SkyReels-Text: Fine-Grained Font-Controllable Text Editing for Poster Design

Supplementary Material

This supplementary material provides further details on the SkyReels-OCR dataset (Section A) and the content-style decoupled dataset (Section B), and presents more qualitative results (Section C) to complement the main manuscript.

A. SkyReels-OCR Dataset

Existing OCR systems are optimized for regular scene text but struggle with ornate, non-standard font geometries. Therefore, we adopt a VLM-based OCR model, Qwen2.5-VL 7B [2], for improved parsing of complex textual patterns. To address the lack of training data across diverse font geometries and artistic variations, we propose the SkyReels-OCR dataset, a meticulously curated collection of high-quality real data sourced from professional design platforms (i.e., publicly accessible images from Pinterest and Amazon, collected solely for academic research) and a public repository, namely the AnyWord-3M [11] dataset. Our dataset covers a diverse range of text styles, including regular, handwritten, calligraphic, printed, and customized artistic fonts. To ensure the quality of the training data, we implement a multi-step filtering pipeline where each raw image is subjected to the following rules:

- The image resolution must be at least 720×720.
- The image must contain at least one text instance.
- The MUSIQ [6] score must be at least 48 and the NRQM [7] score must be at least 4 to ensure visual clarity.
- The Q-Align [12] score must be at least 3 to ensure aesthetic quality.

After applying these strict filtering rules, the resulting images undergo professional OCR annotation. Specifically, ten dedicated domain experts annotate the filtered images, yielding a final collection of 39,986 high-quality images with precise OCR labels to form the SkyReels-OCR dataset. We then partition this collection by randomly sampling 200 images to constitute the SkyReels-OCR benchmark for evaluation, while the remaining images form the training set, denoted as SkyReels-OCR-40K. Note that a dataset size of $\sim$ 40K is empirically sufficient for effective LoRA fine-tuning of Qwen2.5-VL 7B.

SkyReels-OCR-40K. Detailed statistics of text lines and characters for SkyReels-OCR-40K are presented in Table 1, with AnyWord-3M [11] included for comparison. Additionally, examples of annotated images from the dataset are provided in Figure 1.

SkyReels-OCR Benchmark. The SkyReels-OCR benchmark is utilized to evaluate OCR detection and recognition capabilities for regular and non-regular fonts using three metrics: Sen. Acc, NED, and spatial IoU. For comparative analysis, we benchmark our model against several competitive methods: practical models PP-OCRv4 [9] and PP-OCRv5 [4], along with powerful VLM-based approaches including PaddleOCR-VL [3], Qwen2.5-VL-7B-Instruct [2] (Baseline), and Qwen3-VL-8B-Instruct [1]. Quantitative comparisons are summarized in Table 2. Qualitative results are shown in Figure 2, where we present visual comparisons of different OCR models.

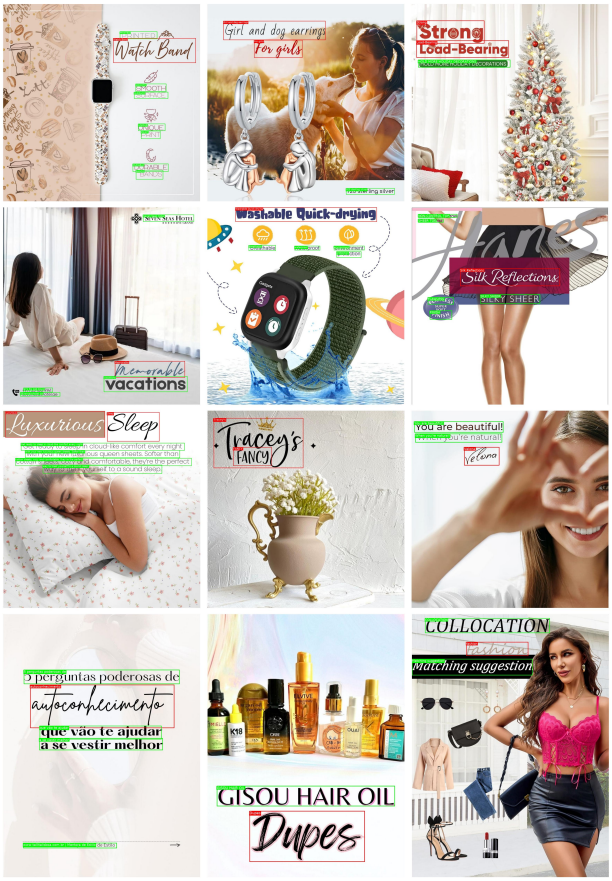

Figure 1. **Sample images from the SkyReels-OCR-40K dataset.** Green boxes denote regular fonts, while red boxes indicate non-regular fonts, including handwritten, calligraphic, and custom-designed artistic typefaces.

B. Content-Style Decoupled Dataset

Pipeline Details. To collect high-quality data with paired target images and reference fonts, we design an automatic data generation pipeline consisting of three steps:

Table 1. Statistics of text lines and characters for SkyReels-OCR-40K and AnyWord-3M.

Dataset	image count	image w/o text	mean lines/img	mean chars/line	#img $\leq=$ 5 lines	#line $\leq=$ 20 chars	#non-regular lines
AnyWord-3M [11]	3.03M	21.7K	3.03	5.35	2.64M, 86.9%	9.15M, 99.6%	-
SkyReels-OCR-40K	39.8K	0	5.57	13.06	23.6K, 59.3%	0.18M, 81.8%	64.0K, 29.1%

Table 2. Performance comparison of different OCR methods for regular and non-regular fonts on the SkyReels-OCR benchmark. **Boldface** and underlining denote the best and second-best results, respectively.

Methods	Regular Fonts			Non-Regular Fonts		
	Sen. Acc↑	NED↑	Spatial↑	Sen. Acc↑	NED↑	Spatial↑
PP-OCRv4 [9]	0.8769	0.9524	0.6543	0.3358	0.6914	0.4333
PP-OCRv5 [4]	0.8635	0.9413	0.6548	0.5569	0.8200	0.5845
PaddleOCR-VL [3]	0.8780	0.9110	0.4393	0.5444	0.6986	0.3920
Qwen3-VL-8B-Instruct [1]	**0.9659**	**0.9906**	0.6887	0.8925	0.9709	0.7831
Baseline	0.8386	0.8921	0.4882	0.6319	0.8453	0.3665
Ours	0.9446	0.9780	**0.7617**	**0.9276**	**0.9824**	**0.8259**

(1) Divergent Text Generation: We employ the Qwen3-8B [13] LLM to generate semantically distinct replacement texts that adhere to the original sequence length and preserve the casing of each word, simulating a constrained text substitution task.

(2) Reference Font Generation: We utilize Nano Banana [5] to perform local text editing with the previously generated divergent texts, aiming for automated reference font generation. This approach maximally preserves the font style and color of the original text. Furthermore, to enable our model to concentrate on font style learning while ignoring background interference, we employ SAM2 [10] to segment the text regions and then blend them with text-free background areas of other images.

(3) Two-pronged Verification Strategy: The final step in our content-style decoupling pipeline involves validating the fidelity of the local text edits. First, we employ our OCR model to confirm the content fidelity of the edits produced by Nano Banana. Specifically, we set a NED threshold of 0.9 between the ground truth replacement text and the edited text to retain only the successful instances. Subsequently, we leverage the DINO v2 [8] model to verify stylistic consistency. This is achieved by setting a DINO similarity threshold of 0.8 to confirm that the edited text maintains the original style, ensuring successful style preservation during the editing process.

Quantitative Analysis. To comprehensively evaluate the quality of the constructed dataset, we report key quantitative statistics in Figure 3. First, we verify the content accuracy and style consistency of the generated pairs by analyzing the NED and DINO distributions. Second, we confirm that the synthetic text lengths closely align with real-world distributions. Furthermore, to quantify the degree of decoupling, we define "Effective Text" as replacements with less than 20% character overlap with the original text. Since the reference fonts merely provide style cues rather than semantic context, strict semantic alignment with the real posters is unnecessary; ensuring content divergence is sufficient. Our statistical analysis validates a high proportion of "Effective Text" in the dataset, demonstrating the success of our decoupling strategy.

Qualitative Check. To qualitatively assess the dataset, we present visual samples of our training pairs in Figure 4. These examples clearly demonstrate the high visual quality, precise content replacement, and accurate font style retention of our generated pairs. Note that the incomplete text in the style reference images is due to random cropping during data augmentation, which encourages the model to learn robust style patterns across varying text lengths.

C. More Qualitative Results

Small-Size Text Editing. In real-world workflows, users frequently need to edit small-size text regions in AI-generated posters, which often suffer from garbled or nonsensical characters. SkyReels-Text proves effective in tackling these highly challenging scenarios. As illustrated in Figure 5, our model can generate or correct small text down to $\sim$20 pixels. Benefiting from the diverse text scales covered in our meticulously curated dataset, SkyReels-Text seamlessly replaces the original content—whether it is standard small text or garbled AI artifacts—with visually coherent target text. Furthermore, it effectively maintains the typographic identity of these small-size text regions and naturally integrates them into the surrounding background. This fine-grained control capability enhances the practical utility of our model.

Paragraph-Level Text Editing. Scaling up from word- and

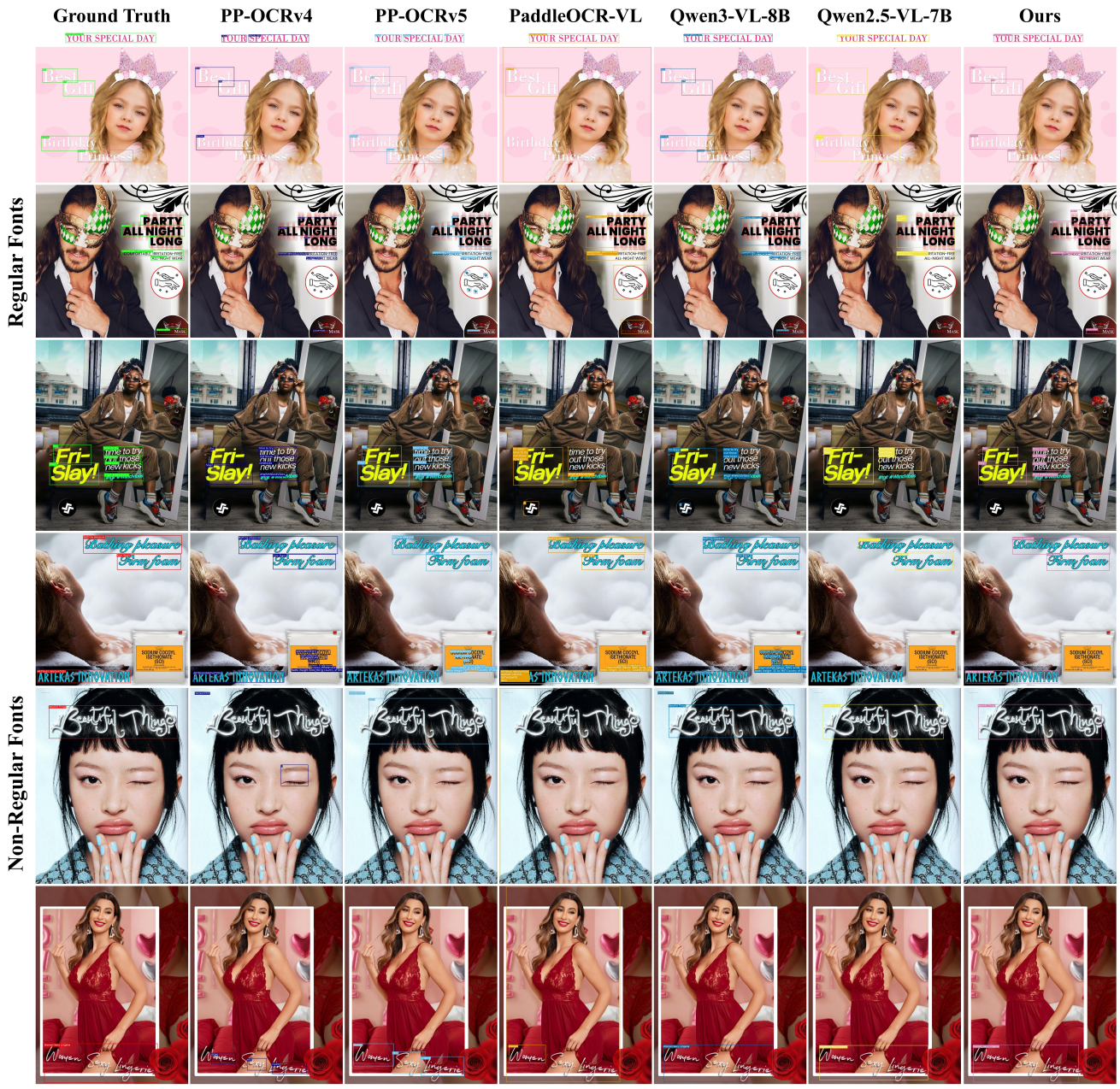

Figure 2. **Sample images of regular and non-regular fonts from the SkyReels-OCR benchmark.** Best viewed at 300% zoom.

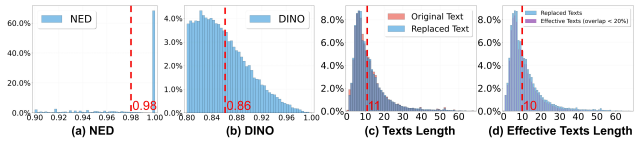

Figure 3. **Distributions of NED, DINO, text length, and effective text length.** The means are marked with red dashed lines.

sentence-level edits, comprehensive poster redesign often necessitates the replacement of long paragraphs. To extend the capability of SkyReels-Text to this scale, we conducted continual training via LoRA on 10K paragraph-level samples (5K from our content-style decoupled dataset and 5K synthetic samples). As shown in Figure 6, our model successfully handles paragraph-level editing for standard fonts. It effectively replaces the original multi-line text (red boxes) with substantial target content (blue boxes) while preserv-

Figure 4. **Examples from the content-style decoupled dataset.** From left to right, the columns display: original target image, edited image with semantically divergent text, extracted font mask, plain-text layout, and font-style reference image. These examples clearly demonstrate precise content replacement while preserving typographic style.

Figure 5. **Small-size text editing.** SkyReels-Text effectively replaces original small text or garbled artifacts (red boxes) with visually coherent target text (blue boxes), with zoomed-in details shown in the center.

ing the original style and layout. Although the model performs well on standard fonts, generating highly unconventional or artistic styles at the paragraph level remains challenging due to the scarcity of such dense, stylized data. We leave the exploration of this limitation to future work.

Text Editing with Single Font Styles. Figure 7 presents more qualitative results of our model guided by a single reference font. Given an input image (leftmost column) and a user-provided glyph patch specifying the desired style (top row), our model accurately synthesizes the new textual content in the target regions. As demonstrated, SkyReels-Text successfully transfers complex stylistic nuances—such as rich material texture, 3D metallic gloss, and multi-chromatic brushwork—to the newly generated text. Furthermore, the model seamlessly blends the stylized text into the original layout while preserving the structural integrity

Figure 6. **Paragraph-level text editing.** The red and blue boxes indicate the original text and our edited results, respectively.

and visual harmony of the unedited background. This highlights its powerful zero-shot font transfer capability, requiring no additional fine-tuning, which meets the rigorous demands of professional artistic design.

Text Editing with Multiple Font Styles. Most existing image or text editing methods require multiple rounds of interaction to apply different font styles across text regions. In contrast, SkyReels-Text can simultaneously edit multiple regions with distinct font styles in a single inference round, achieving precise and efficient multi-style editing. Figure 8 illustrates the capability of our model to edit multiple text regions with diverse font styles, in comparison to SOTA image editing models. The results show that Nano Banana produces severe artifacts and distorts the original image structure, Qwen-Image-Edit tends to leave the image unedited, and Seedream 4.0 fails to preserve either the correct text positioning or the font style. Conversely, SkyReels-Text achieves accurate, style-consistent, and spatially faithful edits in a single inference pass, which takes $\sim$8 seconds on a single A800 GPU.

Handwritten Text Generation. Tables 3 and 4 illustrate the capability of SkyReels-Text to generate English and Chinese handwritten text. Existing methods for handwritten text generation are typically monolingual—requiring separate training for English and Chinese, with one model per language. In contrast, SkyReels-Text enables zero-shot multilingual generation: given only a style reference image, it synthesizes high-fidelity handwritten text without any additional training. This strong generalization capability eliminates the need for language-specific fine-tuning, offering a flexible, efficient, and scalable solution for style-controllable text image synthesis across diverse languages.

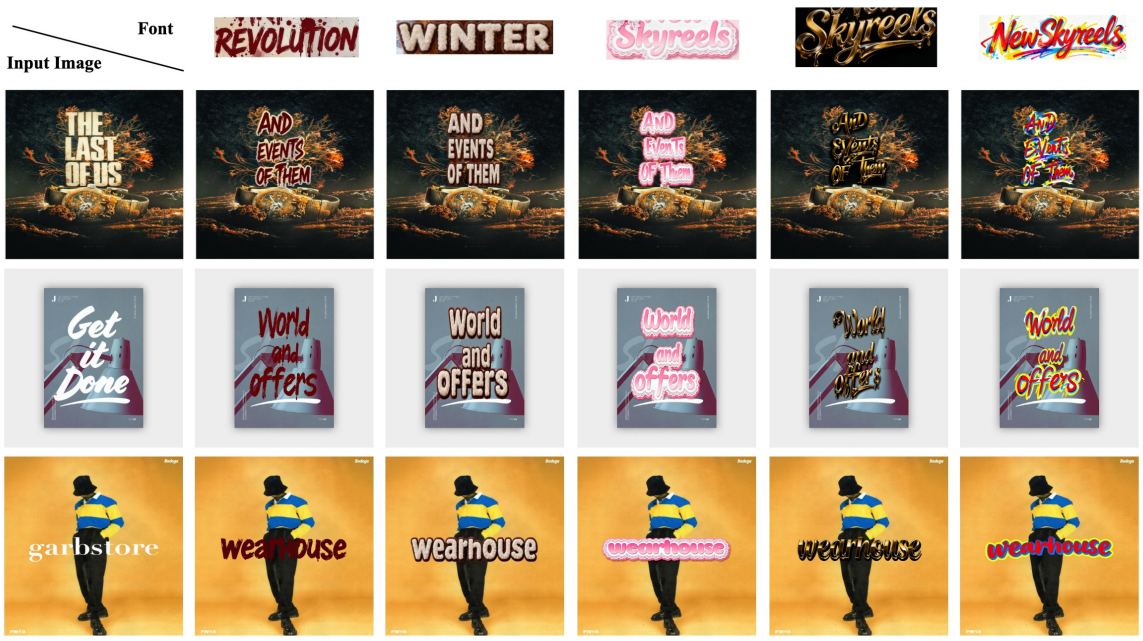

Figure 7. **More qualitative results of our model with single font styles.** The leftmost column shows the input images, and the top row displays the user-provided reference fonts.

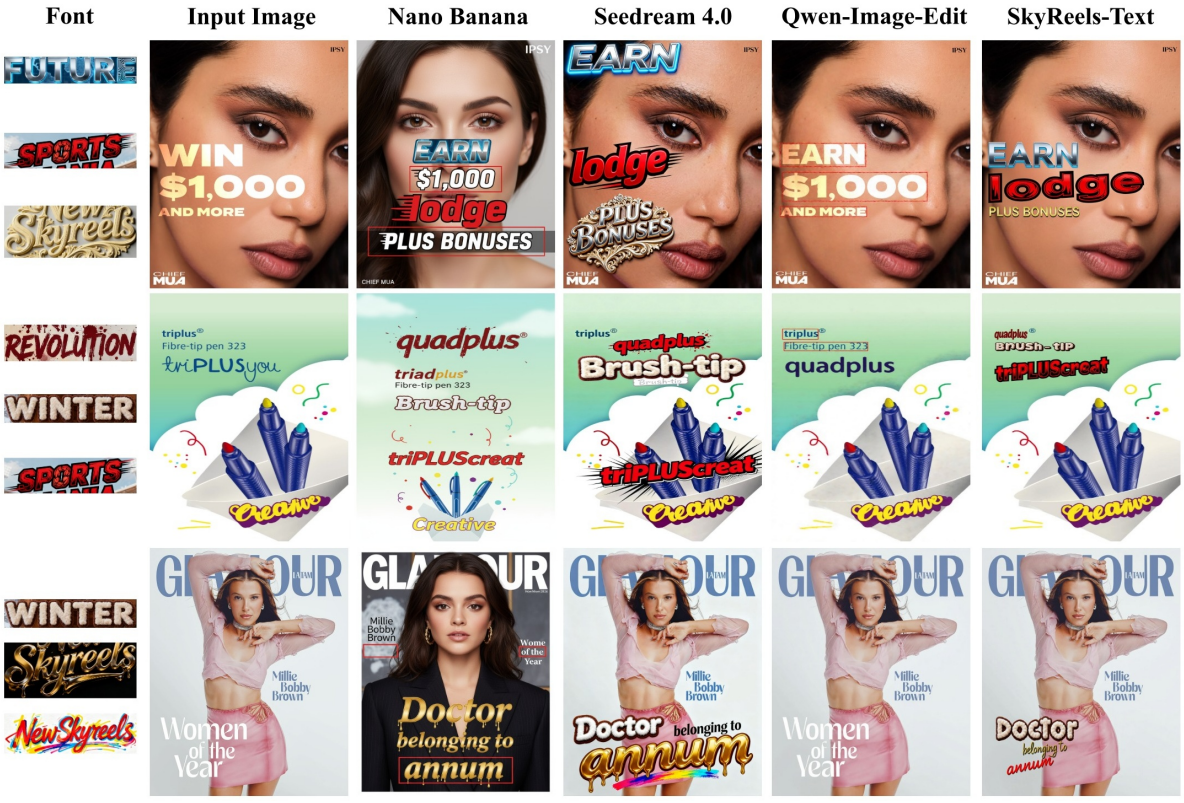

Figure 8. **Comparison with SOTA commercial image editing models in multi-font editing.** The first and second columns display the reference font styles and the input images, respectively. SkyReels-Text accurately applies distinct fonts to specific areas without mutual interference. Note that Flux Pro is unable to process multiple reference images; thus, it is omitted for a fair comparison.

Table 3. English handwritten text generation.

Ground Truth	Become a success with a dise and hey presto! You're a star... Rolly sings with
Style sample	"Bella Bella Marie" (Parlophone), a lively song that change; tempo
SkyReels-Text	Become a success with a disc and hey presto! You're a star.... Rolly sings with
Ground Truth	the work will have to be done by each
Style sample	Lord God with all their visionary power,
SkyReels-Text	the work will have to be done by each
Ground Truth	Levit. 2, B.) Salt preserves and we should have this salt in
Style sample	ourselves and have peace with one- another. And the sweet
SkyReels-Text	Levit. 2, B.) Salt preserves and we should have this salt in
Ground Truth	Hope I die kind of composed, Trout, I mean you can't
Style sample	no matter how bad it felt, the fire you know, or a
SkyReels-Text	Hope I die kind of composed, Trout, I mean you cant
Ground Truth	co-operated, we made the little side chapel a place
Style sample	at Tatsfield to the great modern church of St. Mark's
SkyReels-Text	co-operated, we made the little side chapel a place
Ground Truth	bimpasse which always resulted, and in the
Style sample	In this cruel process which was
SkyReels-Text	bimpasse which always resulted. and in the
Ground Truth	But not for long, for soon pedestrians and cars flocked upon
Style sample	First, it was Sunday morning; and, second, everyone
SkyReels-Text	But not for Long, for soon pedestrians and cars flocked upon

Table 4. Chinese handwritten text generation.

Ground Truth	这一研究成果对准确推算地球生命的历史非常有帮助。
Style sample	磁场进行了直接测量，发现磁气圈早在32亿年前就已
SkyReels-Text	这一研究成果对准确推算地球生命的历史非常有帮助。
Ground Truth	如果是经由法院判决有罪，那么，最重要的应该是法律
Style sample	被抓了，就是哪个被判了，或者是这个案子开庭
SkyReels-Text	如果是经由法院判决有罪那么最重要的应该是法律
Ground Truth	首先登场的是宝瓶座厄塔流星雨，5月6日流星雨极盛，为正在休假的
Style sample	由于3点多宝瓶座才会升起，因此观测这个流星雨需要早起。
SkyReels-Text	首先登场的是宝瓶座厄塔流星雨,5月6日流星雨极盛为正在休假的
Ground Truth	是2006年7月1日青藏铁路的开通，极大地改善了进出藏的交通条件，
Style sample	为西藏旅游业的快速发展提供了良好的外部条件。2006年，西藏夫
SkyReels-Text	是2006年7月1日青藏铁路的开通。极大地改善了进出藏的交通条件。
Ground Truth	区以建设高标准农田为主要任务；中部粮食主产区以保护和提高基
Style sample	标准基本农田建设，并将小型农田水利建设作为重要内容。东部亿
SkyReels-Text	区以建设高标准农田为主要任务；中部粮食主产区以保护和提高基
Ground Truth	提供生源统计情况。确保所有被录取考生都可以在省级招办和高等学校招生信息网络平
Style sample	为保证计划调整工作顺利进行各省级招办应在批次录取开始前，向有关高等学校及时
SkyReels-Text	提供生源统计情况。确保所有被录取考生都可以在省级招办和高台校，向有关高信息网络平
Ground Truth	航空机构合作完成，耗资45亿美元。它的体积是哈勃望远镜的3倍。望远
Style sample	新科技。它有一个红外线照相机，以及一架保持在极低温度下运行
SkyReels-Text	航空机构合作完成耗资45亿美元，它的体积是哈勃望远镜的3倍.望远

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025. 1, 2

[4] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 1, 2

[5] Google. Gemini 2.5 flash image. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image, 2025. 2

[6] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 1

[7] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 1

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[9] PaddlePaddle. Pp-ocrv4. https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.7/doc/doc_ch/PP-OCRv4_introduction.md, 2023. 1, 2

[10] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[11] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[12] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 1

[13] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2