Reading or Reasoning? Format Decoupled Reinforcement Learning for Document OCR

Supplementary Material

6. Implementation Details

In this section, we elaborate on the implementation details of FD-RL. We describe the data construction for both SFT and RL stages, followed by the training configurations and hyperparameters used in each stage.

6.1. Training Data

6.1.1. SFT Data

Our SFT training dataset comprises 566k samples from three complementary sources, as shown in Table 7. Specifically, we collect 240k samples from open-source datasets, 208k samples from real-world PDF documents, and 118k synthetically generated OCR samples. This diverse composition ensures comprehensive coverage of various document types and OCR scenarios.

Table 7. Distribution of training samples across data sources

Source	Samples
Open-Source Dataset	240k
Real-World PDF Data	208k
Synthetic OCR Data	118k
Total	**566k**

Table 8. Distribution of training samples across document categories

Document Category	Samples
Notes	6k
Financial Reports	3k
Slides	10k
Exam Papers	3k
Synthetic Data	118k
Magazines	100k
Academic Papers	136k
Books	108k
Newspapers	82k
Total	**566k**

As detailed in Table 8, the dataset spans nine distinct document categories, each representing common real-world OCR challenges. Academic papers constitute the largest portion with 136k samples, followed by books (108k) and magazines (100k), which collectively provide rich training data for long-form document understanding. Newspapers contribute 82k samples, capturing diverse layout patterns and multi-column structures. The synthetic data category includes 118k samples designed to complement educational materials that are difficult to obtain in real-world scenarios. Additionally, we incorporate slides (10k), notes (6k), financial reports (3k), and exam papers (3k) to ensure coverage of specialized document types with unique structural characteristics. This category-wise distribution enables the model to learn format-specific patterns while maintaining generalization across diverse document scenarios.

6.1.2. RL Data

To ensure the acquisition of format-intensive data for RL training, we manually remove document types that are primarily plain text, such as slides, magazines, books, and newspapers. Among the remaining data, we further filter out plain text samples that do not contain tables or formulas using regular expressions, resulting in a format-intensive candidate dataset of 16k samples. We then use the SFT-trained model to perform inference on all the data. For each sample, we calculate the average output entropy and retain the top 50% of high-entropy samples for RL training. Our ablation studies demonstrate that this proportion is optimal.

6.2. Training Details

6.2.1. SFT

We fine-tune Qwen3-VL-4B [26] into FD-RL(SFT) using the LLaMA Factory framework [40] on two nodes with 8 GPUs each. The training employs the dataset from Section 6.1.1 with the Qwen3-VL template (without thinking tokens) and a maximum sequence length of 8,192 tokens. For efficient data processing, we use 128 preprocessing workers (batch size 256) and 128 dataloader workers. We adopt a learning rate of 1e-5 with cosine scheduling, 10% warmup ratio, per-device batch size of 1, and train for one epoch using BF16 mixed precision. Following Section 3.2.1, we freeze the ViT and MLP modules while applying full parameter fine-tuning to the LLM, enabling comprehensive adaptation to OCR tasks.

6.2.2. RL

For the RL stage, we perform reinforcement learning on the FD-RL(SFT) model obtained from the SFT stage. Specifically, we adopt the EasyR1 [41] framework built upon the veRL project for reinforcement learning training. In terms of batch configuration, both the rollout batch size and training batch size are set to 32. During the generation process, we use a temperature of 1 and generate 8 rollouts per sample, with the maximum response length capped at 10,240 tokens.

Notably, KL divergence is excluded from the loss calculation to allow more flexible policy updates. Regarding optimization settings, we employ BF16 mixed precision training with a learning rate of 1e-6 and weight decay of 1e-2, while keeping all model parameters unfrozen throughout the training process. Finally, the complete RL dataset is trained for one epoch on a single node equipped with 8 GPUs, requiring approximately 10 hours to complete.

7. Experiment Results

In this section, we present evaluations on an additional benchmark, along with ablation studies and qualitative results, to demonstrate the effectiveness of our proposed method.

7.1. Evaluation on Ocean-OCR Benchmark

Ocean-OCR [4] is a bilingual document benchmark consisting of 100 English and 100 Chinese document-level paper images, evaluating text parsing ability with metrics including Normalized Edit Distance, F1-score, BLEU, and METEOR. We validate FD-RL on this benchmark to demonstrate its cross-scenario robustness. As shown in Table 9, FD-RL achieves state-of-the-art results across all metrics in both English and Chinese documents, consistently outperforming pipeline-based and end-to-end methods. This demonstrates that our format-decoupled reward design generalizes well beyond the training distribution. Results of comparison methods are taken from MinerU2.5 [19].

Table 9. Performance comparison on Ocean-OCR benchmark.

Model	Edit Distance↓*	F1-score↑*	BLEU↑*	METEOR↑*
FD-RL	**0.049/0.137**	**0.932/0.950**	**0.892/0.713**	**0.945/0.896**
dots.ocr	0.083/0.179	0.904/0.931	0.849/0.639	0.911/0.842
MonkeyOCR-pro-1.2B	0.064/0.190	0.929/0.934	0.884/0.699	0.941/0.850
PP-StructureV3	0.068/0.210	0.871/0.929	0.796/0.570	0.902/0.802
MinerU2-pipeline	0.099/0.225	0.663/0.919	0.504/0.571	0.670/0.810

* Values are reported as (English/Chinese).

7.2. Ablation Studies

7.2.1. Data Scaling Analysis

We investigate data scaling during the SFT stage. As shown in Table 10, model performance exhibits diminishing returns as the SFT corpus scales up: expanding from 566k to 800k samples yields only marginal gains (+0.18 overall),indicating that our SFT data engineering has largely saturated. This observation underscores the necessity of RL for further improvement, where FD-RL achieves substantial gains of +3.28 overall and +7.19 on format-intensive content with only 8k additional samples.

7.2.2. Comparison of Training Paradigms

To further validate the contribution of our training paradigm, we compare different methods on identical data. As shown in Table 11, transitioning from SFT to RFT [39] yields a moderate improvement (+1.12 overall), while GRPO [28] with a simple NED reward further advances performance to 88.64 (+1.51 over RFT [39]). Our format-decoupled rewards ultimately push performance to 90.41, demonstrating the effectiveness of both RL-based optimization and our reward design.

Table 10. Effect of data scaling on model performance.

Method	Data	Overall↑	Text↓	Formula↑	Table↑	Table*↑	RO↓
SFT	566k	87.13	0.055	85.60	81.27	84.91	0.063
SFT	800k	87.31 (+0.18)	0.054	85.82	81.45	85.09	0.062
FD-RL	566k+8k	**90.41 (+3.28)**	**0.049**	**88.67**	**87.35**	**92.10**	**0.055**

Table 11. Comparison of training paradigms.

Method	Overall↑	Text↓	Formula↑	Table↑	Table*↑	RO↓
SFT	87.13	0.055	85.60	81.27	84.91	0.063
RFT [39]	88.25	0.053	86.20	82.50	86.30	0.060
GRPO [28]	88.64	0.044	87.04	83.27	87.44	0.058
FD-RL	**90.41**	**0.049**	**88.67**	**87.35**	**92.10**	**0.055**

7.2.3. Generalization Across Backbones

To demonstrate that FD-RL is model-agnostic, we apply it to various backbones of different scales and architectures. As shown in Table 12, FD-RL consistently yields substantial gains over SFT across all settings: Qwen3-VL-2B/4B/8B [26] achieve improvements of +3.06/+3.28/+3.32, and InternVL-3.5-4B [33] improves by +3.53. These consistent improvements across scales and architectures demonstrate that FD-RL is a robust, model-agnostic framework.

Table 12. Performance of FD-RL across different backbones.

Backbone	Zero-shot	SFT	FD-RL
Qwen3-VL-2B [26]	25.15	82.36	**85.42**
Qwen3-VL-4B [26]	46.06	87.13	**90.41**
Qwen3-VL-8B [26]	48.34	88.55	**91.87**
InternVL-3.5-4B [33]	40.20	85.12	**88.65**

7.2.4. Comparison with RL-based OCR models

Table 13 compares FD-RL with other RL-based OCR models. Existing methods progress from olmOCR 2 [25] (80.03), Logics-Parsing [5] (83.85), Infinity Parser [31] (88.05), to OCRVerse [42] (89.23). Our FD-RL achieves an overall score of 90.41, setting a new record among RL-based OCR models and demonstrating the effectiveness of our format-decoupled reward design. As a concurrent work, HunyuanOCR [29] achieves a higher score of 94.10 by leveraging more training data, additional training stages, and joint optimization across a broader range of tasks, representing an alternative pathway toward advancing complex document parsing.

7.3. Qualitative Results

Figure 5 compares Qwen3-VL-4B [26] and FD-RL on three types of formatted content: Puzzle, Integral Formula, and

Table 13. Comparison with RL-based OCR models.

Methods	Overall↑	Text↓	Formula↑	Table↑	Table* ↑	RO↓
olmOCR 2 [25]	80.03	0.163	84.00	72.40	76.60	0.145
Logics-Parsing [5]	83.85	0.158	84.36	82.97	87.76	0.108
Infinity Parser [31]	88.05	0.052	88.58	80.78	84.60	0.058
OCRVerse [42]	89.23	0.052	87.13	85.77	90.35	0.068
FD-RL	<u>90.41</u>	<u>0.049</u>	<u>88.67</u>	<u>87.35</u>	**92.10**	**0.055**
HunyuanOCR [29]†	**94.10**	**0.042**	**94.73**	**91.81**	-	-

† Concurrent work.

Table. For Puzzle content, FD-RL accurately recognizes the structure and content of mathematical puzzles, demonstrating superior performance in understanding complex problem layouts. In Integral Formula recognition, FD-RL shows significant advantages in handling complex LaTeX expressions with multi-line alignments and nested symbols. The model maintains better structural integrity when rendering integration formulas. For Table processing, FD-RL accurately preserves both table structures and embedded mathematical notation, which is crucial for scientific document processing. Overall, FD-RL consistently outperforms Qwen3-VL-4B across all three content types, demonstrating that reinforcement learning effectively enhances formatted content understanding and generation.

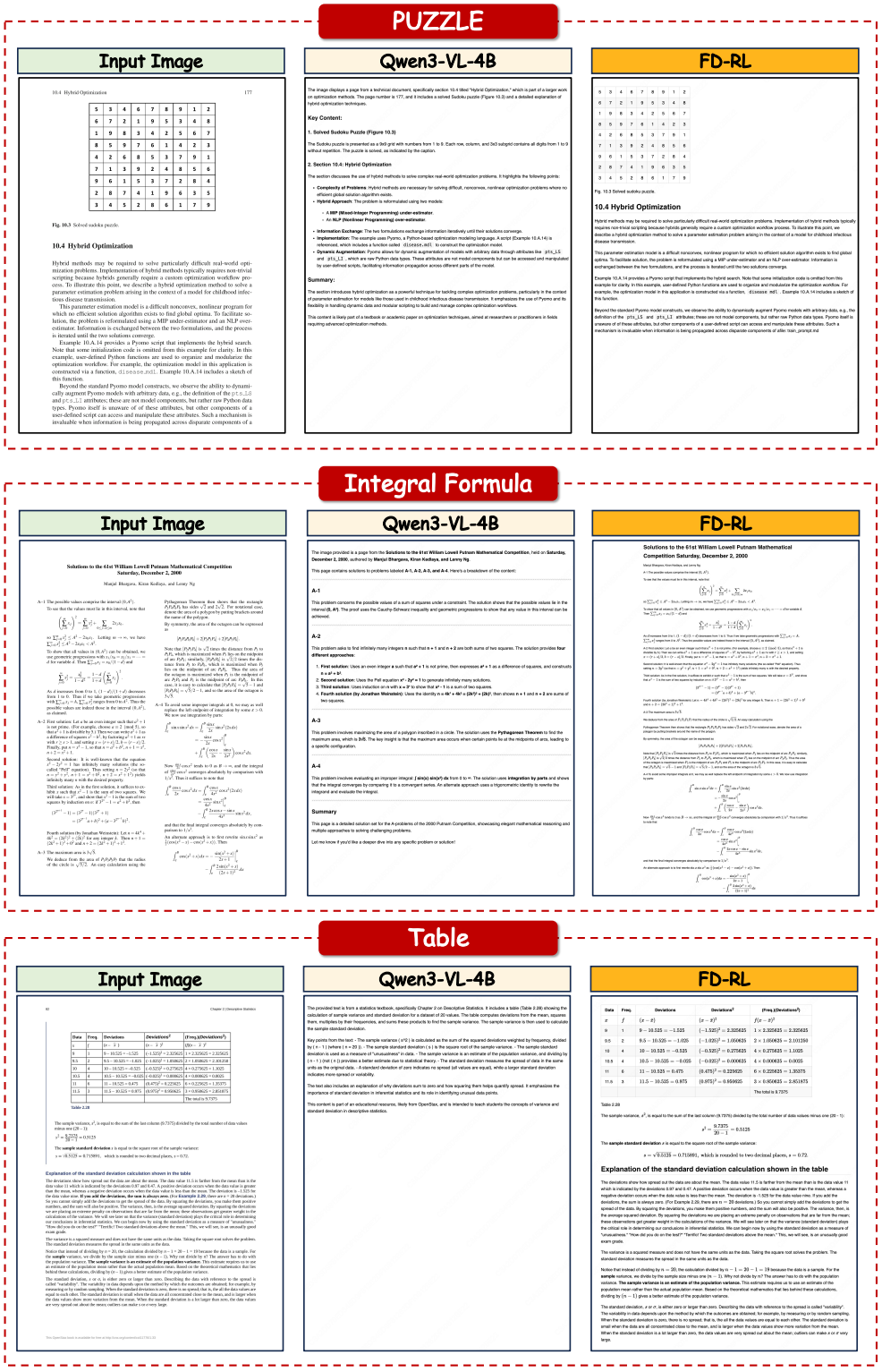

Figure 5. Qualitative comparison of Qwen3-VL-4B and FD-RL on three types of formatted content: Puzzle, Integral Formula, and Table.