HighlightBench: Benchmarking and Diagnosing Markup-Driven Table Reasoning in Scientific Documents

Lexin Wang · Shenghua Liu† · Yiwei Wang · Yujun Cai · Yuyao Ge · Jiayu Yao · Jiafeng Guo · Xueqi Cheng

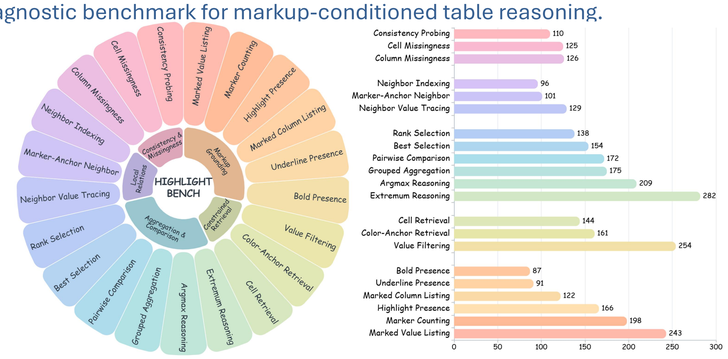
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES · University of Chinese Academy of Sciences · UNIVERSITY OF CALIFORNIA MERCED · THE UNIVERSITY OF QUEENSLAND AUSTRALIA

ECCV 2026 | MALMÖ | SEPT 8–12

Project page

WHY HIGHLIGHTBENCH

Visual markups such as highlights, underlines, and bold text are common in table-centric documents. To systematically evaluate how multimodal models understand and reason with these cues, we construct HighlightBench, a diagnostic benchmark for markup-conditioned table reasoning.



TASK TAXONOMY

T1 grounds visual cues to table regions; T2 preserves scope during constrained retrieval; T3 follows local row–column relations; T4 applies comparison and aggregation over selected subsets; T5 checks evidence consistency across missing or irregular entries.

T1:Markup Grounding
Understand the visual markup itself.
Which entries in column Eff↓ (block Part-B) are underlined?
"1.70"

T2:Constrained Retrieval
Find the correct value in the table under a given constraint.
For the number with value 16.71, list all numbers highlighted with the same color (dark_yellow, hex #EFBC23).
"91.99", "94.08", "92.61", "93.89", "16.71", "19.18", "0.12"

T3:Local Relations
Reason about local structural relations around a target cell.
Locate value 70.07, then return the target value and its four neighbors (up, down, left, right).
"target": "70.07", "up": "1.70", "down": "20.27", "left": "26.66", "right": "80.08"

Condition	Approach	Grp-A Err	Grp-A MetD↑	Part-B Eff↓	Part-B MetB↓	Part-B Qual	Sect-C Qual↓	Sect-C Stat↑
Condition A	Method-E	61.31	60.78	90.25	96.40	**91.99**	16.92	3.80
Condition A	Approach-D	63.82	25.34	70.43	95.37	—	54.14	41.11
Condition A	Baseline-A	86.31	11.63	23.94	87.35	76.05	75.78	44.83
Condition A	Approach-E	71.03	18.06	1.70	28.10	74.96	78.38	27.95
Condition B	Method-X	36.76	26.66	70.07	80.08	—	51.68	**94.08**
Condition B	System-B [88]	**92.61**	55.26	20.27	44.27	37.95	44.23	39.83
Condition B	System-C [52]	28.86	**93.89**	83.34	38.42	N/A	29.17	29.55
Condition B	Model-A	55.24	35.35	66.25	77.30	49.96	**16.71**	39.49
Condition C	Approach-F	77.34	52.15	64.15	**19.18**	40.12	56.64	28.14
Condition C	Approach-Z	75.16	12.14	**0.12**	94.25	—	81.26	39.95

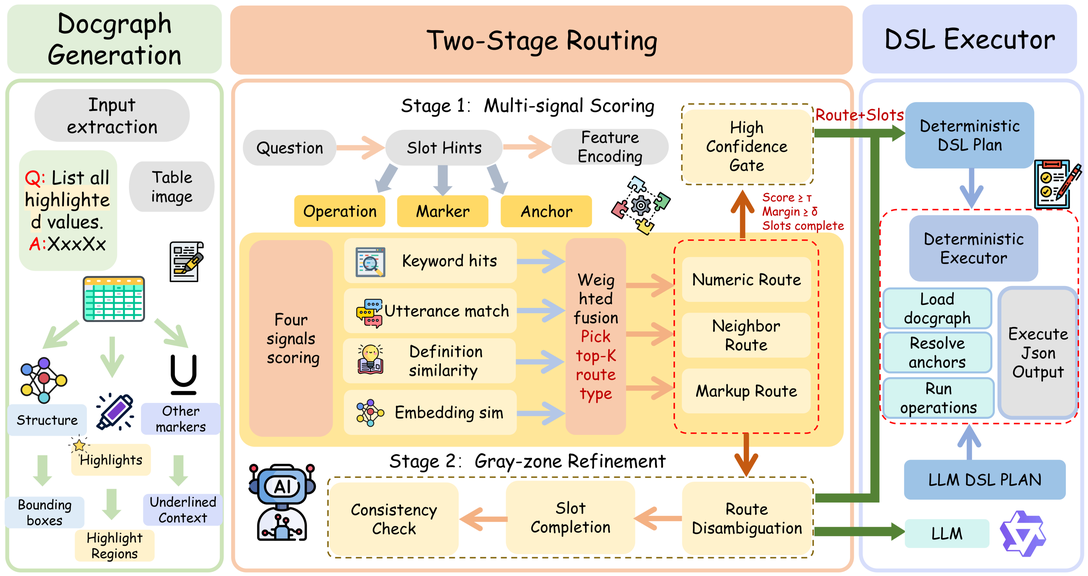
T4:Aggregation & Comparison
Compare, rank, or summarize values within a row, column, or selected group.
In metric Qual (↓) (block Sect-C), which item (main row label in the leftmost column) is optimal? Lower values indicate better performance.
"item": "Condition B", "value": "16.71"

T5:Consistency & Missingness
Handle missing values and make decisions consistently under incomplete data.
List all missing entries in column Qual (block Part-B) (N/A or —).
"Approach-D", "Method-X", "System-C [52]", "Approach-Z"

REFERENCE PIPELINE

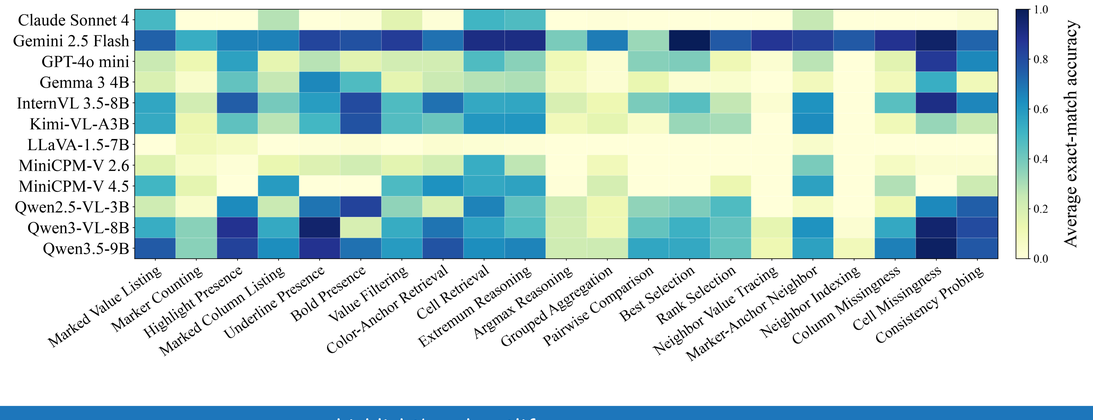
We decompose table reasoning into a three-part reference pipeline for reproducible, stage-specific diagnosis.

Docgraph Generation organizes table content, row–column topology, and detected markups, keeping cue-to-region binding explicit and inspectable. **Two-Stage Routing** selects the operation path, completes marker, anchor, and scope constraints, and directs gray-zone cases through confidence-gated refinement. **DSL Execution** runs the routed plan on the docgraph and returns the structured answer together with route, slot, plan, and execution traces.



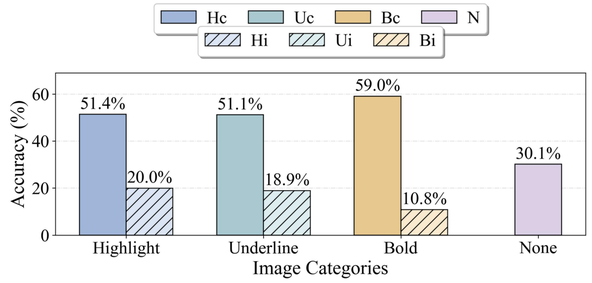
MODEL PERFORMANCE

We evaluate 12 end-to-end models across 21 automatically scorable diagnostic subtasks, covering 266 real-world images with 2,268 QA pairs and 180 synthetic images with 1,015 QA pairs. The results reveal distinct capability profiles across task families: direct retrieval is comparatively stable, while markup grounding, local relations, aggregation, and consistency remain more challenging, exposing different failure modes in markup-conditioned reasoning.

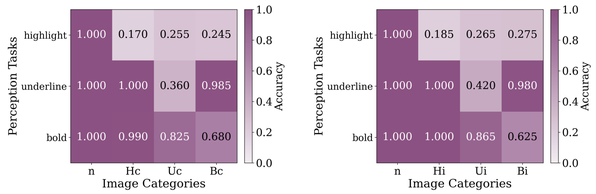


COUNTERFACTUAL EVIDENCE

Controlled marker relocation reveals how visual cues steer model decisions under fixed table content and questions. Congruent and distractor placements produce distinct perception and selection patterns.



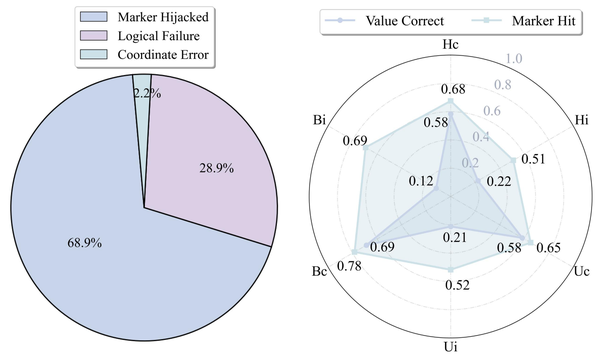
Marker relocation changes answer accuracy across highlight, underline, and bold conditions, revealing cue-dependent decisions.



The matrices separate marker perception from target alignment across counterfactual variants.

MECHANISM READOUT

Secondary analyses connect marker sensitivity with answer selection and error attribution. Marker-hit, value-correct, and failure categories locate how visual cues influence outcomes.



Failure attribution separates logical and coordinate errors; the radar plot compares marker-hit and value-correct behavior across markup types.