that cell (I). By multiplying X times I and adding, the circuit calculates the intensity-weighted average position of the reflected image, with an accuracy of about one-tenth the width of the spot. From the geometry of the situation, it is easy to show that X is proportional to the tangent of 2A (where A is the groove-modulation angle), and for normally small angles it is linearly proportional to the groove-modulation velocity (which is what the mechanical stylus in a conventional magnetic pickup detects).

This method has one obvious limitation. The angle between the projected laser beam and its reflection is over the groove modulation angle. With normal modulations this is fine, but if the groove angle exceeds 45 degree the reflection angle exceeds 90 degrees, meaning that the reflection never makes it back across the groove to the detector. In fact, as Final admitted, the PSD accommodates groove angles only up to 38 degrees; larger modulations are handled by a different technique. (They didn't explain it; this part of the turntable's operation remains a mystery.)

The Final turntable has a full complement of CD-like controls that provide automatic cueing to the start of any band, elapsed- and remaining-time display, programmed random-access equipment, automatic repeat, and other conveniences. And it contains a tick-and-pop suppressor that works very well. The playback speed is continuously adjustable from 33 to 50 rpm. It won't handle 78s, but there is strong demand for a future version that would—if the player ever sees the light of day. Would the laser turntable be popular among audiophiles? Mitchell and Moran guessed that it might get the same reception as the Shure V-15 cartridge, respected for its accuracy but lacking the euphonious qualities of the best MC pickups. (This prompted Micha Shattner to observe that in comparing two pickups that are identical except for stylus shape—conical vs. van den Hul—the main difference, surprisingly, seemed not to be in the highs but in low/midrange definition, perhaps because of a difference in the way the stylus sinks into the groove wall.)

Blind listening tests

David Clark organized blind listening tests at the AES convention. In one, three amplifiers were compared (a Crown pro model, a VTL tube amp, and a Threshold FET amp), driving large infinity speakers. In another, Belden 10-gauge and Monster M1000 speaker cable were compared. In a third, Peter Sutheim switched varying amounts of phase shift and nonlinear distortion into the signal path. About half of the total population of engineers guessed correctly in the amp and cable comparisons, a result consistent with random guessing. But a small subset of listeners scored very high, supporting their own claim to hear cable and amp differences reliably.

At the end of the convention these results were discussed by panelists seated left to right as follows: Noel Lee (Monster), Michael Fremer (The Absolute Sound), Ian Eales (a cable believer), Peter Sutheim, Eugene Pitts, Richard Greiner, and Floyd Toole. Clark said this seating order was accidental, but nobody believed him. The two-hour discussion was argumentative, unrevealing, and inconclusive. Audience "questions" were mostly speeches defending pro or anti ideologies. The only agreement that emerged was that people who report hearing cable and amp differences should become part of a controlled testing program to find out what they are hearing.

One interesting point emerged from the panel discussion: Monster doesn't design cables, it's just a marketing company. Monster hires outside consultants like Bruce Brisson to design cables; Noel Lee listens to prototype designs and makes decisions purely by subjective listening. Apparently he is one of those people acutely sensitive to cable differences. When he hears a design that he likes, Monster farms out the production to a wire-manufacturing company. So it's not surprising that Monster has never tried to document any scientific basis for its performance claims; the company has no engineering staff or measuring labs. Noel Lee knows of no tears; he referred questions on that subject to a consultter, a young Oriental who claimed that audible differences among cables are due to measured differences in time dispersion on the order of 0.1 microsecond. Brisson's ads for his MIT cables make the same claim, but that time-scale corresponds to frequencies around 10 MHz, not to audio. Monster's current marketing campaign for microphone cable is based on testimonials from recognized engineers who have tried the cable and liked it.

Jack Mullin Historical Exhibit
Entering Germany at the end of World War II, Sgt. Jack Mullin of the Army Signal Corps discovered Magnetophon tape decks, the first recorders that used ac bias

The BAS Speaker • volume 17 number 2
page 6

Dolphin-v2

has already been published in other issues of the Spreaker. It won't all be repeated here.

One of the mysteries about the turntable was how it used a beam of light to trace the groove—especially since LPs contain musically significant groove modulations whose amplitude is smaller than the wavelength of light. The Finalia system is not the interferometer that Mitchell had suspected: it simply detects the angle of reflection. A slit-like beam of light is projected across the groove so that the upper part of the beam falls on the \"land\" at the top of the groove, while the lower half falls and used by a servo to maintain the optical system at a position above the surface of the record. The analog audio signal.

The key to the system is the \"position-sensitive de- tectors,\" or PSDs. Each is a long narrow row of CCD el- circuits associated with the CCD array produce two the row (X) and another proportional to the intensity of and adding, the circuit calculates the intensity-weighted omecity of the situation, the width of the spot. From the ge- malo- modulation angle, and for normally small angles it is a magnetic pickup detector. This method has one obvious limit: the projected X-time.

The angle of the reflection and its reflection angle with normal modu \ntomodulates the groove angle exceeds 45 degrees. \nTo the left, the reflection angle exceeds 90 degrees, meaning \nthe modulator is never makes it back up to 30 degrees. \nThe grooves are groove angles, and the PSD act \nas a puzzle. Explain it handed by a different technique. (They \nmake a mystery.) This part of the turntable's operation \nis closed, and the automatic cue system is used. \nThe closed, and the automatic cue system is used. \nThe closed, and the automatic cue system is used. \nThe closed, and the automatic cue system is used.

lack across the g \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled, the \ntrolled .....

main difference, surprisingly, seemed not to be in the highs but in low/mid-range definition, perhaps because of a difference in the way the styls sinks into the grooves wall.

Blind listening tests

David Clark organized blind listening tests at the AES convention. In one, three amplifiers were compared (a Crown pro model, a VTL tube amp, and a Threshold FET amp), driving large infinity speakers. In another, Belden 10-gauge and Monster M1000 speaker cable were compared. In a third, Peter Sutheim switched varying signal path. About half of the total population of engis- isons, a result consistent with random guessing. But, small subset of listeners scored very high, using ably.

At the end of the convention these results were discussed by panelists seated left to right as follows: Noel Lee (Monster), Michael Fremer (The Absolute Sound), Ian Eales (a cable believer), Peter Surtheim, Eugene Pitts, order was accidental, but nobody believed this seating and inconclusive was argumentative, unrevealing, agreements defending pro or anti ideologies. The only hearing cable and amp differences were of a controlled testing program.

One interesting point emerged from the panel discussion: One interesting point emerged from the panel discussion: One interesting point emerged from the panel discussion: One interesting point emerged from the panel discussion: One interesting point emerged from the panel discussion: One interesting point emerged from the panel discussion: One interesting point emerged from .....

eed engineers w Jack Mullin Historical Exhibit Entering Germany at the Jack Mullin of the 19th Ophon tan

the cable a
page 6

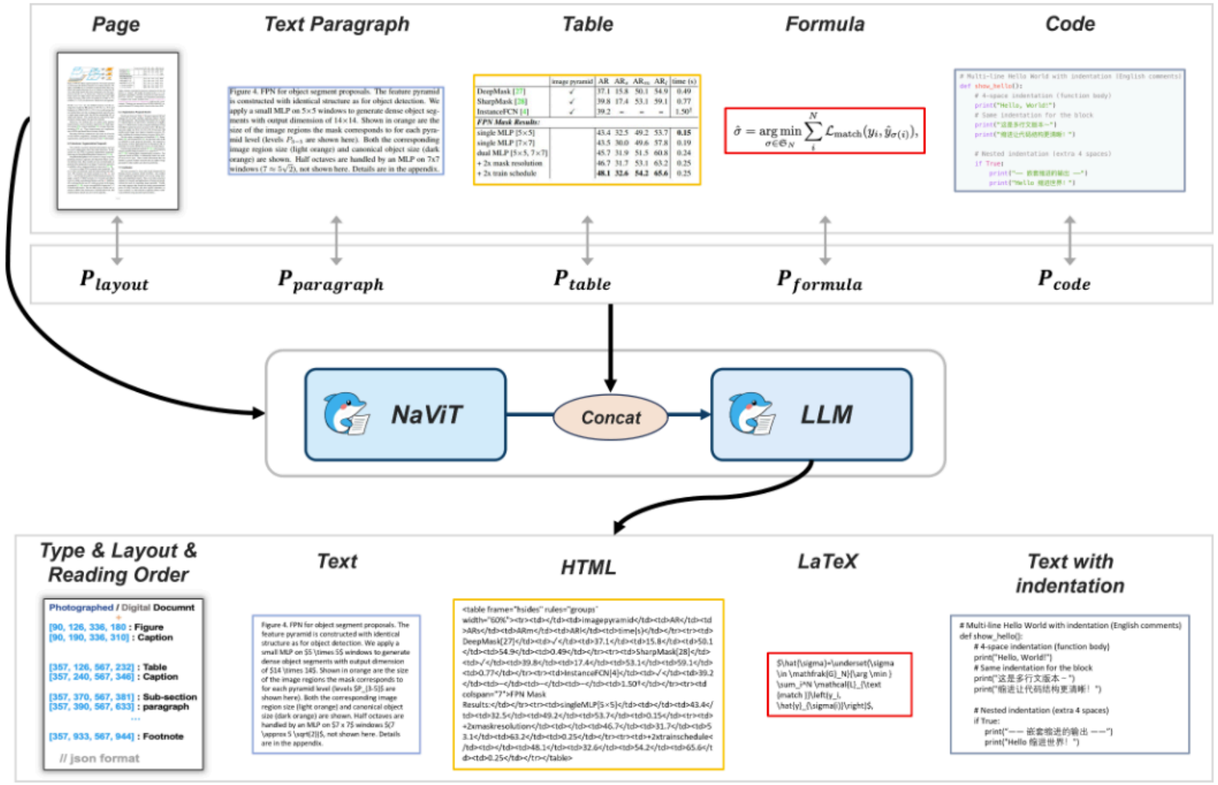
MinerU2.5

촬영 문서 파싱

- 촬영 문서 페이지 전체를 하나의 입력으로 처리
- Stage 1의 시각 특징 재사용
- 전체 내용을 읽기 순서대로 생성

DeepShark Lab

Dolphin - v2 개요



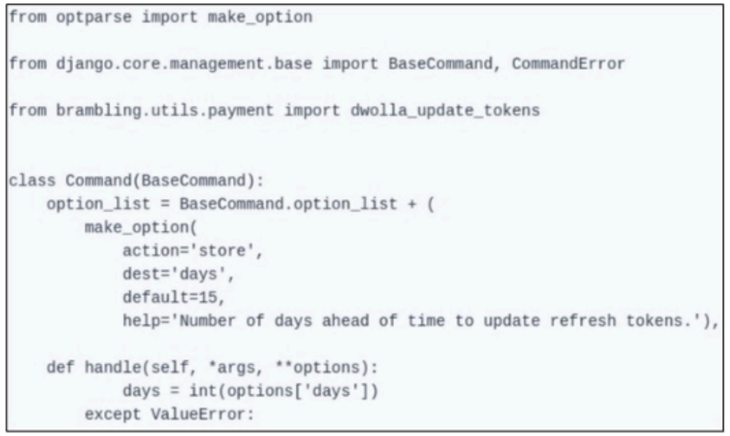
디지털 문서의 요소 단위 병렬 파싱

- Stage 1의 Bounding Box를 기준으로 요소 Crop
- 각 요소의 Local View 생성
- 요소 이미지를 병렬 인코딩
- 유형별 프롬프트 적용하여 디코더에 입력
- 결과를 읽기 순서대로 결합

DeepShark Lab

Dolphin - v2 Datasets

Training Datasets

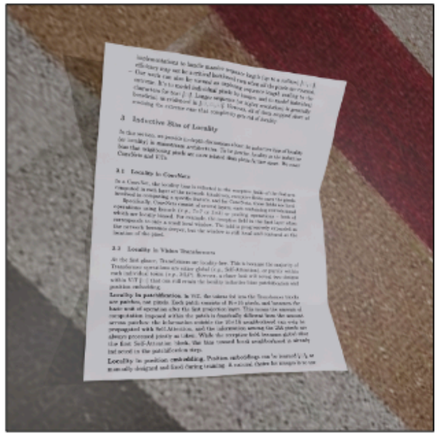


```
from optparse import make_option

from django.core.management.base import BaseCommand, CommandError

from brambling.utils.payment import dwolla_update_tokens


class Command(BaseCommand):
    option_list = BaseCommand.option_list + (
        make_option(
            action='store',
            dest='days',
            default=15,
            help='Number of days ahead of time to update refresh tokens.'),

    def handle(self, *args, **options):
            days = int(options['days'])
        except ValueError:
```

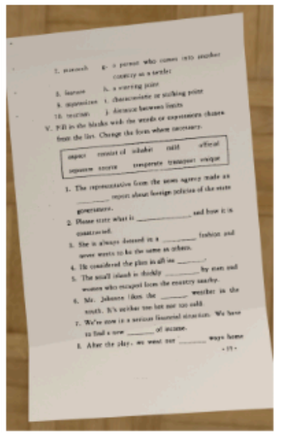
Code Image

Single-column Catalog

Double-column Catalog



Photographed Document



Photographed Document

DeepShark Lab

Dolphin - v2 Datasets

Evaluation Benchmarks

벤치마크	평가 목적	주요 특징
OmniDocBench	범용 문서 파싱	다양한 문서 유형과 요소별 성능
RealDoc-160	모델의 실제 촬영 문서 파싱 성능 평가	영어 80장·중국어 80장
DocPTBench	촬영 문서의 파싱 및 번역 성능 평가, 외부 벤치마크로 성능 검증	1,300장 이상의 고해상도 촬영 문서

DeepShark Lab

Dolphin - v2 Experiments

모델 설정

Backbone model

- Qwen2.5-VL-3B 기반 fine-tuning
- 이미지와 텍스트를 처리하는 VLM

Vision Encoder

- 입력 문서 이미지를 시각 특징으로 변환
- 학습 중 파라미터 고정

Vision-Language Adapter

- 시각 특징을 LLM이 이해할 수 있는 표현으로 변환
- End-to-End 학습

Language Model Decoder

- 변환된 이미지 특징과 프롬프트를 바탕으로 결과를 순차적으로 생성
- End-to-End 학습

학습 설정

- Optimizer: AdamW
- Learning rate: 8e-5
- Weight decay: 0
- Cosine annealing schedule
- Warmup ratio: 0.03
- 10 epochs
- Batch size: GPU당 8
- Gradient accumulation: 4 steps
- Effective batch size: GPU당 32
- Maximum sequence length: 131,072 tokens

DeepShark Lab

Dolphin - v2 Experiment

OmniDocBench Evaluation

	Methods	Single Model	Size	Overall↑	Text^{Edit} ↓	Formula^{CDM} ↑	Table^{TEDS} ↑	Table$^{TEDS\text{-}S}$ ↑	Read Order^{Edit} ↓
Pipeline Systems	PP-StructureV3	*N*	-	86.73	0.073	85.79	81.68	89.48	0.073
	Mineru2-pipeline	*N*	-	75.51	0.209	76.55	70.90	79.11	0.225
	Marker-1.8.2	*N*	-	71.30	0.206	76.66	57.88	71.17	0.250
General VLMs	Qwen3-VL-235B	*Y*	235B	89.15	0.069	88.14	86.21	90.55	0.068
	Gemini-2.5 Pro	—	-	88.03	0.075	85.82	85.71	90.29	0.097
	Qwen2.5-VL	*Y*	72B	87.02	0.094	88.27	82.15	86.22	0.102
	InternVL3.5	*Y*	241B	82.67	0.142	87.23	75.00	81.28	0.125
	InternVL3	*Y*	78B	80.33	0.131	83.42	70.64	77.74	0.113
	GPT-4o	-	-	75.02	0.217	79.70	67.07	76.09	0.148
Specialized VLMs	PaddleOCR-VL	*N*	0.9B	**91.93**	**0.039**	**88.67**	**91.01**	**94.85**	0.048
	MinerU2.5	*Y*	1.2B	90.67	0.047	88.46	88.22	92.38	**0.044**
	MonkeyOCR-pro	*N*	3B	88.85	0.075	87.25	86.78	90.63	0.128
	dots.ocr	*Y*	3B	88.41	0.048	83.22	86.78	90.62	0.053
	MonkeyOCR	*N*	3B	87.13	0.075	87.45	81.39	85.92	0.129
	Deepseek-OCR	*Y*	3B	87.01	0.073	83.37	84.97	88.80	0.086
	MonkeyOCR-pro	*N*	1.2B	86.96	0.084	85.02	84.24	89.02	0.130
	Nanonets-OCR-s	*Y*	3B	85.59	0.093	85.90	80.14	85.57	0.108
	MinerU2-VLM	*Y*	0.9B	85.56	0.078	80.95	83.54	87.66	0.086
	olmOCR	*N*	7B	81.79	0.096	86.04	68.92	74.77	0.121
	POINTS-Reader	*N*	3B	80.98	0.134	79.20	77.13	81.66	0.145
	Mistral-OCR	-	-	78.83	0.164	82.84	70.03	78.04	0.144
	OCRFlux	*N*	3B	74.82	0.193	68.03	75.75	80.23	0.202
	Dolphin	*Y*	0.3B	74.67	0.125	67.85	68.70	77.77	0.124
	Dolphin-v1.5	*Y*	0.3B	85.06	0.085	79.44	84.25	88.06	0.071
	Dolphin-v2 (Ours)	*Y*	3B	89.78	0.054	87.63	87.02	90.48	0.054

DeepShark Lab

Dolphin - v2 Experiment

RealDoc-160 Evaluation

Type	Methods	Size	EN↓	ZH↓	AVG↓
General VLMs	Gemini-2.5 Pro	-	0.0889	**0.0681**	0.0785
	GPT-4o	-	0.0588	0.2694	0.1641
	GPT-4o-mini	-	0.1943	0.7057	0.4500
	Qwen2.5-VL	7B	0.0425	0.1144	0.0785
Specialized VLMs	Dolphin	0.3B	0.3867	0.4859	0.4363
	MonkeyOCR	3B	<u>0.0362</u>	0.1180	<u>0.0771</u>
	MinerU2.5	1.2B	0.3359	0.3641	0.3500
	PaddleOCR-VL	0.9B	0.1616	0.1979	0.1798
	Dolphin-1.5	0.3B	0.1847	0.2614	0.2231
	dots.ocr	3B	0.1480	0.2553	0.2017
	Dolphin-v2	3B	**0.0046**	<u>0.0737</u>	**0.0392**

DeepShark Lab

Dolphin - v2 Experiment

DocPTBench Evaluation

Type	Methods	Overall$^{Edit}\downarrow$		Text$^{Edit}\downarrow$		Formula$^{Edit}\downarrow$		Table$^{TEDS}\uparrow$		Read Order$^{Edit}\downarrow$	
		En	Zh	En	Zh	En	Zh	En	Zh	En	Zh
General VLMs	Qwen2.5-VL-72B [12]	41.5	57.0	36.2	56.6	42.2	61.8	57.0	55.5	28.1	51.3
	Gemini2.5-Pro [75]	**18.2**	**30.4**	**9.8**	**22.0**	**37.1**	**52.2**	**81.3**	**82.9**	**11.2**	**18.1**
	Doubao-1.6-v [76]	54.7	55.4	60.6	58.2	51.5	61.1	27.6	37.9	39.7	40.2
	Qwen-VL-Max [77]	27.7	42.7	15.9	41.5	41.8	57.2	71.1	71.6	16.8	34.4
	GLM-4.5v [78]	36.7	49.6	26.2	47.7	49.9	66.2	58.9	54.0	27.3	35.7
	Kimi-VL [79]	36.5	38.7	17.2	22.0	48.6	52.2	57.1	67.8	14.3	18.1
Specialized VLMs	PaddleOCR-VL [10]	37.5	39.6	29.4	37.7	46.5	52.6	54.2	65.3	28.8	37.9
	MinerU2.5 [13]	37.3	47.4	37.0	53.6	44.3	62.0	54.9	59.8	29.0	40.3
	dots.ocr [60]	33.7	37.3	29.8	35.8	**39.2**	54.4	63.7	**67.6**	32.8	31.8
	MonkeyOCR [14]	46.4	52.8	34.5	43.9	48.7	61.6	33.1	37.4	37.9	44.1
	Dolphin [17]	57.5	71.5	54.9	71.5	65.6	82.8	33.0	19.3	46.2	57.7
	olmOCR [21]	39.1	46.1	19.3	27.2	50.7	66.9	56.5	56.9	20.7	24.4
	OCRFlux [61]	36.2	45.8	30.4	40.4	48.4	81.1	49.5	54.3	22.5	32.1
	SmolDocling [57]	90.1	93.7	89.8	99.2	99.6	99.9	4.4	2.4	72.7	75.9
	Nanonets-OCR [62]	38.6	52.1	21.0	42.0	48.1	67.0	58.5	50.6	21.4	32.7
	DeepSeek-OCR [65]	54.4	57.8	56.7	57.6	54.4	74.1	28.0	35.4	41.7	40.4
	Nanonets-OCR2 [62]	34.2	46.1	25.5	44.6	69.0	76.4	**70.7**	66.0	19.5	31.4
	Dolphin-v2 (Ours)	**30.8**	**37.3**	**21.5**	**35.8**	48.2	**54.4**	59.0	53.6	**19.0**	**29.2**

DeepShark Lab

Dolphin - v2 Experiment

Ablation Study 1. 문서 유형 분류 유무 실험

Model Variant	EN↓	ZH↓	AVG↓
w/o Classification	0.1523	0.2218	0.1871
Dolphin-v2 (Full)	**0.0046**	**0.0737**	**0.0392**

- w/o Classification: 문서 유형 분류 제거
- RealDoc-160 벤치마크 사용
- Edit Distance 점수

DeepShark Lab

Dolphin - v2 Experiment

Ablation Study 2. 수식 전용 파싱

Model Variant	CDM↑
Unified Parsing	83.34
Dedicated Formula Parsing	**86.72**

- 문단과 다른 수식 전용 프롬프트를 사용했을 때의 효과 평가
- 문단과 같은 프롬프트 사용할 경우, 수식이 포함된 일반 문단 전체를 하나의 독립 수식으로 해석할 수 있음

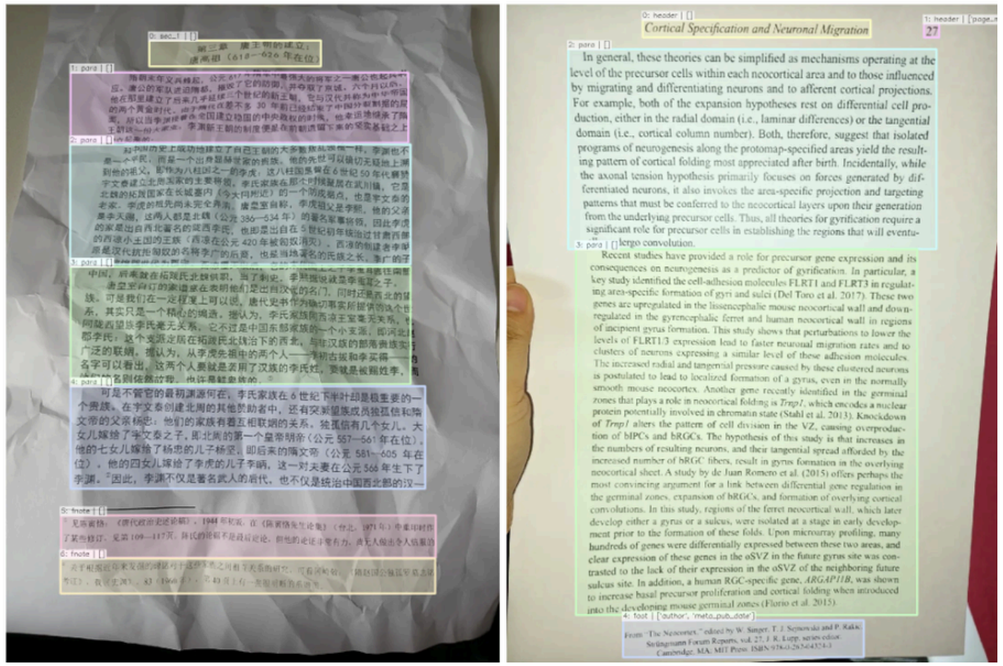
```
and for L^{\circ}, -\Lambda and T (36) holds. From this follows the fact that
u \in W_s^{2,p}(B) solves the Dirichlet problem with right hand side
f \in W_s^{0,p}(B) if and only if for v := (r^{-1/2}u)\circ T with
g := (r^{-3/2}f)\circ T \in W_t^{0,p}(B) and t := 2s + 3p + 3 \in (2p-3,3p-3):
```

```
\begin{array}{l}
\text{and for } L^{\circ}, A \text{ and } T \text{ (36) holds. From this follows the fact that } \\
u \in W_{s}^{2,p}(B) \text{ solves the Dirichlet problem with right hand side } \\
f \in W_{s}^{0,p}(B) \text{ if and only if for } v := (r^{-1/2}u) \circ T \text{ with } \\
g := (r^{-3/2}f) \circ T \in W_{t}^{0,p}(B) \text{ and } t := 2s + \frac{3}{p} + 3 \in (2p-3, 3p-3):
\end{array}
```

$$
\begin{array}{l}
\text{and for } L^{\circ}, A \text{ and } T \text{ (36) holds. From this follows the fact that } \\
u \in W_{s}^{2,p}(B) \text{ solves the Dirichlet problem with right hand side } \\
f \in W_{s}^{0,p}(B) \text{ if and only if for } v := (r^{-1/2}u) \circ T \text{ with } \\
g := (r^{-3/2}f) \circ T \in W_{t}^{0,p}(B) \text{ and } t := 2s + \frac{3}{p} + 3 \in (2p-3, 3p-3):
\end{array}
$$

DeepShark Lab

Dolphin - v2 한계



문서 유형 분류 오류

- 약한 왜곡의 촬영 문서를 디지털 문서로 인식
- 잘못된 파싱 전략 사용

요소 유형 범위 제한

- 화학 구조식, 차트 및 악보 등 미지원

Q & A

감사합니다