

# DeepSeek-OCR: Contexts Optical Compression(2025)

Haoran Wei, Yaofeng Sun, Yukun Li

Paper Review

2025.11.28

Guebeen Lee



# Contents

1. Introduction
2. Methodology
3. Experiments
4. Contributions



# Introduction



## DeepSeek-OCR: Contexts Optical Compression

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

### Abstract

We present DeepSeek-OCR as an initial investigation into the feasibility of compressing long contexts via optical 2D mapping. DeepSeek-OCR consists of two components: DeepEncoder and DeepSeek3B-MoE-A570M as the decoder. Specifically, DeepEncoder serves as the core engine, designed to maintain low activations under high-resolution input while achieving high compression ratios to ensure an optimal and manageable number of vision tokens. Experiments show that when the number of text tokens is within 10 times that of vision tokens (i.e., a

21 Oct 2025

# Introduction

## 1. Introduction

Current Large Language Models (LLMs) face significant computational challenges when processing long textual content due to quadratic scaling with sequence length. We explore a potential solution: leveraging visual modality as an efficient compression medium for textual information. A single image containing document text can represent rich information using substantially fewer tokens than the equivalent digital text, suggesting that optical compression through vision tokens could achieve much higher compression ratios.

# Introduction

1 Token  $\approx$  1 Word

# Introduction



---

## DeepSeek-OCR: Contexts Optical Compression

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

# Introduction

Text Tokens	Vision Tokens =64		Vision Tokens=100		Pages
	Precision	Compression	Precision	Compression	
600-700	96.5%	10.5×	98.5%	6.7×	7
700-800	93.8%	11.8×	97.3%	7.5×	28
800-900	83.8%	13.2×	96.8%	8.5×	28
900-1000	85.9%	15.1×	96.8%	9.7×	14
1000-1100	79.3%	16.5×	91.5%	10.6×	11
1100-1200	76.4%	17.7×	89.8%	11.3×	8
1200-1300	59.1%	19.7×	87.1%	12.6×	4

# Introduction

## Convert

Text Tokens

Image Tokens

Conversation

Images

1,000,000 tokens

100,000 tokens



# Introduction

## Context Window

Historical Conversation   Recent Conversation

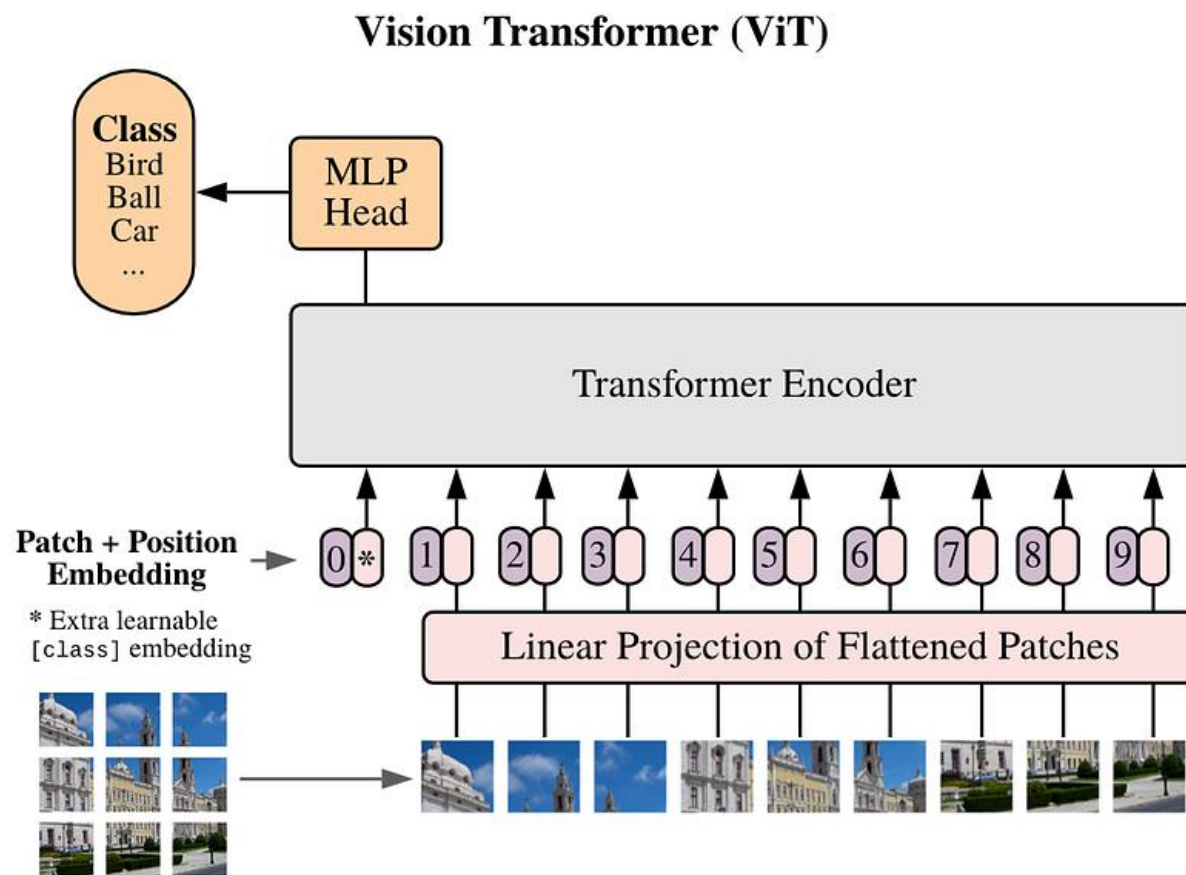


Vision Tokens

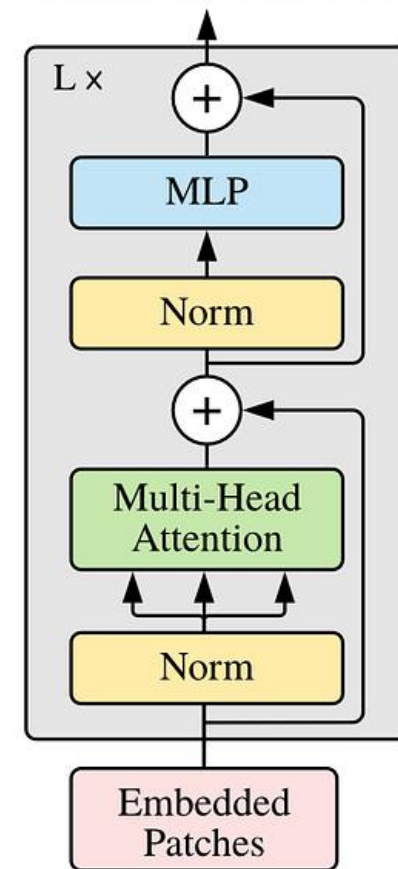
The diagram illustrates the context window for the model. It consists of two main horizontal sections. The top section is labeled 'Historical Conversation' and 'Recent Conversation'. Below these labels are two colored boxes: a green box on the left labeled 'Vision Tokens' and a blue box on the right labeled 'Text Tokens'. The green box is wider than the blue box. The entire diagram is set against a white background with a blue footer bar at the bottom.

Text Tokens

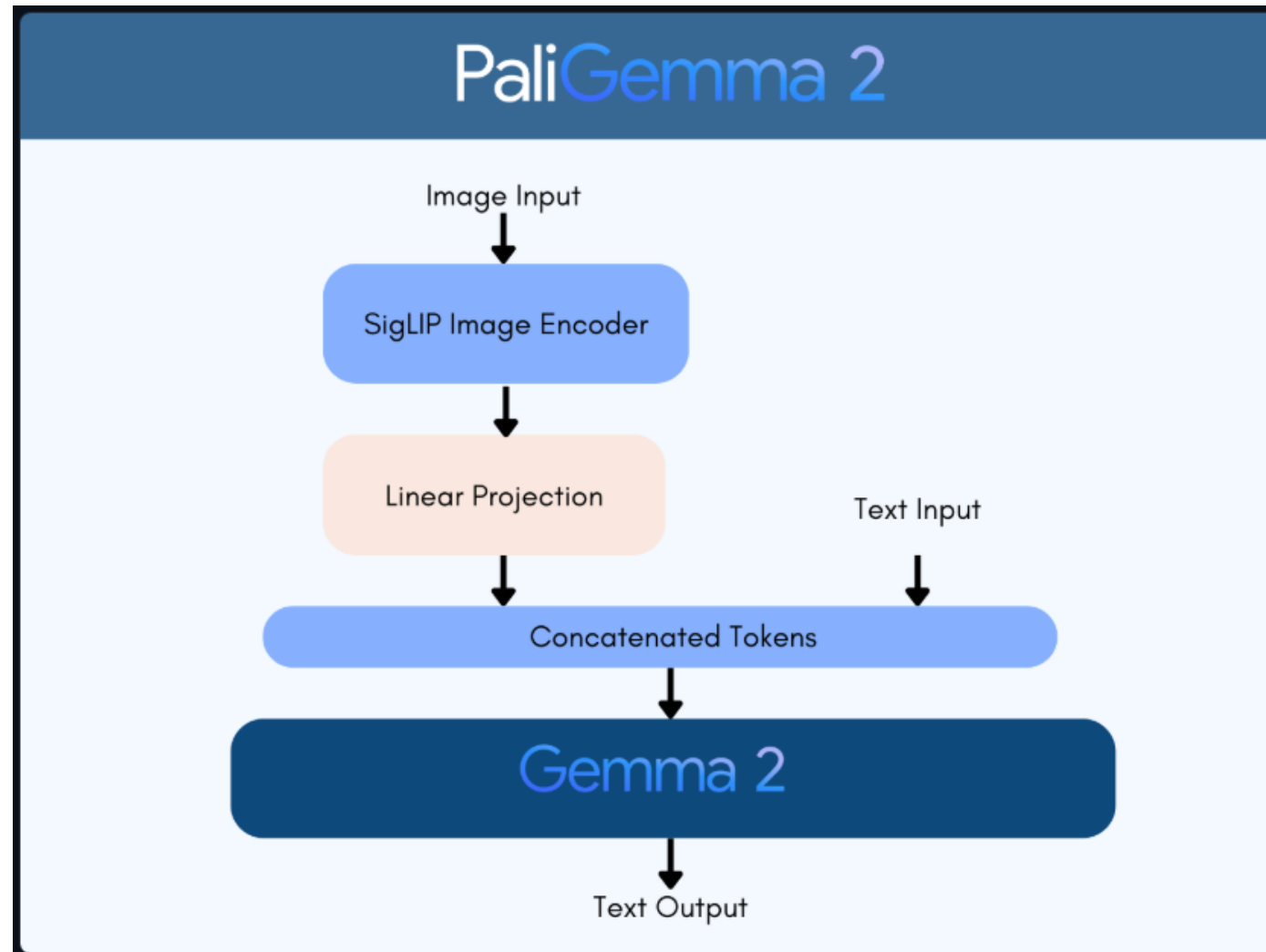
# Methodology



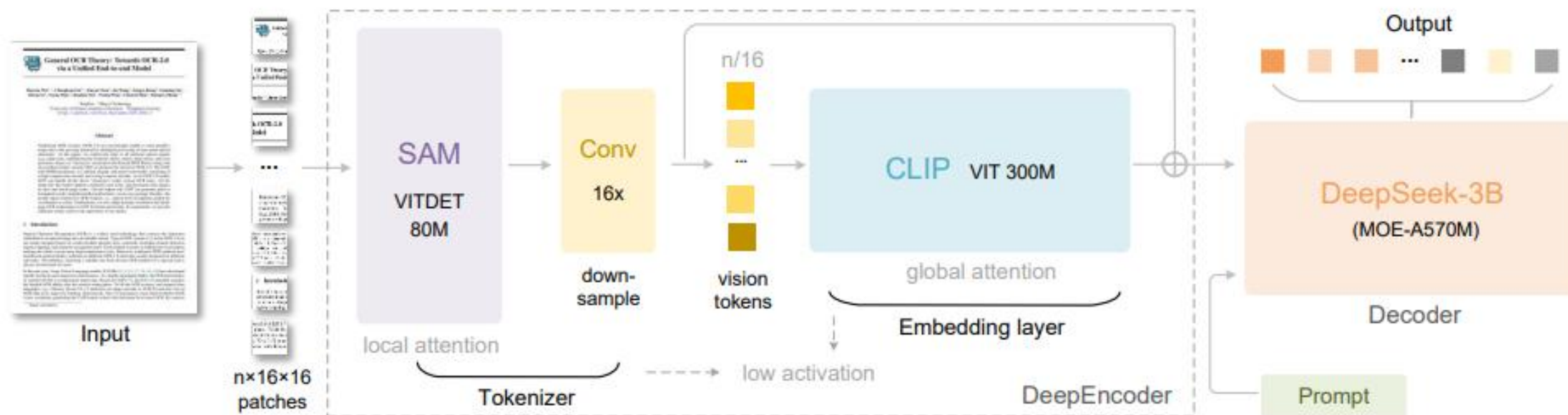
**Transformer Encoder**



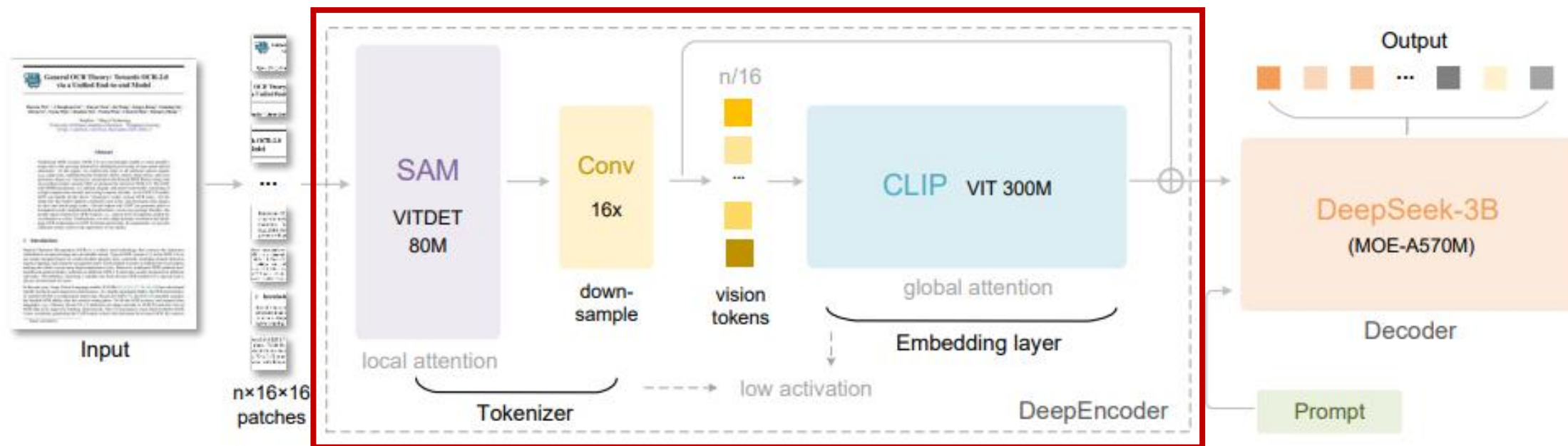
# Methodology



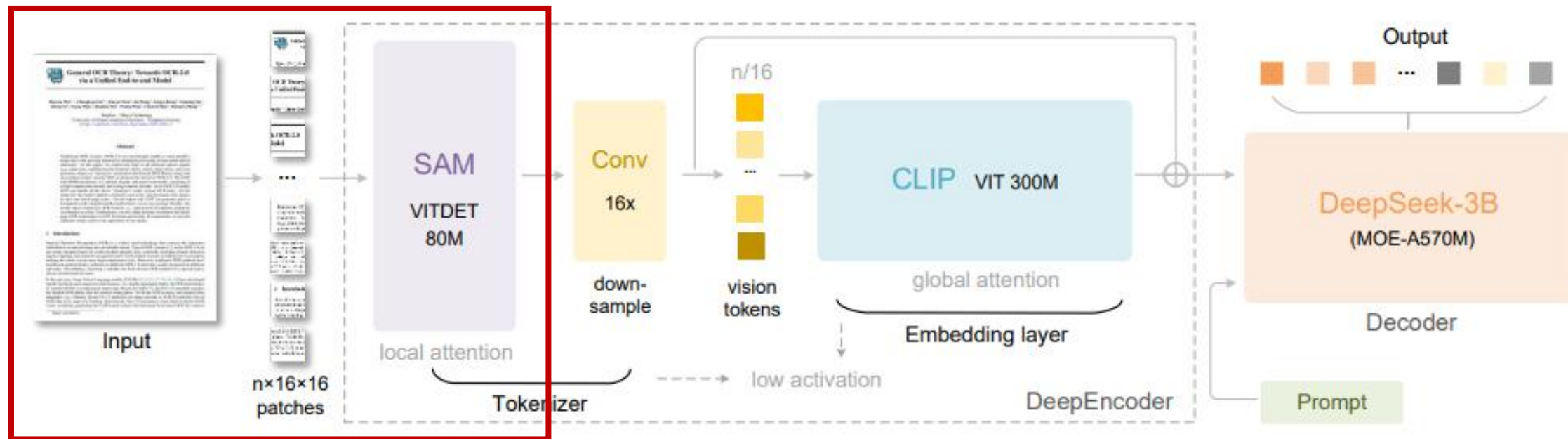
# Methodology



# Methodology



# Methodology



# Methodology

>SAM

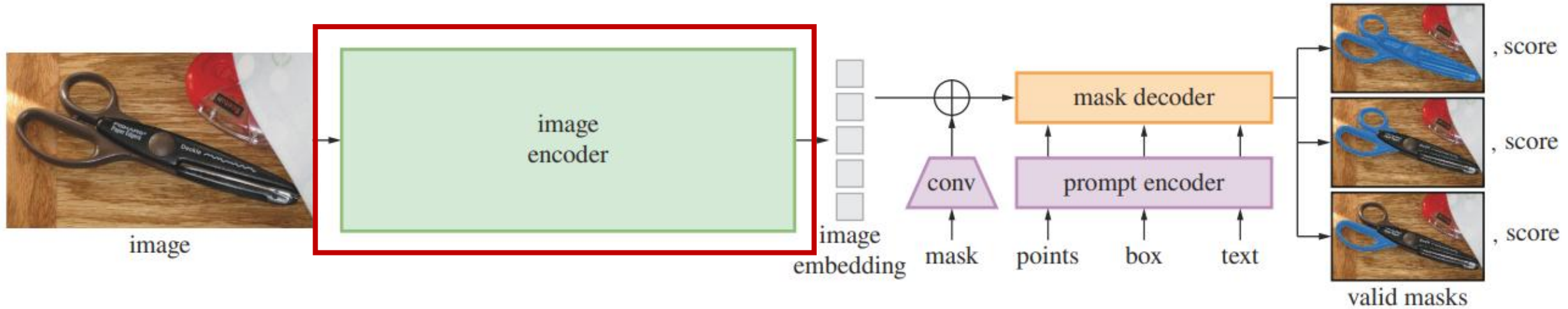
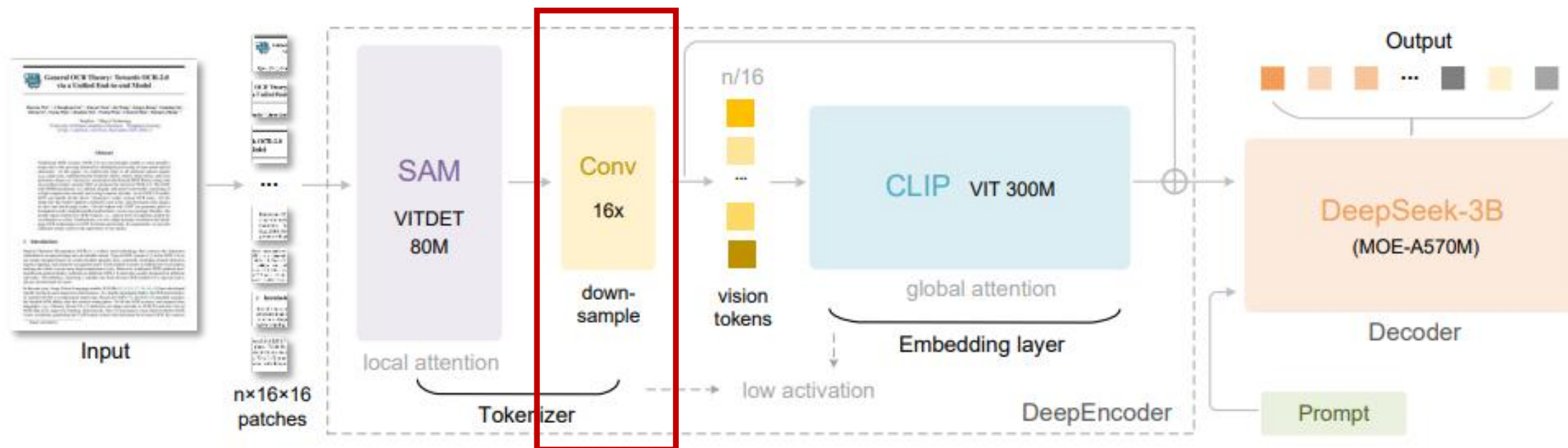
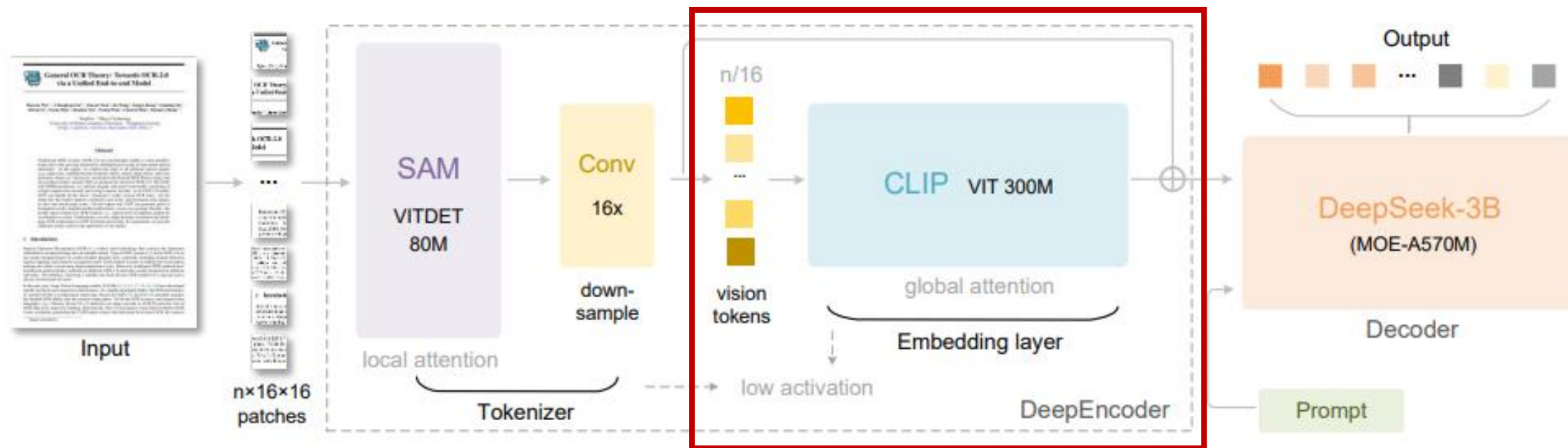


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

# Methodology

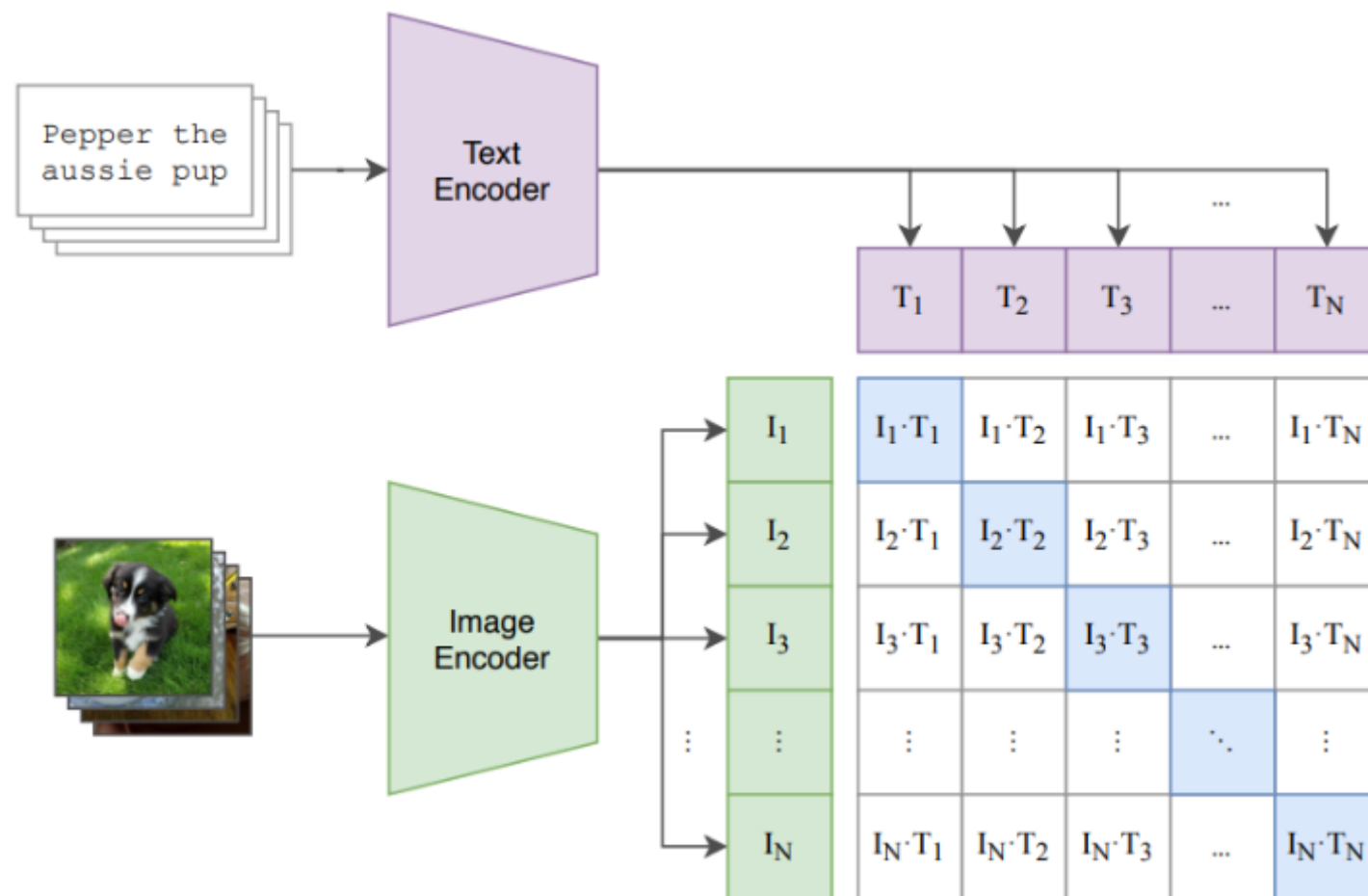


# Methodology

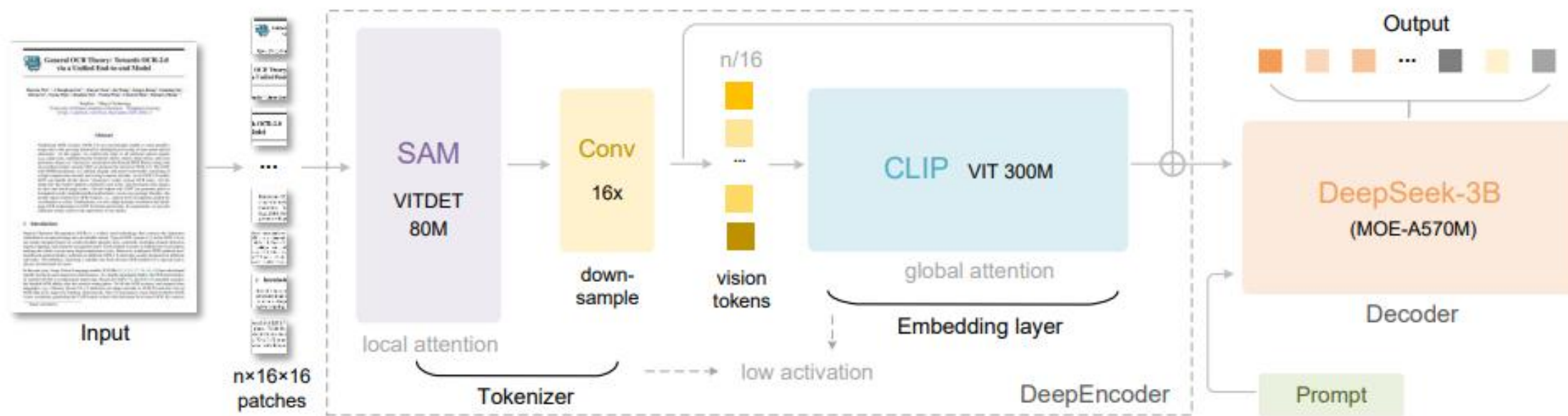


# Methodology

## >CLIP Model



# Methodology



# Experiments

Model	Tokens	English					Chinese				
		overall	text	formula	table	order	overall	text	formula	table	order
Pipeline Models											
Dolphin [11]	-	0.356	0.352	0.465	0.258	0.35	0.44	0.44	0.604	0.367	0.351
Marker [1]	-	0.296	0.085	0.374	0.609	0.116	0.497	0.293	0.688	0.678	0.329
Mathpix [2]	-	0.191	0.105	0.306	0.243	0.108	0.364	0.381	0.454	0.32	0.30
MinerU-2.1.1 [34]	-	0.162	0.072	0.313	0.166	0.097	0.244	0.111	0.581	0.15	0.136
MonkeyOCR-1.2B [18]	-	0.154	0.062	0.295	0.164	0.094	0.263	0.179	0.464	0.168	0.243
PPstructure-v3 [9]	-	0.152	0.073	0.295	0.162	0.077	0.223	0.136	0.535	0.111	0.11
End-to-end Models											
Nougat [6]	2352	0.452	0.365	0.488	0.572	0.382	0.973	0.998	0.941	1.00	0.954
SmolDocling [25]	392	0.493	0.262	0.753	0.729	0.227	0.816	0.838	0.997	0.907	0.522
InternVL2-76B [8]	6790	0.44	0.353	0.543	0.547	0.317	0.443	0.29	0.701	0.555	0.228
Qwen2.5-VL-7B [5]	3949	0.316	0.151	0.376	0.598	0.138	0.399	0.243	0.5	0.627	0.226
OLMOOCR [28]	3949	0.326	0.097	0.455	0.608	0.145	0.469	0.293	0.655	0.652	0.277
GOT-OCR2.0 [38]	256	0.287	0.189	0.360	0.459	0.141	0.411	0.315	0.528	0.52	0.28
OCRFlux-3B [3]	3949	0.238	0.112	0.447	0.269	0.126	0.349	0.256	0.716	0.162	0.263
GPT4o [26]	-	0.233	0.144	0.425	0.234	0.128	0.399	0.409	0.606	0.329	0.251
InternVL3-78B [42]	6790	0.218	0.117	0.38	0.279	0.095	0.296	0.21	0.533	0.282	0.161
Qwen2.5-VL-72B [5]	3949	0.214	0.092	0.315	0.341	0.106	0.261	0.18	0.434	0.262	0.168
dots.ocr [30]	3949	0.182	0.137	0.320	0.166	0.182	0.261	0.229	0.468	0.160	0.261
Gemini2.5-Pro [4]	-	0.148	0.055	0.356	0.13	0.049	0.212	0.168	0.439	0.119	0.121
MinerU2.0 [34]	6790	0.133	0.045	0.273	0.15	0.066	0.238	0.115	0.506	0.209	0.122
dots.ocr <sup>†200dpi</sup> [30]	5545	0.125	<b>0.032</b>	0.329	<b>0.099</b>	<b>0.04</b>	0.16	<b>0.066</b>	0.416	0.092	<b>0.067</b>
DeepSeek-OCR (end2end)											
Tiny	<b>64</b>	0.386	0.373	0.469	0.422	0.283	0.361	0.307	0.635	0.266	0.236
Small	100	0.221	0.142	0.373	0.242	0.125	0.284	0.24	0.53	0.159	0.205
Base	256(182)	0.137	0.054	0.267	0.163	0.064	0.24	0.205	0.474	0.1	0.181
Large	400(285)	0.138	0.054	0.277	0.152	0.067	0.208	0.143	0.461	0.104	0.123
Gundam	795	0.127	0.043	0.269	0.134	0.062	0.181	0.097	0.432	0.089	0.103
Gundam-M <sup>†200dpi</sup>	1853	<b>0.123</b>	0.049	<b>0.242</b>	0.147	0.056	<b>0.157</b>	0.087	<b>0.377</b>	<b>0.08</b>	0.085

# Contributions

