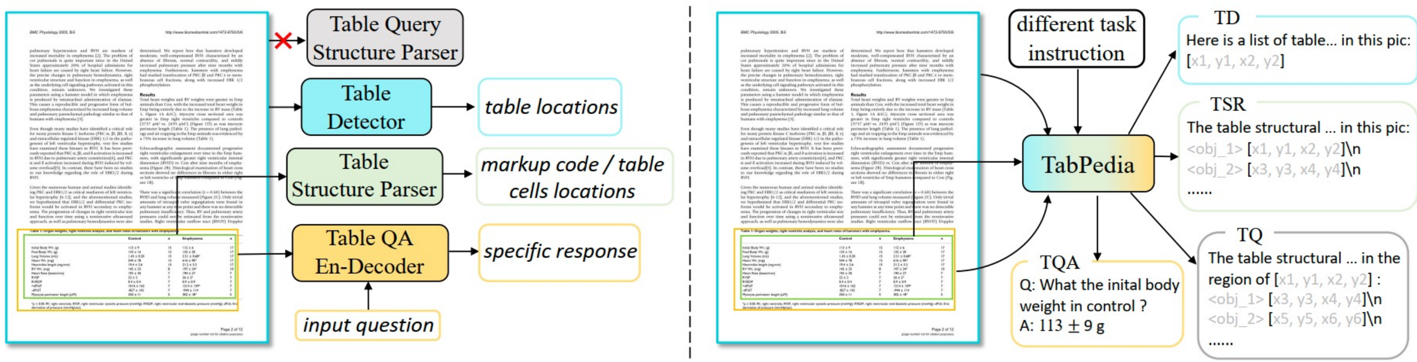
TabPedia: Towards Comprehensive Visual Table Understanding with Concept Synergy

Weichao Zhao[1,2,*] Hao Feng[1,*] Qi Liu[2,*] Jingqun Tang[2] Shu Wei[2] Binghong Wu[2]
Lei Liao[2] Yongjie Ye[2] Hao Liu[2,‡,†] Wengang Zhou[1,†] Houqiang Li[1] Can Huang[2]

[1]University of Science and Technology of China [2]ByteDance

Background

- Tables play a vital role in summarizing facts and quantitative data. The compact yet informative nature of tables makes them advantageous for various applications.
- However, many pioneering works have mainly centered on the specific subtask with various task-specific architectures.



(a) Previous task-specific pipelines (b) Our proposed TabPedia

table detection (TD) table structure recognition (TSR) table query (TQ) table question answering (TQA)

Motivation

- Leverage the generalizable knowledge of the LLMs for comprehensive visual table understanding.
- Two challenges
  - ✓ Discrepancy between the representation formats (two-dimensional structure VS. one-dimensional sequence).
  - ✓ Diverse image resolution (low resolution vision encoder VS. high resolution document images).

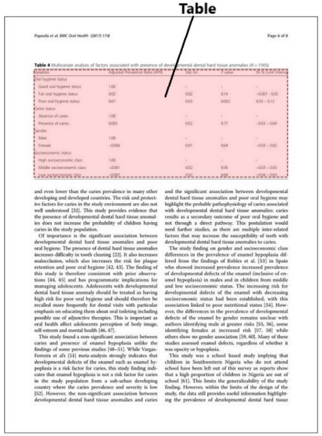


Table Detection

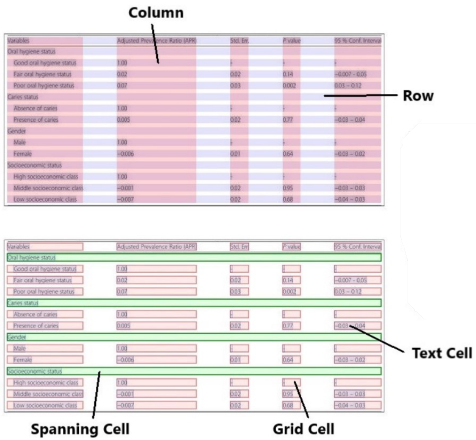
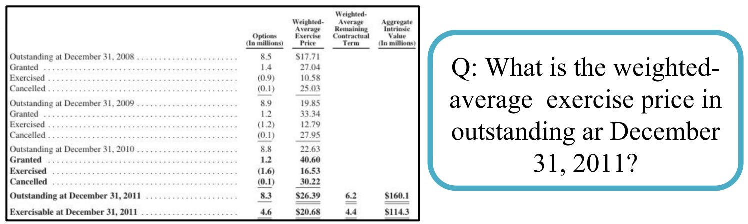


Table Structure Recognition

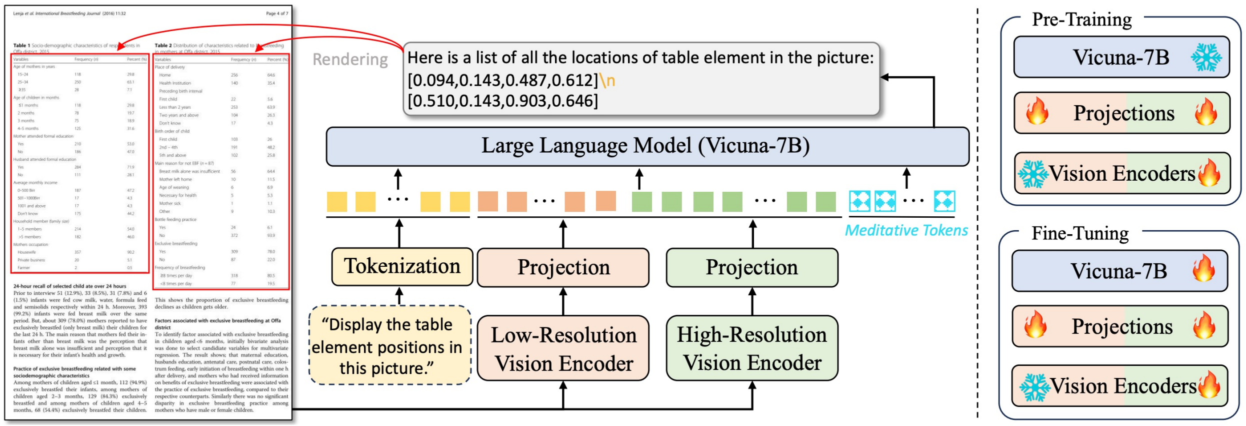


	Options (In millions)	Weighted-Average Exercise Price	Weighted-Average Remaining Contractual Term	Aggregate Intrinsic Value (In millions)
Outstanding at December 31, 2008	8.5	$17.71		
Granted	1.4	27.04		
Exercised	(0.9)	10.58		
Cancelled	(0.1)	25.03		
Outstanding at December 31, 2009	8.9	19.85		
Granted	1.2	33.34		
Exercised	(1.2)	12.79		
Cancelled	(0.1)	27.95		
Outstanding at December 31, 2010	8.8	22.63		
Granted	**1.2**	**40.60**		
Exercised	**(1.6)**	**16.53**		
Cancelled	**(0.1)**	**30.22**		
Outstanding at December 31, 2011	**8.3**	**$26.39**	**6.2**	**$160.1**
Exercisable at December 31, 2011	**4.6**	**$20.68**	**4.4**	**$114.3**

Table Question Answering

Architecture

- Dual-visual encoders
  - ✓ High-resolution encoder: capture local-level fine-grained visual cues
  - ✓ Low-resolution encoder: embed global-level visual signals
- Projectors
  - ✓ Align visual-textual embedding space
- Meditative token
  - ✓ Adaptively aggregate different region of visual tokens



Training Phase

- Pre-training
  - ✓ This phase guides the visual encoder to capture rich textual information and align the input space of the LLMs.
  - ✓ Tasks: text detection, text recognition, text localization, long text reading and image captioning
- Fine-tuning
  - ✓ Guide model to follow instructions to perform different visual table tasks
  - ✓ Tasks: table detection, table structure recognition, table querying, table question answering

Image	Instruction	Task	# Conv
Scene	LLaVA [25]	$\mathcal{C}$	595K
PDF	OCR	$\mathcal{D}, \mathcal{R}, \mathcal{S}, \mathcal{R}_p, \mathcal{R}_f$	325K
PPT	OCR	$\mathcal{D}, \mathcal{R}, \mathcal{S}, \mathcal{R}_p, \mathcal{R}_f$	600K

Table 1. Pre-training data statistics

Dataset	Subset	Task	Num
PubTab1M	PubTab1M-Det	TD	460k
	PubTab1M-Str	TSR,TQA	759k
	PubTab1M-Syn	TQ	381k
FinTabNet	–	TSR,TQA	78k
PubTabNet	–	TSR	434k
WTQ	–	TQA	1k
TabFact	–	TQA	9k

Table 2. Fine-tuning data statistics

Experiment

Quantitative results

Table 3: Comparison with the existing best table detection model TATR [9]. NMS denotes Non-Maximum Suppression.

Method	Backbone	NMS	IoU@0.75		
			Precision	Recall	F1
TATR [9]	Faster R-CNN	✓	92.7	86.6	89.5
	DETR	✓	**98.8**	98.1	**98.4**
TabPedia	LVLM	✗	98.5	**98.4**	**98.4**

Table 1. Table detection task

Table 4: Comparison with end-to-end TSR methods on two datasets. "∗" represents the results reported by [41].

Method	Input Size	PubTabNet	FinTabNet
		S-TEDS	S-TEDS
Donut [43]*	1,280	25.28	30.66
EDD [64]	512	89.90	90.60
OmniParser [41]	1,024	90.45	91.55
TabPedia	2,560	**95.41**	**95.11**

Table 2. Table structure recognition task

Table 5: Quantitative results on two subsets of PubTab1M [9], including PubTab1M-Str and PubTab1M-Syn.

(a) Comparison with the task-specific model, TATR [9] on TSR task. "Cropped" denotes utilizing cropped table-centric images.

Method	Backbone	Image	NMS	PubTab1M-Str			
				$\text{GriTS}_{\text{Top}}$	$\text{GriTS}_{\text{Cont}}$	$\text{GriTS}_{\text{Loc}}$	S-TEDS
TATR [9]	Faster R-CNN	Cropped	✓	86.16	85.38	72.11	–
	DETR	Cropped	✓	**98.46**	**97.81**	**97.81**	**97.65**
TabPedia (TSR)	LVLM	Cropped	✗	96.52	96.73	95.54	95.66

(b) Quantitative results on both TQ and TD+TQ tasks.

Method	Image	NMS	Task	PubTab1M-Syn			
				$\text{GriTS}_{\text{Top}}$	$\text{GriTS}_{\text{Cont}}$	$\text{GriTS}_{\text{Loc}}$	S-TEDS
TabPedia	**Raw**	✗	TQ	96.04	96.23	94.95	95.07
			TD+TQ	94.54	94.63	93.25	93.38

Table 3. Table querying task

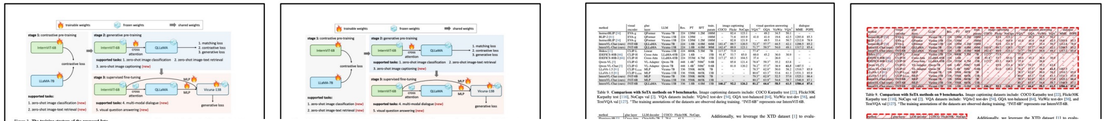
Table 6: Comparison with existing LVLMs on TQA task. "∗" denotes the results obtained through the open-source checkpoint or API of the closed-source model. ComTQA is our released new benchmark. The second best methods are underlined.

Method	Input Size	WTQ	TabFact	ComTQA
		Acc	Acc	Acc
TextMonkey [12]	896	37.9	53.6	13.9*
Monkey [93]	896	25.3*	49.8	–
Cogagent [94]	1,120	30.2*	51.7*	–
DocOwl 1.5 [40]	1,344	39.8	**80.4**	18.5*
GPT4V [95]	645	<u>45.5</u>*	69.3*	27.2*
Gemini Pro [87]	659	32.3*	67.9*	<u>29.3</u>*
Xcomposer2 [96]	511	28.7	62.3	–
TabPedia	2,560	**47.8**	<u>71.3</u>	**53.5**

Table 4. Table question answering task

Experiment

Qualitative results



Model	Image Caption		General VQA				
	Flickr30K	TextCaps	VQAv2	OKVQA	GQA	ScienceQA	VizWiz
Flamingo-80B [1]	67.2	-	56.3	50.6	-	-	31.6
Palm-E-12B [13]	-	-	77.7	60.1	-	-	-
BLIP-2 (Vicuna-13B) [27]	71.6	-	65.0	45.9	32.3	61.0	19.6
InstructBLIP (Vicuna-13B) [12]	82.8	-	-	-	49.5	63.1	33.4
Shikra (Vicuna-13B) [7]	73.9	-	77.4	47.2	-	-	-
mPLUG-Owl2 [56]	85.1	-	79.4	57.7	56.1	68.7	54.5
LLaVA1.5 (Vicuna-7B) [29]	-	-	78.5	-	**62.0**	66.8	50.0
Qwen-VL(Qwen-7B) [3]	85.8	65.1	79.5	58.6	59.3	67.1	35.2
Qwen-VL-Chat [3]	81.0	-	78.2	56.6	57.5	68.2	38.9
Monkey	**86.1**	**93.2**	**80.3**	**61.3**	60.7	**69.4**	**61.2**

Q: Answer the question and give a reason. what's the point of Flamingo-80B in VQAv2?
A: 56.3 Reason: Flamingo-80B is a model in VQAv2. Its VQAv2 score is shown in the second column, which is 56.3.

Method
UReader [49]
Qwen-VL [2]
TextMonkey [31]
Monkey [26]
Cogagent [14]
DocOwl 1.5 [15]
Llava Next 34B [28]
GPT4V [38]
Gemini Pro [8]
Xcomposer2 [9]
TextSquare (ours)

task	#samples	dataset
Captioning	588K	COCO Caption [22], TextCaps [126]
VQA	1.1M	VQAv2 [54], OKVQA [104], A-OKVQA [122], IconQA [99], AI2D [71], GQA [64]
OCR	294K	OCR-VQA [107], ChartQA [105], DocVQA [29], ST-VQA [12], EST-VQA [150], InfoVQA [106], LLaVAR [182]
Grounding	323K	RefCOCO/+/g [103, 170], Toloka [140]
Grounded Cap.	284K	RefCOCO/+/g [103, 170]
Conversation	1.4M	LLaVA-150K [92], SVIT [183], VisDial [36], LRV-Instruction [90], LLaVA-Mix-665K [91]

Q: Answer the question and give a reason. How many datasets utilized in the Conversation task?
A: 5 Reason: There are 5 datasets utilized in the Conversation task, which are LLaVA-150K, SVIT, VisDial, LLaVA-Mix-665K and LRV-Instruction.

(c) In-the-wild cases on TQA task

Method	Arc						
CLIP	ViT-L/14$_{336}$	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION-2B	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION-2B	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4

CLIP	ViT-L/14$_{336}$	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION-2B	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION-2B	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4

(b) In-the-wild cases on TSR task

Dataset

- **ComTQA**
  - ✓ Data sources: PubTab1M + FinTabNet
  - ✓ Data statistic

Table A1: Statistics of ComTQA benchmark.

	PubTab1M	FinTabNet	Total
#images	932	659	1,591
#QA pairs	6,232	2,838	9,070
Avg. per image	6	4	5

Table 2: Effect of varying β on classification accuracy. The effect of varying β was studied for the colon cancer data set. A value of between 0.5 – 1 as suggested by Battiti [21] seems appropriate.

β	0.0	0.2	0.4	0.6	0.8	1.0
accurate	87.1%	88.7%	90.3%	90.3%	90.3%	90.3%

Q: Which beta value has the highest classification accuracy?
A: 0.4 \n 0.6 \n 0.8 \n 1.0

(a) multiple answers

Table 2: Calibration factors for three evolutionary models

	c
Dayhoff	1.3370
JTT	1.2873
MV	1.1775

The raw distance d_r is scaled by the calibration factor c, which was obtained by least squares fitting of 2000 artificial protein sequence

Q: What is the sum of the calibration factors for the three models?
A: 3.8018

(b) mathematical calculation

Table 3: Observed and predicted MTS volume using seven models.

Time(hr)	*Volume*	*H3*	*Weibull*	*H1*	*H2*	*Gompertz*	*Logistic*	*Richards*
0	0.087	0.087	0.087	0.087	0.087	0.087	0.087	0.087
24	0.080	0.080	0.088	0.067	0.089	0.099	0.107	0.108
48	0.082	0.083	0.096	0.093	0.099	0.116	0.132	0.134
72	0.129	0.127	0.125	0.133	0.125	0.140	0.162	0.165
96	0.188	0.189	0.184	0.186	0.182	0.177	0.200	0.202
120	0.255	0.256	0.259	0.251	0.261	0.234	0.245	0.245
144	0.318	0.317	0.317	0.320	0.316	0.327	0.302	0.297

Q: Which model predicts the largest volume at time 72?
A: Richards

(c) logical inference

Thanks ！