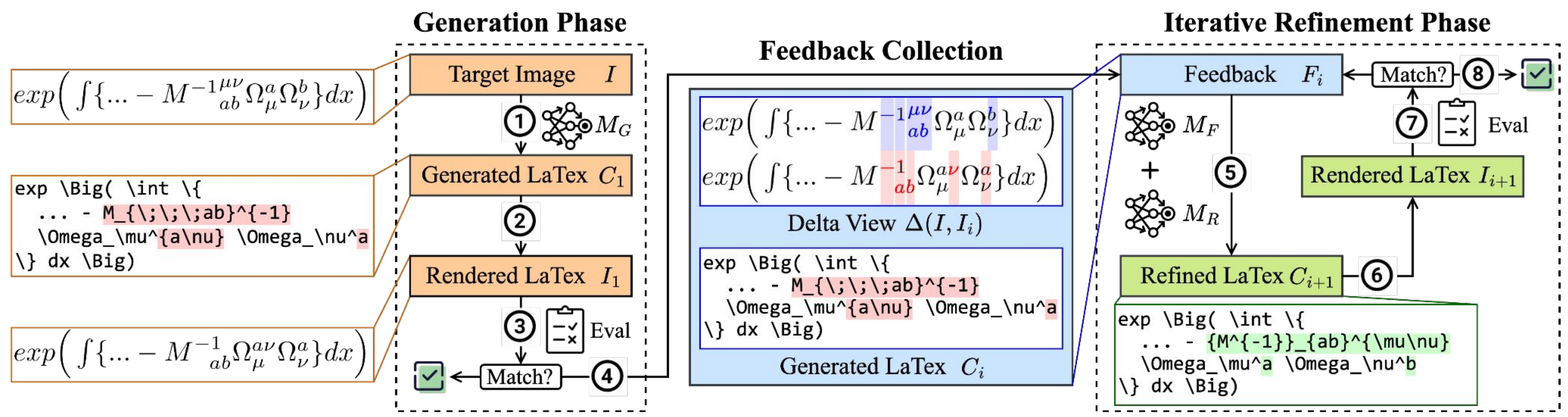
LATTE: Improving LaTeX Recognition for Tables and Formulae with Iterative Refinement

Nan Jiang, Shanchao Liang, Chengxiao Wang, Jiannan Wang, Lin Tan

1. Overview: Generation, Feedback Collection, and Iterative Refinement

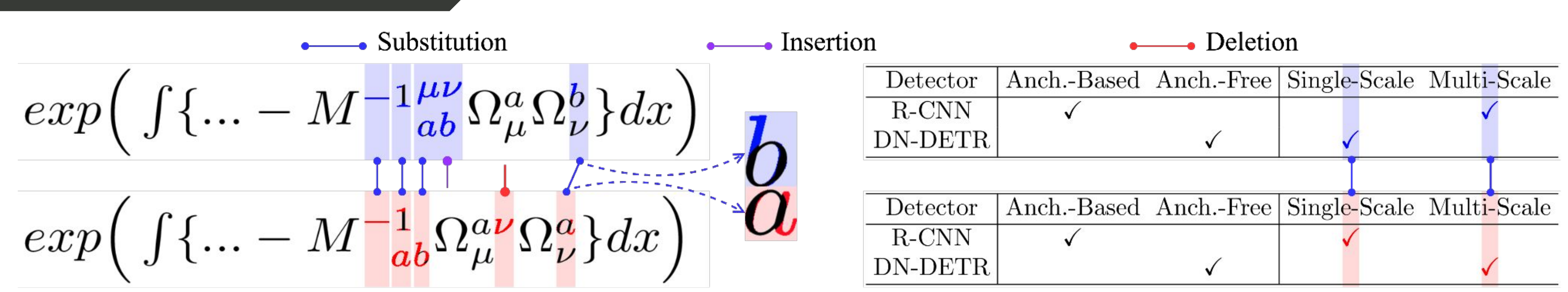


This work propose **LATTE**, a framework to improve LaTex code recognition for formulae and tables.

- LATTE applies **iterative refinement** to enable models revise the incorrect LaTex recognition based on the rendering feedback, similar to the self-debugging in code generation.
- LATTE uses a novel algorithm, **ImageEdit**, to provide visual feedback, **delta-view**, to highlight the visual difference between the rendered LaTex images and the expected ground-truth images.
- With the iterative refinement framework and delta-view feedback, LATTE outperforms existing techniques on formulae recognition accuracy by **7.03%** and table recognition accuracy by **45.28%**.

2. Approach: Delta-View, Fault Localization, and Refinement

Delta-View Feedback

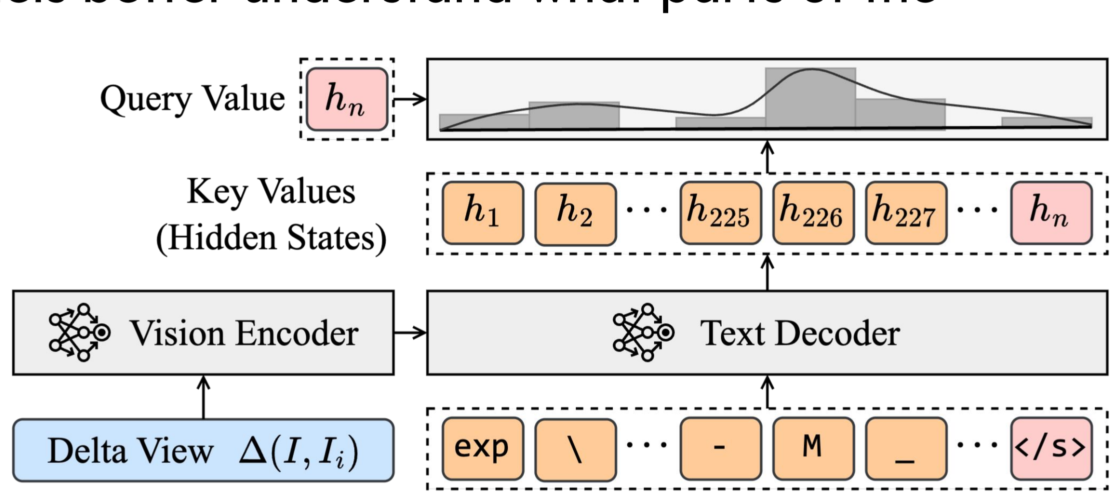


(a) Example of Formula's Delta-View (b) Example of Table's Delta-View

- Delta-view concatenates the expected ground-truth image (top) rendered image from the recognized LaTex code (bottom) , and highlights their differences using a column-wised Wagner–Fischer algorithm.
- Delta-view enables the fault localization and refinement models better understand what parts of the recognized LaTex code need to be fixed.

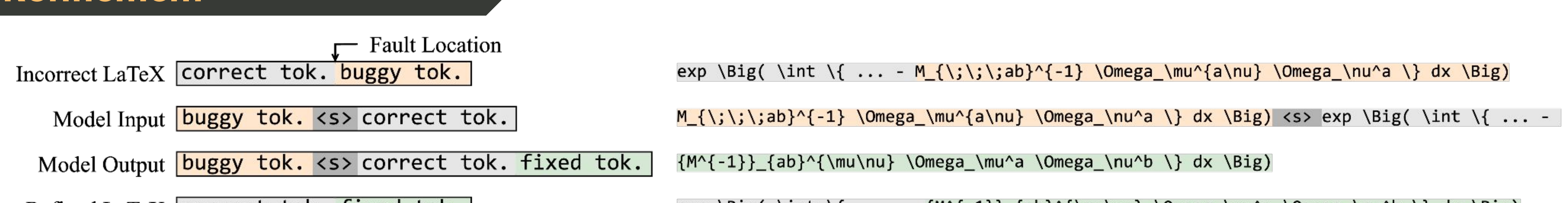
Fault Localization

- The fault localization model includes a vision encoder-decoder model and an attention layer on top of the text decoder. It takes the delta-view and the incorrect LaTex code as input, and predicts the first erroneous token in the incorrect LaTex code.



Fault Localization Model's Architecture

Refinement



Input and Output Format of the Refinement Model

3. Evaluation

Compared With SOTA

LATTE1 refers the initial LaTeX code generation. LATTE2 refers one-round of refinement on the result of LATTE1.

- Tables 1 and 2 show that: LATTE2 achieves **7.03–45.28%** in formulae and tables recognition compared with existing approaches and MLLMs with one round of refinement.
- Table 3 shows that: LATTE2 achieves **56–67%** higher accuracy compared with commercial MLLMs and software.
- Table 3 also shows that: LATTE2 has **stronger refinement ability** by fixing more incorrect LaTex code.

Tools	Match	CW-SSIM	BLEU	Edit	Time
WYGIWYS	77.46	-	87.73	87.60	-
DA	79.81	-	88.42	88.75	-
EDPA	82.07	-	92.31	91.39	-
WAP	82.08	-	88.21	89.58	-
MI2LaTeX	82.33	-	90.28	91.90	-
ConvMath	83.41	-	88.33	90.80	-
Vary-1.8B	11.91	0.7895	69.46	63.47	2.27s
Llava-7B	13.54	0.7548	75.40	64.61	2.29s
LATTE1	82.27	0.9462	92.91	93.11	**0.87s**
LATTE2	**90.44**	**0.9844**	**93.25**	**97.69**	1.53s

Table 1: Evaluation on Formulae Recognition.

Tools	Match	CW-SSIM	BLEU	Edit	Time
Vary-1.8B	6.92	0.6253	62.89	30.50	7.13s
Llava-7B	13.90	0.7278	64.19	39.84	6.13s
LATTE1	45.20	0.8128	79.06	73.82	**2.24s**
LATTE2	**59.18**	**0.8221**	**83.81**	**77.51**	5.34s

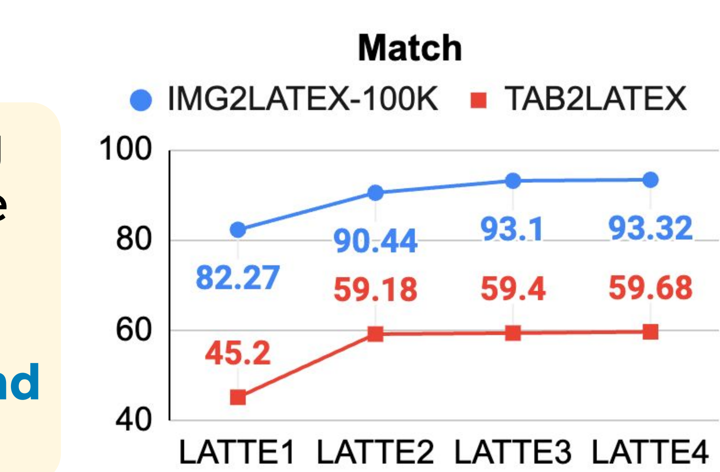
Table 2: Evaluation on Tables Recognition.

Tools	IMG2LATEX-100K				TAB2LATEX			
	Match	CW-SSIM	BLEU	Edit	Match	CW-SSIM	BLEU	Edit
GPT-4V1	3.00	0.7480	52.77	61.25	2.00	0.5189	49.56	8.98
GPT-4V2	7.00	0.7212	50.87	59.46	2.00	0.5059	44.22	5.64
Gemini1	19.00	0.6485	21.47	63.60	0.00	0.3482	35.19	0.94
Gemini2	19.00	0.6191	25.78	61.58	0.00	0.3911	37.58	1.27
Mathpix	20.00	0.8684	20.71	84.44	11.00	0.6749	49.45	28.31
LATTE1	77.00	**0.9878**	92.45	**97.68**	40.00	0.8659	77.53	67.49
LATTE2	**87.00**	0.9778	**93.72**	96.92	**67.00**	**0.8723**	**83.82**	**77.36**

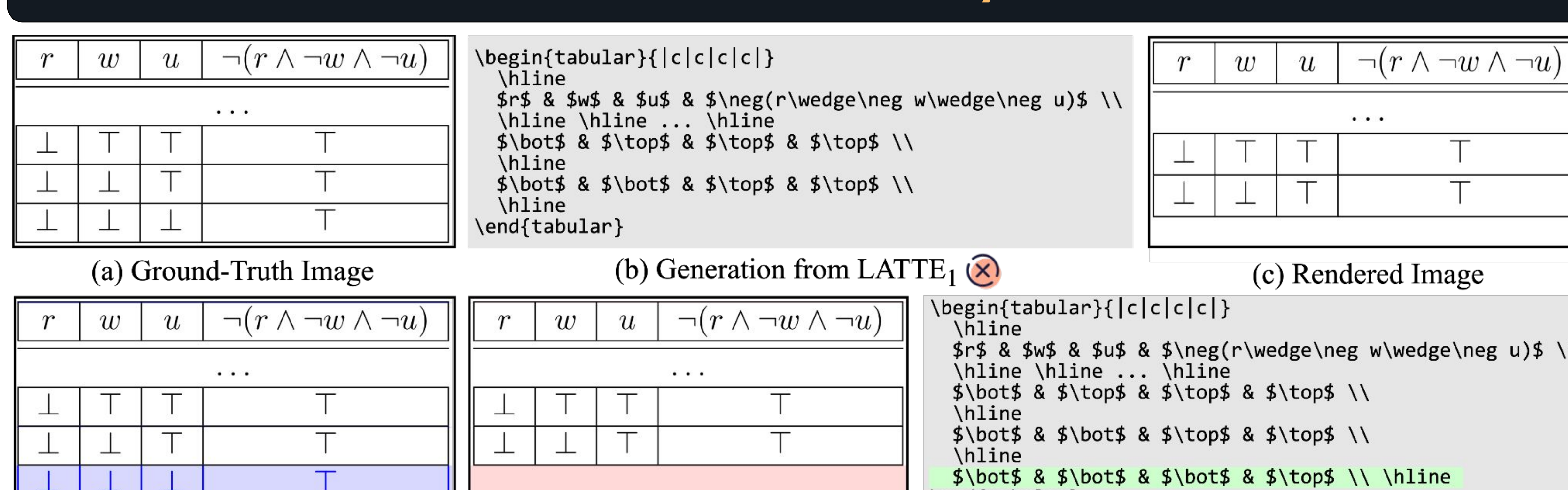
Table 3: Comparison with Commercial Tools on 100 Formula and Table Samples .

Iterative Refinement Ability

- When LATTE refines multiple rounds, the accuracy (Match) keeps increasing for both formula and table recognition. With three rounds of refinement, the accuracy of formula recognition increases from **82.27% to 93.32%**, and the accuracy of table recognition increases from **45.20% to 59.68%**.
- Overall, LATTE shows strong iterative refinement ability, while the **first round of refinement brings the most improvement**.



4. Case Study



(a) Ground-Truth Image (b) Generation from LATTE$_1$ (c) Rendered Image

(d) Delta-View: Highlighted Ground-Truth (left) and Rendered (right) Images (e) Refinement from LATTE$_2$