DOCUMENT POST-TRAIN FOR IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Document page generation with text-to-image models requires accurate rendering of long passages and precise layout control, enabling automated creation of text-rich materials. However, existing models often suffer from local text defects and inaccurate layouts. To address these limitations, we introduce a post-training method covering both data and optimization. On the data side, we construct an automated pipeline turning real PDF pages into OCR-grounded dense captions with layout and visual features. Regarding optimization, we design legibility rewards based on VLM and region-level credit assignment specifically to address local text defects. Evaluations demonstrate that an open-source model trained with our method achieves commercial-level performance on our benchmark, ranking second only to GPT Image 2 overall, and this improvement also generalizes to other text-rich tasks. Our further analysis reveals the ineffectiveness of image-level optimization and OCR rewards, and shows that reinforcement learning reduces sampling needs while retaining diversity for test-time scaling. We believe these findings offer a practical path toward high-quality document generation.

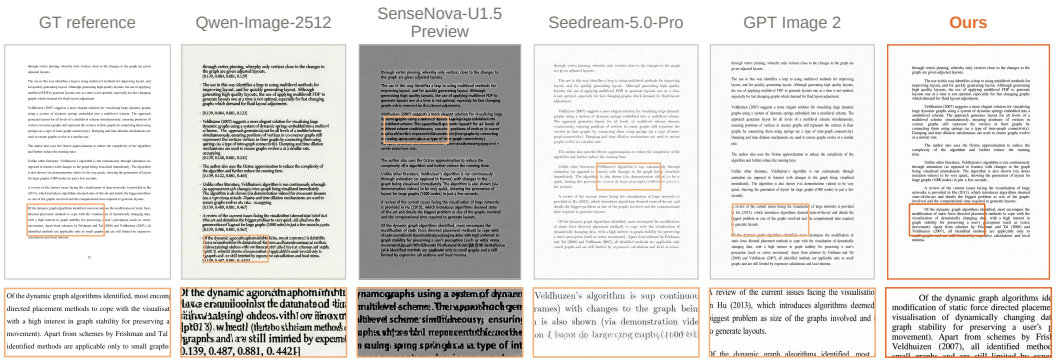


Figure 1: Document generation from the same English request specifying text and bounding boxes. Crops at the same location show that our model preserves the reference paragraph positions more accurately than GPT Image 2. Other crops highlight local text defects in model outputs.

1 INTRODUCTION

Text-to-image models generate images from textual instructions (Rombach et al., 2022). Document page generation combines text, layout, and visual design to create reports, educational materials, and other text-rich documents (Peng et al., 2025; Wang et al., 2025). The challenge is to render long passages completely and accurately (Tuo et al., 2024; Chen et al., 2024) while following spatial constraints (Kang et al., 2026) and maintaining consistent typography throughout a dense page.

However, current models struggle to meet these requirements jointly. Local text defects worsen as text length increases: Figure 1 shows distorted or overlapping glyphs in several models. GPT Image 2 renders clean text in this example but misplaces paragraphs. Existing datasets (Wang et al., 2025; Peng et al., 2025) support learning these requirements through long-text and structured-image supervision, but do not consistently pair real document pages with complete textual content, layout, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Text-rich image datasets. Real denotes collected images or documents; Visual denotes visual design descriptions; Query denotes generation prompts; Multiling. denotes text in two or more languages. ✓ yes, ♦ partial, ✗ no; — unavailable; † subset statistics.

Dataset	Mean text (chars/image)	Image size (pixels)	Real	Visual	Query	Multiling.
MARIO-10M (Chen et al., 2023)	47	512×512	✓	✗	✓	✗
AnyWord-3M (Tuo et al., 2024)	27	512×512	✓	✗	✓	✓
TextAtlas5M (Wang et al., 2025)	$414^{\dagger}$	$512\times512^{\dagger}$	♦	♦	✓	✓
Infographics-650K (Peng et al., 2025)	—	2240×896	✗	✓	✓	✓
OmniDocLayout-1M (Kang et al., 2026)	—	1084×1509	✓	✗	✗	✗
DocPage-1M (ours)	**1,372**	1241×1754	✓	✓	✓	✓

visual descriptions. Complementing this supervision, reward- and preference-based post-training improves image quality and prompt alignment (Black et al., 2024; Wallace et al., 2024). For text, reinforcement learning (RL) with optical character recognition (OCR) rewards improves recognizability (Liu et al., 2026), yet visible artifacts can remain (Shuai et al., 2026). TextAlign assesses these defects with vision-language model (VLM) judgments, but aggregates them into an image-level reward without separate policy feedback for individual text regions (Cui et al., 2026).

To address these limitations, we introduce a document post-training framework combining complete page supervision with regional policy feedback. **DocPage-1M** is a bilingual dataset pairing approximately 1M real PDF pages with OCR-grounded dense captions. Building on detailed caption supervision (Betker et al., 2023), it combines visual descriptions, text, and boxes to support document learning and regional evaluation (Section 3). Supervised fine-tuning (SFT) establishes document generation capability (Section 4.1), but local text defects remain and regions of different quality exist in one page. Our RL addresses both issues. Graded **VLM legibility rewards** distinguish readable but imperfect text from clean typography, providing feedback beyond OCR recognizability (Section 4.2). **Spatial credit assignment** preserves these distinctions in the policy objective by comparing each region with alternative renderings of the same content and mapping its advantage to the corresponding image tokens. Regions within one page can thus receive opposing policy signals, allowing local quality differences to guide targeted refinement (Section 4.3).

We evaluate on **DocPageBench**, our bilingual benchmark for Text, Layout, and Aesthetics. With requests adapted to our training format, post-trained SenseNova-U1.5-8B-MoT-Preview (Diao et al., 2026) outperforms all evaluated open-source baselines on all three, averaging 0.8457 versus 0.5254 for the strongest baseline, Qwen-Image-2512. Among all compared generators, it achieves the highest Layout score and ranks second only to GPT Image 2 overall, although an Aesthetics gap remains. Results on BizGenEval (Li et al., 2026) also show generalization to other text-rich tasks.

Our analyses validate these designs and clarify RL's relationship with test-time scaling (TTS) (Ma et al., 2025). SFT provides a stronger initialization and larger RL gains, while spatial credit assignment uses the same rewards more effectively than image-level updates (Section 6.1). Accurate OCR can coexist with visible defects, while VLM legibility rewards improve both visual quality and OCR coverage more than OCR-reward training (Section 6.2). RL reduces sampling needs with a single output approaching the quality obtained by selecting from 64 SFT candidates. Further TTS gains indicate useful candidate diversity, making RL and test-time selection complementary (Section 6.3).

2 Related Work

Data for Text-Rich Image Generation. Existing datasets provide complementary forms of supervision for text-rich image generation. MARIO-10M (Chen et al., 2023) and AnyWord-3M (Tuo et al., 2024) combine images with OCR text and spatial annotations. TextAtlas5M (Wang et al., 2025) adds structured content and visual design descriptions, while Infographics-650K (Peng et al., 2025) pairs synthetic infographics with layouts and design captions. OmniDocLayout-1M provides layouts across six document types (Kang et al., 2026). Among the datasets in Table 1, DocPage-1M is the first to provide bilingual PDF pages with complete layout, visual, and query annotations.

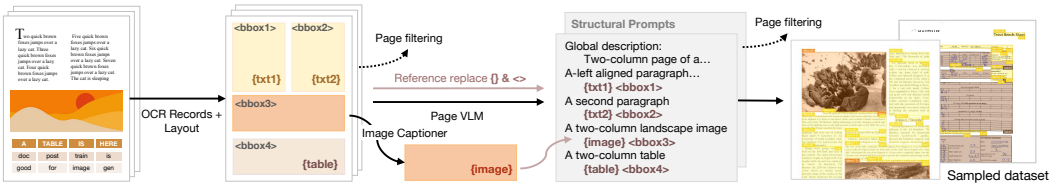


Figure 2: Construction of DocPage-1M. OCR extracts regional content and boxes; an image captioner describes picture regions, and a page VLM produces visual descriptions with indexed references. Refilling content and coordinates forms the dense captions. Braces and angle brackets schematically denote content and box references; example dataset pages appear on the right.

Post-Training for Text-Rich Image Generation. Flow-GRPO (Liu et al., 2026) improves text rendering through online reinforcement learning (RL) with OCR rewards, while MixGRPO (Li et al., 2025) improves training efficiency through windowed stochastic sampling. PosterCraft (Chen et al., 2026) combines text fidelity and aesthetic preferences for poster post-training. GlyphPrinter (Shuai et al., 2026) learns regional glyph preferences through direct preference optimization. TextAlign (Cui et al., 2026) aggregates hierarchical vision-language model (VLM) judgments into a scalar reward for policy and preference optimization. Building on these advances, we use regional legibility rewards to guide local policy updates to individual image tokens for dense document generation.

3 DocPage-1M

We introduce DocPage-1M, a specially constructed bilingual dataset of approximately 1M document page–caption pairs with text, layout, and appearance supervision for post-training image generators. We describe its construction pipeline in Section 3.1 and dataset characteristics in Section 3.2.

3.1 Data Construction

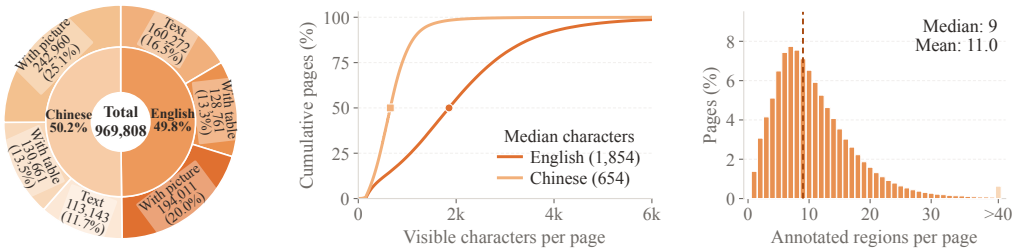
OCR extraction. To obtain text and spatial references for caption construction, we apply `dots.mocr` (Zheng et al., 2026) to each page image I, yielding numbered regional records $\mathcal{O}(I) = \{(i, k_i, b_i, t_i)\}_{i=1}^{m}$ (Figure 2), where m is the number of regions, k_i the block category, $b_i = (x_i^{\min}, y_i^{\min}, x_i^{\max}, y_i^{\max}) \in [0,1]^4$ the normalized bounding box, and t_i the recognized text when available. We filter out pages with short OCR outputs to retain text-rich document examples.

Visual description. To capture picture semantics and page design, we use detailed visual descriptions (Betker et al., 2023). A VLM first describes selected picture regions i, producing v_i. A page-level VLM then receives I, $\mathcal{O}(I)$, and these picture descriptions to describe composition, alignment, and typography. Its description uses indexed references to ordinary text and captioned pictures, allowing the VLM to focus on visual features without retranscribing long passages.

Content refill. To combine visual descriptions with source content, we replace text references with t_i and b_i, and picture references with v_i and b_i. The resulting dense caption begins with an overall composition description, followed by elements in reading order. We clean and filter these captions to form DocPage-1M. The captions condition supervised fine-tuning, while text and boxes provide references for regional rewards (Section 4.2). Appendix A.1 gives construction details.

3.2 Dataset Characteristics

DocPage-1M contains 969,808 document pages from 361,517 unique PDF files. English and Chinese pages are nearly equally represented (Figure 3(a)). *With table* and *With picture* account for 26.7% and 45.1%, respectively. Pages contain 1,372 visible characters on average, and 47.3% contain more than 1,000 (Figure 3(b)). The dense captions contain a median of 9 regional annotations per page and 10.68 million in total (Figure 3(c)). These annotations cover text, tables, formulas, and pictures, providing region-level spatial supervision alongside page-level visual descriptions.



(a) Dataset composition **(b) Text length** **(c) Region count**

Figure 3: Statistics of the 969,808 page–caption pairs in DocPage-1M. (a) Page language in the inner ring and content type in the outer ring. (b) Within-language cumulative distributions of visible characters, shown up to 6,000 characters, with medians marked. (c) Distribution of annotated regions in training captions; counts above 40 annotated regions per page are pooled in the final bin.

4 Document Post-Training

Our framework first learns document generation through supervised fine-tuning (SFT), then refines text with reinforcement learning (RL). Regional legibility rewards assess local quality, and spatial credit assignment directs policy updates to image tokens within each scored region (Figure 4).

4.1 Supervised Fine-Tuning

Our SFT stage fine-tunes all parameters of SenseNova-U1.5-8B-MoT-Preview (Diao et al., 2026) on text-to-image training pairs comprising 70% DocPage-1M and 30% general images. For a caption c and its paired page image I, we optimize the flow-matching objective (Lipman et al., 2023):

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathcal{L}_{\mathrm{FM}}(\theta; I, c). \tag{1}$$

After 20,000 updates, we exponentially average saved checkpoints offline. The averaged model initializes the RL policy and remains frozen as its reference throughout training (Appendix B.1).

4.2 Region-Level Legibility Rewards

Recognizable text can still contain distorted glyphs or inconsistent typography. We therefore use `gemini-3-flash-preview` (Google DeepMind, 2025) to score regions with more than 50 characters: dense text is more prone to defects, while short spans yield noisier judgments. The judge receives a slightly expanded crop and reference text. The reference helps interpret glyph shapes, while the score reflects visual quality rather than content agreement (Appendix B.2).

The judge assigns three quality levels: L1 (unreadable) for regions with contiguous unreadable spans; L2 (readable but imperfect) for isolated illegible characters or visible artifacts without such spans; and L3 (clean typography) for clean, consistent text comparable to print. We request three judgments per crop and average their mapped scores. For generated page g and region i,

$$r_{gi} = \frac{1}{|\mathcal{V}_{gi}|} \sum_{v \in \mathcal{V}_{gi}} f\big(\ell_{gi}^{(v)}\big), \qquad f(1) = 0.1, \quad f(2) = 0.5, \quad f(3) = 1.0, \tag{2}$$

where $\mathcal{V}_{gi}$ contains the successful judgments. We monitor page-level legibility using the character-weighted geometric mean $R_g = \exp(\sum_i n_i \log r_{gi} / \sum_i n_i)$ over valid regions, where n_i is the reference text length. Policy optimization uses each scored text region's reward separately.

4.3 Region-Local Policy Optimization

Regional advantages. We adapt group-relative advantages (Shao et al., 2024) to compare corresponding regions across a group $\mathcal{G}$ of $G = 8$ pages from the same caption. Rollouts use Flow-GRPO (Liu et al., 2026) with MixGRPO-style sampling (Li et al., 2025): 20 denoising steps include a two-step stochastic window sampled independently for each page. Only recorded stochastic transitions enter the policy loss; other steps use deterministic ODE updates. We exclude regions with

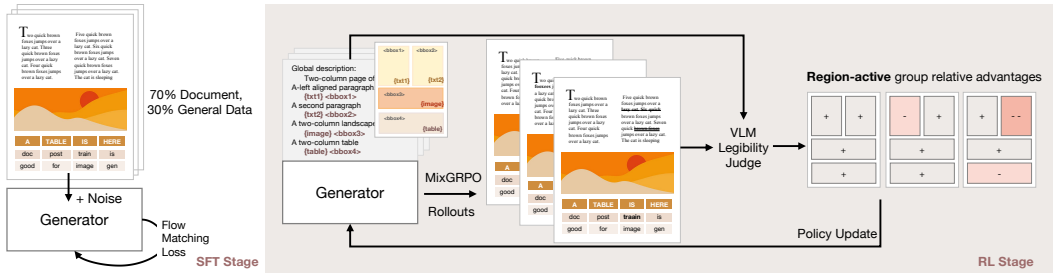


Figure 4: Two-stage document post-training. Left: flow-matching SFT uses 70% document and 30% general image data. Right: a shared caption conditions grouped rollouts with MixGRPO-style sampling; a VLM judges corresponding text regions, and regional advantages guide local policy updates. Plus and minus signs indicate whether a region scores above or below its group mean.

missing scores or negligible within-group variation. The advantage for each active region i is

$$a_{gi} = \frac{r_{gi} - \bar{r}_i}{\sigma_{\text{pool}} + \delta}, \qquad \bar{r}_i = \frac{1}{G}\sum_{g=1}^{G} r_{gi}, \tag{3}$$

where σ_{pool} is the root mean square of centered rewards across all active regions in the batch, and $\delta > 0$ is a stabilizer. This shared scale preserves differences in reward variation between regions.

Spatial credit assignment. We map regional advantages to image tokens. Let $M_i(u)$ indicate whether image token u falls within region i's expanded scoring box. Normalizing each mask as $q_i(u) = M_i(u)/\sum_v M_i(v)$ gives every region equal total spatial weight. Let $\mathcal{A}$ denote the active regions and $d_u = \sum_{i\in\mathcal{A}} q_i(u)$. For covered tokens ($d_u > 0$), the advantage and policy weight are

$$A_{gu} = \text{clip}\left(\frac{\sum_{i\in\mathcal{A}} q_i(u) a_{gi}}{d_u}, -A_{\max}, A_{\max}\right), \qquad w_u = \frac{d_u}{\sum_v d_v}. \tag{4}$$

Here, $A_{\max}$ is the clipping bound for token advantages. Overlapping regions contribute an inverse-area-weighted average of their advantages, giving smaller regions greater weight. Uncovered tokens have zero advantage and weight; groups with no active regions contribute no regional policy loss.

Policy optimization. The advantages and spatial weights enter a clipped policy objective (Schulman et al., 2017). For stochastic step k, define $\rho_{gku} = \exp(\bar{\ell}_{\theta,gku} - \bar{\ell}_{\text{old},gku})$, where $\bar{\ell}$ is the Gaussian transition log-density averaged across channels at token u. The resulting policy loss is

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}_{c,\mathcal{G},k}\left[\frac{1}{G}\sum_{g=1}^{G}\sum_{u} w_u \min\{\rho_{gku}A_{gu},\ \text{clip}(\rho_{gku}, 1-\epsilon, 1+\epsilon)A_{gu}\}\right], \tag{5}$$

where ϵ is the ratio clipping threshold. Both stochastic steps share the spatial maps, and rollout log-densities remain fixed during optimization. We regularize against the frozen SFT model using mean squared differences in conditional and unconditional clean-image predictions. This penalty covers all valid image tokens, including unscored regions. Further details appear in Appendix B.3.

5 Evaluation

We introduce DocPageBench, our constructed benchmark for document page generation (Section 5.1). After describing the models and inference settings (Section 5.2), we examine document generation performance (Section 5.3) and generalization to other text-rich tasks (Section 5.4).

5.1 DocPageBench

Construction. We construct DocPageBench with 400 document pages: 200 English and 200 Chinese. It contains 200 text-only pages, 100 with tables, and 100 with pictures, with a median of

approximately 1,403 recognized characters per page. We exclude source-document directories represented in the document SFT set and build requests without using dense captions in the training format. We vary two aspects of the original request: text is either partially specified or supplied verbatim, and explicit coordinates may be present or absent. Each page is assigned to one of four conditions. Appendix C.1 defines all four conditions (Table 4) and gives construction details.

Metrics. We assess three dimensions. *Text* measures fidelity to supplied text and fulfillment of requirements when content is partially specified. *Layout* assesses adherence to requested positions and spatial relationships. *Aesthetics* assesses visual quality, including typographic consistency, page composition, and rendering artifacts. Scoring rules and validation appear in Appendix C.2.

5.2 Experimental Setup

Models. We evaluate SenseNova-U1.5-8B-MoT-Preview (Diao et al., 2026) (*Base*) and two post-trained checkpoints. *SFT* is the averaged checkpoint from the 20,000-update run in Section 4.1. *SFT + RL* is the checkpoint at RL update 560, initialized from SFT (Section 4.3). The comparison includes 4 commercial models (GPT Image 2 (OpenAI, 2026), Seedream-5.0-Pro (Seedream et al., 2025), Qwen-Image-3-Pro (Alibaba Cloud, 2026), and Nano Banana 2 (Google DeepMind, 2026b)) accessed through APIs, and 4 open-source models (Qwen-Image-2512 (Wu et al., 2025), Z-Image (Cai et al., 2025), FLUX.2-dev (Black Forest Labs, 2025), and HunyuanImage-3 (Team, 2025)) run locally. Results for additional model variants and input settings appear in Appendix C.4.

Prompt adaptation and generation. Requests may be used directly or rewritten through prompt enhancement (PE). Table 2 labels unmodified requests as *Original*. For commercial models, this describes the submitted request without assumptions about internal adaptation. *Official PE* uses the base model's official infographic template, while *Our PE* converts free-form user requests into dense captions in our training format. Both PE use the same `gemini-3.1-pro-preview` model (Google DeepMind, 2026a). Our checkpoints use 50 Euler steps and classifier-free guidance of 4.0 for generation, with benchmark-specific settings in Appendices C.3 and C.5.

5.3 Document Generation Performance

Roles of SFT and RL. To examine the respective benefits of the two training stages, we evaluate Base, SFT, and SFT + RL with Our PE. Table 2 shows that SFT improves Layout from 0.7042 to 0.8629 and Aesthetics from 0.5556 to 0.7525. RL further increases Text from 0.7853 to 0.8260 and Aesthetics to 0.8425, with little change in Layout. The enlarged regions in Figure 5 show glyph artifacts remaining after SFT and more consistent typography after RL. We find that even with original inputs, the final SFT + RL model also improves all three metrics over the Base model.

Comparison with existing models. We use SFT + RL with Our PE as our final configuration. It substantially outperforms all evaluated open-source baselines across the three dimensions, exceeding the strongest open-source baseline, Qwen-Image-2512, by 0.3203 in Avg. (0.8457 versus 0.5254). Compared with commercial models, it achieves the highest Layout score and approaches GPT Image 2 in Text (0.8260 versus 0.8292). Its Avg. ranks second only to GPT Image 2, ahead of Seedream-5.0-Pro, Qwen-Image-3-Pro, and Nano Banana 2. Despite its substantial Aesthetics advantage over open-source baselines, a gap remains to commercial models for improvement.

5.4 Generalization to Other Text-Rich Generation Tasks

We evaluate generalization on BizGenEval (Li et al., 2026), a benchmark of 400 requests for slides, webpages, posters, charts, and scientific figures. It assesses Layout, Attribute, Text, and Knowledge using checklists. With the same Our PE captions, SFT + RL improves the average across these dimensions from Base's 0.6785 to 0.7068 on All checks, and from 0.5790 to 0.6160 on Hard checks (Table 3). Excluding Knowledge, Avg.(3) improves from 0.6740 to 0.7093 on All and from 0.5840 to 0.6213 on Hard. Gains on All checks are concentrated in Attribute and Text, supporting better realization of the visual and textual requirements beyond document pages (Appendix C.5).

Table 2: DocPageBench results. Avg. averages the three reported dimensions. Bold and underlining mark first and second place, excluding GT. Metric components and coverage are in Table 5.

Model	Input	Text ↑	Layout ↑	Aesthetics ↑	Avg. ↑
GPT Image 2	Original	**0.8292**	0.7542	**0.9921**	**0.8585**
Seedream-5.0-Pro	Original	0.7874	0.7396	<u>0.9133</u>	0.8134
Qwen-Image-3-Pro	Original	0.7966	0.7384	0.8952	0.8101
Nano Banana 2†	Original	0.7221	0.7055	0.8669	0.7648
Qwen-Image-2512‡	Original	0.5238	0.5504	0.5020	0.5254
Z-Image‡	Original	0.4765	0.5355	0.3660	0.4593
FLUX.2-dev	Original	0.3534	0.5060	0.6318	0.4971
HunyuanImage-3†	Original	0.4167	0.5266	0.5399	0.4944
Base	Original	0.5818	0.6514	0.5203	0.5845
SFT	Original	0.5869	0.7156	0.4945	0.5990
SFT + RL	Original	0.6270	0.7325	0.5855	0.6483
Base	Official PE	0.6832	0.6708	0.5965	0.6502
Base	Our PE	0.7228	0.7042	0.5556	0.6609
SFT	Our PE	0.7853	<u>0.8629</u>	0.7525	0.8002
SFT + RL	Our PE	<u>0.8260</u>	**0.8686**	0.8425	<u>0.8457</u>
GT reference	—	0.8921	0.9306	0.9730	0.9319

†Resolution limits; ‡prompt-length or positional-encoding extensions (Appendix C.3).

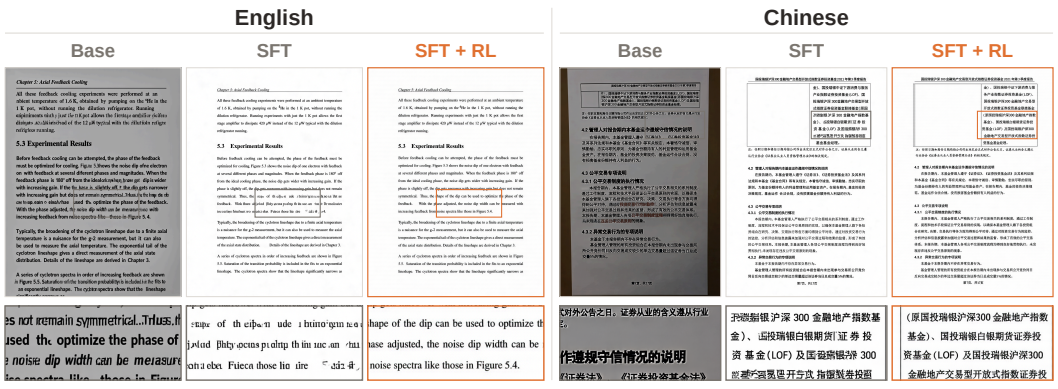


Figure 5: English (left) and Chinese (right) P4 examples, with full pages above matching crops. Base uses Official PE, while SFT and SFT + RL use the same Our PE captions for each example.

Table 3: BizGenEval results. All uses all 20 checks per image; Hard uses the 10 difficult checks on the same images. Avg.(4) averages all four capability dimensions; Avg.(3) averages Layout, Attribute, and Text. Higher scores are better; bold marks the highest reported score in each column.

		All		Hard	
Model	Input	Avg.(4)	Avg.(3)	Avg.(4)	Avg.(3)
Base	Official PE	0.6878	0.7003	0.5765	0.6013
Base	Our PE	0.6785	0.6740	0.5790	0.5840
SFT + RL	Our PE	**0.7068**	**0.7093**	**0.6160**	**0.6213**

6 Analysis

We examine SFT initialization and region-level policy updates (Section 6.1), legibility versus OCR rewards (Section 6.2), and the complementarity between RL and test-time scaling (Section 6.3). Together, these analyses clarify the sources of training gains and the scope for further improvement.

6.1 Necessity of SFT and Credit Assignment

To test whether direct RL can replace the SFT stage, we initialize regional RL from Base and SFT and train both runs for 200 updates using the same legibility reward. As shown in Figure 6, SFT initialization yields a legibility gain of 0.1320 and increases the L3 (clean) share by 15.9%, compared with 0.0739 and 2.9% from Base. Even after RL, Base reaches only 0.3330, below the SFT score of 0.4239 before RL. Document SFT therefore provides both a stronger starting point and larger subsequent RL gains; direct RL does not reproduce these benefits within the current budget.

To evaluate spatial credit assignment, we replace regional updates with page-level updates from the same SFT checkpoint, keeping the legibility reward and other training settings fixed (Appendix D). This variant computes sample-level advantages from the character-weighted geometric mean of regional scores. As shown in Figure 6, page-level updates yield a legibility gain of only 0.0286 after 200 updates, while L3 (clean) remains largely unchanged. These results indicate that reliable regional scores offer limited guidance when reduced to a sample-level reward. Assigning regional feedback to the corresponding image tokens helps guide local refinement more effectively, consistent with earlier uses of spatially resolved visual feedback (Liang et al., 2024; Oh et al., 2026).

(a) Training outcomes

Initialization Updates	SFT Regional	Base Regional	SFT Page-level
Initial reward	0.4239	0.2591	0.4284
Step 200 reward	**0.5559**	0.3330	0.4570
Reward gain	**+0.1320**	+0.0739	+0.0286
Initial L3 (%)	33.5	7.8	35.9
Step 200 L3 (%)	**49.4**	10.7	35.0
ΔL3 (pp)	**+15.9**	+2.9	-0.9

(b) Learning curves

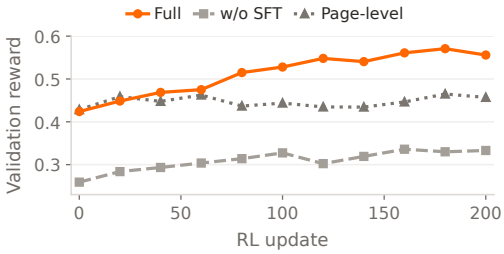


Figure 6: Design ablations over 200 RL updates on the same 256 validation pages. (a) Initial and final rewards and L3 (clean typography) shares, with gains relative to each run's initialization; ΔL3 is in percentage points. (b) Mean VLM validation reward every 20 updates, without smoothing.

6.2 Beyond OCR Recognizability

To evaluate the effectiveness of legibility rewards, we compare them with OCR rewards (Liu et al., 2026) under the same regional RL setup from the SFT checkpoint. Both runs are trained for 200 updates with all other settings fixed and evaluated on the same 256-page validation set. As shown in Figure 7(a,b), the legibility-trained model scores higher on both legibility and OCR coverage. Although OCR coverage is not directly optimized, it reaches 0.9082 versus 0.8997 for OCR-reward training, demonstrating that legibility feedback can outperform direct OCR optimization.

To understand why OCR rewards are less effective, we compare the VLM-assessed regional legibility distributions before and after training (Figure 7(c)). Relative to the initialization, both approaches reduce L1 (unreadable) by ~7%. However, legibility rewards also reduce L2 (imperfect) and increase L3 (clean) by 15.9%, whereas OCR rewards increase L3 by only 6.0% without reducing L2. Further inspection shows that, at update 200, 66.9% of L2 regions in the legibility-trained branch already have OCR recall above 0.95 (Appendix E.3). These findings show that accurate OCR can coexist with visible artifacts, proving OCR alone insufficient to ensure typography approaching real text.

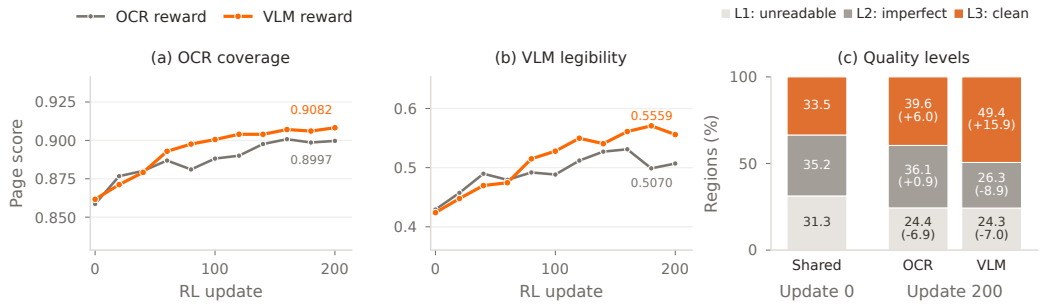


Figure 7: Reward ablation from a common SFT initialization. (a,b) Mean validation scores across pages. (c) Regional legibility distributions (%) from the median of three VLM judgments. Parentheses give percentage-point changes from initialization, computed before rounding (Appendix E).

6.3 Complementarity with Test-Time Scaling

To study how RL interacts with test-time scaling (TTS) (Ma et al., 2025), we compare single outputs during training with best-of-N selection under the same Our PE inputs on DocPageBench, using a fixed VLM selector for both SFT and RL. Figure 8(a) shows that single-output quality progressively approaches that of SFT with a large sampling budget as RL step increases. Under the same TTS protocol, RL@1 reaches 0.8480, close to SFT@64's 0.8507, as SFT's gains taper off at larger N. RL thus achieves comparable quality using only $1/64$ the cost of SFT test-time sampling.

We next examine whether RL substitutes for TTS or the RL model can still benefit from it. As shown in Figure 8(b), increasing the candidate count from 1 to 64 raises the RL model's Overall from 0.8480 to 0.8735. Although additional sampling narrows the SFT–RL gap, RL retains an advantage at every tested budget and leads by 0.0229 at $N = 64$ (Appendix F.2). These gains show useful diversity in the RL model, allowing test-time selection to further improve generation quality.

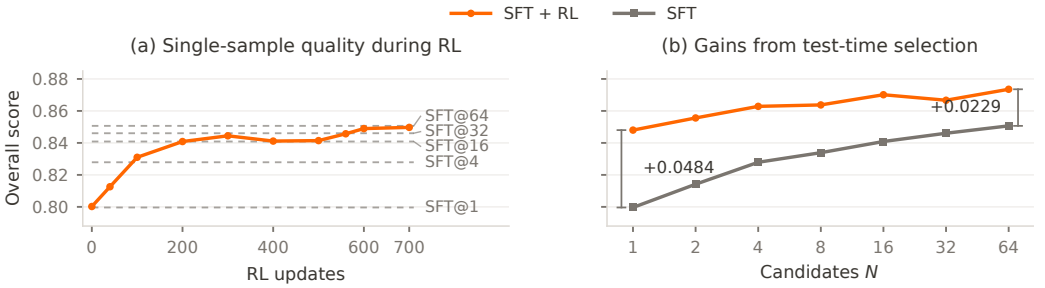


Figure 8: Complementarity between RL and test-time scaling. (a) Single-sample Overall across RL updates; horizontal dashed lines indicate SFT best-of-N selection. (b) SFT and update-560 RL scores with increasing candidate counts; brackets show RL–SFT differences at $N = 1$ and $N = 64$. Selection details and full results for all checkpoints and sampling budgets appear in Appendix F.

7 Conclusion

Our document post-training framework combines OCR-grounded supervision from DocPage-1M with VLM legibility rewards and spatial credit assignment. With adapted inputs, our model ranks second only to GPT Image 2 overall on DocPageBench and generalizes to other text-rich tasks. Analyses highlight the value of SFT initialization and preserving local quality differences in rewards and policy updates. RL reduces repeated sampling while retaining useful diversity for test-time gains. These findings offer a practical route to document generation with text-to-image models.

AI USE STATEMENT

We used AI tools to assist with language polishing, data organization, and the creation and modification of scientific figures and images. These tools helped improve the clarity and fluency of the manuscript and organize experimental results for presentation. They also assisted in preparing plotting code and refining figure layouts, labels, and visual consistency.

REFERENCES

Alibaba Cloud. Qwen-Image-3-Pro. https://modelstudio.console.alibabacloud.com/ap-southeast-1/model/market/detail/qwen-image-3.0-pro, 2026.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Black Forest Labs. FLUX.2 [dev]. https://huggingface.co/black-forest-labs/FLUX.2-dev, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pp. 386–402. Springer, 2024.

Sixiang Chen, Jianyu Lai, Jialin Gao, Tian Ye, Haoyu Chen, Hengyu Shi, Shitong Shao, Yunlong Lin, Song Fei, Zhaohu Xing, Yeying Jin, Junfeng Luo, Xiaoming Wei, and Lei Zhu. Postercraft: Rethinking high-quality aesthetic poster generation in a unified framework. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 64080–64122, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/68d88dcd1e1917c74993902073f08e40-Paper-Conference.pdf.

Mingxuan Cui, Jingpu Yang, Fengxian Ji, Qian Jiang, Zhecheng Shi, Jiaming Wang, Zirui Song, Fajri Koto, and Xiuying Chen. Textalign: Preference alignment for text rendering with hierarchical rewards. *arXiv preprint arXiv:2605.19320*, 2026.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. Sensenova-u1. 5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026.

Google DeepMind. Gemini 3 Flash. https://deepmind.google/models/model-cards/gemini-3-flash/, 2025.

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026a.

Google DeepMind. Gemini 3.1 Flash Image (Nano Banana 2). https://deepmind.google/models/model-cards/gemini-3-1-flash-image/, 2026b.

Hengrui Kang, Zhuangcheng Gu, Zhiyuan Zhao, Zichen Wen, Bin Wang, Weijia Li, and Conghui He. Omnidoclayout: Towards diverse document layout generation via coarse-to-fine llm learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3208–3218, 2026.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, et al. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Yan Li, Zezi Zeng, Ziwei Zhou, Xin Gao, Muzhao Tian, Yifan Yang, Mingxi Cheng, Qi Dai, Yuqing Yang, Lili Qiu, et al. Bizgeneval: A systematic benchmark for commercial visual content generation. *arXiv preprint arXiv:2603.25732*, 2026.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 19401–19411. IEEE, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Yoonjin Oh, Yongjin Kim, Hyomin Kim, Donghwan Chi, and Sungwoong Kim. Ospo: Object-centric self-improving preference optimization for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7620–7631, 2026.

OpenAI. GPT Image 2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026.

Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, and Yuhui Yuan. Bizgen: Advancing article-level visual text rendering for infographics generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23615–23624. IEEE, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xincheng Shuai, Ziye Li, Henghui Ding, and Dacheng Tao. Glyphprinter: Region-grouped direct preference optimization for glyph-accurate visual text rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7674–7683, 2026.

Tencent Hunyuan Foundation Model Team. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, volume 2024, pp. 56783–56799, 2024.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Handong Zheng, Yumeng Li, Kaile Zhang, Liang Xin, Guangwei Zhao, Hao Liu, Jiayu Chen, Jie Lou, Qi Fu, Rui Yang, et al. Multimodal ocr: Parse anything from documents. *arXiv preprint arXiv:2603.13032*, 2026.

A DocPage-1M Documentation

A.1 Annotation Format and Caption Construction

Regional records. Section 3.1 defines the normalized regional records used in construction. For picture regions occupying more than 2% of the page area, a VLM produces a content description v_i. The final dense caption contains a global composition description followed by element descriptions, including content, position, alignment, and typographic attributes such as font style, size, weight, color, and spacing.

References and refill. The page-level VLM uses indexed placeholders for ordinary text and captioned picture regions. Each reference is replaced with the corresponding category marker, content, and normalized bounding box, with coordinates serialized to two decimal places. Ordinary OCR text is cleaned of formatting markup before insertion; picture references receive their crop descriptions. The layout prompt also permits table contents and isolated stylized characters to be described directly. These passages are generated by the VLM and do not use indexed refill.

Cleaning and length filters. The cleaning stage removes model preambles and rejects descriptions without the expected composition heading or content references. It also checks for repeated references and references to empty or missing blocks. Captions longer than 24,000 characters are discarded before SFT. Before caption construction, we remove pages with at most 150 characters in the stored OCR output. This text can include table markup, so the filter uses a different count from the visible-character statistics. Refill preserves the source transcription of referenced text, including any remaining OCR errors.

A.2 Dataset Statistics and Page Categories

Population and text counts. Statistics use all 969,808 examples in DocPage-1M. Source PDFs are identified by the parent directories of their page images, with the renderer's trailing `pages` directory removed. Visible-character counts exclude HTML tags and decode HTML entities in the OCR text. Picture descriptions are excluded from these counts; remaining Markdown and formula notation is retained. Caption length and page-text length are measured separately.

Page language. We infer the page language from its OCR text after HTML cleanup. A page is assigned to English when its number of Latin words is at least its number of CJK ideographs, and to Chinese otherwise. This produces 483,044 English pages and 486,764 Chinese pages. The classification uses the page content rather than the caption-language label or the source collection.

Page content. We assign the three content categories in the order picture, table, text. A page is counted as *With picture* if at least one OCR Picture block occupies more than 2% of the page area. Among the remaining pages, those with an OCR Table block are *With table*; all others are *Text*. Text pages may therefore contain smaller pictures. The 30,466 pages containing both a qualifying picture and a table are assigned to With picture. Table membership is determined from the OCR records.

Regional annotations. Region counts refer to parsed boxes in the training captions, including text, table, formula, and image regions. They measure the spatial annotations provided to the generator. Raw OCR blocks are counted separately and are not used for this statistic.

A.3 Statistics for Dataset Comparison

In Table 1, character counts sum the available text regions per image after removing HTML tags and decoding entities, retaining source whitespace, punctuation, and digits. Mean text counts are rounded to the nearest integer in the table. Missing text annotations are excluded rather than counted as zero; explicitly empty annotations contribute zero. AnyWord counts readable regions without additionally filtering by the `valid` flag, and MARIO counts all supplied OCR entries without a confidence threshold. No OCR is rerun. Measured image sizes are separate medians of stored image width and height. For OmniDocLayout, we report the median page dimensions recorded in the official layout metadata, 1084×1509. Its layout annotations do not provide OCR transcriptions,

so the mean text count is unavailable. For Infographics-650K, the original paper reports a fixed resolution of 2240×896 for all samples, which we include as a paper-reported value. The training images are private, and the paper reports character distributions per text layer rather than a mean per image; its mean text count is therefore also marked unavailable. DocPage-1M uses the full-dataset character statistics described in Appendix A.2; its image width and height are medians from a random sample of 50,000 pages with seed 42.

The Real column describes sample provenance, independently of public image availability. MARIO-10M collects existing LAION images, movie posters, and book covers; AnyWord-3M uses existing web images and OCR datasets. TextAtlas5M combines real-image and synthetic-image splits, so its Real entry is partial. Infographics-650K synthesizes new images through layer replacement and text re-rendering, although its templates come from existing infographics. OmniDocLayout extracts layouts from collected documents, so its Real entry is positive despite its layout-only release. DocPage-1M renders existing PDF pages while preserving their content. Automatically generated annotations do not make the underlying samples synthetic.

Visual refers to explicit descriptions of typography, style, or page composition; generic image captions and aesthetic scores alone do not qualify. Query includes captions or prompts used to condition image generation, without requiring free-form user instructions. For Real, Visual, and Query, a check denotes full coverage, a diamond partial coverage, and a cross absence. Multilingual (Multiling.) denotes target text in two or more languages across the dataset, independently of caption language.

B Training Details

B.1 Supervised Fine-Tuning

SenseNova-U1.5-8B-MoT-Preview is a unified model for understanding and generation with a Mixture-of-Transformers (MoT) architecture. SFT updates all parameters, including the language backbone, generation branch, and pixel head.

Resolution and sequence packing. Training preserves page aspect ratios and limits image area to $[768^2, 2048^2]$ pixels. Each image contributes between 576 and 4,096 image tokens. Image–text sequences are packed into 28,672-token windows, with a global batch of 32 packed sequences per update.

Optimization. SFT uses only text-to-image pairs and the flow-matching objective in Equation 1. We train all model parameters in BF16 for 20,000 updates using Adam with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and no weight decay. The learning rate increases over 600 warmup updates to 5×10^{-5} and then follows cosine decay to 2.5×10^{-5} at the final update. The global gradient norm is clipped to 1.0. The training seed is 42.

Offline checkpoint averaging. We form the SFT policy from the base weights and checkpoints saved every 1,000 updates. Let θ_0 denote the base model and θ_j the checkpoint at update $1000j$, for $j = 1, \ldots, N$ with $N = 20$. With a checkpoint decay factor of $\beta = 0.9999^{1000} \approx 0.904833$, the averaged weights are

$$\theta_{\mathrm{SFT}} = \beta^N \theta_0 + (1 - \beta) \sum_{j=1}^{N} \beta^{N-j} \theta_j. \tag{6}$$

The averaging coefficients sum to one, with the base model contributing approximately 13.53%. We accumulate the average in FP32 and cast it once to BF16 when saving. The resulting model is used for SFT evaluation and serves as both the RL initialization and the frozen reference during RL.

B.2 Regional Reward Implementation

Region selection and crops. We score text blocks with more than 50 characters after stripping surrounding whitespace. Each bounding box is expanded by 12% of its width and height per side, with a minimum margin of 0.012 in normalized coordinates, and clipped to the image boundary. This margin accommodates glyphs extending slightly beyond the specified box. The expanded box is also used for spatial credit assignment. Crops are encoded as PNG images with the longer side capped at 1,600 pixels.

Judge calls and aggregation. The judge uses temperature zero and a zero thinking budget, returning a quality level in JSON format. We request three judgments per crop and retry invalid responses. The regional reward is the mean of the mapped scores from successful judgments, while the lower median of their quality levels is retained for analysis. Regions with no successful judgments are omitted from page aggregation. A page with no scored regions receives a monitoring score of 1.0. Policy optimization uses regional rewards with valid scores for all group members, as described in Section 4.3.

B.3 Reinforcement Learning

RL updates approximately 8.17B parameters in the generation branch and its head and embedding modules. Training uses approximately 100K document prompts, with 256 held-out pages for validation.

Rollouts and stochastic windows. Each update uses 64 captions with eight generated pages per caption, for 512 pages in total. Trajectories use 20 denoising steps. The two-step stochastic window is placed independently for each page within the first 12 steps, with its start sampled uniformly from eligible indices. Stochastic transitions require $t \geq 10^{-2}$, where time runs from noise at $t = 0$ to the

clean image at $t = 1$. The remaining steps use deterministic ODE updates. Classifier-free guidance is fixed at 4.0 during both rollout and transition re-evaluation. The timestep shift is 3.0, the training noise level is 0.3, and the denominator floor in the velocity calculation is 0.02. Validation uses deterministic ODE sampling with zero stochastic noise.

Region selection and spatial maps. A region is retained only if every group member has a valid score and the within-group standard deviation of its rewards exceeds 10^{-3}. The numerical stabilizer in Equation 3 is $\delta = 10^{-4}$. For a page of size $H \times W$, expanded scoring boxes are rasterized onto the $H/32 \times W/32$ image token grid, using the same box boundaries as the regional judge.

Policy updates. We use sharded FP32 AdamW with a constant learning rate of 2×10^{-5} and two optimization epochs per rollout batch. The recorded rollout log-densities remain fixed across both epochs. With $\epsilon = 10^{-3}$, policy ratios are clipped to $[1 - \epsilon, 1 + \epsilon]$. Token advantages are clipped to $[-A_{\max}, A_{\max}]$ with $A_{\max} = 5$ after spatial aggregation and batch-scale normalization.

Reference penalties. For each rollout, two reference-regularization points are sampled uniformly from the full eligible trajectory. At each point, we compute mean squared differences between the current policy and the frozen SFT model's clean-image predictions, separately for the conditional and unconditional branches. Each branch has coefficient 10^{-5}. The penalties are averaged over valid image tokens, rollout samples, and sampled points. The reference weights remain fixed throughout RL.

C Evaluation Protocols

C.1 DocPageBench Construction

Source documents and sampling. We exclude all document directories represented in the document SFT set, then retain at most one page per remaining document. Sampling is stratified by language, page category, and input condition. The 24 combinations contain 25 text-only pages per cell and 12 or 13 picture/table pages per cell, with complementary allocations across languages. Categories are mutually exclusive: a page with pictures is assigned to the picture category; otherwise, a page with tables is assigned to the table category. The resulting benchmark has 400 pages, a median of 1,402.5 OCR characters, and a median of nine blocks per page. Each condition contains 50 English and 50 Chinese pages.

Construction independent of the training caption format. The benchmark pipeline takes page images and raw OCR records as inputs; it never reads the dense captions used for training. A VLM supplies short topics, keywords, summaries, and user-oriented phrasing. A deterministic assembler inserts the text, coordinates, counts, and spatial relations appropriate to each condition from the OCR records and their geometry. Requests use four families of templates: descriptive statements, imperative instructions, specification sheets, and conversational requests. Element phrasing and field order also vary. Thus, the original benchmark requests do not require the generator to follow the training caption template, although both pipelines draw on PDF pages and OCR annotations.

Controlled disclosure. Table 4 summarizes the four request conditions. P3/P4 provide the requested text verbatim. P1/P2 partially disclose it using block-level character-count ranges, topics, keywords, structural descriptions, summaries, or short prefixes. Prefix disclosure is limited to 30% of the source block. P2/P4 additionally supply normalized rectangles. P1/P3 express layout through natural-language positions, alignment, reading order, and relations between elements. The evaluator records the requirements actually included in each request, and does not require recovery of undisclosed source text or geometry.

Table 4: DocPageBench request conditions defined by text disclosure and the availability of explicit coordinates.

Condition	P1	P2	P3	P4
Text disclosure	Partial	Partial	Verbatim	Verbatim
Explicit coordinates	No	Yes	No	Yes
Pages	100	100	100	100

C.2 Scoring Rules and Validation

Text fidelity under full disclosure. For P3/P4, each reference block is matched against the concatenated full-page OCR text using normalized partial-string similarity from RapidFuzz (`fuzz.partial_ratio`). Block scores are weighted by reference character count. This measures OCR-based text fidelity without depending on the OCR system's block segmentation.

Content requirements under partial disclosure. For P1/P2, the score combines character-count agreement, topic agreement, natural-language plausibility, and required-keyword coverage, with nominal weights 0.40, 0.20, 0.15, and 0.25. Only applicable terms enter the weighted mean, and their weights are renormalized. Counts are compared with the rounded quantities stated in the prompt, and keyword coverage checks explicit keyword requirements. The default P1 protocol omits the count term. The overall Text column averages per-page scores across both disclosure settings; it therefore summarizes two distinct tasks.

Position and layout claims. For P2/P4, each specified rectangle is expanded by 10% of its own width and height per side and cropped from the generated image. Crop coordinates are computed from that image's actual dimensions. Scoring checks the information supplied for the region: text or prefix presence, stated character counts, required keywords, or a VLM judgment of the disclosed

topic, structure, or summary. Picture regions use a pixel-content check. For P1/P3, the original request is converted once into a fixed numbered checklist, shared across all model arms. As in question-based image–text evaluation (Lin et al., 2024), a VLM evaluates the generated page against these same layout claims. Position and Claims are averaged over their respective available condition subsets. The main Layout score is their unweighted arithmetic mean, computed before rounding. Each component has weight 0.5, including when the valid page counts differ. Raw components are reported in Table 5; neither they nor the aggregate are divided by or rescaled using GT reference scores.

Aesthetics. The judge checks seven document-quality defects: rendering artifacts, inconsistent fonts or weights, imbalanced margins, disordered alignment, overflow or clipping, unsuitable colors, and an unrealistic typeset appearance. Each violation subtracts 0.15 from a score of 1, with the result clipped below at 0. The same checklist is used for all models.

Reference scores and validation. Source pages are evaluated with the same OCR and judges, yielding GT reference scores of 0.8921 for Text, 0.9306 for Layout, and 0.9730 for Aesthetics; the Layout components are 0.9398 for Position and 0.9214 for Claims. These scores characterize the evaluator's behavior on real pages and are not used to normalize any reported metric. Generated pages can exceed these references, particularly when their appearance is cleaner than that of the source scan. Prompt-achievability checks use synthetic OCR satisfying the disclosed requirements to detect penalties for information absent from the prompt. Displacement checks test whether the spatial scorer responds to errors: in the recorded validation, shifting a region by one block height reduces crop recall from 0.960 to 0.335.

C.3 Inference, Input Adaptation, and Coverage

Checkpoints and generation. Base is `SenseNova-U1.5-8B-MoT-Preview`. SFT is the offline exponential average of saved checkpoints from the 20,000-update full-parameter SFT run in Section 4.1. SFT + RL is the checkpoint at update 560 of the regional RL procedure in Section 4.3, initialized from SFT. All three use 50 Euler steps, classifier-free guidance 4.0, timestep shift 3.0, and seed 42. With this seed convention, examples of the same dimensions share the initial noise realization. Each request produces one image for the main comparison.

Our prompt adaptation. We use `gemini-3.1-pro-preview` at temperature 0.3 with one English and one Chinese demonstration. The instruction asks for a caption in the request's language, with an overall composition followed by element-level descriptions. It is instructed to preserve supplied text, complete missing content according to the request, and write all coordinates to two decimal places. When no coordinates are provided, the adapter chooses a layout and supplies boxes. Generated captions are checked for the expected caption structure, language, and coordinate coverage, and retried when checks fail. The same saved captions are used for Base, SFT, and SFT + RL.

Official prompt enhancement. Official PE follows the base model's `-enhance` path with the infographic style. We use its released `templates/infographic_system.md` verbatim as the system instruction and supply the original request as the user message. The template requests a 400–600-word prose description. Since it does not prescribe an executor model, we use the same `gemini-3.1-pro-preview` and temperature 0.3 as Our PE. Generation settings are otherwise unchanged, including the default `cfg_norm=none`. We report this comparison on Base.

Dimensions and baseline constraints. Target dimensions follow the source page's aspect ratio, subject to a maximum edge of 3,840 pixels and a pixel area between 0.7 and 7.9 million, then are rounded to each model's required granularity. Our checkpoints use multiples of 32; GPT Image 2 uses multiples of 16 and high-quality mode. The target linear dimensions from these two granularities differ by at most 1.4%. API and resolution-bucket restrictions can change actual output dimensions: Nano Banana 2 outputs approximately 39% of the target area; HunyuanImage-3 has a reported linear-size ratio of approximately 0.587, with aspect-ratio error exceeding 1% on 299 of 400 pages. Seedream-4.5 enlarges 213 of 400 target pages to meet its minimum area of 3.69 million pixels.

Long-input settings. For Qwen-Image and Qwen-Image-2512, the prompt token limit is increased from 1,024 to 3,900 using the existing positional encoding range. This avoids truncation on all 400 requests. For Z-Image and Z-Image-Turbo, the caption rotary-position table is extended from 1,536 to 3,264 entries to accommodate the benchmark requests. These settings apply throughout the main and appendix comparisons.

OCR recovery. Malformed OCR outputs are retried at temperature 0.1, followed by up to two attempts at 0.9 if parsing still fails. Remaining JSON errors are handled through field-level extraction or truncation of a repeating tail. Valid outputs are retained unchanged. No available page in the reported results is excluded because of an OCR parsing failure. Across the main and additional model comparisons, 15 pages require repetition truncation and one requires field-level recovery. Repetition truncation retains only the transcription before the loop.

Scoring coverage. The evaluator uses `dots.mocr` for OCR and `gemini-3.8-flash` for VLM judgments. Table 5 reports the number of available valid scoring records for each metric. Means are computed over those records; missing records are not assigned zero. Five API baselines return fewer than 400 images, with coverage ranging from 392 to 399 pages. Seedream-4.5 also has two images without valid Aesthetics scores. Separately, layout-claim judgments remain unavailable for four Z-Image pages and one page each for Qwen-Image, SFT with Our PE, and SFT + RL with Our PE. For crop-based scoring, a region without a valid judgment is omitted from that page's aggregation.

C.4 Score Breakdowns

All scores are reported to four decimal places using round-half-up. Text and Aesthetics aggregate subgroup means using the valid counts in Table 5; Layout equally weights its two components. Avg. averages the three reported dimension scores.

Additional baselines. Table 6 reports the remaining model variants under the same evaluation protocol. The main table retains Qwen-Image-3-Pro, Qwen-Image-2512, Seedream-5.0-Pro, and Z-Image as representatives of their respective model lines. All reported variants remain in the component, coverage, and subgroup tables.

Text under different disclosure conditions. The similar overall Text scores of SFT + RL with Our PE and GPT Image 2 combine different outcomes across conditions. On P1/P2, their mean content-requirement scores are 0.7097 and 0.6783, respectively. On P3/P4, their mean verbatim-fidelity scores are 0.9422 and 0.9801. Each comparison covers 200 pages per model. Our model therefore scores higher under partial disclosure, while GPT Image 2 has higher verbatim fidelity. Table 7 reports all four conditions.

Layout components. Table 5 separates Position from Claims. SFT + RL with Our PE scores 0.8115 in Position, compared with GPT Image 2's 0.5824, while their Claims scores are numerically close (0.9258 and 0.9259). The advantage in aggregate Layout therefore comes primarily from the coordinate-based component, which checks requested content at specified locations.

Language and page categories. Tables 8 and 9 report language and content breakdowns. SFT + RL with Our PE scores 0.8170 in English Text and 0.8350 in Chinese Text. Its visual quality is less balanced: Aesthetics is 0.7818 for English and 0.9033 for Chinese. This difference remains despite RL improving both languages over SFT, whose corresponding scores are 0.6828 and 0.8223.

Additional inference references. Table 10 reports two supplementary settings. Base with `-think` uses the original requests with reasoning enabled. We report this setting only for Base. The GT-caption setting supplies SFT + RL with dense captions derived from the source pages. It scores higher in Text than Our PE (0.8471 versus 0.8260), with similar aggregate Layout. Since these captions use source-page information beyond the user request, this setting serves as a reference and is excluded from the main model comparison.

Table 5: Layout components and valid scoring counts. Layout equally averages Position and Claims, regardless of their valid counts. The lower panel lists subgroup counts for incomplete configurations. All others have 100 Text records per P1–P4 condition, 200 Text and Aesthetics records per language, and 200/100/100 Text records for text-only/table/picture pages. EN/ZH denotes English/Chinese.

		Layout components		Valid pages		
Model	Input	Position	Claims	Text / Aesthetics	Position	Claims
GPT Image 2	Original	0.5824	0.9259	400 / 400	200	200
Seedream-5.0-Pro	Original	0.5635	0.9157	398 / 398	200	198
Qwen-Image-3-Pro	Original	0.5520	0.9248	392 / 392	194	198
Qwen-Image-3	Original	0.5398	0.9289	393 / 393	194	199
Nano Banana 2	Original	0.5003	0.9106	399 / 399	200	199
Seedream-4.5	Original	0.3942	0.7941	399 / 397	200	199
Qwen-Image-2512	Original	0.3388	0.7621	400 / 400	200	200
Qwen-Image	Original	0.2325	0.6666	400 / 400	200	199
Z-Image	Original	0.3357	0.7352	400 / 400	200	196
Z-Image-Turbo	Original	0.3135	0.6619	400 / 400	200	200
FLUX.2-dev	Original	0.2947	0.7174	400 / 400	200	200
HunyuanImage-3	Original	0.2929	0.7603	400 / 400	200	200
Base	Original	0.4611	0.8417	400 / 400	200	200
SFT	Original	0.6164	0.8148	400 / 400	200	200
SFT + RL	Original	0.6516	0.8135	400 / 400	200	200
Base	Official PE	0.4706	0.8710	400 / 400	200	200
Base	Our PE	0.5136	0.8948	400 / 400	200	200
SFT	Our PE	0.7975	0.9282	400 / 400	200	199
SFT + RL	Our PE	0.8115	0.9258	400 / 400	200	199
Base	Original + think	0.4397	0.8288	400 / 400	200	200
SFT + RL	GT caption	0.8288	0.9069	400 / 400	200	200
GT reference	—	0.9398	0.9214	400 / 400	200	200

Model	Text P1/P2/P3/P4	Text EN/ZH	Text Text/Table/Picture	Aesthetics EN/ZH
Seedream-5.0-Pro	99/100/99/100	199/199	199/100/99	199/199
Qwen-Image-3-Pro	99/97/99/97	197/195	195/99/98	197/195
Qwen-Image-3	100/97/99/97	198/195	195/99/99	198/195
Nano Banana 2	99/100/100/100	199/200	199/100/100	199/200
Seedream-4.5	99/100/100/100	199/200	199/100/100	199/198

Table 6: Additional baseline results on DocPageBench. Definitions and aggregation follow Table 2; inference constraints and valid scoring counts are detailed in Appendix C.3.

Model	Input	Text	Layout	Aesthetics	Avg.
Qwen-Image-3	Original	0.7549	0.7344	0.8458	0.7784
Seedream-4.5	Original	0.6828	0.5941	0.6592	0.6454
Qwen-Image	Original	0.3910	0.4496	0.4209	0.4205
Z-Image-Turbo	Original	0.4416	0.4877	0.5880	0.5058

Table 7: Text scores by original request condition. P1/P2 partially specify text; P3/P4 supply it verbatim. P2/P4 include explicit coordinates. Each condition has 100 target pages; scores average available records. The supplementary input references are described below.

Model	Input	P1	P2	P3	P4
GPT Image 2	Original	0.7332	0.6233	0.9810	0.9791
Seedream-5.0-Pro	Original	0.6960	0.5674	0.9470	0.9400
Qwen-Image-3-Pro	Original	0.6901	0.6103	0.9407	0.9444
Qwen-Image-3	Original	0.6432	0.5506	0.9062	0.9199
Nano Banana 2	Original	0.5518	0.4175	0.9681	0.9495
Seedream-4.5	Original	0.5734	0.4376	0.8744	0.8447
Qwen-Image-2512	Original	0.3696	0.2926	0.7754	0.6574
Qwen-Image	Original	0.3035	0.2607	0.5430	0.4569
Z-Image	Original	0.3062	0.2620	0.6913	0.6464
Z-Image-Turbo	Original	0.3062	0.2603	0.6094	0.5905
FLUX.2-dev	Original	0.2986	0.2609	0.4541	0.4002
HunyuanImage-3	Original	0.3333	0.3087	0.5394	0.4854
Base	Original	0.3131	0.2814	0.8713	0.8615
SFT	Original	0.2969	0.3074	0.8594	0.8838
SFT + RL	Original	0.3066	0.3377	0.9328	0.9307
Base	Official PE	0.5868	0.4108	0.8642	0.8709
Base	Our PE	0.6126	0.5537	0.8644	0.8607
SFT	Our PE	0.6567	0.6608	0.9208	0.9029
SFT + RL	Our PE	0.7228	0.6967	0.9528	0.9316
Base	Original + think	0.4094	0.3739	0.8261	0.8300
SFT + RL	GT caption	0.7418	0.7647	0.9409	0.9411
GT reference	—	0.7767	0.8124	0.9881	0.9912

Table 8: Text scores by language and page category. Each language has 200 target pages; text-only, table, and picture categories contain 200, 100, and 100 pages, respectively. Scores average available records.

Model	Input	English	Chinese	Text-only	Table	Picture
GPT Image 2	Original	0.8278	0.8305	0.8308	0.8302	0.8249
Seedream-5.0-Pro	Original	0.7846	0.7902	0.7882	0.7885	0.7847
Qwen-Image-3-Pro	Original	0.7996	0.7935	0.7856	0.8171	0.7977
Qwen-Image-3	Original	0.7817	0.7277	0.7204	0.8039	0.7739
Nano Banana 2	Original	0.7395	0.7049	0.7348	0.7257	0.6934
Seedream-4.5	Original	0.6716	0.6939	0.7034	0.6552	0.6694
Qwen-Image-2512	Original	0.5279	0.5196	0.5274	0.5317	0.5085
Qwen-Image	Original	0.4297	0.3523	0.3704	0.4329	0.3904
Z-Image	Original	0.5208	0.4322	0.4766	0.4906	0.4621
Z-Image-Turbo	Original	0.4796	0.4035	0.4263	0.4640	0.4496
FLUX.2-dev	Original	0.4158	0.2911	0.3510	0.3631	0.3487
HunyuanImage-3	Original	0.4298	0.4036	0.3992	0.4688	0.3996
Base	Original	0.5739	0.5898	0.5751	0.5833	0.5938
SFT	Original	0.5832	0.5906	0.5776	0.5964	0.5960
SFT + RL	Original	0.6306	0.6233	0.6243	0.6326	0.6267
Base	Official PE	0.6705	0.6958	0.6869	0.6568	0.7021
Base	Our PE	0.7373	0.7084	0.7133	0.7434	0.7214
SFT	Our PE	0.7818	0.7889	0.7768	0.8064	0.7812
SFT + RL	Our PE	0.8170	0.8350	0.8339	0.8170	0.8192
GT reference	—	0.8969	0.8872	0.8908	0.8846	0.9020

Table 9: Aesthetics scores by language, averaged over available records. Each language has 200 target pages.

Model	Input	English	Chinese
GPT Image 2	Original	0.9888	0.9955
Seedream-5.0-Pro	Original	0.8666	0.9601
Qwen-Image-3-Pro	Original	0.8645	0.9262
Qwen-Image-3	Original	0.8250	0.8669
Nano Banana 2	Original	0.8402	0.8935
Seedream-4.5	Original	0.6013	0.7174
Qwen-Image-2512	Original	0.4563	0.5478
Qwen-Image	Original	0.3843	0.4575
Z-Image	Original	0.3508	0.3813
Z-Image-Turbo	Original	0.5653	0.6108
FLUX.2-dev	Original	0.5808	0.6828
HunyuanImage-3	Original	0.5035	0.5763
Base	Original	0.4380	0.6025
SFT	Original	0.4218	0.5673
SFT + RL	Original	0.5303	0.6408
Base	Official PE	0.5230	0.6700
Base	Our PE	0.4923	0.6190
SFT	Our PE	0.6828	0.8223
SFT + RL	Our PE	0.7818	0.9033
GT reference	—	0.9723	0.9738

Table 10: Supplementary inference settings on 400 pages each. GT caption denotes a source-page dense caption used as model input; GT reference elsewhere denotes scoring the source image itself.

Model	Input	Text	Layout	Aesthetics	Avg.
Base	Original + think	0.6098	0.6343	0.5346	0.5929
SFT + RL	GT caption	0.8471	0.8678	0.8568	0.8572

C.5 BizGenEval Protocol and Additional Results

Data and scoring. We use all 400 BizGenEval requests, with 80 per document type and 100 per capability dimension. Each request has 20 binary checklist questions, including 10 designated as hard. Following the official implementation, the per-image score is $\max(0, 1 - \alpha e)$, where e is the number of failed checks and α is 0.1 for All and 0.2 for Hard. The score is zero when at most half the checks are satisfied. We average within each capability dimension, then report their macro-average as Avg.(4). Avg.(3) averages Layout, Attribute, and Text, covering 300 requests. All systems use `google/gemini-3-flash-preview` through OpenRouter at temperature 0.0 with the official checklists. Table 11 gives the complete dimensional scores.

Generation and input adaptation. All three systems use 50 Euler steps, classifier-free guidance 4.0, timestep shift 3.0, and seed 42. Requested aspect ratios are mapped to the model's resolution buckets, with the same dimensions used across systems. We use 4:3 for the 120 chart or scientific-figure requests without an aspect ratio. Official PE uses the released infographic template described in Appendix C.3; Our PE uses the same system instruction as on DocPageBench, with the target aspect ratio added to the user message. Both run on `gemini-3.1-pro-preview` at temperature 0.3. This shared PE model supports comparing Knowledge across the complete generation pipelines. Base and SFT + RL with Our PE additionally share identical saved captions, controlling the supplied content when comparing checkpoints.

Table 11: BizGenEval scores by capability dimension. All systems have complete coverage: 400 images and 100 requests per dimension.

Model	Input	Layout	Attribute	Text	Knowledge	Avg.(4)	Avg.(3)
All							
Base	Official PE	0.7880	0.5450	0.7680	0.6500	0.6878	0.7003
Base	Our PE	0.8220	0.4730	0.7270	0.6920	0.6785	0.6740
SFT + RL	Our PE	0.8060	0.5610	0.7610	0.6990	0.7068	0.7093
Hard							
Base	Official PE	0.6860	0.4360	0.6820	0.5020	0.5765	0.6013
Base	Our PE	0.7420	0.3820	0.6280	0.5640	0.5790	0.5840
SFT + RL	Our PE	0.7460	0.4360	0.6820	0.6000	0.6160	0.6213

D Design Ablations

D.1 Comparison and Measurement

We compare the first 200 updates of the full recipe with two variants: regional RL initialized from Base, and page-level RL initialized from the same SFT EMA checkpoint as the full recipe. The Base variant changes the weights while preserving the model configuration, tokenizer, and input format. All runs use the same training data, 64 GPUs, and training hyperparameters from Appendix B. Each variant completes 200 updates. There is one training run per configuration.

All three runs use the same regional VLM judge, three requested judgments per crop, and character-weighted geometric mean for the page score. The page-level variant turns off spatial credit assignment and uses the page score to compute a scalar GRPO advantage. The full recipe instead uses regional advantages and spatial weights as defined in Section 4.3. Thus, the comparison keeps the reward service fixed while changing the optimization scheme, including its weighting and loss reduction.

Validation uses deterministic ODE sampling on 256 fixed pages every 20 updates. We verify that the prompts, reference strings, and region geometries match across all runs and updates. Each evaluation contains 1,273 graded regions, with no completely ungraded regions. Across intermediate updates, 16 region evaluations in the Base run and two in the page-level run retain only one or two successful judgments; all regions at updates 0 and 200 have three. We use the saved scores, which average the available mapped judgments before regional aggregation; quality levels use the lower median of the available votes.

D.2 Learning Curves and Regional Quality

Figure 6(b) shows all measured validation points. The regional SFT-initialized run improves substantially over the comparison window. The page-level run fluctuates after its early improvement, while the Base-initialized run improves from a lower starting point. These trajectories describe the observed budget and do not establish the eventual performance of longer runs.

The quality distributions in Table 12 distinguish reductions in unreadability from increases in clean typography. Base-initialized RL reduces the L1 share but leaves most regions in L2. Page-level RL also reduces L1, with little change in L3. The full recipe reduces both L1 and L2 while increasing L3. These are changes in pooled region proportions, rather than measured transitions of individual regions between levels. Regions have equal weight in this analysis, whereas page scores use character-weighted aggregation.

Table 12: Regional quality at initialization and update 200. Every row contains 1,273 regions. L1: unreadable; L2: readable but imperfect; L3: clean typography.

Initialization	Updates	Step	L1 (%)	L2 (%)	L3 (%)
SFT	Regional	0	31.3	35.2	33.5
SFT	Regional	200	24.3	26.3	49.4
Base	Regional	0	52.8	39.4	7.8
Base	Regional	200	38.1	51.2	10.7
SFT	Page-level	0	33.0	31.1	35.9
SFT	Page-level	200	28.7	36.3	35.0

D.3 Optimization Diagnostics and Scope

Training diagnostics averaged over 200 updates appear in Table 13. They are measured on training rollouts (64 prompts with eight candidates each per update), separately from the 256-page validation set. The Base-initialized run has a higher fraction of active regions and greater regional reward spread than the full recipe, so its weaker improvement is not accompanied by an absence of within-group variation. The page-score standard deviations are similar for the regional and page-level SFT runs (0.1326 and 0.1322), but their gradient norms differ. These observations support examining

how rewards enter the update, rather than interpreting the outcome solely through the magnitude of page-score variation.

Table 13: Mean training diagnostics over 200 updates. Active regions are those retained for spatial updates; regional SD measures within-group reward spread. These spatial diagnostics do not apply to page-level RL.

Initialization	Updates	Active regions	Regional SD	Gradient norm
SFT	Regional	73.8%	0.162	4.16×10^{-4}
Base	Regional	83.3%	0.172	4.30×10^{-4}
SFT	Page-level	—	—	9.30×10^{-5}

The regional scheme concentrates weights on text regions and uses a different loss reduction from the page-level scheme, including a window-length factor. These effects have not been separately ablated; the comparison supports the regional update scheme as a whole, rather than a numerical attribution to spatial localization alone. Likewise, the initialization comparison includes both task-specific rendering ability and familiarity with the training input format.

The quality analysis uses the same VLM as the training reward. Unlike the reward comparison in Appendix E, the two design variants have no complete OCR cross-evaluation here. Training logs also do not retain per-page judge failures. Although every reported validation dump has a score for every region, a page whose crops all fail during training receives the configured maximum-score fallback. Such a page has no regional payload and contributes no local policy-gradient term, but its scalar reward can affect page-level updates. We cannot exclude this training-time effect from the available logs.

E Reward Analysis

E.1 Controlled Comparison and Scoring

Both branches start from the SFT checkpoint in Section 4.1 and use the same training data and region-local optimization described in Section 4.3. The comparison changes only the regional reward, retaining the same text-block filter, expanded crop boxes, and spatial credit assignment. We evaluate every 20 updates over the common 0–200-update window. Each branch has one training run; the OCR-reward branch was stopped at update 200.

For the OCR reward, `dots.mocr` recognizes each crop using temperature 0 and a maximum of 4,096 output tokens. We strip HTML tags and whitespace from the reference and recognized text, and use up to 2,000 reference characters. Regional recall is the length of their longest common subsequence divided by the reference length. This penalizes missing characters while allowing extra recognized text, including neighboring lines admitted by the crop margin. The regional reward is floored at 0.1. VLM regional rewards average three mapped judgments as defined in Section 4.2. For both metrics, page scores use a character-weighted geometric mean of regional scores; we then average these page scores over the available validation pages. These validation reward scores are distinct from the DocPageBench Text and Aesthetics metrics.

E.2 Measurement and Coverage

Each branch's training metric comes from its saved validation records. The other metric was obtained by scoring the saved generated images: OCR scores for the VLM-trained branch and VLM scores for the OCR-trained branch. For the latter, we consistently use the completed offline scores, including their regional judgments. These records supersede incomplete auxiliary VLM measurements made during training. All primary and offline scores used here report zero failed regions.

The validation set contains 256 pages. Available cross-evaluation records cover 252–256 pages per branch and update (Table 14); the offline evaluation skips pages whose saved region count does not align with the filtered reference blocks. We use the saved scores without imputing omitted pages. Figure 7(a,b) plots the mean score over each branch's available records.

Table 14: Reward cross-evaluation. VLM and OCR in the column headers identify the training reward. Each metric is averaged over the available pages of that branch.

	Valid pages		OCR coverage		VLM legibility	
Update	VLM	OCR	VLM	OCR	VLM	OCR
0	256	255	0.8616	0.8586	0.4239	0.4292
20	255	255	0.8712	0.8767	0.4477	0.4574
40	254	254	0.8792	0.8801	0.4697	0.4898
60	254	256	0.8929	0.8869	0.4745	0.4793
80	256	256	0.8976	0.8811	0.5149	0.4919
100	256	256	0.9006	0.8882	0.5279	0.4883
120	255	255	0.9040	0.8900	0.5497	0.5119
140	256	256	0.9040	0.8976	0.5405	0.5271
160	256	256	0.9071	0.9008	0.5610	0.5310
180	256	254	0.9061	0.8986	0.5707	0.4988
200	256	252	0.9082	0.8997	0.5559	0.5070

E.3 Regional Quality and OCR Recall Distributions

Quality levels are the median of three VLM judgments for each region. We pool regions without character weighting when computing level proportions, unlike the page-score aggregation above. Both branches generate the same images at update 0, so we use the VLM branch's judgments to define one shared initial distribution for Figure 7(c) and all reported changes in level proportions. It contains 1,273 regions: 31.3% L1, 35.2% L2, and 33.5% L3. The final VLM and OCR distributions contain 1,273 and 1,248 regions, respectively. Parenthesized values in the figure give changes from

this shared baseline in percentage points, calculated before rounding. These are changes in pooled proportions, not individual region transitions.

Figure 9 groups regional OCR recall by VLM quality level at initialization and update 200 in the VLM-trained branch. We use recall before the 0.1 reward floor and normalize each empirical cumulative distribution by its own level count. Already at initialization, 314 of 448 level-2 regions (70.1%) have OCR recall above 0.95. At update 200, this holds for 224 of 335 level-2 regions (66.9%). Levels 2 and 3 both contain many regions with high OCR recall, illustrating that character coverage does not fully distinguish the visual quality levels assigned by this judge.

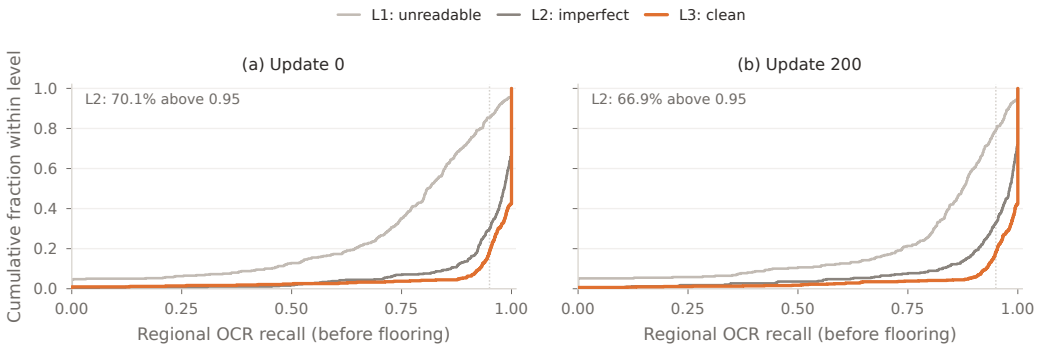


Figure 9: Empirical cumulative distributions of regional OCR recall conditioned on VLM quality level, using the VLM-trained branch. The dotted line marks recall 0.95. Level counts at update 0 are 398/448/427 for L1/L2/L3; at update 200 they are 309/335/629. Many level-2 regions have high OCR recall despite remaining rendering defects according to the VLM judge.

Figure 10 illustrates this discrepancy on an English validation page at update 200.

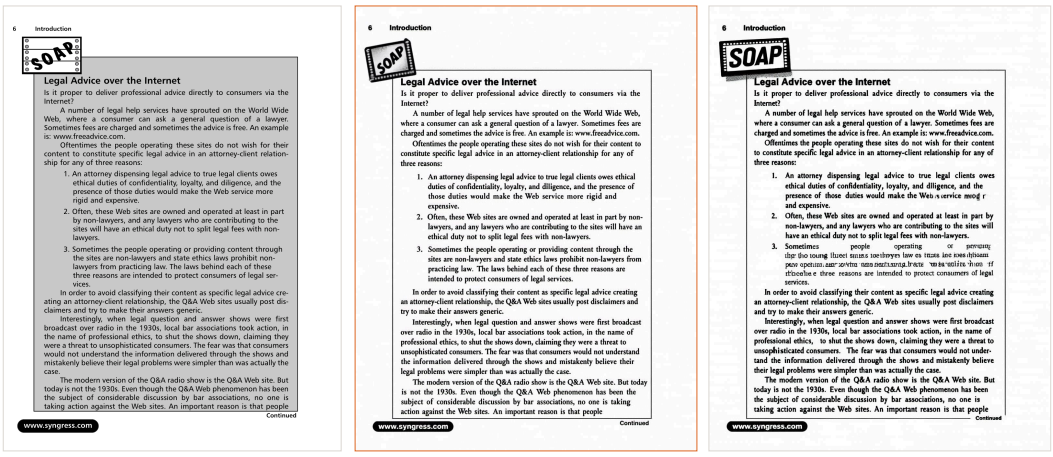


Figure 10: Full-page reward comparison on an English validation example after 200 RL updates. All nine scored regions are L3 with legibility rewards. With OCR rewards, one region is L2 and one is L1; the remaining seven are L3. The L2 region (the first numbered paragraph) has OCR recall 0.9716 despite visible glyph defects.

The quality labels are produced by the same VLM used for the legibility reward, so this analysis characterizes its relation to OCR coverage rather than establishing independently assessed human preferences. The OCR cross-evaluation provides complementary evidence on a metric not directly optimized by the VLM-trained branch. Conclusions are restricted to the evaluated model, data, and 200-update comparison window.

F TEST-TIME SCALING

F.1 GENERATION AND SELECTION PROTOCOL

The checkpoint sweep evaluates RL updates 0, 40, 100, 200, 300, 400, 500, 560, 600, and 700, with update 0 corresponding to the SFT initialization. It uses one generated image per DocPageBench request. The TTS experiment evaluates SFT and the update-560 checkpoint, generating 64 candidates per request by varying the initial noise. Both experiments share the Our PE inputs and generation settings in Section 5.2. The checkpoint sweep uses the original seed-42 generation runs; TTS varies seeds across requests and candidates. Its $N = 1$ output is the first candidate in a request-specific shuffled entry order. Consequently, the single-sample estimates in these two experiments need not be numerically identical.

The selector is `gemini-3-flash-preview`, with temperature 0.2. It compares two images on glyph quality and resemblance to a printed document, without receiving the original request, coordinates, rubric, or reference page. Images retain their native dimensions and are encoded as JPEG at quality 95 for transmission. The first candidate is the incumbent; each later candidate replaces it only if preferred by the selector. Ties and failed comparisons retain the incumbent. Candidate order and left/right presentation are determined by request-specific seeded permutations and coin flips. The incumbent after candidates 1, 2, 4, 8, 16, 32, and 64 provides the evaluated outputs.

Benchmark scores are computed after selection and are not used to select candidates. Candidate count describes the sampling budget: selecting from N images requires N generations and up to $N - 1$ pairwise comparisons, in addition to the shared prompt adaptation. It does not include the one-time RL training cost.

F.2 SCORES AND COVERAGE

We retain the unnormalized metric definitions from Appendix C.2. Let T, P, C, and A denote the mean Text, Position, Claims, and Aesthetics scores, respectively. We report Layout as $(P + C)/2$ and Overall as $(T + (P + C)/2 + A)/3$. The two spatial components keep equal weight even when their valid record counts differ. All calculations use unrounded scores. We report scores and differences to four decimal places using round-half-up.

Table 15 reports the checkpoint sweep using available scores at each update. Text has 400 valid pages and Position has 200 at every checkpoint. Aesthetics has 399 valid pages at update 500; Claims has 199 at updates 0 and 560. The remaining checkpoints have 400 Aesthetics and 200 Claims scores. Missing scores are not imputed. Text scores incorporate the OCR recovery described in Appendix C.2. For appended layout-judge retries, we retain the last successful record for each request, matching the main evaluation tables.

Table 15: Checkpoint-sweep scores with Our PE. Counts n_A and n_C denote valid Aesthetics and Claims records. Text and Position have 400 and 200 records at every update.

Update	Text	Layout	Aesthetics	Overall	n_A	n_C
0	0.7853	0.8629	0.7525	0.8002	400	199
40	0.8014	0.8605	0.7759	0.8126	400	200
100	0.8117	0.8632	0.8181	0.8310	400	200
200	0.8207	0.8657	0.8361	0.8409	400	200
300	0.8247	0.8631	0.8455	0.8444	400	200
400	0.8257	0.8612	0.8365	0.8411	400	200
500	0.8252	0.8675	0.8316	0.8414	399	200
560	0.8260	0.8686	0.8425	0.8457	400	199
600	0.8283	0.8667	0.8519	0.8490	400	200
700	0.8324	0.8712	0.8455	0.8497	400	200

For TTS, the stored records contain Text and Aesthetics scores for all 400 requests and Position scores for all 200 applicable requests at every candidate count in both models. Claims is missing for two RL requests at $N = 4$, one of which is also missing at $N = 8$. We use the common set of 198

Claims requests across both models and all candidate counts, retaining all valid Text, Aesthetics, and Position records. This fixes the request set within each component while preserving the specified component weights. Table 16 uses this common coverage throughout.

Table 16: TTS scores under fixed component coverage: 400 Text, 200 Position, 198 Claims, and 400 Aesthetics records for each row. Layout includes both spatial components.

Model	N	Text	Position	Claims	Layout	Aesthetics	Overall
SFT	1	0.7867	0.8006	0.9231	0.8618	0.7503	0.7996
SFT	2	0.7996	0.8023	0.9237	0.8630	0.7799	0.8142
SFT	4	0.8093	0.8078	0.9244	0.8661	0.8084	0.8279
SFT	8	0.8162	0.8142	0.9258	0.8700	0.8155	0.8339
SFT	16	0.8229	0.8228	0.9268	0.8748	0.8249	0.8409
SFT	32	0.8269	0.8234	0.9209	0.8721	0.8391	0.8460
SFT	64	0.8238	0.8231	0.9251	0.8741	0.8541	0.8507
SFT + RL	1	0.8303	0.8099	0.9262	0.8681	0.8456	0.8480
SFT + RL	2	0.8361	0.8127	0.9237	0.8682	0.8625	0.8556
SFT + RL	4	0.8403	0.8193	0.9262	0.8727	0.8755	0.8629
SFT + RL	8	0.8362	0.8154	0.9264	0.8709	0.8841	0.8638
SFT + RL	16	0.8410	0.8224	0.9249	0.8736	0.8958	0.8701
SFT + RL	32	0.8416	0.8179	0.9299	0.8739	0.8845	0.8667
SFT + RL	64	0.8431	0.8157	0.9307	0.8732	0.9044	0.8735

Table 16 shows that RL retains an Overall advantage at every tested budget. At $N = 64$, the gap is 0.0229, while additional sampling improves RL by 0.0255 over $N = 1$.

F.3 Component Trends and Interpretation

Figure 11 separates Text and Aesthetics during training and test-time selection. Both improve during early RL updates, while later changes are smaller. The TTS comparison uses update 560 to match the main evaluation.

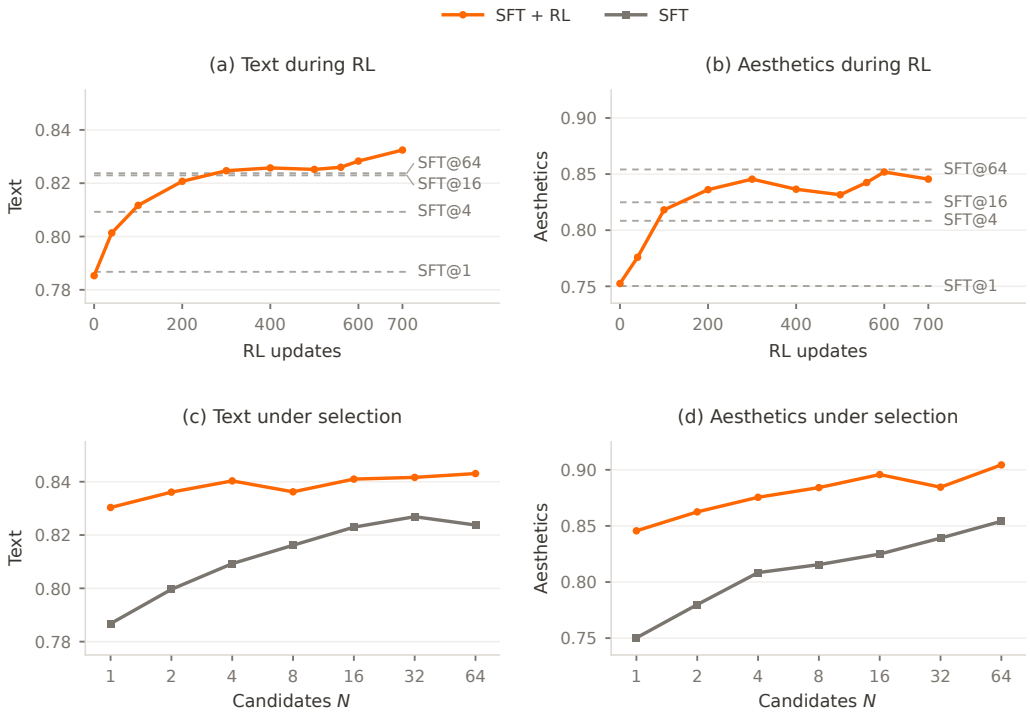


Figure 11: Text and Aesthetics under training and test-time selection. (a,b) Single-sample scores versus RL updates, with SFT best-of-N references. (c,d) SFT and update-560 RL scores as the candidate count increases. All curves connect observed means without smoothing. Coverage follows Tables 15 and 16.

Increasing the sampling budget from $N = 32$ to $N = 64$ mainly improves Aesthetics, by 0.0150 for SFT and 0.0199 for RL, while Text changes are small. The RL Aesthetics score recovers from its $N = 32$ dip to the highest observed mean at $N = 64$. These gains show that useful quality differences remain among the RL model's candidates.