Detect Any AI-Counterfeited Text Image

Supplementary Material

Contents

1. More Details about the DanceText Dataset

1.1. Original Image

Our collected original text images can be roughly categorized as the following types: Photographed Contract, Concert Ticket, Web Page Screenshot, Scanned Invoice, Handwritten Diary, Bus Stop Sign, Academic Paper, Food Packaging, E-commerce App GUI, Utility Bill Notice, Scanned ID Card from 30+ countries, Sheet Music, Newspaper, Product Warranty Card, Purchase Order Form, Graffiti, Digital Menu, Class Schedule, Photographed Receipt, Insurance Policy, Travel Guide, Device Nameplate, Printed Exam Paper, Commemorative Silk Banner, Admission Letter, Statistical Chart, Scanned Passport from 30+ countries, User Manual, Mobile Capture of a Computer Screen, Comic, Postcard, Stock Market Data, Legal Document, Design Drawing, Boarding Pass, Class Notes, Magazine, Shopping History, Calligraphy Practice Sheet, Photographed ID Card from 30+ countries, Cheque, Movie Poster, Scanned Contract, Flowchart, Digital Exam Paper, Clothing Hang Tag, Map, Credit Card Statement, Source Code Snippet, Letter of Recommendation, Scanned Sales Slip, Business License, Presentation Slides, Medicine Instructions, Photographed Passport from 30+ countries, Photographed Phone Screen, Construction Blueprint, WeChat Chat History, Printed Menu, Supermarket Poster, Ingredients List, Ship Ticket, User Agreement, Bank Passbook, Proof of Payment, Express Bill, Scanned Bid Award Notice, Product Review, Dashboard, Calendar, Book Cover, XR Application Interface, Bus Ticket, Badge, Sesame Credit Certificate, Logistics Tracking Screenshot, License Plate, Greeting Card, Photographed Invoice, Award Certificate, Stock-in, Survey, Antique Book, Museum Exhibit Description, Leave of Absence Request, Data Analysis Report, Bid Award Letter, Calligraphy Work, Lottery Ticket, Email, Travel Itinerary, Certificate of Conformity, Patent Certificate, Answer Sheet, Price List Sign, Sticky Note, Movie Ticket, Delivery Note, Application Form, Schedule / Plan, Navigation Map, Transcript, License, Coupon, Blackboard Writing, Memo, Scene Text Signboards. Half of the image types above are occupied by cross-image-style test sets.

Ethics Statement: These original images were legally acquired from both open-source [2–4, 14, 22–25, 30] and proprietary internal datasets. Collection was performed with explicit consent from data owners, ensuring no sensitive personal information was disclosed.

1.2. Per-Generator Statistics

The per-generator breakdown of the DanceText dataset is detailed in Table 1. The number of images allocated to each generator is intentionally non-uniform, reflecting factors such as the generator's specific capabilities and the success rate of producing high-quality, plausible edits. For instance, the RSSTE [8] model is designed to modify large, clean English scene text. Consequently, it was applied exclusively to the scene text category, resulting in a smaller number of samples compared to more versatile models like Qwen-Image-Edit-2509.

1.3. Size and Ratio Statistics

Figure 1 presents a statistical overview of the DanceText dataset, showing the distributions of image dimensions (left) and the area ratio of counterfeited text (right). The diverse range of image shapes and the prevalence of small counterfeit regions are notable characteristics.

1.4. More Dataset Samples

Fully-generated fake image samples are provided in Figures 4 and 5. Regionally edited fake images are provided in Figures 6 and 7. Multi-lingual test images are provided in Figure 8. These images are in custom shapes. We resized

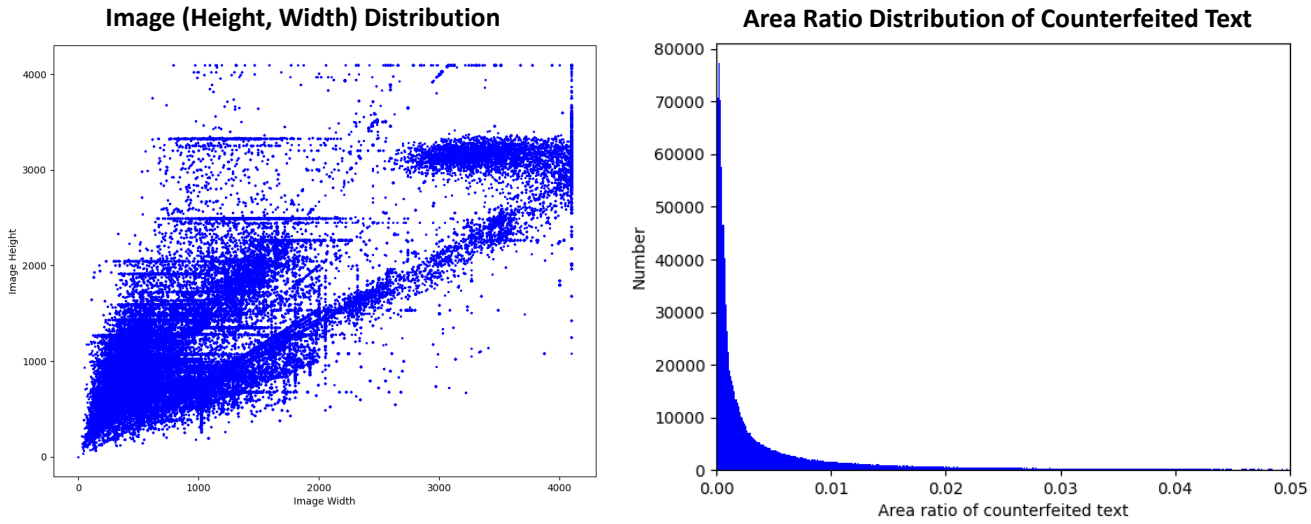

Figure 1. Distributions of image height-width (truncated at 4096 for better visualization) and area ratio of counterfeited texts.

them to squares for a better layout. These samples demonstrate the realism and diversity of our DanceText dataset.

2. More Details about the Creative Proposer Pipeline

2.1. Automated Data Cleaning Pipeline

To ensure the quality of our regionally edited text images, we propose a novel and effective data cleaning pipeline. This automated process is applied prior to manual inspection to efficiently filter out erroneously generated samples. The pipeline comprises three sequential steps:

1. Visual Marker Verification: To unambiguously define the target region for editing, a visual marker of a unique color (e.g., blue in RGB(0,0,255)) is first inserted into the original image. This initial step of the pipeline verifies that the marker has been completely removed during the editing process. Any image where remnants of the marker color are detected is discarded.

2. Change Detection via Image Binarization: This step confirms that the content within the target region has been substantially altered. We apply adaptive binarization to the target region in both the original and edited images and compute the mean absolute difference between the two resulting binary maps. If this difference falls below a predefined threshold (e.g., 32), it indicates negligible change, and the sample is discarded.

3. Semantic Content Verification using OCR: The final step uses Optical Character Recognition (OCR) to validate the semantic content of the edit. The requirements depend on the editing mode: For replacement edits, we verify that the OCR output from the edited region matches the target text. For removal edits, we verify that the OCR output is empty. Samples that fail to meet their respective criteria are filtered out.

Steps 2 and 3 are complementary; the binarization check ensures a structural change occurred, while the OCR check validates the semantic outcome. Employing both provides redundancy and robustness against potential failures in either method. This automated pipeline significantly reduces the manual effort required for data curation and improves the overall quality of the final dataset.

2.2. Detailed Prompts

The prompt in our image-to-text-to-image pipeline is ”Describe this image in detail so that I can prompt an image-generation AI model to generate the very same image with your description. Please only respond with the final answer and avoid unrelated words such as ’Sure! This is the prompt’ or ’To re-create it through an AI model, include the following details in your prompt’, so that I can directly use your entire text as the prompt.”

For our text-to-text-to-image pipeline, two example text prompts in the prompt pool are (1): ”A handwritten journal entry in flowing handwriting with slight right slant, black ink on cream paper: September 15, 2024. Today marked my first week in Tokyo, and the city continues to amaze me at every turn. The morning began with a visit to the Tsukiji Outer Market, where the narrow alleys were already buzzing with activity by 7 AM. The aroma of grilled seafood and the calls of vendors created an atmosphere that felt both chaotic and perfectly orchestrated. I managed to try tamago on a stick - a sweet Japanese omelet that melted in my mouth. The vendor, an elderly man with kind eyes, showed me how they carefully roll the eggs layer by layer.

It's these small interactions that make traveling so meaningful. In the afternoon, I explored the Yanaka district, one of Tokyo's oldest neighborhoods. The area survived the wartime bombings, preserving its traditional architecture and atmosphere. Small temples are tucked between modern homes, and cats roam freely through the quiet streets. I stopped at a local coffee shop where the owner has been roasting beans for over 40 years. Must remember to visit: - Sensoji Temple at sunrise - Shimokitazawa for vintage shopping - Try the ramen place recommended by Mari - Book tea ceremony for next week"; (2): "Bookstore window display. A sign displays 'New Arrivals This Week'. Below, a shelf tag with the text 'Best-Selling Novels Here'. To the side, a colorful poster advertises 'Author Meet And Greet on Saturday' with a central portrait of the author. There are four books on the bookshelf, namely 'The light between worlds' 'When stars are scattered' 'The silent patient' 'The night circus' ." Our query prompt for MLLMs is "You are an expert with rich imagination. I plan to use AI models to generate diverse text images. Therefore, I need diverse prompts for the generative AI models' input. I will provide three example prompts, and I would like you to design ten more with completely different content. The characteristics of the examples are that they describe a detailed scene and determine the text content of that scene. Since I am going to generate diverse prompts, please avoid repetition in your responses. Please begin each answer with a numbered index, such as '1. ' Here are examples: ..."

The prompts used for regional editing are detailed in Figure 2 of the paper. While our goal was to maintain semantic plausibility, we found this was consistently challenging for certain image types. In practice, to achieve a balance between the data scale, diversity and semantic plausibility, we permitted a small, controlled proportion of semantically implausible edits. The allowable ratio of such edits was adjusted for each image type based on the inherent difficulty of generating plausible modifications.

After obtaining the proposals for regional editing, we draw a visual marker (mostly a blue box) and prompt the image generator with a query "Replace the text ... to ... within the blue bounding box, then remove the blue box; change the text content while keeping the font color and style unchanged." For regional removal, the query is "Remove both the text ... within the blue bounding box and the blue bounding box itself."

3. More Details about the DS-Net Model

During training, our Artifact Decouple Network (ADN) generates predictions via a 1×1 conv-layer. Within the Forensic Decoupling Encoder, channel-attention-based feature fusion is achieved through three steps: (1) concatenating the two feature maps along the channel dimension, (2) subsequently passing the concatenated features through a channel attention layer [9], and (3) finally reducing the output's feature dimension by half using a 3×3 conv-layer.

Our Synergy Denoising Decoder is built upon the DINO structure [29], specifically incorporating an additional Global Forensic Query vector with a feature dimension of 256. This vector is initialized using the same methodology as the other object queries within DINO. Both the transformer encoder and decoder within our SDD comprise six transformer layers, each with a hidden dimension of 256 and a feedforward dimension of 2048. The positional embeddings have a feature dimension of 128.

To ensure full transparency and reproducibility, our model code will be made publicly available.

4. More Experiments and Analyses

4.1. T-SNE Analysis of ADN's Extracted Features

Figure 2 presents a t-SNE [13] visualization comparing the feature spaces learned by our ADN with and without the proposed loss functions (using the same set of samples). The points represent features extracted from different image regions: (i) non-text forgeries (red), (ii) in-domain text forgeries (green), (iii) cross-generator text forgeries (yellow), and (iv) real image text regions (blue).

Baseline (Without Proposed Losses): The left panel shows the ADN model trained on both text and non-text data but without our losses. Features from text and non-text forgeries form distinct, isolated clusters, with minimal overlap between each other and with cross-generator samples (yellow). This demonstrates that simple joint training is insufficient to bridge the domain gap, as the model fails to learn a shared representation due to content and task misalignment, leading to ineffective usage of the non-text data.

Our Method: In contrast, the right panel shows the effect of our method (Section 5.1 of the paper). The features from non-text (red) and in-domain text (green) forgeries merge into a shared, unified cluster with much larger area. Crucially, features from cross-generator forgeries (yellow) are also pulled into the full region of this shared space, indicating that ADN has learned a general, task- and content-agnostic representation of forgery artifacts. Furthermore, the positioning of diverse non-text features (red) along the boundary with the real feature space (blue) helps to define a more robust decision boundary.

As a result, our method establishes a clearer separation between real and fake features, which minimizes classification ambiguity and significantly enhances generalization to forgeries from unseen generators.

4.2. Extending Detection to Non-AI Text Forgeries through Joint Training

While our primary contribution is advancing the detection of AI-generated text forgeries, our dataset and approach can

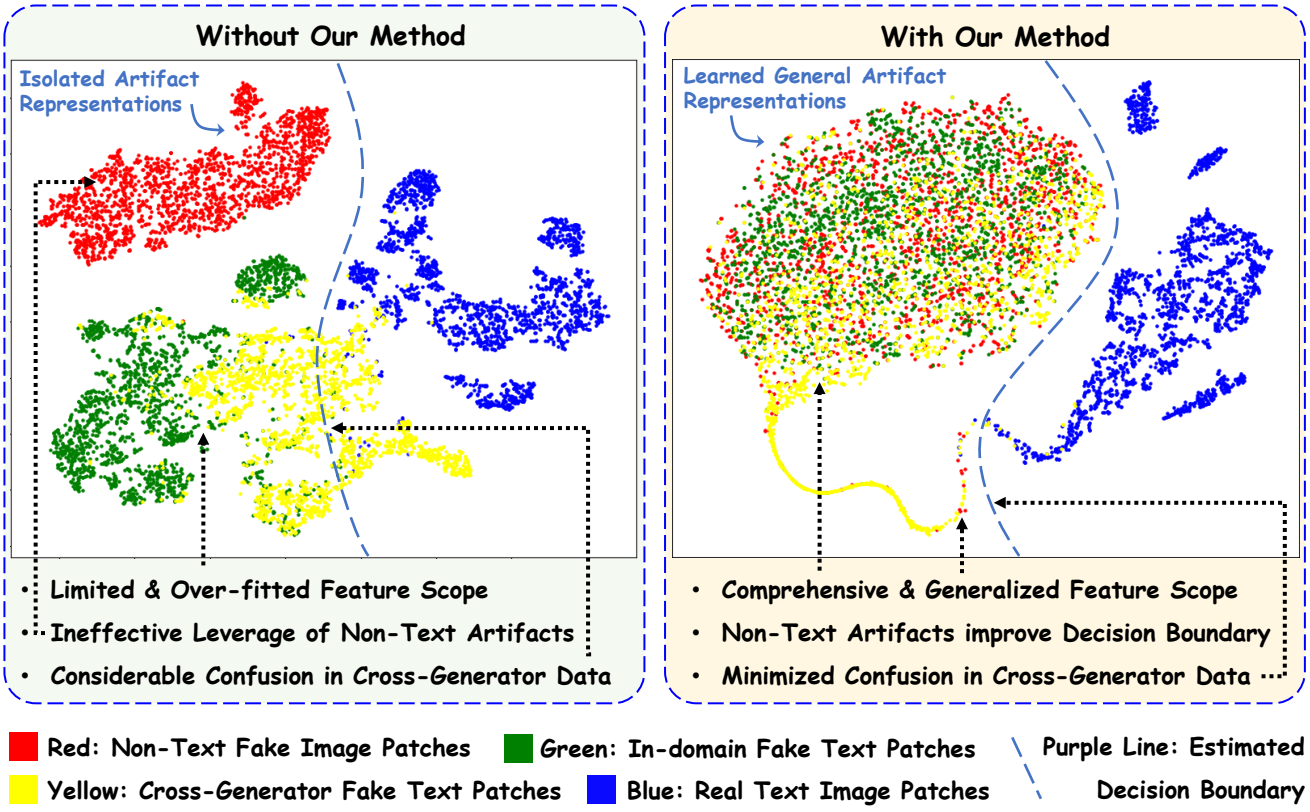

Figure 2. T-SNE visualization of features extracted by the Artifact Decouple Network (ADN). (Left) Without our method, jointly training on text and non-text forgeries yields an isolated and overfit feature space. (Right) In contrast, our method enables the ADN to learn a unified and highly generalizable feature space for forgery artifacts, significantly improving robustness to unseen generators.

Train Data	Test		CT		CG		CTG		CL		CLG		RW		RWT		DanceText Avg.	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
DanceText	93.8	80.1	**94.9**	**83.9**	79.0	65.5	65.4	51.2	**92.9**	**67.2**	**62.6**	**46.5**	64.5	12.3	66.2	24.3	77.4	53.9
DanceText +DocTamper [15]	**94.2**	**80.5**	94.7	83.5	**79.4**	**67.3**	**65.6**	**52.6**	92.2	65.6	61.6	45.7	**65.3**	**13.8**	**66.7**	**25.0**	**77.5**	**54.3**

Table 1. DS-Net model performance on DanceText dataset using different training data.

Method	Training Data	DocTamper F1		
		Test	FCD	SCD
ManTra-Net [27]	DocTamper [15]	15.3	20.9	15.7
MVSS-Net [6]	DocTamper [15]	43.1	42.4	41.4
PSCC-Net [11]	DocTamper [15]	38.4	42.0	37.4
BEiT-UPer [1]	DocTamper [15]	50.1	48.7	40.2
Swin-UPer [12]	DocTamper [15]	63.8	54.6	57.4
CAT-Net [10]	DocTamper [15]	70.0	55.3	63.1
DTD [15]	DocTamper [15]	79.2	81.6	75.4
DS-Net (Ours)	DocTamper [15]	80.6	83.7	76.1
DS-Net (Ours)	DanceText+ DocTamper [15]	**80.8**	**84.3**	**76.8**

Table 2. DS-Net model performance on DocTamper dataset using different training data.

Training Data of DS-Net	DanceText-RW		DanceText-RWT	
	Acc.	F1	Acc.	F1
DanceText-Train +50% DanceText-RW +50% DanceText-RWT	95.7	90.5	96.0	92.2

Table 3. DS-Net performance when the training set is augmented with a random 50% of the DanceText-RW and DanceText-RWT images, and evaluation is performed on the remaining half.

be extended to a joint training paradigm that also incorporates traditional, non-AI manipulations such as copy-move and splicing [5, 7, 15–18, 20, 21, 28]. As shown in Tables 1 and 2, this joint training strategy yields benefits for both

Training Data	Type		Test F1	
	Train	Full	OSTF	DanceText
OSTF [19]	✓	×	78.2	12.1
OSTF [19]	×	✓	-	15.8
DanceText	✓	×	81.0	53.9
DanceText	×	✓	96.5	-

Table 4. Cross dataset evaluation between OSTF and DanceText.

domains: it slightly improves performance on AI-generated data while also significantly enhancing the detection of non-AI forgeries. This mutual improvement stems from a fundamental synergy between the two forgery types. First, artifacts from imperfect AI generation, such as inconsistent text styles or edge anomalies, often resemble the characteristic traces of traditional tampering methods. Second, non-AI tampering methods can be regarded as novel, out-of-distribution "generators." The diverse artifacts they produce help regularize the model and mitigate overfitting to the signatures of specific AI generators. Therefore, our work not only addresses critical gaps in detecting forgeries created by advanced AI models, but also contributes to the development of a generalist forensic system capable of detecting text images counterfeited by any method.

4.3. Practical Defense Against Text Forgeries from Real-World Software

To rigorously analyze model generalization in an academic context, our DanceText dataset's training protocol involves training models solely on data from a limited number of open-source generators and evaluating them against data from real-world software and Apps. Under this setup, model performance is consistently low, attributable to significant discrepancies in generator characteristics and the varied pre- and post-processing pipelines of these software.

However, for real-world deployment, forensic systems can be effortlessly strengthened by augmenting their training sets with examples of forgeries from these real-world software, thereby preparing them against such attacks. To investigate this, we randomly split the DanceText-RW and DanceText-RWT data into two halves: one half was integrated into the training set, and the remainder was reserved for evaluation. Table 3 demonstrates that this modified training protocol enables the effective detection of these software-produced forgeries. This improvement is attributed to the expanded training data, which significantly broadens the models' generalization capabilities.

Without our DanceText dataset, these real-world challenges posed by software-produced forgeries would be exceedingly difficult to expose and defend against. In contrast, our DanceText dataset provides a solid data foundation, enabling both in-depth academic analyses and comprehensive defense against a wide spectrum of real-world attacks.

PatchSize	8	16	32	64
Acc.	76.5	77.2	**77.4**	76.9
F1	50.1	53.6	**53.9**	51.2

Table 5. Ablation study of the shuffle patch size.

Model	CRCNN+DAF	DS-Net	FFDN	MVSS-Net
Parameters	78M	120M	128M	143M
FPS	7.8	5.1	1.5	6.6

Table 6. Complexity comparison. 'FPS': Frame Per Second

4.4. Cross-Dataset Evaluation

This evaluation assesses the comprehensiveness and diversity of our DanceText dataset. Given that OSTF [19] integrates Tampered-IC13 [26] (a subset of SRNet), making it inherently more diverse than Tampered-IC13 alone, OSTF was selected as the comparative benchmark. Table 4 presents the results. Models trained exclusively on either the OSTF training set or the full OSTF dataset achieve F1-scores below 20 when evaluated on DanceText. Conversely, a model trained on DanceText-Train achieves an F1-score exceeding 80 on OSTF, and one trained on the full DanceText dataset surpasses 96 F1-score on OSTF. These findings confirm DanceText's significantly greater comprehensiveness and diversity compared to OSTF, which was previously considered the leading public dataset for AI-counterfeited text image detection.

4.5. More Ablation Studies

Table 5 presents an ablation study on the impact of patch size for the internal and external shuffle operations in ADN. A patch size of 32 yields the optimal performance. This result highlights a critical trade-off: Overly large patches fail to sufficiently disrupt high-level semantic context and are not very effective in encouraging the model to focus on fine-grained, local analysis. Overly small patches, conversely, risk destroying the very low-level artifact patterns essential for effective forgery detection. Thus, a patch size of 32 strikes the ideal balance, preserving local forgery traces while breaking higher-level contextual dependencies.

4.6. Comparison against Segmentation Models

We benchmarked DS-Net against several leading open-source, segmentation-based forensic models. To ensure a fair comparison, all models were retrained on the DanceText-Train set using their official implementations and identical training configurations. For our detection-based DS-Net, the predicted bounding boxes were converted into binary segmentation masks to enable a direct, pixel-level evaluation. Specifically, we highlight the predicted bounding boxes on a zero map. The results, summarized in Table 6, demonstrate that DS-Net substantially

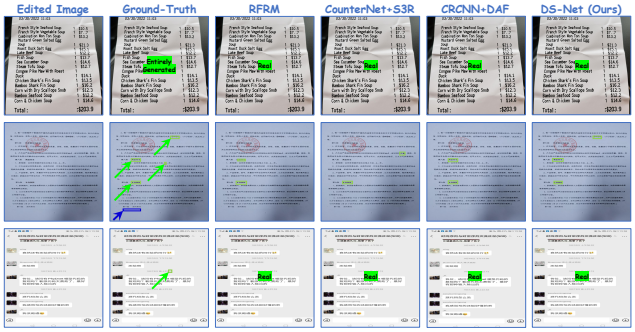

Figure 3. Failure cases for existing models.

outperforms these segmentation-focused approaches.

4.7. Robustness Evaluation

We evaluated our model's robustness by applying common distortions to the DanceText dataset. Table 8 confirms its enhanced robustness against these real-world distortions.

4.8. More Qualitative Comparisons

More qualitative comparisons are presented in Figures 9, 10 and 11, further confirming the effectiveness of our method.

5. Limitations

Although our proposed DS-Net significantly outperforms existing methods, room for improvement remains in challenging scenarios, particularly cross-generator and cross-real-world software evaluation (Table 2 of the paper). Some of the failure cases of our model are presented in Figure 3, illustrating that even advanced models can misclassify realistic images from unseen generators as authentic, or overlook small-sized counterfeits with minimal visual anomalies, especially when originating from novel generative approaches. It is notable that such real-world challenges are rarely exposed or effectively studied using existing public datasets. We believe that our DanceText dataset and DS-Net can serve as a foundation and inspire future research to better address these real-world challenges. Additionally, DS-Net is optimized primarily for performance rather than complexity, resulting in a slightly larger size (Table 6).

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 4

[2] Konstantin Bulatov, Daniil Matalov, and Vladimir V Arlazarov. Midv-2019: challenges of the modern mobile-based document ocr. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, pages 717–722. SPIE, 2020. 1

[3] Bulatov Konstantin Bulatovich, Emelianova Ekaterina Vladimirovna, Tropin Daniil Vyacheslavovich, Skoryukina Natalya Sergeevna, Chernyshova Yulia Sergeevna, Ming Zuheng, Burie Jean-Christophe, and Luqman Muhammad Muzzamil. Midv-2020: A comprehensive benchmark dataset for identity document analysis. , 46(2):252–270, 2022.

[4] Maria Jose Castro-Bleda, Salvador España-Boquera, Joan Pastor-Pellicer, and Francisco Zamora-Martínez. The noisy-office database: A corpus to train supervised machine learning filters for image processing. *The Computer Journal*, 63 (11):1658–1667, 2020. 1

[5] Xinhong Chen, Bangdong Chen, Chenfan Qu, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Dtsm: Toward dense table structure recognition with text query encoder and adjacent feature aggregator. In *International Conference on Document Analysis and Recognition*, pages 438–452. Springer, 2024. 4

[6] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[7] Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Jizhe Zhou. Forensichub: A unified benchmark codebase for all-domain fake image detection and localization, 2025. 4

[8] Zhengyao Fang, Pengyuan Lyu, Jingjing Wu, Chengquan Zhang, Jun Yu, Guangming Lu, and Wenjie Pei. Recognition-synergistic scene text editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13104–13113, 2025. 1

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[10] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, pages 1875–1895, 2022. 4

[11] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 4

[12] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 4

[13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 3

[14] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, pages 1454–1459. IEEE, 2017. 1

[15] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust

Figure 4. Example images from our DanceText dataset, fully-generated in image-level by text-to-image models using pure textual prompts.

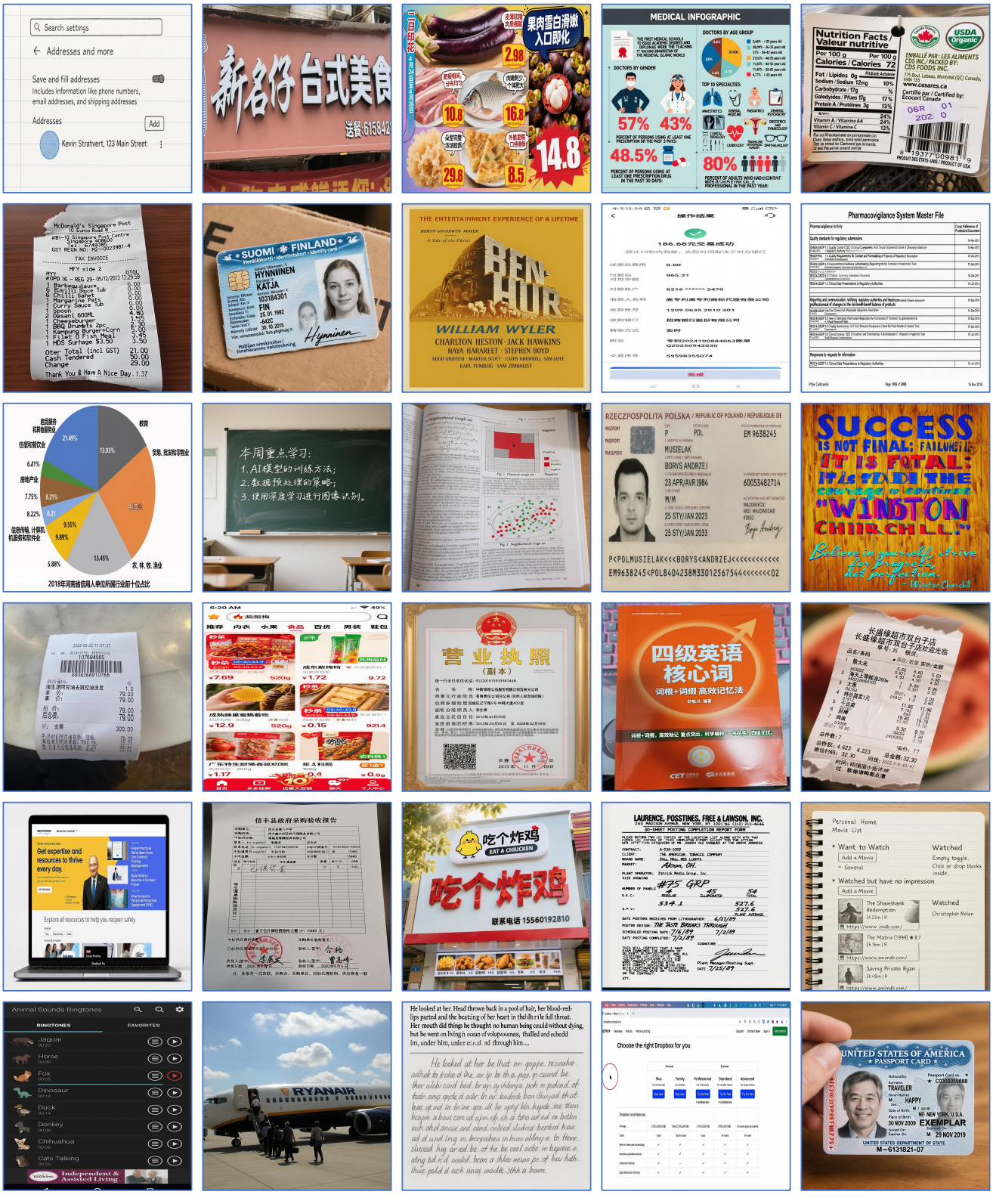

Figure 5. Example images from our DanceText dataset, fully-generated in image-level by text-to-image models using pure textual prompts.

Figure 6. Examples of regionally edited images from our DanceText dataset. Edited images are shown in the odd rows. The even rows display the corresponding annotations, where green boxes indicate regions of text replacement and blue boxes indicate text removal.

Figure 7. Examples of regionally edited images from our DanceText dataset. Edited images are shown in the odd rows. The even rows display the corresponding annotations, where green boxes indicate regions of text replacement and blue boxes indicate text removal.

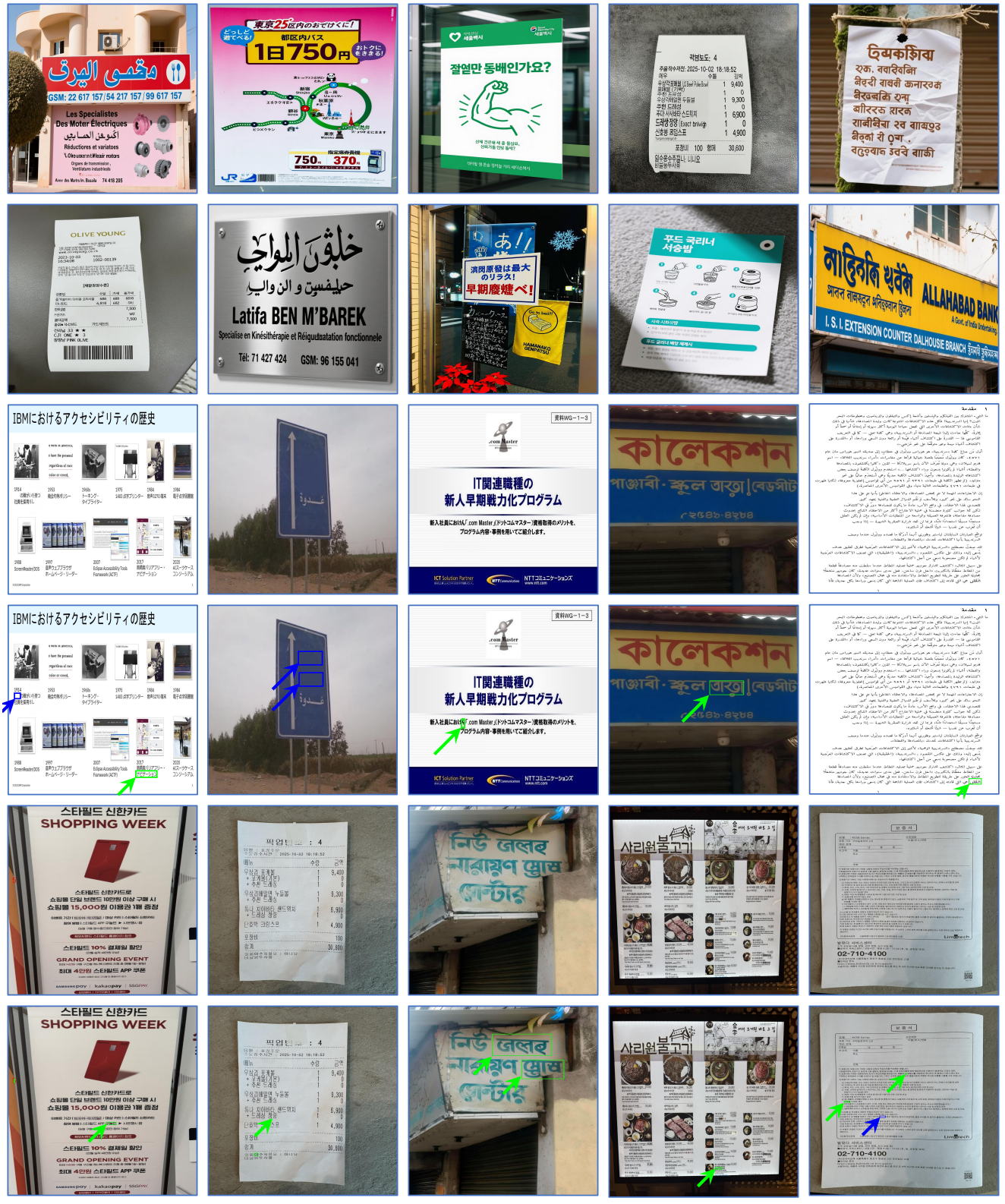

Figure 8. Examples of cross-lingual test images from our DanceText dataset. First and second rows: fully-generated images. Third and fifth rows: regionally edited images. Fourth and sixth rows: the corresponding annotations, where green boxes indicate regions of text replacement and blue boxes indicate text removal.

tampered text detection in document image: new dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5946, 2023. 4

[16] Chenfan Qu, Jian Liu, Haoxing Chen, Baihan Yu, Jingjing Liu, Weiqiang Wang, and Lianwen Jin. Textsleuth: Towards explainable tampered text detection. *arXiv preprint arXiv:2412.14816*, 2024.

[17] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Omni-iml: towards unified image manipulation localization. *arXiv preprint arXiv:2411.14823*, 2024.

[18] Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2024. 4

[19] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 694–702, 2025. 5

[20] Chenfan Qu, Yiwu Zhong, Huiguo He, Bin Li, and Lianwen Jin. Webly-supervised image manipulation localization via category-aware auto-annotation. *arXiv preprint arXiv:2508.20987*, 2025. 4

[21] Chenfan Qu, Yiwu Zhong, Jian Liu, Xuekang Zhu, Bohan Yu, and Lianwen Jin. Textshield-r1: Reinforced reasoning for tampered text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8621–8629, 2026. 4

[22] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 1

[23] Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*, 2021.

[24] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019.

[25] HuaWei Team. Huawei cloud visual information extraction competition. 2022. 1

[26] Yuxin Wang, Hongtao Xie, Mengting Xing, Jing Wang, Shenggao Zhu, and Yongdong Zhang. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, pages 215–232. Springer, 2022. 5

[27] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 4

[28] Jiaruo Yu, Dagong Lu, Xingyue Shi, Chenfan Qu, and Fengjun Guo. Unified face attack detection with micro disturbance and a two-stage training strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 960–969, 2024. 4

[29] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[30] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019. 1

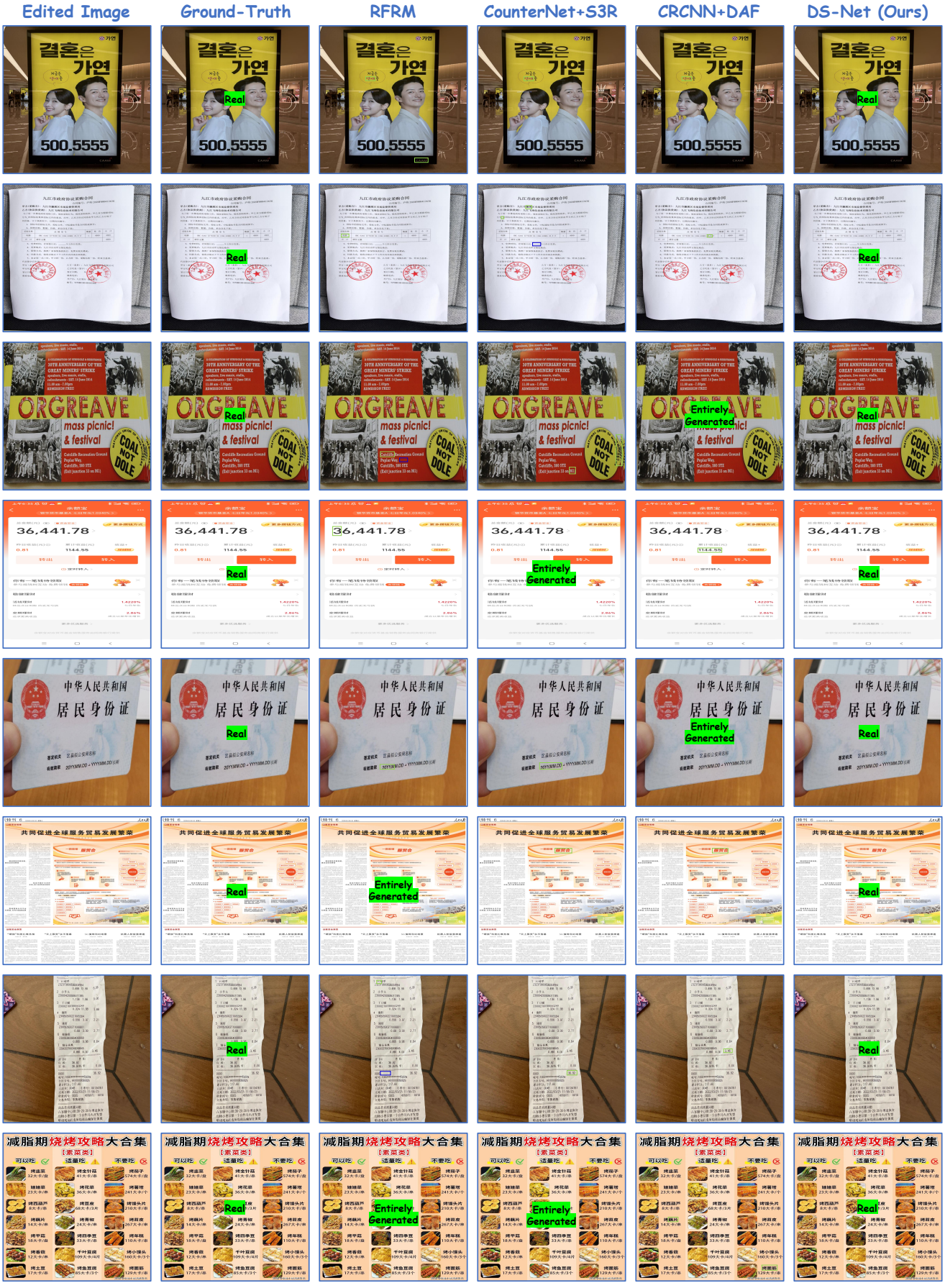

Figure 9. Qualitative comparison on real images.

Figure 10. Qualitative comparison on fully-generated images.

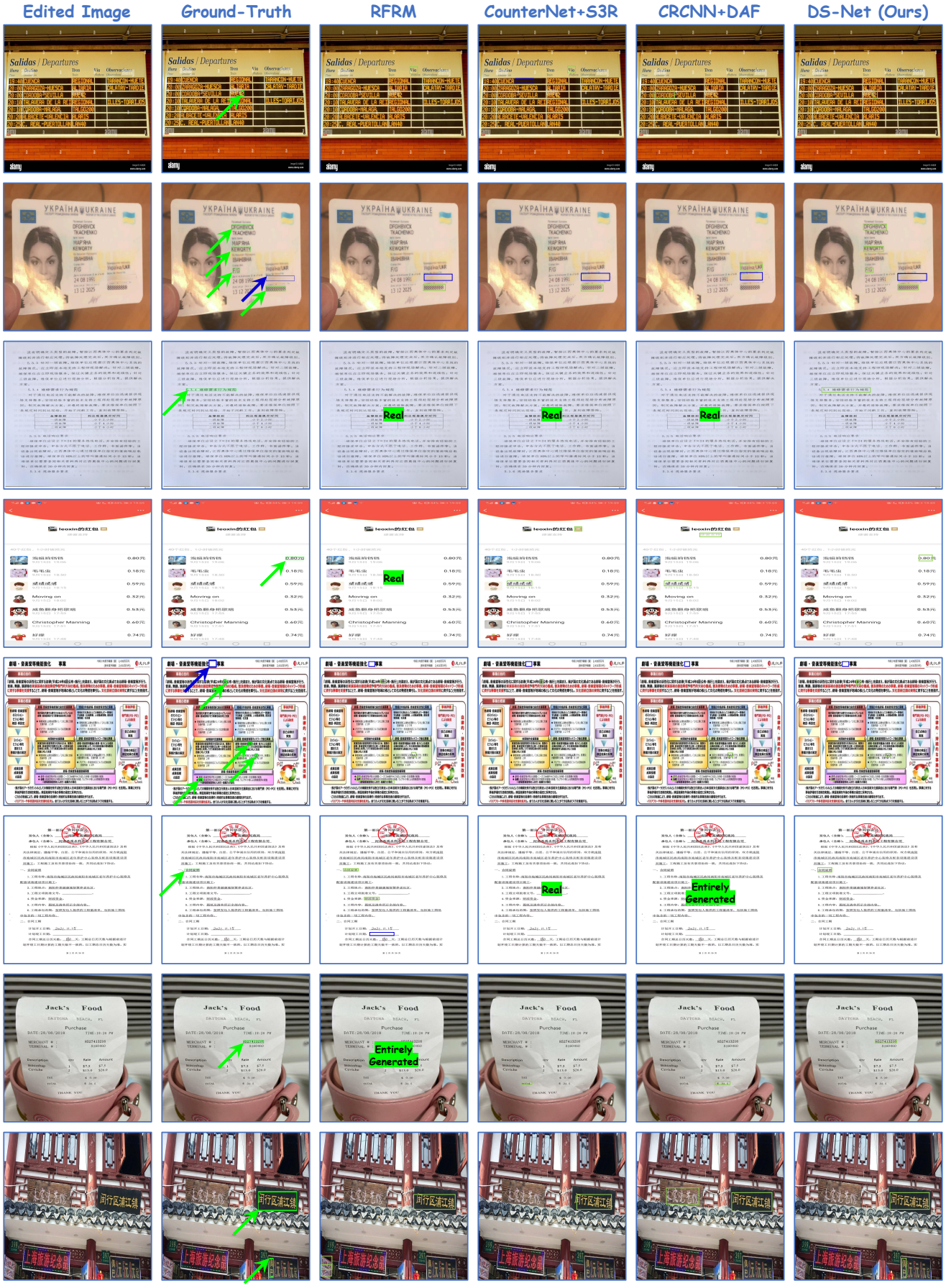

Figure 11. Qualitative comparison on regionally edited images. Green boxes: regions of text replacement. Blue boxes: text removal.