VDE Bench: Evaluating The Capability of Image Editing Models to Modify Visual Documents

Anonymous ACL submission

Abstract

In recent years, multimodal image editing models have achieved substantial progress, enabling users to manipulate visual content through natural language in a flexible and interactive manner. Nevertheless, an important yet insufficiently explored research direction remains visual document image editing, which involves modifying textual content within images while faithfully preserving the original text style and background context. Existing approaches, including AnyText, GlyphControl, and TextCtrl, predominantly focus on English-language scenarios and documents with relatively sparse textual layouts, thereby failing to adequately address dense, structurally complex documents or non-Latin scripts such as Chinese. To bridge this gap, we propose **V**isual **D**oc **E**dit Bench(VDE Bench), a rigorously human-annotated and evaluated benchmark specifically designed to assess image editing models on multilingual and complex visual document editing tasks. The benchmark comprises a high-quality dataset encompassing densely textual documents in both English and Chinese, including academic papers, posters, presentation slides, examination materials, and newspapers. Furthermore, we introduce a decoupled evaluation framework that systematically quantifies editing performance at the OCR parsing level, enabling fine-grained assessment of text modification accuracy. Based on this benchmark, we conduct a comprehensive evaluation of representative state-of-the-art image editing models. Manual verification demonstrates a strong consistency between human judgments and automated evaluation metrics. VDE Bench constitutes the first systematic benchmark for evaluating image editing models on multilingual and densely textual visual documents.

1 Introduction

In recent years, the capabilities of multimodal image editing models have continuously advanced (Wu et al., 2025; Xu et al., 2025; Gao et al., 2025; Cao et al., 2025; Zhang et al., 2023b; Rombach et al., 2022). Image editing models allow users to iteratively manipulate image content through natural language, and this simple and intuitive editing paradigm greatly enhances the flexibility and interactivity of visual creation, quickly becoming a core tool in the design field.

However, one important category is often overlooked in the field of image editing: **visual document image editing**. This task involves modifying text content on input images while preserving the style of both text and background. Although some works have explored text-to-visual document generation or the modification of text in visual documents, such as AnyText (Tuo et al., 2024), GlyphControl (Yang et al., 2024), and TextCtrl (Zeng et al., 2024), these studies have notable limitations. First, the vast majority of research focuses on English text modification, whereas the real challenge in this field lies in modifying non-Latin scripts such as Chinese (Tuo et al., 2024; Eli et al., 2025). Second, current studies mainly address visual document editing in scenarios with a small amount of text, such as posters, while neglecting more complex documents with dense text, such as papers and exams, which are the most challenging to edit. Although some benchmarks exist to evaluate the ability of image editing models to modify visual documents (Gui et al., 2025; Fu et al., 2025; Wang et al., 2025), the lack of research on multilingual and dense-text visual document editing means that there is currently no comprehensive benchmark to assess image editing models' performance on complex visual documents. To systematically evaluate the usability of different image editing models on multilingual and densely textual documents, it is necessary to construct a challenging benchmark addressing these problem.

To tackle these issues, this paper proposes **VDE Bench**, a rigorously human-evaluated benchmark

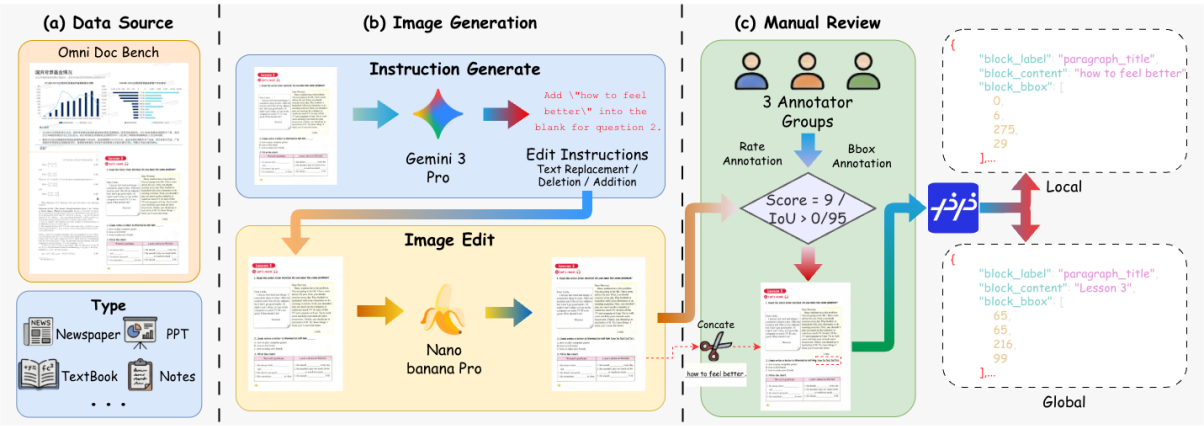


Figure 1: Overview of the VDE Bench pipeline, including document data sourcing, instruction generation and image editing, followed by rigorous manual review and OCR-based global and local evaluation.

designed to diagnose the actual performance of image editing models on multilingual and complex visual document editing tasks. The main contributions of this paper are as follows:

1. Multilingual complex text document dataset: We constructed high-quality evaluation sets in English and Chinese, covering various densely textual document types such as paper, poster, PPT, exam, and newspaper. The dataset construction involved extensive human participation and rigorous manual curation to ensure quality and reliability.

2. Decoupled and robust evaluation methodology: We propose a decoupled diagnostic evaluation framework to comprehensively assess the ability of image editing models to edit textual content in complex visual documents. The evaluation is conducted at the OCR parsing level, enabling a more comprehensive and fine-grained analysis of the models' document editing capabilities.

3. Systematic diagnostic analysis of mainstream image editing models: Based on the proposed benchmark, we systematically evaluate a variety of mainstream image editing models, revealing their real-world performance. Furthermore, the evaluation results were manually verified, showing high consistency with automated metrics, further demonstrating the effectiveness and reliability of VDE Bench.

2 Related Work

Image Editing Models. Qwen-Image-edit (Wu et al., 2025) has been specifically optimized for visual document editing and demonstrates particularly strong performance in modifying Chinese visual documents. The Nano Banana (LLC, 2025) series of models have attracted considerable attention due to their robust capabilities in both general image and visual document generation and editing. More recently, Longcat-Image-edit (Team et al., 2025) was introduced as an image editing model with optimizations targeting Chinese image editing tasks. Additionally, models such as Step1X (Liu et al., 2025) and Instruct-pix2pix (Brooks et al., 2023) serve as representative examples of widely adopted general-purpose image editing frameworks.

Image Editing Benchmarks. The vast majority of existing image editing benchmarks focus on entity-level modifications (Sheynin et al., 2024; Basu et al., 2023; Zhou et al., 2025). For example, I2EBench (Ma et al., 2024) explicitly distinguishes *high-level* and *low-level* editing tasks through a hierarchical design; CompBench (Jia et al., 2025) supports multi-turn editing; and ImgEdit-Bench (Ye et al., 2025) emphasizes the evaluation of content memory, content understanding, and version rollback capabilities. While these benchmarks cover a wide range of testing scenarios, none of them consider densely textual images such as visual documents.

Visual Text Generation and Modification Benchmarks. AnyText-Bench (Tuo et al., 2024) is the first widely recognized large-scale visual document editing benchmark, primarily focusing on text editing on regular images. CVTG-2K (Du et al., 2025) is a recent English visual document generation benchmark, mainly concentrating on the generation of long-text visual documents. Qwen-Image-edit also recently proposed a Chinese visual document generation benchmark called Chinese-Word (Wu et al., 2025). However, all these existing benchmarks overlook testing with multi-lingual

and complex text documents.

3 VDE Bench

Constructing VDE Bench requires extensive manual annotation and processing. To reduce annotation time, we utilized **Nano Banana Pro** for assistance. Overall, our processing pipeline is illustrated in the overview framework in Figure 1, which consists of three main stages: data collection, groud truth generation, and manual review.The main comparisons between our VDE Bench and existing open-source image editing benchmarks in the community are summarized in Table 1.

3.1 Data Collection Stage

To collect high-quality multilingual complex text document data, we sourced English and Chinese complex text documents from **Omni Doc Bench** (Ouyang et al., 2024). Omni Doc Bench is a high-quality OCR benchmark with extensive manually annotated complex text documents. Therefore, we directly adopt this benchmark as our seed data.

3.2 Image Generation Stage

In the Image generation stage, modification instructions mainly encompass three types: text replacement, text addition, and text deletion. To accelerate data generation, we employed **Nano Banana Pro** and **Gemini3-Pro** for assistance. Specifically, we first used Gemini3-Pro to generate modification instructions for 1,355 document page images from Omni Doc Bench, with three randomly generated instructions per image covering the aforementioned types. This process yielded a total of 4,065 modification instructions.

Next, the original images along with their corresponding modification instructions were input to Nano Banana Pro to produce the edited images. Due to some modification instructions triggering safety checks, the final output consisted of 3,989 successfully edited images.

3.3 Manual Review Stage

After image generation, we conducted a rigorous manual review stage. The review workflow differs for table modifications and text modifications.

Human Annotation Rules Alignment: Our annotators include 15 master's and PhD students in computer science. Before annotation, we trained them to understand the annotation rules and establish unified standards. The annotation process includes the following metrics:

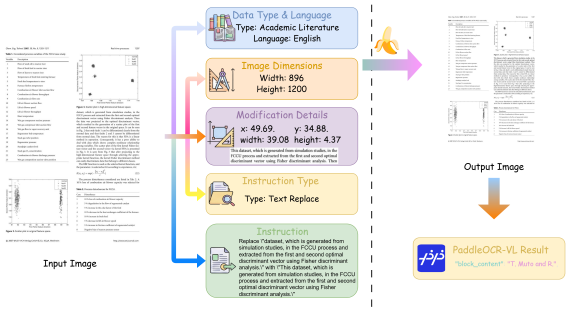


Figure 2: An example data sample from VDE Bench.

- **Instruction Compliance:** Whether the modified image accurately follows the modification instructions.
- **Modification Authenticity:** Whether the visual document appears realistic after modification, e.g., no artifacts in the text.
- **Subjective Satisfaction:** Annotators provide subjective scores, allowing them to handle issues such as unclear images or incorrect aspect ratios.

Each metric is scored from 1 to 3, with 1 being the worst and 3 being the best.

Text Modification Annotation: Since text modification often involves full-image edit and dense background text, the annotation workflow is somewhat more complex:

(1) To obtain the groud truth for text-modified images, annotators first segment the foreground and background. Annotators are divided into three groups and share a common data pool. They initially score the text modification regions, filtering out high-quality images with a total score of 9.

(2) For the high-quality filtered data, annotators draw bounding boxes around the modified text regions. If a piece of data is annotated by all three groups and the average pairwise IOU of the boxes exceeds 0.95, it is considered approved; otherwise, the data is re-initialized and returned to the pool. This process continues until all data are approved.

(3) Once the bounding boxes are completed, the largest box in each annotation is selected as the groud truth. The modified image region within the box is extracted and pasted back onto the original image, producing a new image that

Table 1: The advantages of Visual Doc Bench compared with other image editing benchmarks.

Benchmark	Human Verified	Mask	OCR Based	Text Edited
I2EBench	✓	✓	✗	✗
EditBench (Wang et al., 2023)	✓	✓	✗	✗
EditVal (Basu et al., 2023)	✓	✗	✗	✗
EmuEdit (Sheynin et al., 2024)	✓	✗	✗	✗
AnyEdit (Yu et al., 2025)	✗	✓	✗	✗
CompBench	✓	✓	✗	✗
MagicBrush (Zhang et al., 2023a)	✓	✗	✗	✗
ImgEdit-Bench	✓	✗	✗	✗
MuCIE (Zhou et al., 2025)	✗	✗	✗	✗
AnyText	✗	✗	✗	✓
VDE Bench (Ours)	✓	✓	✓	✓

preserves the original background while incorporating the modifications.

(4) Annotators perform a final screening of the generated images. This step is entirely subjective. An image is considered approved only if all groups agree; otherwise, it is rejected.

(5) After completing the image-level annotations, we further apply PaddleOCR-VL (Cui et al., 2025a) to perform OCR recognition on both the entire generated image and the modified regions. During this step, manual verification of the OCR results in the modified regions is also required. If the OCR results for the modified regions are correct, the data sample is approved.

Upon completion of the aforementioned annotation process, we obtained a dataset consisting of 674 instruction-modified images, along with cropped regions corresponding to the modified areas. In addition, the dataset includes structured JSON annotation files generated using PaddleOCR-VL, containing both global annotation information for the entire image and localized annotation information for the modified regions.

3.4 Statistical Analysis

For the final set of 674 generated samples, we conducted a detailed statistical analysis, as shown in Figure 3. With respect to document type, the samples are categorized into nine distinct classes, which are approximately evenly distributed across the dataset. In terms of language, the samples encompass three categories: Chinese, English, and mixed Chinese-English. The distribution of these language categories is also largely balanced, indicating that the dataset provides good representativeness in both language and document type, and thus

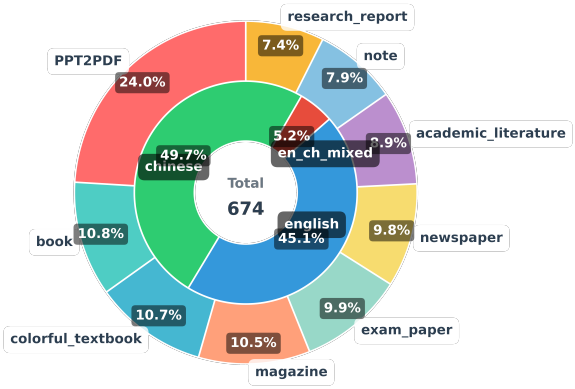


Figure 3: Overview of the data distribution.

offers a reliable basis for subsequent evaluation and analysis.

4 Evaluation Method and Metrics

To systematically evaluate the performance of different models in modifying complex text documents, we introduce the following evaluation method and metrics. These metrics consider both spatial localization accuracy and textual content correctness, providing a comprehensive reflection of model performance in real-world document editing scenarios.

4.1 Spatial Matching Algorithm

4.1.1 Matching Strategy

We adopt a greedy distance-based matching algorithm to establish one-to-one correspondences between GT (Ground Truth) blocks and predicted blocks. Let the center of groud truth block i be (c_x^{gt}, c_y^{gt}) and the center of predicted block j be (c_x^{pred}, c_y^{pred}). The Euclidean distance is defined as:

$$d(i,j) = \sqrt{(c_x^{gt} - c_x^{pred})^2 + (c_y^{gt} - c_y^{pred})^2}. \quad (1)$$

The rationale for using a distance-based greedy matching strategy is that complex text documents often contain multiple text or table blocks with irregular spatial distribution. By computing the Euclidean distance between block centers, we can efficiently establish the most reasonable correspondences, ensuring that each groud truth block is paired with the nearest predicted block. This provides a reliable basis for subsequent spatial and textual metric calculations.

4.1.2 Matching Result Categories

Matched Pairs are groud truth and predicted block pairs that were successfully matched; **Unmatched groud truth Blocks** are groud truth blocks with no corresponding prediction; and **Unmatched Pred Blocks** are predicted blocks with no corresponding groud truth block. In this work, text-related metrics are computed exclusively on successfully matched text blocks, while IoU is evaluated over all three categories: matched pairs, unmatched groud truth blocks, and unmatched predicted blocks.

4.2 Spatial Localization Metrics

4.2.1 Intersection over Union

IoU measures the spatial overlap between two bounding boxes $B_a = (x_1^a, y_1^a, x_2^a, y_2^a)$ and $B_b = (x_1^b, y_1^b, x_2^b, y_2^b)$:

$$\text{IoU}(B_a, B_b) = \frac{\text{Area}(B_a \cap B_b)}{\text{Area}(B_a \cup B_b)}. \tag{2}$$

When computing the overall IoU, all three matching scenarios need to be considered. The formula is therefore expressed as follows:

$$\text{IoU}_{\text{all}} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}(B_i^{\text{GT}}, B_i^{\text{pred}}), \tag{3}$$

where N denotes the number of groud truth bounding boxes, and B_i^{GT} and B_i^{pred} represent the i-th groud truth bounding box and the i-th bounding box detected in the image generated by the model, respectively.

IoU is chosen as the spatial localization metric because it directly quantifies the overlap between predicted and groud truth blocks in terms of position and size. Reporting both matched-pair and all-block mean IoU allows evaluation of precision on successfully matched blocks as well as overall spatial prediction performance, including missed and false detections.

4.3 Text Content Metrics

After completing the matching of bounding boxes, in addition to computing spatial performance metrics, it is more important to evaluate the correctness of the text modification content.These metrics are computed only for matched block pairs to ensure that content evaluation is based on correctly localized blocks.

4.3.1 Character Distance Metric

The Character Distance Metric (CDM) is defined based on the Levenshtein distance d_{lev} (Levenshtein, 1965), which measures the minimum number of character-level edit operations (insertions, deletions, and substitutions) required to transform one string into another:

$$\text{CDM}(c, r) = 1 - \frac{d_{\text{lev}}(c, r)}{\max(|c|, |r|)}. \tag{4}$$

By normalizing the edit distance with respect to the maximum length of the compared strings, CDM yields a similarity score bounded between 0 and 1, where higher values indicate greater character-level consistency between the generated text c and the reference text r. This formulation enables fair comparison across text blocks of varying lengths.

CDM provides a fine-grained evaluation of textual differences at the character level, making it highly sensitive to minor modification errors, typographical mistakes, or symbol mismatches that may not be adequately captured by word based metrics.

4.3.2 BLEU-4

BLEU-4 (Papineni et al., 2002) is computed using modified n-gram precision for $n = 1$ to 4, combined with additive smoothing and a brevity penalty (BP) to account for length discrepancies between the generated and reference texts:

$$\text{BLEU-4} = \text{BP} \cdot \exp\left(\frac{1}{4} \sum_{n=1}^{4} \log P_n\right). \tag{5}$$

BLEU-4 evaluates the degree of consistency between the generated text and the reference text at the n-gram level, thereby capturing local word order and phrase-level correctness. By combining BLEU-4 and CDM, the model's text editing capability can be evaluated both at the overall level and at the character-level.

Table 2: Performance of image editing models on full-image settings with average scores. AVG represents the mean of the four metrics for each language. Bold indicates the best value in each column.

Model	IoU		CDM		TEDS		BLEU		AVG	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
16B+ Models										
Step1x	**0.721**	0.808	0.817	0.829	0.757	0.696	0.546	0.230	**0.710**	**0.641**
Qwen	0.593	0.630	0.828	**0.868**	0.767	**0.743**	**0.568**	**0.246**	0.689	0.622
6B-16B Models										
ICEdit	0.021	0.032	0.184	0.153	0.106	0.087	0.100	0.070	0.103	0.086
3B-6B Models										
Longcat	0.688	**0.823**	0.737	0.801	0.648	0.600	0.472	0.217	0.636	0.610
Pix2pix	0.584	0.506	0.273	0.123	0.161	0.088	0.147	0.085	0.291	0.201
Closed-source Models										
Gemini	0.692	0.593	**0.852**	0.797	**0.773**	0.621	0.316	0.218	0.658	0.558

4.3.3 TEDS-like Similarity

Token-level Levenshtein distance is calculated between token sequences t_c and t_r:

$$\text{TEDS-like}(c, r) = 1 - \frac{d_{\text{lev}}^{\text{token}}(t_c, t_r)}{\max(|t_c|, |t_r|)}. \quad (6)$$

The TEDS-like metric (Zhong et al., 2020) combines token-level text similarity and edit distance to better reflect structural and sequential fidelity. In complex documents, it captures cross-word or cross-line editing errors, focusing more on textual logic and semantic completeness than purely character-level metrics.

5 Benchmarks

Based on the aforementioned evaluation metrics, we decompose the overall evaluation framework into two complementary components to ensure the systematicity and reliability of the evaluation process.

(1) Pipeline Tools. We adopt PaddleOCR (Cui et al., 2025b) as the core evaluation tool to perform OCR-based text extraction on images generated by image editing models. The extracted textual content is then quantitatively compared with the corresponding ground-truth annotations at both the global image level and the local edited-region level. This pipeline enables an objective assessment of textual correctness and consistency, forming the foundation of the automated evaluation process.

(2) Image Editing Models. We evaluate a series of mainstream image editing models, including both open-source and closed-source approaches, on the Visual Document Editing Benchmark. By conducting comparative analyses under a unified evaluation pipeline and consistent metric set, we are able to systematically reveal the strengths and limitations of different models in handling multilingual and densely textual visual document editing tasks.

Overall, the proposed evaluation framework can be summarized by the procedural steps illustrated in Figure 4.

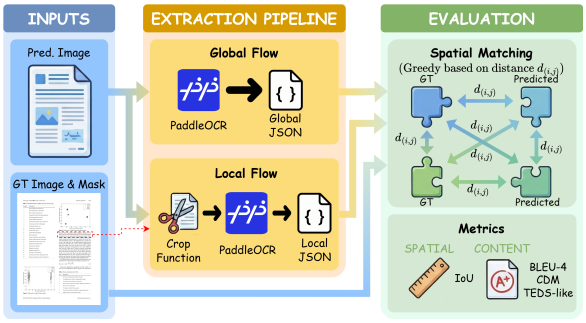


Figure 4: Overview of the evaluation pipeline. The model-generated images are cropped according to the edited region boxes provided in the groud truth data to obtain the local regions. OCR recognition is then performed on both the global and local regions using PaddleOCR-VL, and the discrepancies between the OCR results and the groud truth are subsequently calculated.

We primarily evaluate both open-source and closed-source image editing models that support Chinese and English image editing, including Step1X (Liu et al., 2025), Qwen-Image-Edit (Wu et al., 2025), ICEdit (Zhang et al., 2025), Longcat-Image-Edit (Team et al., 2025), Instruct Pix2Pix

Table 3: Performance of image editing models on w/ Crop setting with average scores. AVG is the mean of the four metrics for each language. Bold indicates the best value in each column.

Model	IoU		CDM		TEDS		BLEU		AVG	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
16B+ Models										
Step1x	0.671	0.725	0.528	0.480	0.348	0.178	0.163	0.039	0.428	0.356
Qwen	**0.693**	0.725	**0.731**	0.683	**0.647**	0.411	**0.352**	0.036	**0.606**	0.464
6B-16B Models										
ICEdit	0.135	0.162	0.329	0.292	0.232	0.036	0.139	0.016	0.209	0.127
3B-6B Models										
Longcat	0.667	**0.734**	0.681	**0.707**	0.586	**0.424**	0.321	0.038	0.564	**0.476**
Pix2pix	0.534	0.563	0.272	0.096	0.136	0.013	0.117	0.013	0.265	0.171
Closed-source Models										
Gemini	0.644	0.608	0.705	0.476	0.600	0.175	0.322	**0.040**	0.568	0.325

(Brooks et al., 2023), and Gemini 2.5-Flash-Image. Notably, Nano Banana corresponds to Gemini 2.5-Flash-Image; in the remainder of this paper, we refer to these models using their abbreviated names.

5.1 Overall Performance

We conducted an evaluation of the capabilities of several representative image editing models, including Qwen-Image-Edit, LongCat-Image-Edit, Step1x, Instruct Pix2Pix, and ICEdit. The evaluation was carried out from both Chinese and English perspectives to assess the models' ability to perform image editing under instructions in different languages. In addition, we evaluated these models from global and local perspectives, focusing on their ability to preserve the overall image content and to accurately modify the targeted regions as specified by the editing instructions.

As shown in Table 2, for English text, Step1X, Qwen-Image-Edit, and LongCat-Image-Edit all demonstrate strong capabilities in preserving global textual content. Although Qwen-Image-Edit achieves high scores in text editing performance, it performs poorly in maintaining the overall page layout, resulting in a relatively low IoU score. In contrast, models such as Instruct Pix2Pix and ICEdit perform significantly worse than more recent open-source models, which indirectly indicates that the ability of image editing models to handle dense text is steadily improving.

On the other hand, all models exhibit noticeable shortcomings in Chinese text editing. Compared with their performance on English text, the evaluation metrics for Chinese show a consistent decline across all models.

Evaluating image editing models solely based on metrics over the entire image is insufficient, because some models have poor instruction-following ability but exhibit strong background preservation. This can lead to artificially inflated performance scores. Therefore, it is necessary to evaluate metrics specifically on the edited regions only.

As shown in Table 3, the table reports the evaluation results of different models on the locally modified regions. It can be observed that Qwen-Image-Edit achieves the best overall performance, while Step1X—despite performing best on global metrics—shows a significant performance drop under local-region evaluation. This indicates that Step1X has relatively weak instruction-following capability and cannot adequately meet the requirements for editing dense text documents.

5.2 Performance By Instruction Type

Different types of modification instructions inherently pose varying levels of difficulty for image editing models. For instance, text deletion instructions typically require only the removal of specific text segments, whereas text replacement instructions involve modifying limited portions of text. Both of these instruction types generally induce minor changes to the overall textual structure, thus representing moderate challenges for the models. In contrast, text addition instructions are considerably more complex, as they often impact the overall text layout. For example, inserting several words into the middle of a sentence can expand it from a

single line to multiple lines, potentially affecting the alignment and formatting of surrounding content. Existing image editing models demonstrate pronounced difficulties in handling such modifications, frequently producing low-quality results. Table 4 summarizes the evaluation outcomes of various models for these three instruction types. To isolate the effect of layout variations, the metrics reported here are restricted to the modified regions only. It is evident that performance metrics for text addition are significantly lower than those for text deletion and text replacement, indicating that text addition represents a more challenging scenario and remains a critical bottleneck in current visual document image editing systems.

Table 4: Evaluation results of image editing models across different instruction types, measured using IOU, BLEU-4, CDM, and TEDS-like metrics. The evaluated models are categorized into open-source and closed-source groups for systematic comparison.

Model	Type	IOU	BLEU	CDM	TEDS
16B+ Models					
Qwen	replace	0.815	0.188	0.794	0.586
	delete	0.798	0.207	0.574	0.532
	add	0.631	0.148	0.457	0.256
Step1X	replace	0.679	0.093	0.572	0.269
	delete	0.522	0.119	0.345	0.299
	add	0.499	0.095	0.326	0.176
6B-16B Models					
ICEdit	replace	0.678	0.093	0.572	0.269
	delete	0.522	0.119	0.345	0.299
	add	0.499	0.095	0.236	0.176
3B-6B Models					
LongCat	replace	0.755	0.172	0.798	0.584
	delete	0.666	0.169	0.421	0.326
	add	0.553	0.144	0.495	0.284
Pix2pix	replace	0.622	0.064	0.194	0.067
	delete	0.347	0.025	0.086	0.065
	add	0.442	0.075	0.171	0.082
Closed-source Models					
Gemini	replace	0.562	0.180	0.639	0.404
	delete	0.547	0.148	0.504	0.422
	add	0.405	0.160	0.448	0.250

5.3 Human Eval

To ensure the validity and reliability of VDE Bench, we incorporate a dedicated human evaluation stage. In this stage, human annotators carefully assess the quality and accuracy of the edited visual documents. The results of this manual evaluation are then systematically compared with the outcomes obtained from automated evaluation metrics. By quantifying the discrepancies between human judgments and algorithmic measurements, we are able to verify the fidelity of the benchmark, assess the consistency of automated metrics, and demonstrate that VDE Bench provides a trustworthy and rigorous framework for evaluating image editing models on multilingual and densely textual documents.

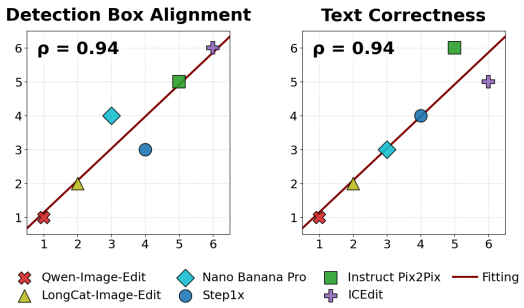


Figure 5: Correlation between human rankings and automated rankings.The horizontal axis represents the human ranking results, and the vertical axis represents the automated ranking results.

Specifically, we randomly sampled 20 instances from VDE Bench and collected the corresponding outputs generated by each image editing model. Human annotators then ranked the models according to two criteria: detection box alignment and text correctness, assigning scores from 6 for the highest-ranked model down to 1 for the lowest. Bounding box alignment corresponds to the IoU metric, while text correctness is quantified as the mean of BLEU-4, CDM, and TEDS-like metrics. The averaged rankings across all annotators were then compared with the rankings derived from automated evaluation metrics. As shown in Figure 5, the human annotation results exhibit a strong correlation with the automated rankings, validating the reliability and effectiveness of the automated evaluation protocol.

6 Conclusion

This paper introduces VDE Bench, a benchmark designed to systematically evaluate the direct editing capabilities of image editing models on complex text-based documents. It assesses model performance in text recognition, text modification, format preservation, and layout consistency, providing an accurate reflection of real-world document editing performance. Based on VDE Bench, we evaluate multiple open-source and closed-source image editing models, offering insights into their strengths, limitations, and directions for future optimization.

Limitations

While VDE Bench provides a comprehensive evaluation framework for multilingual and densely textual visual documents, there are several limitations that should be noted. First, although our dataset covers English and Chinese, other languages and scripts, such as Arabic or Hindi, are not included, which limits the generalizability of the benchmark to truly multilingual scenarios. Second, our current evaluation primarily focuses on static document images and does not address dynamic or interactive document content, such as editable PDFs or slides with animations. Third, while human verification ensures high-quality ground truth, it is labor-intensive and may introduce subtle biases despite rigorous annotation protocols. Finally, the benchmark emphasizes text modification and table edits, but other document elements, such as figures, charts, or complex layouts, are less represented, which may limit the evaluation of models' holistic document editing capabilities. Future work could extend the benchmark to more languages, dynamic content, and richer document elements to address these limitations.

References

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. 2023. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Duojun Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, and 55 others. 2025. Hunyuanimage 3.0 technical report. *Preprint*, arXiv:2509.23951.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025a. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *Preprint*, arXiv:2510.14528.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025b. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. 2025. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *Preprint*, arXiv:2503.23461.

Elham Eli, Dong Wang, Wenting Xu, Hornisa Mamat, Alimjan Aysa, and Kurban Ubul. 2025. A comprehensive review of non-latin natural scene text detection and recognition techniques. *Engineering Applications of Artificial Intelligence*, 156:111107.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, and 12 others. 2025. Seedream 3.0 technical report. *Preprint*, arXiv:2504.11346.

Rui Gui, Yang Wan, Haochen Han, Dongxing Mao, Fangming Liu, Min Li, and Alex Jinpeng Wang. 2025. Texteditbench: Evaluating reasoning-aware text editing beyond rendering. *Preprint*, arXiv:2512.16270.

Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, and 1 others. 2025. Compbench: Benchmarking complex instruction-guided image editing. *arXiv preprint arXiv:2505.12200*.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, and 5 others. 2025. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Google LLC. 2025. Gemini 2.5 flash image (a.k.a. "nano banana")–image generation editing model. https://cloud.google.com/blog/products/ai-machine-learning/

gemini-2-5-flash-image/. Accessed: 2025-11-12.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. 2024. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems*, 37:41494–41516.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2024. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *Preprint*, arXiv:2412.07626.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2024. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. 2025. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2024. Anytext: Multilingual visual text generation and editing. *Preprint*, arXiv:2311.03054.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. 2025. Textatlas5m: A large-scale dataset for dense text image generation. *Preprint*, arXiv:2502.07870.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, and 1 others. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. 2025. Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. *Preprint*, arXiv:2507.16193.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. 2024. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. 2025. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2025. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26125–26135.

Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. 2024. Textctrl: Diffusion-based scene text editing with prior guidance control. *Preprint*, arXiv:2410.10133.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. *Preprint*, arXiv:2302.05543.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. 2025. In-context edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2504.20690.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. *Preprint*, arXiv:1911.10683.

Zijun Zhou, Yingying Deng, Xiangyu He, Weiming Dong, and Fan Tang. 2025. Multi-turn consistent image editing. *arXiv preprint arXiv:2505.04320*.

A Prompts

To balance diversity, consistency, and controllability in the design of modification instructions, we construct and adopt a systematic set of predefined instruction templates. This framework is centered on three representative and fundamental document editing operations—text deletion, text replacement, and text addition—thereby enabling comprehensive coverage of practical document modification scenarios. By constraining instruction generation through a template-based formulation, semantic ambiguity is effectively reduced, while structural consistency across samples and the reproducibility of experimental results are enhanced.

During the dataset construction stage, each document image is systematically assigned the complete set of modification instructions. This design ensures a balanced distribution of different editing operations within the dataset and allows the generated modification tasks to exhibit structured diversity in terms of edited content, target locations, and operation sequences. The prompts used in the instruction generation process are provided below.

Text Addition Instruction

```
Generate a text addition instruction for the input image. You are not allowed to modify any text inside tables or within images; only titles and body text may be modified.
1. Your response must contain only the editing instruction itself, with no additional content.
2. Your response must be plain text, without any Markdown formatting.
3. The instruction you provide must clearly specify which text in the image is to be added.
4. The language of your instruction must match the primary language used in the image. For example, if the main language in the image is Chinese, respond in Chinese; if it is English, respond in English.
5. Modify only one location.
```

Text Deletion Instruction

```
Generate a text deletion instruction for the input image. You are not allowed to modify any text inside tables or within images; only titles and body text may be modified.
1. Your response must contain only the editing instruction itself, with no additional content.
2. Your response must be plain text, without any Markdown formatting.
3. The instruction you provide must clearly specify which text in the image is to be deleted.
4. The language of your instruction must match the primary language used in the image. For example, if the main language in the image is Chinese, respond in Chinese; if it is English, respond in English.
5. Modify only one location.
```

Text Replacement Instruction

```
Generate a text modify instruction for the input image. You are not allowed to modify any text inside tables or within images; only titles and body text may be modified.
1. Your response must contain only the editing instruction itself, with no additional content.
2. Your response must be plain text, without any Markdown formatting.
3. The instruction you provide must clearly specify which text in the image is to be deleted.
4. The language of your instruction must match the primary language used in the image. For example, if the main language in the image is Chinese, respond in Chinese; if it is English, respond in English.
5. Modify only one location.
```

B Model Edit Example

This section provides a series of qualitative examples of image editing results generated by various models, illustrating their respective capabilities and differences in handling complex editing tasks. By examining these examples, we aim to offer a deeper

understanding of each model's strengths and limitations in practical image manipulation scenarios.

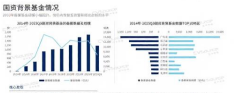
Example 1

Instruction:
将标题“国资背景基金情况”替换为“国资背景基金分析”

GT

Gemini

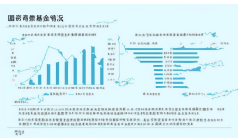
ICEdit



Qwen

Longcat

Pix2pix

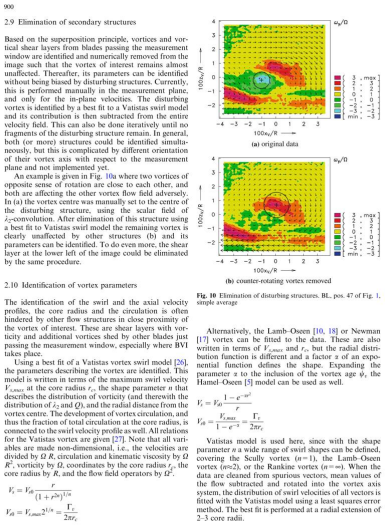


Step1x

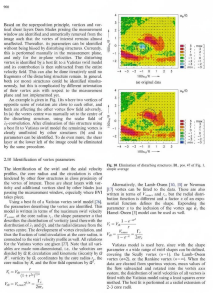
Example 2

Instruction:
Change "808" to "900"

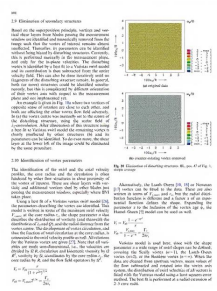
GT



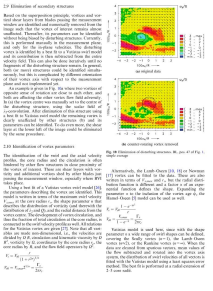
Gemini



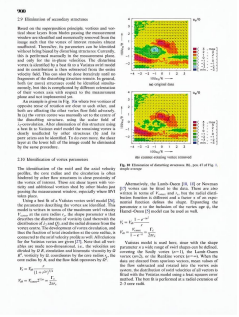
ICEdit



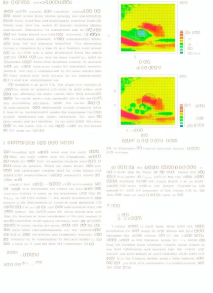
Qwen



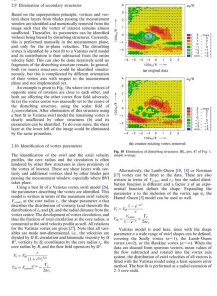
Longcat



Pix2pix



Step1x



C GT Example

This section presents several representative examples of the ground-truth data, providing insight into the characteristics and complexity of the dataset. These examples serve as a reference for evaluating model performance and understanding the types of tasks and scenarios encompassed within the data.

Example 1

Instruction:
Replace \"This is a chart of a car's fuel consumption during a certain day.\" with \"This chart illustrates a car's fuel consumption over a specific day.\"

Input

Output

Example 2

Instruction:
将标题“篇目跟踪练习”修改为“《劝学》篇目跟踪练习”。

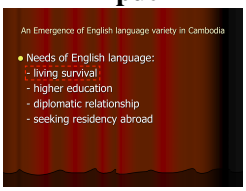
Input

Output

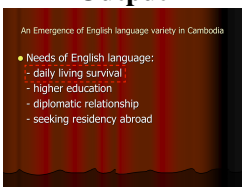
Example 3

Instruction:
Replace \"living survival\" with \"daily living survival\"

Input



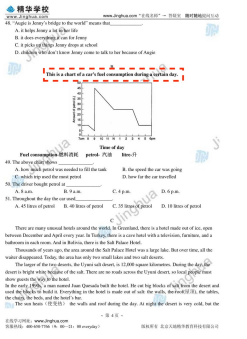
Output



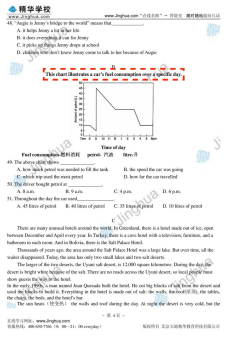
Example 4

Instruction:
将标题“江海横流显本色 人间正道是沧桑”替换为“新的时代新的征程 伟大复兴谱新篇”。

Input



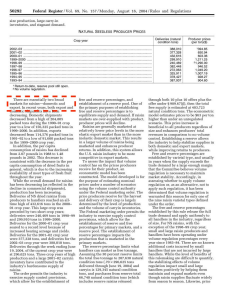
Output



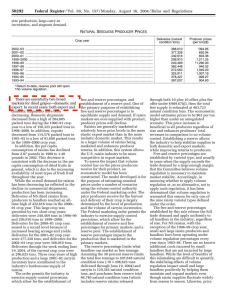
Example 5

Instruction:
Replace \"There are essentially two broad markets for raisins—domestic and export.\" with \"There are essentially two broad markets for dried grapes—domestic and export.\"

Input



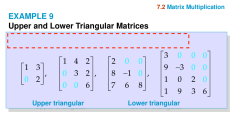
Output



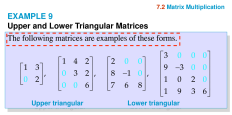
Example 6

Instruction:
Add the text 'The following matrices are examples of these forms.' after the title 'Upper and Lower Triangular Matrices'.

Input



Output



826

Example 8

Instruction:
Add \"The issue of academic dishonesty, particularly in scientific publications, has become a significant concern in the country's rapidly expanding research sector.\" after the heading \"Fake research is rampant in China. The government is trying to stop it\".

Input

Output

828

Example 7

Instruction:
Replace \"Beyond traditional corporate governance: What's the role of the board in fostering sustainability and innovation?\" with \"Beyond modern corporate governance: What's the role of the board in fostering sustainability and innovation?\"

Input

Output

827

Example 9

Instruction:
Delete the text: \"S. Korea court issues new arrest warrant for Yuon\" and the associated body text below it, starting with \"DELHI – An Indian court issued a fresh arrest\" and ending with \"supported by the opposition.\"

Input

Output

829

Example 10

Instruction:
Delete \"The 70th William Lowell Putnam Mathematical Competition\".

Input

Output