P2TPO: STRUCTURE-ALIGNED PAGE-TO-TOKEN CREDIT ASSIGNMENT FOR FULL-PAGE DOCUMENT PARSING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end full-page document parsing must preserve the content and structure of heterogeneous elements, yet otherwise accurate outputs can contain local errors and omissions. A single page-level advantage does not explicitly identify which elements differ in quality across sampled outputs or where their errors occur. We introduce Page-to-Token Policy Optimization (P2TPO), a structure-aligned credit-assignment framework that operates on full-page policy trajectories. Reference-element alignment first associates block-level quality contrasts with generated spans. Unified Anchor Pricing (UAP) incorporates matched and missing element outcomes and provides feedback relative to a fixed quality ceiling when sampled outputs receive identical but imperfect scores for a reference element. Within matched blocks, Span Credit Redistribution (SCR) redistributes existing block credit according to aligned correct and erroneous spans, using boundary proxies for within-block omissions while preserving the total block-derived advantage across its tokens. Under a fixed-weight additive page utility, we show that element-indexed contrasts retain local quality differences that page-level aggregation can cancel out. In the main experimental comparison, P2TPO outperforms all evaluated policy optimization baselines in seed-averaged scores on each of the six document parsing benchmarks. Additional experiments demonstrate gains across model scales and initializations.

1 INTRODUCTION

Document parsing converts page images from digital documents, scans, and photographs into machine-readable content while preserving structure and reading order (Ouyang et al., 2025; Du et al., 2025). Two common paradigms are multi-stage pipelines and end-to-end generation. Pipelines separate layout analysis from specialized recognition of text, formulas, and tables, then assemble the regions in reading order. This decomposition enables local processing but introduces dependencies between detection, recognition, and assembly (Zhang et al., 2025). End-to-end parsers directly generate structured content, reducing coordination across stages while jointly reconstructing the full page (Blecher et al., 2024; Taghadouini et al., 2026).

We study full-page transcription into a single Markdown document in natural reading order: text is reproduced faithfully, formulas use LaTeX, and tables use HTML. Dense heterogeneous content, skew, warping, and uneven illumination make recognition and structural recovery challenging (Ouyang et al., 2025; Zhou et al., 2026). Errors are often local: an otherwise accurate output may corrupt a formula, omit words, or skip an entire element.

Reference-based metrics make reinforcement learning (RL) useful for optimizing content accuracy, structural fidelity, and completeness beyond supervised next-token learning. Recent parsers demonstrate this potential through layout-aware rewards and verifiable unit tests (Wang et al., 2025; Chen et al., 2025; Poznanski et al., 2025). Yet olmOCR 2 aggregates local test outcomes into a page-level reward before optimization (Poznanski et al., 2025). For RL post-training of an SFT-initialized parser, we study how these measurements become element-specific feedback within full-page outputs (Table 5).

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.

With page-level outcome rewards, GRPO assigns one response advantage to every generated token (Shao et al., 2024). GDPO normalizes reward dimensions separately but then combines them into a shared response coefficient (Liu et al., 2026b). Although evaluation can detect missing blocks or characters, this aggregation does not explicitly localize their feedback. Sampled outputs of the *same page* can make different local errors yet receive similar overall scores, as illustrated in Figure 1(a); page quality alone cannot identify which content needs correction.

We address four obstacles between evaluation and token feedback, summarized in Figure 1(b). Pooling can cancel opposing quality differences between blocks, including instances of the same category. Group centering removes feedback when scores tie below perfect quality. Uniform block credit obscures error locations, while omitted elements have no generated span. Our motivation is therefore to preserve *element identity*, *missing outcomes*, and *where credit acts* within full-page trajectories.

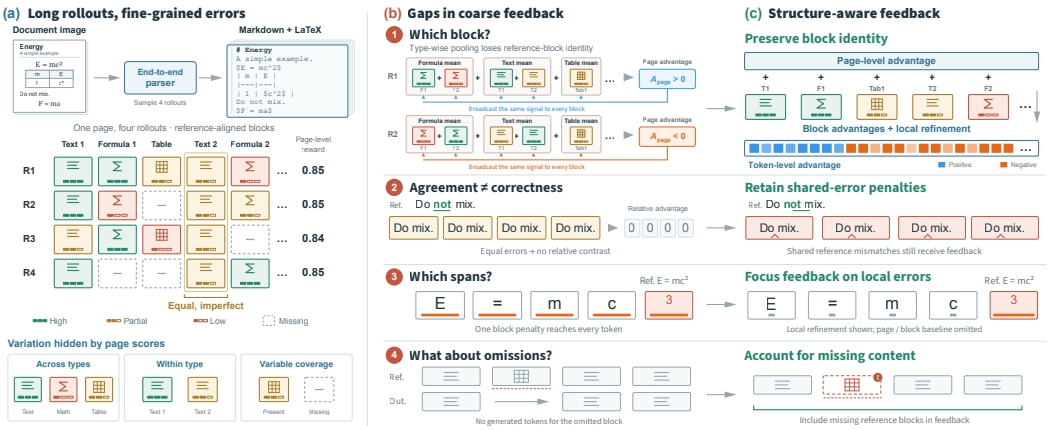


Figure 1: Schematic motivation for structure-aware feedback. (a) Similar page rewards hide local differences. (b) Coarse feedback obscures block identity, shared errors, and error/omission locations. (c) Structure-aware feedback retains element identity and assigns local credit. Rewards and serialized content are illustrative; training outputs use HTML tables.

We introduce *Page-to-Token Policy Optimization* (P2TPO), which constructs local feedback while retaining full-page autoregressive trajectories and global page feedback (see Fig. 1(c)). Its base, Page-to-Block Policy Optimization (P2BPO), aligns predictions to reference elements and compares corresponding matched blocks across rollouts. *Unified Anchor Pricing* (UAP) completes these comparisons with missing outcomes, supplies ceiling-gap feedback at imperfect consensus, and routes coefficients to matched spans or omission-boundary proxies. *Span Credit Redistribution* (SCR) then directs more negative credit to aligned errors and more positive credit to correct positions within eligible matched blocks, conserving each block's coefficient budget. Together, these stages connect element identity, outcome availability, and local allocation.

Our contributions are fourfold:

- **Reference-element credit assignment.** P2TPO retains the identity of individual text, formula, and table blocks across full-page rollouts. Under a fixed-weight additive page utility, we characterize local quality contrasts that aggregation can erase.
- **Complete outcome feedback.** UAP includes matched and missing outcomes for each reference element, retains feedback at imperfect consensus, and maps omission feedback to generated boundary proxies.
- **Budget-preserving intra-block redistribution.** SCR reallocates signed credit within matched blocks. We establish budget conservation and conditions on mask precision under which more credit reaches target positions.
- **Empirical effectiveness and generality.** P2TPO exceeds four response-level baselines in mean score on all 13 primary splits across six benchmarks. It achieves the highest reported mean in 37 of 39 additional comparisons covering SFT initialization, model scale, and backbone.

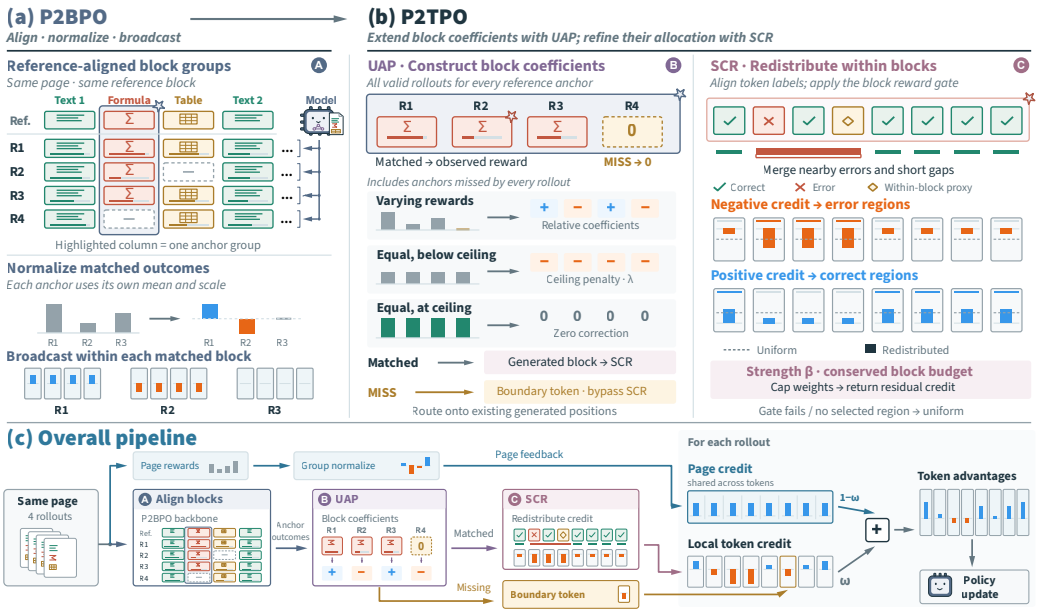


Figure 2: P2TPO overview. (a) P2BPO aligns reference elements, normalizes matched outcomes, and broadcasts block credit. (b) UAP incorporates missing outcomes and imperfect-consensus feedback; SCR redistributes matched-block credit under a conserved budget. (c) The complete pipeline combines local token credit with page feedback. Missing-block coefficients reach boundary tokens directly, bypassing SCR.

2 METHOD

P2TPO maps reference-element quality to token feedback through alignment, outcome completion, and within-block redistribution. Local credit and page feedback jointly drive a full-page policy update (Figure 2(c)); generation and optimization retain complete autoregressive trajectories.

2.1 REFERENCE-ELEMENT COMPARISONS

Given a page image x and prompt p, we follow GRPO's group sampling protocol (Shao et al., 2024) and draw G full-page outputs $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, p)$. Predictions and references are parsed into ordered text, formula, and table blocks. Content-based sequence alignment permits insertions and deletions while preserving reading order. Each matched prediction is indexed by its corresponding reference block $y_k^\star$, rather than its position in the generated sequence. Thus, different instances of the same category remain distinct across rollouts, even when preceding elements are omitted.

Let i, k, and t index rollouts, reference elements, and tokens. A matched block has a type-specific quality score $s_{i,k} \in [0, c_k]$ and a generated-token support $\mathcal{B}_{i,k}$ of length $L_{i,k}$. The implementation uses normalized scores with $c_k = 1$ and maps generated block boundaries to disjoint token spans. For each reference k with a valid match, let $\mathcal{I}_k$ contain its valid matched rollouts, and let μ_k^{B} and σ_k^{B} be the mean and population standard deviation of their scores.

Page-to-Block Policy Optimization (P2BPO) normalizes each matched reference group independently:

$$A_{i,k}^{\text{B}} = \begin{cases} (s_{i,k} - \mu_k^{\text{B}})/\sigma_k^{\text{B}}, & \sigma_k^{\text{B}} > \epsilon, \\ 0, & \sigma_k^{\text{B}} \le \epsilon, \end{cases} \qquad i \in \mathcal{I}_k. \tag{1}$$

Here $\epsilon \ge 0$ is a numerical tolerance. Each coefficient is broadcast over $\mathcal{B}_{i,k}$ (see Fig. 2(a)), so blocks within one response can receive different signs and magnitudes. Reference-wise comparisons retain contrasts that page- or category-level pooling can cancel. However, unmatched references have no coefficient, while a singleton or tied matched group produces zero credit. UAP addresses these two gaps. Appendix B.1 summarizes the algorithm inputs.

2.2 UAP: COMPLETE OUTCOMES AND RETAIN QUALITY DEFICITS

Unified Anchor Pricing (UAP) completes each reference element's outcomes across valid rollouts $\mathcal{V}_k$: $Q_{i,k} = s_{i,k}$ when matched and $Q_{i,k} = 0$ when missing. With mean μ_k and population standard deviation σ_k, its coefficient uses relative comparison when outcomes differ and ceiling-gap feedback when dispersion is negligible:

$$A^{\mathrm{U}}_{i,k} = \begin{cases} (Q_{i,k} - \mu_k)/\sigma_k, & \sigma_k > \epsilon, \\ -\lambda h(\mu_k, c_k), & \sigma_k \le \epsilon, \end{cases} \qquad h(u, c) = \max(0, c - u). \tag{2}$$

Here $\lambda \ge 0$ weights the ceiling gap; the same tolerance ϵ separates the two branches. For $\lambda > 0$, both fully matched imperfect ties and all-missing groups receive negative coefficients, whereas a ceiling-quality tie receives zero. Thus a single element-indexed rule retains feedback for missing and uniformly imperfect outcomes.

In a mixed group, zero-valued omissions enter the same mean and dispersion as observed block scores. Outcome completion therefore makes missing outputs part of the comparison rather than reducing the group to matched members, and can change the coefficients for both matched and missing outcomes.

Outcome completion adds missing scores to the comparison without inserting tokens into the sampled response. Matched coefficients act on generated blocks; missing-element coefficients reach existing boundary tokens through $\nu_{i,k}(t) \ge 0$, where $\sum_t \nu_{i,k}(t) = 1$, bypassing SCR. Appendix B.2 details the coefficient construction.

2.3 SCR: REDISTRIBUTE CREDIT UNDER A CONSERVED BUDGET

Span Credit Redistribution (SCR) reallocates an available block coefficient while preserving its budget. Uniform credit treats correct and erroneous tokens alike; SCR instead concentrates negative credit on error regions and positive credit on correct regions.

Quality gating restricts redistribution to high-scoring, type-matched blocks, focusing local allocation on residual defects within otherwise accurate elements. Within an eligible block, alignment identifies correct and erroneous spans, using boundary proxies for within-block omissions. A binary mask $u_{i,t}$ selects errors for negative credit and correct positions for positive credit (see Fig. 2(b)). For a nonempty mask of size $N_{i,k} = \sum_{t \in \mathcal{B}_{i,k}} u_{i,t}$, the redistribution has the form

$$d_{i,t} = 1 - \eta_{i,k} + \eta_{i,k} \frac{L_{i,k} u_{i,t}}{N_{i,k}}, \qquad \sum_{t \in \mathcal{B}_{i,k}} d_{i,t} = L_{i,k}. \tag{3}$$

The implementation mixes uniform and selected-token allocation, applies a pre-rebate cap to concentrated weights, and returns the removed mass uniformly. This determines the effective strength $\eta_{i,k} \in [0, 1]$ in Appendix B.2.2. Nonnegative weights retain the coefficient's sign, while their unit mean preserves both signed and absolute block-credit budgets. Blocks ineligible for redistribution or with empty masks retain uniform weights $d_{i,t} = 1$. Appendix D.3 establishes the resulting increase in target allocation for informative masks.

2.4 PAGE-TO-TOKEN COMPOSITION

Let $\mathcal{K}_i$ and $\mathcal{M}_i$ index matched and missing reference elements, respectively, and let A^{P}_i be the group-normalized mean of the applicable page-reward dimensions. Combining the two local contributions with page feedback gives

$$\begin{aligned} C^{\mathrm{US}}_{i,t} &= \sum_{k \in \mathcal{K}_i} A^{\mathrm{U}}_{i,k} \mathbf{1}[t \in \mathcal{B}_{i,k}] d_{i,t} + \sum_{k \in \mathcal{M}_i} A^{\mathrm{U}}_{i,k} \nu_{i,k}(t), \\ A^{\mathrm{US}}_{i,t} &= (1 - \omega) A^{\mathrm{P}}_i + \omega C^{\mathrm{US}}_{i,t}, \qquad \omega \in [0, 1]. \end{aligned} \tag{4}$$

The page term is shared across generated positions, while local terms vary with reference identity and within-block location. Tokens outside local supports retain the page term. The mixing weight ω balances these scopes of feedback before the policy update. Alignments and the resulting $A^{\mathrm{US}}_{i,t}$ remain fixed during the clipped policy update; token probabilities retain the full autoregressive page context. Appendix B.2 provides the complete coefficient construction, component variants, and policy loss; Appendix C.1.3 lists parameter settings.

3 Related Work

Document parsing and reinforcement post-training. End-to-end parsers such as Donut and Nougat generate structured documents directly from images (Kim et al., 2022; Blecher et al., 2024). Recent systems extend this setting to diverse layouts and capture conditions (Cai et al., 2026; Lu et al., 2026). Reinforcement post-training introduces rewards for transcription, layout, reading order, output validity, and structural fidelity (Wang et al., 2025; Huang et al., 2026; Chen et al., 2025; Poznanski et al., 2025; Wang et al., 2026a; Hunyuan Vision Team et al., 2025; Zhang et al., 2026). Table- and element-specific approaches further tailor the prediction unit and reward to local content (Zhang et al., 2025; Li et al., 2026a; Wang et al., 2026c). These systems motivate heterogeneous quality measures, while P2TPO studies how feedback for individual reference elements reaches tokens within a full-page rollout. Appendix A.1 compares reward units and optimization supports.

Group-relative optimization and reward normalization. GRPO and RLOO construct response advantages from sampled groups without a learned critic (Shao et al., 2024; Ahmadian et al., 2024). GSPO uses sequence-level importance weighting and clipping, while DAPO and Dr. GRPO revise sampling, clipping, reduction, or normalization (Zheng et al., 2025; Yu et al., 2025a; Liu et al., 2025). GDPO, RDPO, and GD^{2}PO address the normalization and interaction of multiple reward dimensions (Liu et al., 2026b; Bai et al., 2026; Liu et al., 2026a). NGRPO combines virtual-maximum calibration with asymmetric clipping to learn from homogeneous failures (Nan et al., 2025). UAP shares ceiling-informed feedback, while unifying missing outcomes and imperfect reward consensus within reference-element comparisons and routing their feedback to generated tokens. Appendix C.1.4 details the evaluated baselines; Appendix A.1 compares reward interfaces.

Structured and local credit assignment. Process supervision evaluates reasoning steps (Lightman et al., 2024; Wang et al., 2024); GiGPO compares repeated environment states while retaining trajectory feedback (Feng et al., 2025). CSPO extends GRPO with table-wide functional-attribute rewards whose normalized advantages are routed to corresponding tokens alongside global feedback (Yang et al., 2026). P2TPO performs local comparisons for individual reference-block instances in full-page rollouts, preserving quality differences between elements of the same type and supporting omission feedback and intra-block redistribution. Functional-attribute assessment and reference-instance comparison are complementary; Appendix A.2 discusses possible adaptations.

4 Experiments

We first evaluate cross-benchmark parsing quality, then examine component contributions and credit diagnostics. We next assess generality across initializations, model scales, and backbones, followed by hyperparameter sensitivity and performance under a longer training budget.

4.1 Experimental setup

Baselines. We compare P2TPO with four response-level policy optimization baselines: GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), GDPO (Liu et al., 2026b), and RDPO (Bai et al., 2026). P2BPO and its +UAP and +SCR variants serve as component ablations. Appendix C.1.4 details the comparison methods.

Benchmarks. We evaluate on six benchmarks with 13 splits: OmniDocBench v1.6 (Ouyang et al., 2025; OmniDocBench contributors, 2026) for heterogeneous full-page parsing; PureDocBench (Li et al., 2026c) for Clean, Digital, and Real conditions; FinixDocBench (Wang et al., 2026b) for digital and photographed documents; Real5-OmniDocBench (Zhou et al., 2026) for five capture conditions; DocPTBench (Du et al., 2025) for photographed and dewarped pages; and Wild-OmniDocBench (Li et al., 2026b) for real-world captures. Scores compare methods within each benchmark: five benchmarks use the shared OmniDocBench evaluator, while FinixDocBench pools its Digital and Photo tracks using its official evaluator. Real5 and Wild reuse OmniDocBench source pages and assess capture robustness. Appendix C.1.5 specifies the evaluated scope and score definitions, including the adapted DocPTBench composite.

Table 1: Cross-benchmark Overall scores (↑, 0–100) from Qwen3.5-2B initialized by SFT on 700K pages. Init reports pre-RL scores; RL entries report mean ± standard deviation over five seeds after 100 updates.

Benchmark	Split	Init	GRPO	GSPO	GDPO	RDPO	P2BPO	+UAP	+SCR	**P2TPO**
OmniDocBench	v1.6	92.17	92.25 ± 0.34	92.38 ± 0.18	92.31 ± 0.24	92.32 ± 0.35	92.48 ± 0.15	92.84 ± 0.30	92.56 ± 0.21	**93.12** ± 0.23
PureDocBench	Clean	81.28	83.19 ± 0.70	83.44 ± 0.31	82.68 ± 0.65	83.10 ± 0.54	84.16 ± 0.62	83.98 ± 0.37	83.78 ± 0.40	**84.68** ± 0.74
	Digital	76.66	79.89 ± 0.48	79.50 ± 0.33	79.24 ± 0.50	79.26 ± 0.50	80.31 ± 0.48	80.26 ± 0.39	79.79 ± 0.31	**81.03** ± 0.60
	Real	63.29	68.89 ± 0.64	67.64 ± 0.38	68.23 ± 0.69	68.27 ± 0.28	69.36 ± 0.46	69.87 ± 0.49	69.08 ± 0.35	**70.33** ± 0.70
FinixDocBench	—	65.30	69.36 ± 0.49	69.24 ± 0.44	69.21 ± 1.42	68.15 ± 0.74	70.03 ± 0.77	70.29 ± 0.62	69.37 ± 0.44	**70.92** ± 0.92
Real5-OmniDocBench	Scanning	91.06	92.07 ± 0.20	92.00 ± 0.17	92.00 ± 0.21	91.90 ± 0.28	92.44 ± 0.29	**92.52** ± 0.16	92.25 ± 0.31	92.40 ± 0.12
	Warping	86.50	90.45 ± 0.53	89.36 ± 0.29	90.11 ± 0.15	89.92 ± 0.37	90.44 ± 0.57	90.59 ± 0.30	90.31 ± 0.46	**90.77** ± 0.45
	Skew	81.77	88.26 ± 0.48	88.47 ± 0.34	88.09 ± 0.39	88.47 ± 0.42	88.82 ± 0.64	88.78 ± 0.38	88.56 ± 0.61	**89.15** ± 0.52
	Illumination	91.23	92.41 ± 0.12	91.98 ± 0.26	92.22 ± 0.20	92.35 ± 0.09	92.59 ± 0.20	92.62 ± 0.19	92.41 ± 0.28	**92.83** ± 0.17
	Screen photo	90.32	92.04 ± 0.17	91.76 ± 0.31	91.75 ± 0.30	91.52 ± 0.28	92.33 ± 0.30	92.28 ± 0.12	92.20 ± 0.50	**92.62** ± 0.29
DocPTBench	SynReal	62.27	73.21 ± 0.67	72.85 ± 0.65	72.59 ± 0.54	72.80 ± 0.26	73.61 ± 0.56	73.60 ± 0.45	**73.81** ± 1.14	73.79 ± 0.39
	Dewarp	79.66	83.57 ± 0.29	82.85 ± 0.36	83.12 ± 0.30	83.29 ± 0.82	83.95 ± 0.44	**83.99** ± 0.42	83.36 ± 0.50	83.76 ± 0.51
Wild-OmniDocBench	—	80.04	84.67 ± 0.18	83.48 ± 0.20	84.28 ± 0.20	84.50 ± 0.29	84.72 ± 0.47	84.68 ± 0.54	84.55 ± 0.56	**85.02** ± 0.21

Note. Init is the shared SFT-initialized checkpoint; no standard deviation is reported for Init. P2BPO is the matched-element reference configuration; +UAP and +SCR enable one component at a time. Bold and underlined values mark the best and second-best RL means, respectively. — denotes an aggregate benchmark score.

Implementation details. We use converted Infinity-Doc2-5M annotations (Huang et al., 2026) for document SFT followed by RL. SFT provides the initial checkpoint: Qwen3.5-2B trained on 700K pages in the primary setting, with additional settings covering 50K-page checkpoints, Qwen3.5-4B, and Qwen3-VL-2B-Instruct. Within each setting, all RL methods start from the same checkpoint and share the update budget. Init reports its measured benchmark score before RL. Table 1 reports primary RL Overall scores as means ± standard deviations over five seeds. Appendices C.1.1, C.1.3, and C.4 provide further details.

4.2 Cross-benchmark performance

Table 1 reports Overall scores for the primary Qwen3.5-2B/700K setting across all six benchmarks and 13 evaluation splits.

On the primary OmniDocBench v1.6 benchmark, P2TPO reaches 93.12 ± 0.23, a gain of 0.95 points from its pre-RL score and 0.74 points over the strongest response-level baseline. It also exceeds the strongest response-level mean on all 12 external splits, with margins from 0.19 points on DocPTBench Dewarp to 1.56 points on FinixDocBench. The next-largest margins occur on PureDocBench Real (+1.44) and Clean (+1.24), and the gains hold across all five Real5-OmniDocBench capture conditions.

4.3 Component ablations and credit diagnostics

Reference-element routing. Table 1 first isolates matched-element comparisons: P2BPO exceeds the strongest response-level baseline on 12 of 13 splits, with the largest gains on PureDocBench Clean (+0.72) and FinixDocBench (+0.67). The exception is Real5-OmniDocBench Warping (-0.01). All differences below are in mean Overall-score points.

SCR is more consistently beneficial with UAP. Adding SCR alone improves P2BPO on only two splits, OmniDocBench (+0.08) and DocPTBench SynReal (+0.20). With UAP enabled, adding SCR improves 11 of 13 splits, led by PureDocBench Digital (+0.77) and Clean (+0.70). Conversely, adding UAP to P2BPO + SCR improves 12 splits. Full P2TPO leads the four component configurations on ten splits; +UAP leads on Scanning and Dewarp, and +SCR on SynReal. These results show a conditional benefit of SCR: redistribution is more consistently effective with UAP's completed and calibrated block feedback. On OmniDocBench, the joint gain exceeds the sum of the standalone gains by 0.20 points; this interaction in means has not been tested for statistical significance.

Internal UAP interventions. We retain outcome completion and SCR while separately disabling consensus feedback ($\lambda = 0$) or omission routing. Table 2(a) shows that full P2TPO exceeds the former in mean score on eight of 13 splits and the latter on 11. Full P2TPO also has lower observed

Table 2: UAP ablations and SCR allocation controls across six benchmarks and 13 splits. Entries report mean Overall score (↑, 0–100) ± standard deviation over five runs; settings follow Table 1.

Benchmark	Split	Full P2TPO	(a) UAP interventions		w/o UAP	(b) SCR allocation controls		w/o SCR
			w/o consensus feedback	w/o omission routing		Circular shift	Random support	
OmniDocBench	v1.6	**93.12** ± 0.23	92.92 ± 0.45	92.86 ± 0.23	92.56 ± 0.21	92.73 ± 0.17	92.82 ± 0.25	92.84 ± 0.30
PureDocBench	Clean	**84.68** ± 0.74	84.38 ± 0.94	83.74 ± 0.35	83.78 ± 0.40	83.85 ± 0.79	83.84 ± 0.16	83.98 ± 0.37
	Digital	**81.03** ± 0.60	80.79 ± 0.88	80.15 ± 0.36	79.79 ± 0.31	80.08 ± 0.73	80.43 ± 0.45	80.26 ± 0.39
	Real	**70.33** ± 0.70	70.05 ± 0.93	69.67 ± 0.64	69.08 ± 0.35	69.69 ± 0.78	69.33 ± 0.47	69.87 ± 0.49
FinixDocBench	—	70.92 ± 0.92	70.42 ± 1.36	70.05 ± 0.80	69.37 ± 0.44	**70.98** ± 1.22	70.03 ± 1.09	70.29 ± 0.62
Real5-OmniDocBench	Scanning	92.40 ± 0.12	92.48 ± 0.35	92.45 ± 0.19	92.25 ± 0.31	92.26 ± 0.41	92.36 ± 0.23	**92.52** ± 0.16
	Warping	**90.77** ± 0.45	90.68 ± 0.77	90.45 ± 0.39	90.31 ± 0.46	90.41 ± 0.35	90.25 ± 0.48	90.59 ± 0.30
	Skew	**89.15** ± 0.52	88.92 ± 0.81	88.80 ± 0.31	88.56 ± 0.61	88.70 ± 0.46	88.42 ± 0.28	88.78 ± 0.38
	Illumination	92.83 ± 0.17	**92.84** ± 0.38	92.59 ± 0.16	92.41 ± 0.28	92.52 ± 0.45	92.45 ± 0.16	92.62 ± 0.19
	Screen photo	92.62 ± 0.29	**92.67** ± 0.64	92.45 ± 0.14	92.20 ± 0.50	92.19 ± 0.27	92.52 ± 0.56	92.28 ± 0.12
DocPTBench	SynReal	73.79 ± 0.39	**74.49** ± 1.23	74.10 ± 0.55	73.81 ± 1.14	74.21 ± 1.17	73.33 ± 0.60	73.60 ± 0.45
	Dewarp	83.76 ± 0.51	83.92 ± 0.61	83.69 ± 0.13	83.36 ± 0.50	83.94 ± 1.09	83.83 ± 0.48	**83.99** ± 0.42
Wild-OmniDocBench	—	**85.02** ± 0.21	84.79 ± 0.57	84.81 ± 0.75	84.55 ± 0.56	84.74 ± 0.37	84.28 ± 0.38	84.68 ± 0.54

Note. Bold marks the highest mean per split. w/o UAP: P2BPO + SCR; w/o SCR: P2BPO + UAP. — denotes an aggregate benchmark score.

across-seed standard deviations than the $\lambda = 0$ configuration on all 13 splits. These interventions assess UAP's feedback and delivery rules; Appendix C.2.1 gives their definitions and detailed comparisons.

SCR allocation controls. To test where credit acts, we retain each eligible block's coefficient values and selected-token count while changing their positions. Table 2(b) shows that aligned SCR exceeds random support on 12 of 13 splits and circular shift on ten. The randomized controls improve over uniform allocation on only two and three splits, respectively, supporting the contribution of alignment-informed placement. Appendix C.2.2 reports the protocols and additional analysis.

Table 3: Component diagnostics on OmniDocBench v1.6 after 100 RL updates from the shared 700K-page SFT initialization. Parsing metrics report five-seed means ± standard deviations; best and second-best values are bold and underlined, respectively.

Method	Parsing metrics					Omission rates (%)	
	Text ED↓	Formula CDM↑	Formula ED↓	Table TEDS↑	Order ED↓	Element↓	Local↓
P2BPO	0.0360 ± 0.0018	0.9340 ± 0.0028	0.1279 ± 0.0020	0.8765 ± 0.0046	0.0475 ± 0.0025	1.09	6.72
P2BPO + UAP	**0.0333** ± 0.0017	<u>0.9398</u> ± 0.0042	<u>0.1261</u> ± 0.0041	0.8786 ± 0.0059	<u>0.0459</u> ± 0.0020	<u>1.01</u>	<u>5.59</u>
P2BPO + SCR	0.0364 ± 0.0008	0.9325 ± 0.0046	0.1275 ± 0.0031	<u>0.8807</u> ± 0.0080	0.0500 ± 0.0017	1.08	5.99
P2TPO	<u>0.0343</u> ± 0.0015	**0.9446** ± 0.0027	**0.1212** ± 0.0023	**0.8833** ± 0.0061	**0.0423** ± 0.0024	**0.95**	**5.53**

Note. ED: edit distance. Parsing metrics use $[0, 1]$; omission rates use $[0, 100]$. Element omission counts missing eligible reference elements; local omission counts omitted characters within matched text blocks and excludes whole-element omissions. Omission-rate standard deviations are unavailable. Appendix C.1.5 gives the complete definitions.

Parsing accuracy and omissions. Table 3 compares the four component configurations on content-specific errors on OmniDocBench. UAP reduces element omission from 1.09% to 1.01% and local omission from 6.72% to 5.59%. Full P2TPO has the lowest observed omission rates and the best formula, table, and reading-order scores, while P2BPO + UAP has the lowest Text ED. SCR improves Formula CDM and reading order with UAP, but worsens both w/o UAP, consistent with its conditional Overall-score gains.

Figure 3 compares five methods on two frozen eight-response groups (no policy update; Appendix C.3). Its selected profiles show block-dependent magnitudes and mixed-sign coefficients under P2BPO, whereas GRPO is uniform within each response.

4.4 Generality across initialization, scale, and backbone

Table 4 complements the primary Qwen3.5-2B/700K results with three additional SFT initialization settings: Qwen3.5-2B/50K, Qwen3.5-4B/50K, and Qwen3-VL-2B-Instruct/700K. Within each setting,

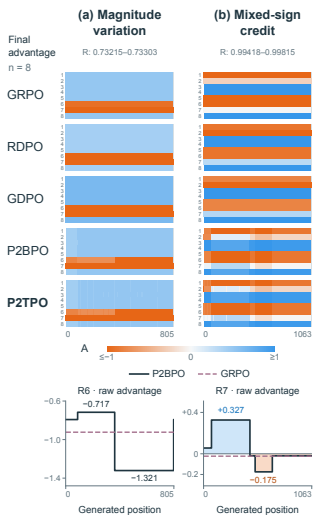


Figure 3: Frozen-response credit assignment ($n = 8$ per case; no policy update). Heatmaps saturate beyond ± 1. Lower profiles compare GRPO and P2BPO for (a) R6 and (b) R7; y-scales differ.

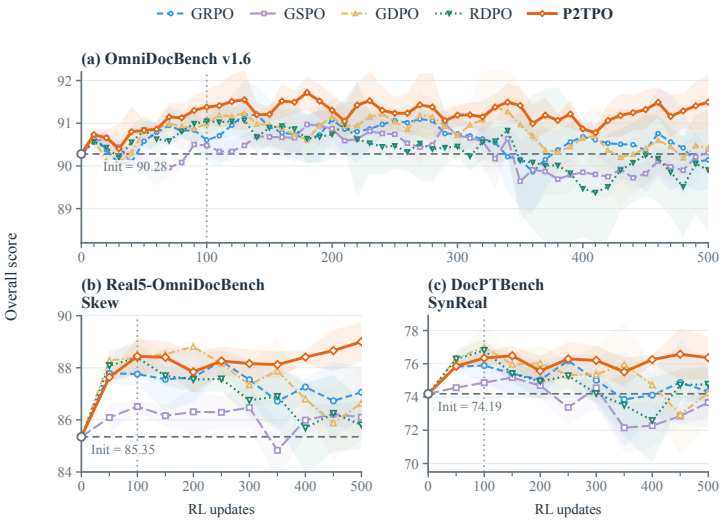


Figure 4: Extended Qwen3.5-2B/50K training with 64,000 RL records. Lines show unsmoothed five-run means; bands denote ± 1 SD. Markers and ticks indicate evaluations every ten updates in (a) and every 50 in (b,c). Each panel starts at its measured Init (dashed line), with SD bands beginning at the first evaluation. Panel-specific y-axis ranges truncate some SD bands; vertical guides mark update 100.

all RL methods share the starting checkpoint and 100-update budget, with five runs using distinct RL data seeds. The table reports means on the same six benchmarks and 13 splits as Table 1; Appendix C.4 provides standard deviations and model-setting details.

Table 4: Cross-benchmark RL mean scores (↑) across three SFT-initialized checkpoints. Benchmark names and splits follow Table 1.

Initial Conditions		**Qwen3.5-2B** SFT data: 50K pages					**Qwen3.5-4B** SFT data: 50K pages					**Qwen3-VL-2B-Instruct** SFT data: 700K pages				
Benchmark	Split	GRPO	GSPO	GDPO	RDPO	**P2TPO**	GRPO	GSPO	GDPO	RDPO	**P2TPO**	GRPO	GSPO	GDPO	RDPO	**P2TPO**
OmniDocBench	v1.6	91.00	90.39	91.05	90.85	**91.62**	92.13	91.99	92.07	91.54	**92.53**	90.45	90.55	90.57	88.25	**91.29**
PureDocBench	Clean	80.83	81.01	81.02	79.75	**81.98**	84.41	84.28	84.07	82.90	**84.95**	81.81	81.80	81.75	79.67	**81.95**
	Digital	78.26	78.11	77.85	77.27	**79.16**	81.80	81.06	81.47	80.14	**82.34**	76.08	76.29	**76.57**	73.98	76.51
	Real	69.82	68.66	69.58	69.20	**71.01**	73.55	72.52	73.61	71.86	**74.48**	65.45	65.44	65.76	63.53	**66.25**
FinixDocBench	—	69.45	70.11	70.95	70.95	**73.01**	74.33	74.43	73.59	69.38	**76.74**	67.68	67.75	67.95	65.18	**68.23**
Real5-OmniDocBench	Scanning	90.66	90.23	90.97	90.97	**91.29**	92.04	91.89	91.81	91.52	**92.23**	90.07	89.42	89.89	87.79	**90.46**
	Warping	89.56	89.06	89.47	89.53	**89.86**	91.02	90.84	91.18	90.22	**91.45**	85.19	84.17	85.62	84.46	**86.37**
	Skew	88.02	87.04	88.11	87.61	**88.44**	90.95	90.40	90.78	90.02	**91.33**	76.83	77.74	77.82	76.66	**79.79**
	Illumination	90.81	90.35	91.10	90.83	**91.23**	92.48	92.33	92.59	91.71	**92.80**	89.26	88.79	89.38	87.15	**89.65**
	Screen photo	89.79	89.36	90.43	90.24	**90.50**	91.91	91.71	91.87	91.08	**92.16**	88.46	88.09	88.39	86.01	**89.33**
DocPTBench	SynReal	75.57	75.21	**76.47**	76.23	76.31	77.86	76.51	77.76	76.37	**78.19**	67.80	67.18	67.23	65.91	**68.94**
	Dewarp	82.43	82.47	83.26	82.90	**83.63**	84.02	83.98	84.11	83.23	**84.88**	78.28	77.38	77.97	76.16	**79.34**
Wild-OmniDocBench	—	83.67	82.99	83.62	84.27	**84.48**	86.04	85.34	85.83	85.10	**86.28**	79.20	78.58	79.65	77.78	**80.32**

Note. Best and second-best RL means within each initial condition and split are bold and underlined, respectively. All scores are measured after 100 RL updates. Pre-RL scores (Init) and full mean ± standard deviation results appear in Table 9. — denotes an aggregate benchmark score.

Across the three additional model/initialization settings, P2TPO attains the highest reported mean in 37 of 39 split–setting comparisons and exceeds GSPO in all 39. It leads all 13 comparisons with Qwen3.5-4B/50K; the two exceptions are Qwen3.5-2B/50K on DocPTBench SynReal (-0.16 versus GDPO) and Qwen3-VL-2B-Instruct/700K on PureDocBench Digital (-0.06 versus GDPO).

4.5 HYPERPARAMETER SENSITIVITY AND EXTENDED TRAINING

Hyperparameter sensitivity. Across one-at-a-time sweeps, all configurations stay within 0.95 points of the default on each split and exceed the strongest response-level baseline on at least 11 of 13 splits, indicating robustness within the tested ranges (Table 10, Appendix C.5).

Extended training horizon. The 500-update extension uses Qwen3.5-2B/50K, 64,000 RL records, and five runs per method. We evaluate OmniDocBench v1.6 every ten updates and the other 12 splits every 50. P2TPO attains the highest mean on all 13 splits at update 500. On Real5-OmniDocBench Skew and DocPTBench SynReal, its changes from updates 100 to 500 are $+0.56$ and $+0.02$ points, while all four response-level baselines decline (Figure 4). Across the 12 added splits, P2TPO improves on seven and declines on five. Appendix C.6 reports the setup and all additional checkpoints.

5 THEORETICAL ANALYSIS

What is lost in aggregation? P2BPO compares individual reference blocks; element-group feedback pools multiple blocks, such as those of the same type, and page feedback pools the groups. For common matched rollouts, let $\mathbf{Q}$ contain scores of K blocks, $\mathbf{H}$ center across rollouts, and $\mathbf{B} = \mathbf{HQ}$. With fixed positive normalized weights, $\mathbf{M}$ pools blocks into C nonempty groups and $\boldsymbol{\alpha}$ combines their scores into an additive page utility:

$$\underbrace{\mathbf{B}}_{\text{reference blocks}} \longmapsto \underbrace{\mathbf{BM}}_{\text{element groups}} \longmapsto \underbrace{\mathbf{BM}\boldsymbol{\alpha}}_{\text{page}}. \tag{5}$$

When $1 < C < K$, both maps are non-invertible: grouping can cancel opposing block contrasts, while page aggregation can additionally cancel contrasts between groups. Retaining reference-block identity preserves distinctions that these aggregations can erase before normalization. Appendix D.1 proves this result (Proposition D.1) and quantifies the coefficient-reconstruction error imposed by coarser routing (Lemma D.2).

Appendix D.2 establishes how UAP activates and routes feedback for omitted or uniformly imperfect outcomes. Appendix D.3 proves SCR's source-block budget conservation and the conditions on redistribution strength and mask precision under which more credit reaches the intended targets.

6 DISCUSSION AND LIMITATIONS

Our theory establishes coefficient-level properties under fixed alignments, masks, and a fixed-weight additive utility; it does not guarantee gradient conservation, unbiased policy optimization, policy improvement, or omission recovery. Routing also depends on reference annotations, element alignment, and boundary proxies, which can be unreliable under annotation noise, repeated text, or alternative valid serializations. Quality gating restricts where SCR is applied but does not establish localization accuracy. Finally, the ablations, frozen-response diagnostics, and selected output examples provide complementary evidence, but independently annotated recipient correctness, omission recovery, and mask precision are still needed to verify the recovery mechanism. Reported means and standard deviations describe the observed runs; paired analyses are needed to assess statistical significance.

7 CONCLUSION

We introduced P2TPO for structure-aligned RL post-training of full-page document parsers. It preserves reference-element identity, unifies feedback for missing outcomes and imperfect consensus through UAP, and redistributes credit within blocks through budget-preserving SCR. Our analysis characterizes local quality contrasts that aggregation can erase. Across six benchmarks, P2TPO exceeds four response-level baselines in mean score on all 13 primary splits, with gains across additional initializations, model scales, and backbones. Ablations further support the value of combining complete outcome feedback with alignment-informed token allocation.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval and discovery, drafting sections, improving the clarity and language of the manuscript, and developing proofs of mathematical claims. The authors reviewed all AI-assisted content, checked the cited references against the original sources, and verified the mathematical arguments and their assumptions. The authors are responsible for the final text, references, mathematical arguments, and reported results.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Yang Bai, Kaiyuan Liu, Ziyuan Zhuang, Jiahong Zhou, Rongxiang Weng, Xin Chen, Jingang Wang, and Xunliang Cai. Multi-objective and mixed-reward reinforcement learning via reward-decorrelated policy optimization. *arXiv preprint arXiv:2605.13641*, 2026. URL https://arxiv.org/abs/2605.13641.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fUtxNAKpdV.

Peng Cai, Zhaofan Zou, Shifa Liu, Yikun Wang, Jiawei Tang, Kaicheng Yang, Meng Tong, Zhongjiang He, and Hao Sun. NaviDC-OCR: Navigating document parsing across digital and camera-captured documents. *arXiv preprint arXiv:2608.12898v2*, 2026. URL https://arxiv.org/abs/2608.12898.

Xiangyang Chen, Shuzhao Li, Xiuwen Zhu, Yongfan Chen, Fan Yang, Cheng Fang, Lin Qu, Xiaoxiao Xu, Hu Wei, and Minggang Wu. Logics-Parsing technical report. *arXiv preprint arXiv:2509.19760*, 2025. URL https://arxiv.org/abs/2509.19760.

DocPTBench authors. Official document-parsing evaluation instructions, 2025. URL https://github.com/Topdu/DocPTBench/blob/main/docs/parsing.md. Accessed September 19, 2026.

Yongkun Du, Pinxuan Chen, Xuye Ying, and Zhineng Chen. DocPTBench: Benchmarking end-to-end photographed document parsing and translation. *arXiv preprint arXiv:2511.18434*, 2025. URL https://arxiv.org/abs/2511.18434.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. *arXiv preprint arXiv:2505.10978*, 2025.

Zuming Huang, Jun Huang, Kexuan Ren, Baode Wang, Weizhen Li, Jianming Feng, Yu Wang, Yichen Yao, Shijun Lin, Yige Tang, Cheng Peng, Weidi Xu, Wei Chu, Yinghui Xu, and Yuan Qi. Infinity-Parser2 technical report. *arXiv preprint arXiv:2607.07836*, 2026. URL https://arxiv.org/abs/2607.07836.

Hunyuan Vision Team et al. HunyuanOCR technical report. *arXiv preprint arXiv:2511.19575*, 2025.

inclusionAI. FinixDocBench public release: Dataset card and Markdown evaluation protocol, 2026. URL https://huggingface.co/datasets/inclusionAI/FinixDocBench. Accessed September 19, 2026.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *European Conference on Computer Vision*, pp. 155–172. Springer, 2022. doi: 10.1007/978-3-031-19815-1_9. URL https://arxiv.org/abs/2111.15664.

Gengluo Li et al. StrucTab: A structured optimization framework for table parsing. *arXiv preprint arXiv:2606.29905*, 2026a.

Gengluo Li et al. Towards real-world document parsing via realistic scene synthesis and document-aware training. *arXiv preprint arXiv:2603.23885*, 2026b. URL https://arxiv.org/abs/2603.23885.

Zhiheng Li et al. How far is document parsing from solved? PureDocBench: A source-traceable benchmark across clean, degraded, and real-world settings. *arXiv preprint arXiv:2605.07492*, 2026c. URL https://arxiv.org/abs/2605.07492.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Haotian Liu et al. GD^{2}PO: Mitigating multi-reward conflicts via group-dynamic reward-decoupled policy optimization. *arXiv preprint arXiv:2606.16771*, 2026a. URL https://arxiv.org/abs/2606.16771.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=Y1rHxA58If.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Shiyin Lu, Yinglun Li, Yu Xia, Yuhui Chen, An-Yang Ji, Jun-Peng Jiang, Qing-Guo Chen, Jianshan Zhao, En Lin, Haijun Li, Cheng Qin, Zhao Xu, and Weihua Luo. OvisOCR2 technical report. *arXiv preprint arXiv:2607.13639*, 2026. URL https://arxiv.org/abs/2607.13639.

Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, and Xingzhong Xu. NGRPO: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*, 2025. URL https://arxiv.org/abs/2509.18851.

OmniDocBench contributors. OmniDocBench official repository and evaluation protocol, 2026. URL https://github.com/opendatalab/OmniDocBench. Accessed September 19, 2026.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ouyang_OmniDocBench_Benchmarking_Diverse_PDF_Document_Parsing_with_Comprehensive_Annotations_CVPR_2025_paper.html.

Jake Poznanski, Luca Soldaini, and Kyle Lo. olmOCR 2: Unit test rewards for document OCR. *arXiv preprint arXiv:2510.19817*, 2025.

PureDocBench authors. Official inference and scoring instructions, 2026. URL https://github.com/zhihengli-casia/PureDocBench/blob/main/docs/INFERENCE_SCORING.md. Accessed September 19, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Said Taghadouini, Adrien Cavaillès, and Baptiste Aubertin. LightOnOCR: A 1b end-to-end multilingual vision-language model for state-of-the-art OCR. *arXiv preprint arXiv:2601.14251*, 2026.

Baode Wang, Biao Wu, Weizhen Li, Meng Fang, Zuming Huang, Jun Huang, Haozhe Wang, Yanjie Liang, Ling Chen, Wei Chu, and Yuan Qi. Infinity Parser: Layout aware reinforcement learning for scanned document parsing. *arXiv preprint arXiv:2506.03197*, 2025. URL https://arxiv.org/abs/2506.03197.

Chengye Wang, Lin Fu, Zexi Kuang, and Yilun Zhao. TexOCR: Advancing document OCR models for compilable page-to-LaTeX reconstruction. *arXiv preprint arXiv:2604.22880*, 2026a.

Hang Wang et al. FinixDoc: Rethinking financial document parsing beyond saturated benchmarks. *arXiv preprint arXiv:2608.22842*, 2026b. URL https://arxiv.org/abs/2608.22842.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9423–9439. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.510/.

Yunhao Wang, Binghong Wu, Zhenyu Huang, Jiacheng Shi, Shuo Huang, Tinghao Yu, and Feng Zhang. DocPO: Advancing document policy optimization via tailored step-aware rewards. *arXiv preprint arXiv:2608.00536*, 2026c.

Junyu Xiong, Yonghui Wang, Weichao Zhao, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. DocR1: Evidence page-guided GRPO for multi-page document understanding. *arXiv preprint arXiv:2508.07313*, 2025. URL https://arxiv.org/abs/2508.07313.

Yunfan Yang, Cuiling Lan, Jitao Sang, and Yan Lu. CSPO: Alleviating reward ambiguity for structured table-to-LaTeX generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25374–25391. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1163. URL https://aclanthology.org/2026.acl-long.1163/.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025a. URL https://arxiv.org/abs/2503.14476.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. DocThinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. *arXiv preprint arXiv:2508.08589*, 2025b. URL https://arxiv.org/abs/2508.08589.

Jiarui Zhang et al. MonkeyOCR v1.5 technical report: Unlocking robust document parsing for complex patterns. *arXiv preprint arXiv:2511.10390*, 2025.

Zelun Zhang et al. PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *arXiv preprint arXiv:2606.03264*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. URL https://arxiv.org/abs/2507.18071.

Changda Zhou, Ziyue Gao, Xueqing Wang, Tingquan Gao, Cheng Cui, Jing Tang, and Yi Liu. Real5-OmniDocBench: A full-scale physical reconstruction benchmark for robust document parsing in the wild. *arXiv preprint arXiv:2603.04205*, 2026. URL https://arxiv.org/abs/2603.04205.

APPENDIX OVERVIEW

The supplementary material presents reward-interface comparisons, method details and algorithms, experimental protocols and results, theoretical analysis, and qualitative case studies, in that order. Section titles and page numbers below link directly to their destinations.

Contents **Page**

A Supplementary Related Work and Reward Interfaces 14
A.1 Reward targets and optimizer signals 14
A.2 Complementary mechanisms and possible adaptations 14
A.3 Scope of reported parsing recipes 15

B Method Extensions and Algorithms 15
B.1 Algorithmic inputs and credit flow 15
B.2 Credit construction and policy update 15

C Experimental Details and Supplementary Results 17
C.1 Experimental protocols 17
C.2 Component ablations 21
C.3 Credit-assignment diagnostics 22
C.4 Generality across initialization, scale, and backbone 22
C.5 Robustness to hyperparameter choices 23
C.6 Extended training over 500 updates 24
C.7 Computational cost 25

D Supplementary Theoretical Analysis 29
D.1 What is lost in aggregation? 29
D.2 How does UAP restore missing feedback? 30
D.3 When does SCR direct feedback to effective targets? 31

E Qualitative Case Studies 31

A SUPPLEMENTARY RELATED WORK AND REWARD INTERFACES

This section supports the Introduction and Related Work by comparing reward targets and optimizer signals, relating P2TPO to CSPO and NGRPO, and clarifying the scope of the reported parsing recipes. The evaluated baselines and component interventions are specified in Appendix C.1.4; the possible adaptations discussed here are not part of the reported experiments.

A.1 REWARD TARGETS AND OPTIMIZER SIGNALS

Table 5 separates the RL prediction target, algorithm, reward construction, and resulting optimizer signal. The target describes the reported RL task rather than every capability of a model. These classifications concern the reported reward interfaces, not an intrinsic restriction of GRPO to response-level feedback.

Table 5: RL targets and optimizer signals in document parsing. The signal column distinguishes response aggregation, functional-attribute routing, and reference-block credit assignment. Target and signal labels are defined below.

Work	Target	RL algorithm	Reward construction	Optimizer signal
Infinity-Parser (Wang et al., 2025)	end2end	GRPO	Edit similarity + paragraph count + reading order	Sum → Scalar
Logics-Parsing (Chen et al., 2025)	end2end	GRPO	Text similarity + box localization + reading order	Sum → Scalar
olmOCR 2 (Poznanski et al., 2025)	end2end	GRPO	Unit-test pass rate + termination and metadata checks	Sum → Scalar
HunyuanOCR (Hunyuan Vision Team et al., 2025)	end2end	GRPO	Parsing: NED-based similarity; format/length validity gate	Scalar
LightOnOCR (Taghadouini et al., 2026)	end2end	GRPO	OCR: unit-test pass rate with repetition, termination, math, and format checks	Scalar
TexOCR (Wang et al., 2026a)	end2end	GRPO-style	Pass rate of binary tests derived from nine metrics	Scalar
Infinity-Parser2 (Huang et al., 2026)	end2end; table; formula	GRPO	Task-selected EDS/TEDS/CDM; doc2json sums spatial mIoU and text EDS	Scalar; Sum → Scalar (doc2json)
MonkeyOCR v1.5 (Zhang et al., 2025)	table	GRPO	Rule check and learned render-and-compare VLM reward	Scalar
StrucTab (Li et al., 2026a)	table	GRPO	Weighted validity + prefix structure + destylized content	Sum → Scalar
CSPO (Yang et al., 2026)	table	GRPO-based	Table-wide functional-attribute rewards; global TEDS + compilation	Attribute → Token (+ global)
DocPO (Wang et al., 2026c)	text; formula; table	GRPO-style	Type-specific patch rewards with training-step sharpening	Scalar
PaddleOCR-VL-1.6 (Zhang et al., 2026)	element	GRPO	Task-specific $\mathrm{Valid} \times \mathrm{Struct} \times \mathrm{Similarity}$	Product → Scalar
P2TPO (ours)	end2end	P2TPO	Type-specific scores for individual matched/missing reference blocks; global page reward	Element type → Ref. block → Token (UAP/SCR; + global)

Target. **end2end**: complete heterogeneous page transcription; **table**, **formula**, **text**: recognition of the specified content type. **element** is reserved for a shared, type-prompted region recognizer within a staged page parser; each call transcribes only that region. It does not denote P2TPO's internal reference elements. A single-type RL recipe retains its specific target label, e.g., MonkeyOCR's table recipe.

Optimizer signal. **Scalar**: one normalized response advantage broadcast over its tokens; a test pass rate still belongs here. **Sum/Product → Scalar**: aggregate reward components before normalization. **Attribute → Token**: CSPO normalizes one reward per functional attribute class across rollouts and routes its advantage to that class's code tokens. **Element type → Ref. block → Token**: type selects the scoring rule; each reference-block identity defines its own comparison group; UAP completes missing outcomes and routes their credit; SCR redistributes matched-block credit within its span. P2TPO does not split a pooled type advantage among blocks. Both token-routing methods retain global feedback; Infinity-Parser2's task selection only chooses a reward function.

Here ANLS is average normalized Levenshtein similarity, RoI is region of interest, IoU is intersection over union, mIoU is mean IoU, NED is normalized edit distance, EDS is edit-distance similarity, TEDS is tree-edit-distance similarity, and CDM is character detection matching; VLM denotes a vision-language model.

A.2 COMPLEMENTARY MECHANISMS AND POSSIBLE ADAPTATIONS

Attribute-informed block scoring. A table-block scorer could use CSPO-inspired functional attributes (Yang et al., 2026) to assess content and structure, then supply the resulting block scores to

P2TPO's reference-instance comparisons. This would refine evaluation within aligned table blocks while retaining UAP and SCR.

Possible controlled comparisons. A separate full-page baseline inspired by CSPO could normalize rewards for text, formula, and table categories and route each category advantage to its generated tokens. This adaptation would require a parser and rewards for the full-page output format. Comparing it with P2BPO would examine routing granularity while controlling the scoring functions, missing-outcome handling, normalization rule, and policy objective. To compare advantage-calibration rules, the UAP coefficient in Equation 2 could instead be replaced by NGRPO's virtual-maximum normalization (Nan et al., 2025), retaining outcome completion, omission routing, SCR, and the policy objective. These are feasible controlled adaptations; neither is included in the reported experiments.

A.3 Scope of reported parsing recipes

Recipe scope and auxiliary tasks. HunyuanOCR's row describes its document-parsing RL recipe, not every task supported by the model. LightOnOCR's row covers its OCR recipe; its separate bounding-box recipe uses IoU-based localization rewards and is not element transcription. PaddleOCR-VL-1.6 uses layout detection followed by task-prompted region recognition; its RL tasks include text, formulas, and tables as well as charts, seals, and spotting. DocPO reports mixed text/formula/table patches w/o establishing a two-stage full-page parsing pipeline, so we retain those specific targets. Its step-aware sharpening follows the training step and still yields one response coefficient.

Document question answering. DocR1 (Xiong et al., 2025) and DocThinker (Yu et al., 2025b) optimize answers and reasoning rather than page transcriptions, so they are outside the table's parsing targets. DocR1's EviGRPO sums format, answer ANLS, and evidence-page F1 rewards. DocThinker combines format, answer accuracy, RoI IoU, and question-rephrasing rewards before GRPO. Both use a Sum $\rightarrow$ Scalar interface.

B Method Extensions and Algorithms

This appendix specifies the procedures underlying Section 2: coefficient construction (Algorithm 1), SCR redistribution (Algorithm 2), and the policy update (Algorithm 3). Appendix C.1.3 provides the training settings.

B.1 Algorithmic inputs and credit flow

For each page's rollout group, the evaluation interface supplies page rewards r_i, reference-indexed scores $s_{i,k} \in [0, 1]$, matched and missing index sets $\mathcal{K}_i$ and $\mathcal{M}_i$, and generated-token supports $\mathcal{B}_{i,k}$. SCR additionally receives disjoint correct- and error-span masks for eligible blocks. Missing-element contributions use the normalized routing weights $\nu_{i,k}(t)$ defined in Section 2.4.

B.2 Credit construction and policy update

B.2.1 Page-to-token coefficients

All comparisons are performed within a page's rollout group. Binary switches U and S enable UAP and SCR: $(U, S) = (0, 0)$ gives P2BPO, $(1, 0)$ adds UAP, $(0, 1)$ adds SCR, and $(1, 1)$ gives full P2TPO. With UAP, the comparison set $\mathcal{S}_k$ contains all valid rollouts; otherwise it contains only rollouts matching reference k. Setting $\lambda = 0$ removes consensus feedback while retaining outcome completion and routing; setting $\beta = 0$ gives uniform block allocation.

B.2.2 Span credit redistribution

Algorithm 2 selects the error-span mask for negative coefficients and the correct-span mask for positive coefficients. It mixes uniform and selected-token allocation with strength $\beta \in [0, 1]$, applies a pre-rebate cap $D > 1$, and uniformly restores the removed coefficient mass. For a nonempty selected

Algorithm 1 Construct page-to-token coefficients

Input: Page-group rewards, reference outcomes, token supports, masks, and routing weights
Input: Switches U, S; weights ω, λ; dispersion tolerance ϵ
Output: Token advantages A for the sampled responses

for each page's rollout group **do**
  Compute page advantages A_i^{P} by group normalization of $(r_i)_i$
  Initialize local contributions $C_{i,t} \leftarrow 0$
  for each reference k with nonempty comparison set $\mathcal{S}_k$ **do**
    $Q_{i,k} \leftarrow s_{i,k}$ if matched, otherwise 0, for $i \in \mathcal{S}_k$
    $(\mu_k, \sigma_k) \leftarrow$ mean and population standard deviation of $(Q_{i,k})_{i \in \mathcal{S}_k}$
    if $\sigma_k > \epsilon$ **then**
      $a_{i,k} \leftarrow (Q_{i,k} - \mu_k)/\sigma_k$ for $i \in \mathcal{S}_k$
    else
      $a_{i,k} \leftarrow -U\lambda \max(0, 1 - \mu_k)$ for $i \in \mathcal{S}_k$
    end if
    for each $i \in \mathcal{S}_k$ **do**
      if $k \in \mathcal{K}_i$ **then**
        $d \leftarrow \mathrm{SCR}((i,k), a_{i,k}, S)$ ▷ Algorithm 2
        $C_{i,t} \leftarrow C_{i,t} + a_{i,k} d_t$ for $t \in \mathcal{B}_{i,k}$
      else
        $C_{i,t} \leftarrow C_{i,t} + a_{i,k} \nu_{i,k}(t)$ for response tokens t
      end if
    end for
  end for
  $A_{i,t} \leftarrow (1 - \omega) A_i^{\mathrm{P}} + \omega C_{i,t}$ for response tokens t
end for
return A, held fixed during policy optimization

fraction $p_{i,k} = N_{i,k}/L_{i,k}$, this is equivalent to Equation 3 with

$$\eta_{i,k} = \min\{\beta, p_{i,k}(D - 1 + \beta)\}. \tag{6}$$

The resulting weights have unit mean and preserve the block's signed and absolute credit budgets. Proposition D.4 in Appendix D.3 establishes this equivalence and the conservation identities.

Algorithm 2 Redistribute matched-block credit with SCR

Input: Block (i, k) of length L, coefficient a, and switch S
Input: Eligibility, correct/error masks, strength β, and pre-rebate cap D
Output: Nonnegative weights d with $\sum_t d_t = L$

function $\mathrm{SCR}((i,k), a, S)$
  if $S = 0$, $a = 0$, or the block is ineligible **then**
    return $\mathbf{1}_L$
  end if
  Select error mask u if $a < 0$, otherwise select correct mask u
  $N \leftarrow \sum_{t=1}^{L} u_t$
  if $N = 0$ **then**
    return $\mathbf{1}_L$
  end if
  $d_t^0 \leftarrow 1 - \beta + \beta L u_t / N$ for each t
  $v_t \leftarrow \min(d_t^0, D)$ for each t
  $d_t \leftarrow v_t + 1 - L^{-1} \sum_{j=1}^{L} v_j$ for each t
  return d
end function

B.2.3 Policy objective and update

The policy update uses the clipped PPO surrogate (Schulman et al., 2017). For sampled tokens, the importance ratio is $\rho_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, p, y_{i,<t}) / \pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid x, p, y_{i,<t})$. With clipping width ε and

dual-clip factor γ, define

$$f(\rho, A) = \min\{\rho A, \mathrm{clip}(\rho, 1-\varepsilon, 1+\varepsilon)A\},$$
$$f_{\mathrm{dual}}(\rho, A) = \begin{cases} \max\{f(\rho, A), \gamma A\}, & A < 0, \\ f(\rho, A), & A \geq 0. \end{cases} \tag{7}$$

For minibatch $\mathcal{D}$ and response-token mask $m_{i,t}$, the objective averages over active response tokens:

$$\widehat{\mathcal{J}}(\theta) = \frac{1}{N_{\mathrm{act}}} \sum_{i \in \mathcal{D}} \sum_{t=1}^{T_i} m_{i,t} f_{\mathrm{dual}}\big(\rho_{i,t}(\theta), A_{i,t}^{\mathrm{US}}\big), \qquad N_{\mathrm{act}} = \sum_{i \in \mathcal{D}} \sum_{t=1}^{T_i} m_{i,t}. \tag{8}$$

Advantages and old-policy probabilities remain fixed while differentiating this objective. Component variants use the same loss with their corresponding token coefficients; Table 6 gives the optimization settings.

Algorithm 3 Optimize the shared clipped policy objective

Input: Sampled responses, old-policy probabilities, token advantages A, and masks m
Input: Current policy π_θ, clipping width ε, and dual-clip factor γ
Output: Updated policy parameters θ

1: **for** each optimizer minibatch $\mathcal{D}$ **do**
2: Evaluate current-policy probabilities and form importance ratios $\rho_{i,t}(\theta)$
3: Compute $f_{\mathrm{dual}}(\rho_{i,t}, A_{i,t})$ with fixed advantages and old probabilities
4: $\mathcal{L} \leftarrow -\widehat{\mathcal{J}}(\theta)$ using the response-token mean in Equation 8
5: Update θ by an optimizer step using $\nabla_\theta \mathcal{L}$
6: **end for**
7: **return** θ

C Experimental Details and Supplementary Results

Appendix C.1 collects the shared data, training, comparison, and evaluation protocols. The supplementary results then follow Section 4: component interventions and credit diagnostics, generality, sensitivity, and extended training, followed by computational cost. Reward-interface comparisons appear in Appendix A.

C.1 Experimental protocols

C.1.1 Training data construction and selection

Source data and serialization. We constructed the training corpus by converting the original Infinity-Doc2-5M `Document_Parsing/doc2json` annotations into full-page Markdown (Huang et al., 2026). Equations are written in LaTeX, and tables are written in HTML. Appendix C.1.2 provides the shared SFT and RL parsing prompt.

SFT data selection. We randomly sampled 700K and 50K page-level records from the transformed corpus as subsets for supervised fine-tuning. These subsets provide the SFT initialization settings reported in Tables 1 and 4.

RL data selection. We used an expert model to re-annotate the transformed corpus and compared its annotations with the corresponding original labels. We ranked records by annotation agreement as a proxy for label quality and retained the 100,000 records with the highest agreement. Expert annotations were used solely for quality assessment and sample selection; RL training retained the original annotations. We sampled 12,800 and 64,000 records from this pool for the 100-update and 500-update experiments, respectively.

C.1.2 Document parsing prompt

The following prompt is used for both supervised fine-tuning and reinforcement-learning rollouts. Our preprocessing converts the `Document_Parsing/doc2json` annotations into targets that follow this serialization.

Listing 1: Full-page document parsing prompt.

```
DOC_PARSE_PROMPT = (
    "\nExtract all readable content from the image in natural human
    reading "
    "order and output the result as a single Markdown document. Format
    formulas "
    "as LaTeX. Format tables as HTML: <table>...</table>. Transcribe all
    other "
    "text as standard Markdown. Preserve the original text without
    translation "
    "or paraphrasing."
)
```

C.1.3 Training configuration and run provenance

Training rewards construct optimization coefficients; benchmark metrics evaluate the resulting transcriptions. Table 6 specifies the configuration for P2TPO and its component variants. Within each setting, methods share the initialization, data and evaluation pipelines, sampling settings, and training budget. Baseline-specific normalization, importance weighting, and clipping follow Appendix C.1.4; the P2TPO policy objective is specified in Appendix B.

Table 6: Training configuration for P2TPO and its component variants, which share optimizer and loss settings. Baseline-specific policy objectives are described in Appendix C.1.4.

Setting	Specification
Training framework	veRL with Megatron actor training and asynchronous vLLM generation; BF16.
RL budget	100 updates; 128 page prompts per update; $G = 8$ responses per prompt.
PPO minibatch	128 prompts, expanded to 1,024 responses; one PPO epoch.
Sequence limits	8,192 prompt tokens (including image tokens) and 8,192 response tokens; total model length 16,384.
Rollout sampling	Temperature 1.0, top-$p = 1$, top-k disabled.
Optimizer	Megatron Adam; learning rate 10^{-6}, moments $(0.9, 0.999)$, weight decay 0.01, constant schedule, no warmup.
Clipping and reduction	Gradient norm limit 1.0; PPO ratio bounds $[0.8, 1.2]$ and dual-clip factor 3; global response-token mean.
Numerical ratio safeguard	Clamp the log importance ratio to $[-20, 20]$ before exponentiation.
Regularization	No KL reward or loss; entropy coefficient 0.
Memory and parallelism	Offloading disabled; vLLM GPU-memory utilization 0.6; tensor, context, and pipeline parallel sizes 1.
Hardware	High-performance computing cards: 8 for RL and 64 for SFT; also used for evaluation.
Default credit parameters	$(\lambda, \beta, \omega) = (0.1, 0.5, 0.2)$; pre-rebate cap $D = 8$; dispersion tolerance $\epsilon = 10^{-8}$; SCR thresholds $0.8/0.8/0.8$ for text/formula/table.

Run aggregation. Primary RL Overall scores and type-specific parsing metrics are summarized across five seeds. Each additional initialization, model-scale, and backbone setting also uses five RL runs, varying the RL data seed. The two UAP ablations in Table 2 use the same five paired seeds as full P2TPO.

Study configurations. Table 7 distinguishes the experimental series. All trained comparisons use 128 page prompts per update and the common data, sampling, and optimizer settings in Table 6, with method-specific objectives and the stated interventions. The 100- and 500-update studies share an SFT initialization where indicated, but use separate RL datasets and runs.

C.1.4 Comparison methods

The evaluated response-level baselines are GRPO, GSPO, GDPO, and RDPO; P2BPO, +UAP, and +SCR are component controls (Table 1).

Table 7: Configuration of each experimental series. SFT counts specify initialization data; RL counts specify training records. "All splits" denotes the six benchmarks and 13 splits in Table 1. Overall scores are summarized by means and standard deviations across five runs.

Study	Initialization	RL records / updates	Summary	Evaluation
Primary comparison and UAP/SCR combinations	Qwen3.5-2B; 700K SFT	12,800 / 100	$\mu \pm \sigma$; 5 runs	All splits; content and omission diagnostics on OmniDocBench
Internal UAP interventions	Qwen3.5-2B; 700K SFT	12,800 / 100	$\mu \pm \sigma$; 5 paired runs	All splits
SCR allocation controls	Qwen3.5-2B; 700K SFT	12,800 / 100	$\mu \pm \sigma$; 5 runs	All splits
Initialization and scale	Qwen3.5-2B and 4B; same 50K SFT manifest	12,800 / 100	$\mu \pm \sigma$; 5 data seeds	All splits
Additional backbone	Qwen3-VL-2B-Instruct; 700K SFT	12,800 / 100	$\mu \pm \sigma$; 5 data seeds	All splits
Parameter sensitivity	Qwen3.5-2B; 700K SFT	12,800 / 100	$\mu \pm \sigma$; 5 runs	All splits
Extended training	Qwen3.5-2B; 50K SFT	64,000 / 500	$\mu \pm \sigma$; 5 runs	OmniDocBench v1.6 every ten updates; other 12 splits every 50

Note. The extended-training series is separate from the 100-update comparisons. Frozen-response diagnostics involve no training update and are specified separately in Appendix C.3.

Response-level baselines. **GRPO** (Shao et al., 2024) standardizes scalar page rewards within each prompt's rollout group and broadcasts the resulting advantage across each response. **GSPO** (Zheng et al., 2025) retains group-relative response advantages while using length-normalized sequence importance ratios and sequence-level clipping. **GDPO** (Liu et al., 2026b) normalizes each reward dimension separately within the rollout group, aggregates the normalized dimensions, and applies batch-wise advantage normalization. **RDPO** (Bai et al., 2026) applies magnitude-aware quantile normalization followed by Mahalanobis whitening within active reward subspaces before aggregation. Each evaluated interface produces one advantage per sampled response.

Component ablations. P2BPO assigns relative advantages to matched reference blocks alongside page feedback. Its +UAP variant adds outcome completion, consensus feedback, and omission routing; +SCR redistributes matched-block credit w/o UAP. Full P2TPO combines both components (Appendix B.2).

C.1.5 Benchmark protocols and score conventions

We evaluate frozen checkpoints w/o benchmark-specific fine-tuning. OmniDocBench v1.6, Real5-OmniDocBench, Wild-OmniDocBench, DocPTBench, and PureDocBench all use the repository's OmniDocBench v1.6 evaluation engine. FinixDocBench alone uses its official Markdown evaluator through the repository adapter. Prompts, decoding, preprocessing, and post-processing are fixed across methods within each benchmark. All headline scores range from 0 to 100, with higher values better.

Datasets and evaluated scope. We evaluate the complete published DocPTBench Photographed (SynReal) and Unwarping (Dewarp) sets, each containing 1,381 images (Du et al., 2025; DocPTBench authors, 2025). SynReal comprises 981 simulated captures and 400 real photographs of 100 source documents under four conditions; Dewarp rectifies the entire Photographed set. Only parsing is evaluated: translation metrics such as BLEU, chrF, METEOR, and STEDS are not used.

FinixDocBench combines the 242 Digital pages and 300 Photo pages into one 542-page evaluation set; Track 3 Huge is excluded. We report pooled metrics for this combined set, rather than separate track scores or an equal-weight mean of two track-level scores. Real5-OmniDocBench and Wild-OmniDocBench reuse OmniDocBench source pages and should be interpreted as related robustness evaluations.

Table 8: Benchmark scope and evaluation engines. Page counts specify the evaluated sets. Engine names identify the shared OmniDocBench v1.6 engine or the official FinixDocBench Markdown evaluator.

Benchmark	Pages	Evaluated Scope	Engine
OmniDocBench v1.6 (Ouyang et al., 2025; OmniDocBench contributors, 2026)	1,651	Official v1.6 ground truth, including 1,355 v1.5 pages.	OmniDocBench v1.6
Real5-OmniDocBench (Zhou et al., 2026)	1,355 per condition	Scanning, Warping, Skew, Illumination, Screen photo; shared v1.5 ground truth.	OmniDocBench v1.6
Wild-OmniDocBench (Li et al., 2026b)	1,350	Real captures of a subset of v1.5 pages.	OmniDocBench v1.6
DocPTBench SynReal (Du et al., 2025; DocPTBench authors, 2025)	1,381	Photographed set: simulated and real captures.	OmniDocBench v1.6
DocPTBench Dewarp (Du et al., 2025; DocPTBench authors, 2025)	1,381	Unwarping set: rectified Photographed images.	OmniDocBench v1.6
PureDocBench (Li et al., 2026c; PureDocBench authors, 2026)	1,475 per condition	Clean, Digital, Real; ground truth uses the v1.6 page schema.	OmniDocBench v1.6
FinixDocBench (Wang et al., 2026b; inclusionAI, 2026)	542	Track 1 Digital (242) and Track 2 Photo (300).	FinixDocBench

Metric definitions and aggregation. For each benchmark evaluated with the OmniDocBench v1.6 engine, the headline composite is

$$\mathrm{Overall}_{\mathrm{OmniDocBench}} = \frac{100}{3}\left[(1 - E_{\mathrm{text}}) + S_{\mathrm{CDM}} + S_{\mathrm{TEDS}}\right], \tag{9}$$

where all three input metrics lie in $[0, 1]$ (OmniDocBench contributors, 2026). Text edit distance first pools edit operations within an eligible page, dividing by the sum of maximum prediction/reference lengths, then averages across eligible pages. CDM and TEDS first average element scores within a page and then macro-average across eligible pages. The report reads the evaluator's page-level CDM and TEDS outputs, rather than the raw global element averages. Ground-truth formula/table pages w/o scored records receive zero page similarity.

DocPTBench images and annotations are passed through this same engine. Its Overall is the composite in Eq. (9), not a numerical conversion of a native DocPTBench headline score. The $1 - E_{\mathrm{text}}$ term is part of the OmniDocBench composite definition. Formula edit distance and reading-order edit distance retain their lower-is-better directions and are reported separately; they do not enter this composite. Table 3 presents CDM and TEDS on the normalized $[0, 1]$ scale.

For FinixDocBench, the official evaluator and adapter use

$$S_{\mathrm{FinixDocBench}} = \frac{100(1 - E_{\mathrm{text}}) + 100(1 - E_{\mathrm{order}}) + S_{\mathrm{TEDS},100}}{3}, \tag{10}$$

with the adapter clipping the result to $[0, 100]$ (inclusionAI, 2026). Text and text-only reading-order edit distances are macro-averaged over applicable page groups after within-page pooling; TEDS is averaged over table samples and reported on a $[0, 100]$ scale. FinixDocBench therefore replaces the formula component with reading-order accuracy. Headline scores are compared between methods within a benchmark, not directly across benchmarks with different composites.

Omission diagnostics and uncertainty. Element omission uses eligible reference elements as its denominator. Local omission uses reference characters within matched text blocks and excludes whole-element omissions, so its interpretation depends on matching coverage. Coverage and omission-rate uncertainty are unavailable for Table 3. Across-seed standard deviations describe variability in final evaluation scores; paired document-bootstrap intervals would additionally require source-document identity to be preserved across methods and related capture conditions. The formulas and aggregation rules above specify the inspected evaluation implementation.

C.2 COMPONENT ABLATIONS

Table 1 compares reference-element routing and the UAP/SCR combinations; Table 2 reports the complete intervention results as means ± standard deviations under the configurations in Table 7. This section supplements those results with intervention definitions, protocols, and detailed comparisons.

Reference configurations. Full P2TPO, w/o UAP (P2BPO + SCR), and w/o SCR (P2BPO + UAP) reproduce Table 1. The w/o UAP configuration retains SCR while removing outcome completion, consensus feedback, and omission routing; all other configurations retain outcome completion. Scores and standard deviations are rounded to two decimal places, while score differences are computed before display rounding.

C.2.1 INTERNAL UAP INTERVENTIONS

Ablation protocol. The UAP ablations in Table 2 follow the primary configuration in Table 7, using full P2TPO's five paired seeds and keeping the mixing weight ω fixed.

- **P2TPO w/o consensus feedback ($\lambda = 0$).** Setting $\lambda = 0$ removes the advantage assigned at imperfect reward consensus. Outcome completion, relative normalization for non-consensus groups, omission routing, and SCR are retained.
- **P2TPO w/o omission routing.** Missing reference elements retain zero-valued outcomes in the group statistics, while their token-level advantage contributions are omitted from Equation 4. The removed contributions are not redistributed; consensus feedback and SCR remain enabled.

Scope of the comparison. Section 4.3 summarizes the cross-split effects. The exceptions are Real5-OmniDocBench Scanning, Illumination, and Screen photo and both DocPTBench splits, where removing consensus feedback gives higher means. Removing omission routing gives higher means on Scanning and SynReal.

C.2.2 SCR ALLOCATION CONTROLS

Removing SCR ($\beta = 0$) eliminates both nonuniform weighting and its alignment with token correctness. We therefore compare aligned SCR with two controls that retain the redistribution strength while changing where credit is assigned. The existing full P2TPO and P2TPO w/o SCR (P2BPO + UAP) results in Table 1 provide the aligned and uniform reference configurations.

Experimental protocol. Both controls follow the primary five-run configuration in Table 7, retaining UAP, page feedback, and the remaining training settings. Perturbations operate only on valid tokens within eligible matched blocks, after SCR has computed its final coefficients. The original mask selects error spans for negative block credit and correct positions for positive block credit.

- **Circular shift.** We jointly rotate the selected-token mask and its coefficient sequence by an offset sampled uniformly from $\{0, \ldots, L-1\}$, where L is the valid-token count. This preserves cyclic order and spacing while randomizing alignment with the original token labels. The zero offset is included.
- **Random support.** We sample the same number of valid positions uniformly w/o replacement. Existing selected and unselected coefficient values are copied to the new support and its complement, respectively. This randomizes token correspondence and disrupts the original cyclic span structure.

Both controls preserve the selected-token count and complete multiset of final coefficient values within each block, including the cap-and-rebate results. Hence $\sum_t \widetilde{d}_t = \sum_t d_t = L$, and the signed and absolute source-block credit budgets are unchanged. Reference-derived correct/error labels and the response-validity mask remain fixed; only the SCR support and its associated coefficients are reassigned among valid tokens within the same block. A fixed mask seed makes assignments repeatable for each rollout block at a given training step, using a separate random stream from rollout sampling.

Comparing aligned SCR with these controls examines the contribution of alignment-informed placement at the same per-block coefficient budget; comparing circular shifts with random support examines the role of cyclic span structure. The uniform reference measures the effect of removing redistribution altogether.

Results and analysis. The SCR columns of Table 2 compare the two controls with aligned SCR and uniform allocation. Differences below are in mean Overall-score points.

- **Alignment-informed placement.** The largest gains over random support occur on PureDocBench Real (+1.00), FinixDocBench (+0.89), and PureDocBench Clean (+0.84); over circular shift, the largest gains occur on PureDocBench Digital/Clean (+0.95/+0.83).
- **Comparison with uniform allocation.** Together with the cross-split counts in Section 4.3, these comparisons show that preserving coefficient values and selected-token counts generally does not reproduce aligned SCR's gains. Alignment-informed placement contributes beyond nonuniform weighting alone.
- **Cyclic structure and exceptions.** Circular shift exceeds random support on nine splits, suggesting that preserving cyclic span structure moderates the effect of randomization on most splits. Circular shift also exceeds aligned SCR on FinixDocBench (+0.06) and DocPTBench SynReal/Dewarp (+0.42/+0.18); random support exceeds aligned SCR only on Dewarp (+0.07). Uniform allocation leads all four configurations on Real5-OmniDocBench Scanning and DocPTBench Dewarp. The results therefore favor aligned allocation on most evaluated splits, with dataset-dependent exceptions.

C.3 Credit-assignment diagnostics

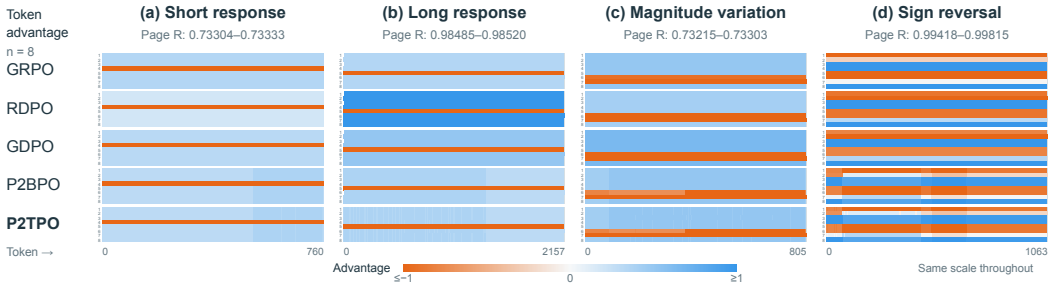


Figure 5: **Frozen responses / no policy update.** Final token-wise advantages on four selected saved rollout groups. Methods are evaluated on common responses; each method row contains eight rollout bands. Horizontal position indexes generated tokens, and panel headers give page-reward ranges. Red/blue denote negative/positive advantages on a shared zero-centered scale, with colors saturating beyond ± 1. Panels show (a,b) shorter/longer responses, (c) magnitude variation across rollouts, and (d) within-response sign changes.

The responses in Figure 5 were sampled from a frozen checkpoint. Its panels (c,d) supply the heatmaps in Figure 3(a,b), respectively; the corresponding lower profiles show responses R6 and R7. P2TPO uses the default credit parameters in Table 6, and RDPO's covariance is fitted to each source batch.

Output-level diagnostics. Table 3 evaluates trained outputs, complementing the coefficient comparisons on fixed outputs in Figure 5. Appendix C.1.5 defines the content and omission metrics and their uncertainty reporting.

C.4 Generality across initialization, scale, and backbone

Table 9 supplements Table 4 with pre-RL scores and standard deviations for the three additional model and SFT initialization settings. Each method is evaluated over five RL runs using distinct data seeds under the configurations in Table 7.

Initialization and model scale. The Qwen3.5-2B experiments use initial checkpoints obtained by SFT on 50K or 700K pages. For the model-scale experiments, Qwen3.5-2B and Qwen3.5-4B checkpoints are prepared using the same 50K-page SFT manifest. Each checkpoint defines a separate setting for comparing RL algorithms, assessing their applicability across initial policies and model scales.

Generality across backbones. Qwen3-VL-2B-Instruct receives document SFT on 700K pages to obtain its initial checkpoint, followed by the same 100-update RL budget. Comparing RL methods from this shared checkpoint evaluates their applicability to another backbone.

Table 9: Cross-benchmark RL evaluation across model scales and SFT-initialized checkpoints. Metrics and splits follow Table 1. Within each panel, Init reports pre-RL scores and RL methods share the starting checkpoint and 100-update budget.

Benchmark	Split	Init	GRPO	GSPO	GDPO	RDPO	P2TPO
(a) **Qwen3.5-2B** SFT data: **50K pages**							
OmniDocBench	v1.6	90.28	91.00 ± 0.44	90.39 ± 0.22	91.05 ± 0.36	90.85 ± 0.25	**91.62** ± 0.30
PureDocBench	Clean	80.60	80.83 ± 1.19	81.01 ± 0.53	81.02 ± 1.00	79.75 ± 1.15	**81.98** ± 0.62
	Digital	76.20	78.26 ± 1.30	78.11 ± 0.44	77.85 ± 0.77	77.27 ± 1.08	**79.16** ± 0.59
	Real	64.64	69.82 ± 1.42	68.66 ± 0.70	69.58 ± 0.81	69.20 ± 0.86	**71.01** ± 0.72
FinixDocBench	—	69.46	69.45 ± 1.93	70.11 ± 0.36	70.95 ± 1.12	70.95 ± 2.89	**73.01** ± 1.72
Real5-OmniDocBench	Scanning	89.27	90.66 ± 0.37	90.23 ± 0.12	90.97 ± 0.23	90.97 ± 0.30	**91.29** ± 0.25
	Warping	87.66	89.56 ± 0.33	89.06 ± 0.26	89.47 ± 0.50	89.53 ± 0.42	**89.86** ± 0.19
	Skew	85.35	88.02 ± 0.75	87.04 ± 0.35	88.11 ± 0.48	87.61 ± 0.68	**88.44** ± 0.29
	Illumination	89.51	90.81 ± 0.49	90.35 ± 0.33	91.10 ± 0.28	90.83 ± 0.24	**91.23** ± 0.27
	Screen photo	88.45	89.79 ± 0.44	89.36 ± 0.33	90.43 ± 0.37	90.24 ± 0.33	**90.50** ± 0.21
DocPTBench	SynReal	74.19	75.57 ± 0.54	75.21 ± 0.39	**76.47** ± 0.48	76.23 ± 1.02	76.31 ± 0.43
	Dewarp	81.72	82.43 ± 0.79	82.47 ± 0.36	83.26 ± 0.71	82.90 ± 0.79	**83.63** ± 0.40
Wild-OmniDocBench	—	80.60	83.67 ± 0.22	82.99 ± 0.37	83.62 ± 0.41	84.27 ± 0.62	**84.48** ± 0.31
(b) **Qwen3.5-4B** SFT data: **50K pages**							
OmniDocBench	v1.6	91.45	92.13 ± 0.41	91.99 ± 0.29	92.07 ± 0.21	91.54 ± 0.70	**92.53** ± 0.26
PureDocBench	Clean	83.50	84.41 ± 0.41	84.28 ± 0.44	84.07 ± 0.37	82.90 ± 1.32	**84.95** ± 0.48
	Digital	79.46	81.80 ± 0.62	81.06 ± 0.38	81.47 ± 0.20	80.14 ± 1.45	**82.34** ± 0.48
	Real	68.77	73.55 ± 0.53	72.52 ± 0.44	73.61 ± 0.26	71.86 ± 1.77	**74.48** ± 0.48
FinixDocBench	—	71.55	74.33 ± 1.65	74.43 ± 0.49	73.59 ± 1.37	69.38 ± 7.14	**76.74** ± 1.49
Real5-OmniDocBench	Scanning	91.18	92.04 ± 0.34	91.89 ± 0.22	91.81 ± 0.18	91.52 ± 0.60	**92.23** ± 0.10
	Warping	89.41	91.02 ± 0.36	90.84 ± 0.43	91.18 ± 0.25	90.22 ± 0.73	**91.45** ± 0.44
	Skew	89.64	90.95 ± 0.23	90.40 ± 0.31	90.78 ± 0.30	90.02 ± 0.67	**91.33** ± 0.33
	Illumination	91.13	92.48 ± 0.39	92.33 ± 0.13	92.59 ± 0.14	91.71 ± 0.66	**92.80** ± 0.30
	Screen photo	90.15	91.91 ± 0.26	91.71 ± 0.23	91.87 ± 0.23	91.08 ± 0.77	**92.16** ± 0.28
DocPTBench	SynReal	73.86	77.86 ± 0.86	76.51 ± 0.73	77.76 ± 0.72	76.37 ± 1.22	**78.19** ± 0.40
	Dewarp	82.55	84.02 ± 0.65	83.98 ± 0.28	84.11 ± 0.66	83.23 ± 1.25	**84.88** ± 0.58
Wild-OmniDocBench	—	82.93	86.04 ± 0.78	85.34 ± 0.30	85.83 ± 0.37	85.10 ± 0.77	**86.28** ± 0.20
(c) **Qwen3-VL-2B-Instruct** SFT data: **700K pages**							
OmniDocBench	v1.6	90.27	90.45 ± 0.24	90.55 ± 0.16	90.57 ± 0.23	88.25 ± 1.17	**91.29** ± 0.09
PureDocBench	Clean	78.32	81.81 ± 0.20	81.80 ± 0.24	81.75 ± 0.30	79.67 ± 0.89	**81.95** ± 0.33
	Digital	71.77	76.08 ± 0.29	76.29 ± 0.13	**76.57** ± 0.20	73.98 ± 1.07	76.51 ± 0.28
	Real	59.72	65.45 ± 0.45	65.44 ± 0.30	65.76 ± 0.25	63.53 ± 0.96	**66.25** ± 0.29
FinixDocBench	—	64.55	67.68 ± 0.63	67.75 ± 0.61	67.95 ± 0.43	65.18 ± 1.41	**68.23** ± 0.24
Real5-OmniDocBench	Scanning	89.11	90.07 ± 0.21	89.42 ± 0.21	89.89 ± 0.32	87.79 ± 0.80	**90.46** ± 0.32
	Warping	84.41	85.19 ± 0.75	84.17 ± 0.47	85.62 ± 0.67	84.46 ± 0.43	**86.37** ± 0.52
	Skew	76.94	76.83 ± 1.48	77.74 ± 0.62	77.82 ± 0.95	76.66 ± 1.52	**79.79** ± 1.24
	Illumination	87.81	89.26 ± 0.35	88.79 ± 0.36	89.38 ± 0.45	87.15 ± 0.55	**89.65** ± 0.23
	Screen photo	87.42	88.46 ± 0.46	88.09 ± 0.30	88.39 ± 0.46	86.01 ± 1.05	**89.33** ± 0.35
DocPTBench	SynReal	62.70	67.80 ± 0.67	67.18 ± 0.85	67.23 ± 0.97	65.91 ± 1.41	**68.94** ± 0.59
	Dewarp	77.02	78.28 ± 0.54	77.38 ± 0.23	77.97 ± 0.67	76.16 ± 0.42	**79.34** ± 0.53
Wild-OmniDocBench	—	76.47	79.20 ± 0.58	78.58 ± 0.13	79.65 ± 0.40	77.78 ± 0.38	**80.32** ± 0.19

Note. Init reports the measured pre-RL score of the shared SFT-initialized checkpoint; no standard deviation is reported. RL entries are mean ± standard deviation over five runs with distinct RL data seeds; best and second-best RL means are bold and underlined, respectively. All panels use 100 RL updates with 128 page prompts per update (12,800 page-prompt presentations). The Qwen3.5-2B/700K setting is reported in Table 1.

C.5 Robustness to hyperparameter choices

We vary one parameter at a time around the default P2TPO configuration $(\lambda, \beta, \omega) = (0.1, 0.5, 0.2)$ while holding the other two fixed (Figure 6; Table 10). Here λ controls imperfect-consensus feedback, β controls SCR concentration, and ω controls the local contribution relative to page feedback. Entries report means ± standard deviations over five seeds; the default column reproduces Table 1.

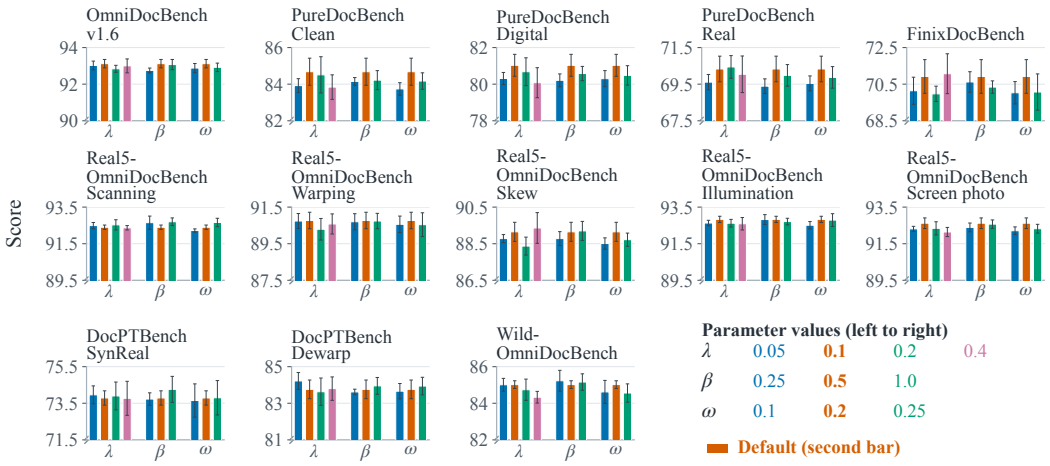


Figure 6: One-at-a-time sensitivity of P2TPO to the imperfect-consensus strength λ, span-concentration strength β, and page–local mixing weight ω. In each sweep, the other two parameters remain at their default values. Bars show five-seed means; error bars show standard deviations.

Table 10: One-at-a-time parameter sensitivity. Benchmark names and splits follow Table 1. The default column uses $(\lambda, \beta, \omega) = (0.1, 0.5, 0.2)$. The λ, β, and ω columns vary the indicated parameter while keeping the other two at their default values.

		Default	λ sweep			β sweep		ω sweep	
Benchmark	Split	(0.1, 0.5, 0.2)	0.05	0.2	0.4	0.25	1.0	0.1	0.25
OmniDocBench	v1.6	93.12 ± 0.23	93.02 ± 0.24	92.84 ± 0.19	93.00 ± 0.38	92.75 ± 0.13	93.07 ± 0.28	92.88 ± 0.25	92.92 ± 0.23
PureDocBench	Clean	84.68 ± 0.74	83.92 ± 0.38	84.51 ± 0.99	83.84 ± 0.67	84.15 ± 0.21	84.22 ± 0.52	83.74 ± 0.34	84.17 ± 0.45
	Digital	81.03 ± 0.60	80.31 ± 0.33	80.68 ± 0.76	80.08 ± 0.82	80.21 ± 0.35	80.58 ± 0.39	80.30 ± 0.44	80.48 ± 0.53
	Real	70.33 ± 0.70	69.61 ± 0.42	70.44 ± 0.62	70.04 ± 1.00	69.39 ± 0.40	69.98 ± 0.59	69.54 ± 0.41	69.87 ± 0.59
FinixDocBench	—	70.92 ± 0.92	70.14 ± 0.74	69.97 ± 0.42	71.07 ± 1.09	70.62 ± 0.56	70.34 ± 0.34	70.03 ± 0.61	70.07 ± 0.99
Real5-OmniDocBench	Scanning	92.40 ± 0.12	92.49 ± 0.17	92.53 ± 0.28	92.37 ± 0.12	92.65 ± 0.36	92.70 ± 0.21	92.22 ± 0.09	92.66 ± 0.23
	Warping	90.77 ± 0.45	90.74 ± 0.41	90.29 ± 0.59	90.58 ± 0.55	90.70 ± 0.44	90.74 ± 0.42	90.56 ± 0.45	90.54 ± 0.65
	Skew	89.15 ± 0.52	88.78 ± 0.22	88.37 ± 0.49	89.36 ± 0.84	88.78 ± 0.38	89.20 ± 0.52	88.51 ± 0.31	88.72 ± 0.36
	Illumination	92.83 ± 0.17	92.63 ± 0.15	92.62 ± 0.21	92.59 ± 0.34	92.82 ± 0.27	92.71 ± 0.18	92.51 ± 0.20	92.79 ± 0.35
	Screen photo	92.62 ± 0.29	92.31 ± 0.14	92.34 ± 0.36	92.14 ± 0.25	92.39 ± 0.24	92.57 ± 0.23	92.21 ± 0.21	92.32 ± 0.24
DocPTBench	SynReal	73.79 ± 0.39	73.96 ± 0.49	73.90 ± 0.76	73.77 ± 0.93	73.73 ± 0.34	74.26 ± 0.71	73.65 ± 0.91	73.80 ± 0.94
	Dewarp	83.76 ± 0.51	84.22 ± 0.46	83.63 ± 0.75	83.80 ± 0.64	83.62 ± 0.15	83.95 ± 0.46	83.66 ± 0.42	83.94 ± 0.48
Wild-OmniDocBench	—	85.02 ± 0.21	85.01 ± 0.35	84.74 ± 0.58	84.33 ± 0.32	85.23 ± 0.57	85.16 ± 0.45	84.62 ± 0.63	84.56 ± 0.50

The sweeps cover $\lambda \in \{0.05, 0.1, 0.2, 0.4\}$, $\beta \in \{0.25, 0.5, 1.0\}$, and $\omega \in \{0.1, 0.2, 0.25\}$. For any fixed split, the maximum difference between tested values is at most 1.10 points for λ and 0.94 points for either β or ω; every reported mean remains within 0.95 points of its default. All eight distinct RL configurations improve upon their shared pre-RL scores on every split, and each exceeds the strongest response-level baseline on at least 11 of 13 splits. The only exceptions to the latter comparison are Real5-OmniDocBench Warping and Skew at $\lambda = 0.2$, and Wild-OmniDocBench at $\lambda = 0.4$ or $\omega \in \{0.1, 0.25\}$. These bounded score changes and retained cross-benchmark gains indicate robustness within the tested one-at-a-time ranges.

C.6 Extended training over 500 updates

GRPO, GSPO, GDPO, RDPO, and full P2TPO (UAP and SCR) each start from the same Qwen3.5-2B checkpoint obtained by SFT on 50K pages and undergo 500 RL updates. The extension uses 64,000 RL records, 128 page prompts per update, and five runs per method with distinct RL data seeds. All 13 splits use this initialization and training protocol; OmniDocBench v1.6 is evaluated every ten updates, and the other 12 splits every 50. The measured pre-RL Overall scores are 90.28 on OmniDocBench v1.6, 85.35 on Real5-OmniDocBench Skew, and 74.19 on DocPTBench SynReal; these provide the Init references in Figure 4. As summarized in Table 7, this series uses separate RL runs and data

from the 100-update experiments in Tables 1 and 4; changes below use its own checkpoint means. Figure 4(a) shows the OmniDocBench v1.6 trajectory; Table 11 reports selected checkpoints and temporal averages. The additional evaluations span updates 50 through 500. Tables 12–14 report all 120 additional split–checkpoint results, with means and standard deviations across five runs. Metrics and split definitions follow Table 1.

Table 11: Five-run checkpoint scores (mean ± standard deviation) and temporal averages. Δ is update 500 minus update 100, computed before rounding. Early and late windows average the ten checkpoint means at updates 110–200 and 410–500, respectively; these summarize repeated evaluations of the same runs.

Method	Update 100	Update 500	Δ	Early window	Late window
GRPO	90.60 ± 0.48	90.14 ± 0.70	−0.46	90.89	90.45
GSPO	90.48 ± 0.28	90.31 ± 0.10	−0.17	90.66	89.95
GDPO	90.99 ± 0.32	90.42 ± 0.65	−0.57	90.99	90.41
RDPO	91.05 ± 0.26	89.91 ± 1.42	−1.14	90.86	89.86
P2TPO	**91.38** ± 0.42	**91.49** ± 0.65	+0.12	**91.44**	**91.24**

Performance retention. P2TPO peaks at 91.72 at update 180, falls to 90.78 at update 410, and recovers to 91.49 at update 500: 1.21 points above the pre-RL score and 0.22 below its peak (Figure 4(a)). All four baselines also fluctuate. Between the early and late windows, P2TPO decreases by 0.20 points, versus 0.44, 0.71, 0.58, and 0.99 for GRPO, GSPO, GDPO, and RDPO (Table 11). GSPO recovers to 90.31 after reaching 89.64 at update 350, so its smaller endpoint decline does not describe the full trajectory. P2TPO has the highest mean at 45 of 50 checkpoints. These comparisons support better retention of gains despite fluctuations; they do not establish monotonic improvement or uniformly lower variability. Differences and checkpoint rankings are computed before display rounding.

Performance across additional splits. P2TPO has the highest mean at update 500 on all 12 additional splits, with margins of 0.25–1.94 points over the strongest response-level baseline in each split. It strictly leads at 78 of the 120 evaluated split–checkpoint pairs and ties for the highest mean at two more. Figure 4(b,c) presents complementary patterns: Real5-OmniDocBench Skew improves from 88.44 to 89.00 between updates 100 and 500, whereas DocPTBench SynReal remains nearly unchanged at 76.34 versus 76.36. Their final margins over the strongest baseline are 1.94 and 1.60 points, respectively. Skew also rises across the final four evaluations (88.12, 88.41, 88.66, and 89.00), while SynReal fluctuates around its earlier performance.

Across all 12 additional splits, P2TPO improves from update 100 to 500 on seven and decreases on five: PureDocBench Clean (-0.44), Digital (-0.99), and Real (-0.11), Real5-OmniDocBench Illumination (-0.53), and DocPTBench Dewarp (-0.54). Baseline degradation is also task-dependent: all five methods improve on FinixDocBench over this interval, and GSPO improves on all three PureDocBench splits. Thus the extension supports stronger comparative performance at the final checkpoint, with neither universal baseline degradation nor monotonic P2TPO improvement.

Mechanistic interpretation. Element-indexed contrasts, feedback at imperfect consensus, and intra-block redistribution may preserve weak local signals and reduce uninformative updates late in training. This is a plausible explanation for the observed retention, rather than a demonstrated reduction in optimization noise. Benchmark variability does not measure gradient noise, and repeated checkpoints are not independent samples for significance tests.

C.7 Computational cost

The primary 100-update P2TPO run requires approximately 60 GPU-hours on eight high-performance computing cards. The full credit-assignment mechanism takes 1.215 s/update, compared with 0.707 s/update for GRPO, adding 0.51 s/update, equivalent to approximately 0.19% of per-update wall time. The recorded advantage-computation shares are 0.465% for P2TPO and 0.265% for GRPO. Generation and actor-update timing distributions are comparable to GRPO.

Table 12: PureDocBench and FinixDocBench during 500-update Qwen3.5-2B/50K training. Entries are Overall scores (mean ± standard deviation over five runs), evaluated every 50 updates. Bold denotes the highest mean at each checkpoint, determined before rounding.

Split	Update	GRPO	GSPO	GDPO	RDPO	P2TPO
PureDocBench						
Clean	50	**80.69** ± 0.60	79.86 ± 0.30	80.29 ± 0.44	80.03 ± 0.67	80.51 ± 0.52
	100	80.64 ± 1.14	79.97 ± 0.63	81.03 ± 2.25	80.68 ± 0.48	**81.72** ± 0.80
	150	80.54 ± 0.60	80.24 ± 0.49	81.09 ± 1.23	80.58 ± 0.79	**81.45** ± 0.28
	200	79.88 ± 0.90	80.49 ± 0.59	**81.95** ± 0.78	80.42 ± 1.04	80.36 ± 0.82
	250	80.23 ± 0.48	80.17 ± 0.83	80.68 ± 0.63	80.94 ± 1.35	**81.20** ± 0.39
	300	79.99 ± 1.06	80.93 ± 1.02	79.93 ± 1.63	79.68 ± 0.78	**81.24** ± 0.61
	350	80.26 ± 1.58	79.70 ± 1.39	80.96 ± 1.53	79.81 ± 0.42	**81.06** ± 0.71
	400	80.49 ± 0.79	79.80 ± 0.96	80.87 ± 0.81	78.91 ± 1.30	**81.69** ± 0.77
	450	79.93 ± 2.10	78.99 ± 1.16	78.51 ± 1.36	78.85 ± 2.06	**82.02** ± 1.50
	500	79.59 ± 1.37	80.34 ± 0.42	79.37 ± 1.23	78.97 ± 0.46	**81.28** ± 0.38
Digital	50	77.41 ± 0.67	76.63 ± 0.50	77.56 ± 0.45	77.42 ± 0.87	**77.58** ± 0.79
	100	77.62 ± 1.35	76.65 ± 0.73	78.67 ± 2.21	78.56 ± 0.40	**79.26** ± 0.93
	150	77.82 ± 0.66	77.06 ± 0.43	78.18 ± 1.36	77.69 ± 0.62	**78.82** ± 0.23
	200	77.75 ± 0.80	77.79 ± 0.41	**79.09** ± 0.66	77.46 ± 1.10	77.80 ± 0.80
	250	77.77 ± 0.64	77.43 ± 0.90	77.94 ± 0.54	78.26 ± 1.34	**78.64** ± 0.46
	300	77.37 ± 1.23	78.28 ± 0.98	77.17 ± 1.41	77.20 ± 0.54	**78.66** ± 1.06
	350	77.40 ± 1.63	76.18 ± 1.54	**78.34** ± 1.76	76.48 ± 0.60	78.15 ± 0.53
	400	77.53 ± 1.01	76.94 ± 1.04	78.24 ± 0.72	75.98 ± 1.21	**78.77** ± 0.72
	450	77.59 ± 1.83	76.35 ± 1.05	75.43 ± 1.36	76.05 ± 1.93	**79.09** ± 1.44
	500	77.06 ± 1.19	77.43 ± 0.40	77.00 ± 1.11	76.56 ± 0.56	**78.27** ± 0.34
Real	50	69.26 ± 0.43	67.52 ± 0.42	**69.32** ± 0.36	68.65 ± 0.00	69.16 ± 0.42
	100	69.77 ± 0.85	67.41 ± 0.66	69.98 ± 1.96	69.77 ± 0.93	**70.60** ± 0.75
	150	69.93 ± 0.55	68.50 ± 0.57	70.19 ± 0.95	69.96 ± 1.11	**70.74** ± 0.42
	200	69.23 ± 0.52	69.12 ± 0.26	**71.31** ± 0.70	69.22 ± 1.46	70.23 ± 1.03
	250	70.25 ± 0.22	68.82 ± 0.80	70.16 ± 0.67	70.30 ± 1.42	**70.78** ± 0.51
	300	69.71 ± 0.87	69.24 ± 1.16	70.06 ± 1.39	69.60 ± 0.60	**70.99** ± 1.05
	350	69.03 ± 1.91	67.65 ± 1.46	**70.70** ± 1.55	68.57 ± 0.77	70.52 ± 0.91
	400	69.86 ± 0.59	68.20 ± 1.18	69.75 ± 0.77	67.93 ± 1.26	**70.94** ± 0.84
	450	69.73 ± 2.15	68.19 ± 0.69	67.44 ± 1.58	68.15 ± 1.45	**70.99** ± 1.09
	500	69.08 ± 1.40	68.64 ± 0.54	68.86 ± 1.14	68.94 ± 0.75	**70.49** ± 0.53
FinixDocBench						
—	50	69.09 ± 0.45	68.74 ± 0.55	69.35 ± 0.51	**69.46** ± 0.80	**69.46** ± 0.42
	100	69.71 ± 0.70	68.87 ± 0.06	69.18 ± 1.66	69.76 ± 1.28	**71.13** ± 0.72
	150	70.47 ± 0.56	69.31 ± 0.50	69.39 ± 1.28	69.18 ± 1.05	**71.86** ± 0.89
	200	69.88 ± 1.18	70.13 ± 0.43	**71.71** ± 1.06	69.74 ± 1.59	70.60 ± 1.02
	250	69.90 ± 1.01	70.64 ± 0.94	69.35 ± 2.11	**71.64** ± 0.73	71.12 ± 0.35
	300	69.44 ± 1.56	70.87 ± 1.19	69.79 ± 1.93	71.18 ± 1.47	**72.18** ± 1.18
	350	70.22 ± 0.45	68.65 ± 0.86	70.63 ± 1.27	68.29 ± 1.48	**72.18** ± 1.87
	400	70.84 ± 0.48	69.13 ± 1.39	70.41 ± 0.86	69.41 ± 0.61	**70.95** ± 0.50
	450	**71.25** ± 1.20	68.77 ± 1.25	69.29 ± 0.28	69.74 ± 1.27	71.19 ± 1.19
	500	71.38 ± 1.20	69.60 ± 1.09	69.59 ± 1.19	69.80 ± 0.88	**71.63** ± 1.28

Table 13: Real5-OmniDocBench: Scanning, Warping, Skew, and Illumination during 500-update Qwen3.5-2B/50K training. Entries are Overall scores (mean ± standard deviation over five runs), evaluated every 50 updates. Bold denotes the highest mean at each checkpoint, determined before rounding.

Split	Update	GRPO	GSPO	GDPO	RDPO	P2TPO
Real5-OmniDocBench						
Scanning	50	90.83 ± 0.22	89.98 ± 0.22	**91.44** ± 0.16	90.98 ± 0.22	90.87 ± 0.32
	100	90.63 ± 0.25	90.10 ± 0.21	**91.28** ± 0.43	91.25 ± 0.31	91.21 ± 0.41
	150	90.67 ± 0.48	90.36 ± 0.14	**91.17** ± 0.21	90.68 ± 0.31	90.83 ± 0.27
	200	90.53 ± 0.39	90.07 ± 0.66	90.98 ± 0.36	90.06 ± 0.48	**91.18** ± 0.60
	250	90.70 ± 0.26	90.17 ± 0.57	90.68 ± 0.38	89.87 ± 0.66	**90.95** ± 0.21
	300	90.40 ± 0.25	89.88 ± 0.52	90.44 ± 0.65	89.29 ± 0.82	**90.77** ± 0.24
	350	89.13 ± 1.08	88.83 ± 0.97	90.33 ± 0.61	89.65 ± 0.79	**90.42** ± 0.24
	400	90.41 ± 0.38	89.27 ± 0.61	90.16 ± 0.30	89.70 ± 0.54	**90.74** ± 0.08
	450	89.75 ± 1.08	89.63 ± 0.15	89.76 ± 0.87	89.70 ± 0.46	**90.78** ± 0.57
	500	89.77 ± 0.71	89.42 ± 0.36	89.94 ± 0.91	89.27 ± 0.35	**91.40** ± 0.75
Warping	50	89.09 ± 0.35	87.84 ± 0.41	**89.47** ± 0.06	88.93 ± 0.52	89.25 ± 0.29
	100	89.08 ± 0.58	88.43 ± 0.41	**89.93** ± 0.31	89.39 ± 0.26	89.82 ± 0.49
	150	89.40 ± 0.67	88.78 ± 0.28	**89.76** ± 0.31	89.01 ± 0.38	89.75 ± 0.38
	200	88.99 ± 0.50	88.41 ± 0.75	**90.03** ± 0.34	88.81 ± 0.71	89.49 ± 0.61
	250	89.93 ± 0.35	88.51 ± 0.49	89.68 ± 0.64	88.98 ± 0.40	**89.99** ± 0.61
	300	89.07 ± 0.47	89.00 ± 0.48	88.85 ± 0.95	88.10 ± 0.53	**89.96** ± 0.31
	350	88.53 ± 1.19	87.01 ± 0.91	89.32 ± 1.00	88.35 ± 0.77	**89.99** ± 0.57
	400	89.07 ± 0.37	87.59 ± 0.76	89.31 ± 0.25	87.53 ± 1.23	**89.68** ± 0.25
	450	88.59 ± 1.36	87.95 ± 0.75	87.99 ± 1.40	88.21 ± 0.89	**89.71** ± 0.66
	500	88.43 ± 0.93	87.64 ± 0.42	88.52 ± 1.25	88.01 ± 0.45	**90.00** ± 0.88
Skew	50	87.79 ± 0.41	86.09 ± 0.56	**88.28** ± 0.20	88.08 ± 0.59	87.64 ± 0.20
	100	87.76 ± 0.43	86.51 ± 0.24	88.39 ± 0.78	88.39 ± 0.57	**88.44** ± 0.69
	150	87.56 ± 1.22	86.16 ± 0.42	**88.52** ± 0.12	87.69 ± 0.37	88.40 ± 0.61
	200	87.63 ± 0.95	86.31 ± 0.90	**88.81** ± 0.25	87.55 ± 0.49	87.84 ± 0.79
	250	**88.26** ± 0.43	86.29 ± 0.62	88.16 ± 0.43	87.57 ± 0.47	**88.26** ± 0.55
	300	87.54 ± 0.78	86.47 ± 0.81	87.35 ± 0.85	86.75 ± 0.64	**88.16** ± 0.55
	350	86.72 ± 1.20	84.83 ± 1.11	87.88 ± 1.25	86.90 ± 1.12	**88.12** ± 0.50
	400	87.26 ± 0.35	85.99 ± 0.56	86.80 ± 0.38	85.69 ± 1.74	**88.41** ± 0.26
	450	86.73 ± 1.51	86.22 ± 0.41	85.88 ± 1.42	86.25 ± 1.48	**88.66** ± 0.79
	500	87.06 ± 0.96	86.10 ± 0.62	86.64 ± 1.42	85.81 ± 0.92	**89.00** ± 0.77
Illumination	50	90.22 ± 0.30	89.42 ± 0.20	**90.79** ± 0.19	90.46 ± 0.17	90.32 ± 0.36
	100	90.47 ± 0.39	89.58 ± 0.18	90.84 ± 0.35	91.00 ± 0.48	**91.38** ± 0.42
	150	90.79 ± 0.63	90.36 ± 0.41	90.90 ± 0.38	90.75 ± 0.27	**91.28** ± 0.18
	200	90.24 ± 0.67	90.03 ± 0.46	**91.10** ± 0.39	90.56 ± 0.57	90.97 ± 0.59
	250	90.85 ± 0.21	90.31 ± 0.49	91.22 ± 0.61	90.78 ± 0.49	**91.32** ± 0.49
	300	90.36 ± 0.50	90.47 ± 0.43	90.09 ± 0.98	90.17 ± 0.21	**90.90** ± 0.26
	350	90.09 ± 1.07	88.68 ± 0.74	90.49 ± 1.02	90.32 ± 0.61	**90.87** ± 0.47
	400	**90.64** ± 0.47	88.99 ± 0.71	90.37 ± 0.21	89.15 ± 0.44	90.44 ± 0.27
	450	89.65 ± 0.94	89.75 ± 0.16	89.49 ± 0.94	89.78 ± 0.36	**90.84** ± 0.55
	500	89.87 ± 0.85	89.57 ± 0.35	90.35 ± 0.66	90.01 ± 0.21	**90.85** ± 0.55

Table 14: Real5-OmniDocBench: Screen photo; DocPTBench; Wild-OmniDocBench during 500-update Qwen3.5-2B/50K training. Entries are Overall scores (mean ± standard deviation over five runs), evaluated every 50 updates. Bold denotes the highest mean at each checkpoint, determined before rounding.

Split	Update	GRPO	GSPO	GDPO	RDPO	P2TPO
Real5-OmniDocBench						
Screen photo	50	89.94 ± 0.27	88.74 ± 0.41	**90.18** ± 0.25	89.90 ± 0.33	89.74 ± 0.14
	100	89.82 ± 0.40	89.19 ± 0.22	**90.44** ± 0.64	90.06 ± 0.15	90.33 ± 0.31
	150	89.97 ± 0.57	89.59 ± 0.26	89.98 ± 0.40	89.89 ± 0.43	**90.29** ± 0.22
	200	89.75 ± 0.54	89.58 ± 0.41	**90.55** ± 0.30	89.66 ± 0.64	90.31 ± 0.48
	250	90.04 ± 0.15	89.16 ± 0.24	90.18 ± 0.23	89.75 ± 0.47	**90.74** ± 0.60
	300	89.81 ± 0.48	89.29 ± 0.29	89.65 ± 0.95	89.28 ± 0.47	**90.35** ± 0.29
	350	89.17 ± 1.20	87.88 ± 0.82	89.92 ± 0.83	89.02 ± 0.85	**90.25** ± 0.40
	400	90.06 ± 0.39	88.67 ± 0.71	89.80 ± 0.10	88.25 ± 0.93	**90.19** ± 0.24
	450	89.39 ± 1.19	88.60 ± 0.36	88.53 ± 1.24	89.22 ± 0.52	**90.10** ± 0.31
	500	89.66 ± 0.86	88.74 ± 0.22	89.97 ± 1.04	89.34 ± 0.18	**90.49** ± 0.33
DocPTBench						
SynReal	50	75.83 ± 0.39	74.56 ± 0.42	**76.33** ± 0.11	76.28 ± 0.59	75.85 ± 0.24
	100	75.89 ± 0.97	74.86 ± 0.54	**76.99** ± 1.39	76.80 ± 0.69	76.34 ± 0.45
	150	75.41 ± 1.33	75.17 ± 0.89	75.97 ± 0.61	75.40 ± 0.51	**76.48** ± 0.58
	200	75.03 ± 1.09	74.69 ± 0.83	**76.03** ± 1.09	74.94 ± 1.03	75.57 ± 0.93
	250	76.18 ± 0.35	73.37 ± 1.04	75.31 ± 0.82	75.27 ± 0.80	**76.29** ± 1.01
	300	75.01 ± 1.50	74.54 ± 1.23	75.36 ± 1.74	74.21 ± 0.53	**76.20** ± 0.89
	350	73.85 ± 2.83	72.15 ± 1.48	**75.95** ± 2.47	73.51 ± 1.88	75.51 ± 0.85
	400	74.11 ± 1.34	72.28 ± 1.31	74.73 ± 1.21	72.58 ± 2.58	**76.25** ± 0.29
	450	74.87 ± 2.50	72.85 ± 0.57	72.93 ± 2.45	74.72 ± 0.98	**76.57** ± 1.53
	500	74.32 ± 1.57	73.67 ± 1.50	74.26 ± 1.75	74.76 ± 1.03	**76.36** ± 1.28
Dewarp	50	82.87 ± 0.34	80.99 ± 0.64	**83.22** ± 0.21	82.85 ± 0.37	82.93 ± 0.59
	100	82.31 ± 0.79	81.41 ± 0.52	83.48 ± 0.77	83.28 ± 0.52	**83.78** ± 0.78
	150	82.46 ± 1.26	81.89 ± 0.37	**83.49** ± 0.54	82.71 ± 0.66	83.13 ± 0.43
	200	82.43 ± 0.79	82.62 ± 1.05	**83.63** ± 0.42	82.02 ± 0.68	83.01 ± 0.82
	250	82.88 ± 0.52	82.44 ± 0.39	82.69 ± 0.42	82.21 ± 1.35	**82.95** ± 0.84
	300	82.02 ± 1.06	**82.46** ± 0.68	82.09 ± 1.67	81.16 ± 0.88	82.41 ± 0.38
	350	81.51 ± 1.86	80.61 ± 1.35	**82.74** ± 1.35	80.77 ± 1.05	82.65 ± 0.54
	400	81.77 ± 0.48	80.55 ± 1.17	81.93 ± 0.54	80.26 ± 1.63	**82.54** ± 0.30
	450	80.58 ± 2.03	80.65 ± 0.37	80.71 ± 1.73	81.56 ± 0.93	**83.15** ± 1.32
	500	81.62 ± 1.57	80.99 ± 0.78	80.97 ± 1.61	80.76 ± 1.38	**83.24** ± 1.09
Wild-OmniDocBench						
—	50	83.77 ± 0.36	81.70 ± 0.28	**84.32** ± 0.44	83.71 ± 0.11	83.75 ± 0.16
	100	83.36 ± 0.33	82.35 ± 0.45	**84.30** ± 0.76	84.25 ± 0.60	84.00 ± 0.48
	150	83.44 ± 0.95	83.05 ± 0.12	**84.47** ± 0.33	83.22 ± 0.87	84.31 ± 0.35
	200	83.49 ± 0.63	82.63 ± 0.96	**84.40** ± 0.55	83.13 ± 0.35	84.10 ± 0.60
	250	**84.38** ± 0.79	82.64 ± 0.56	83.98 ± 0.49	83.20 ± 0.46	84.29 ± 0.48
	300	83.45 ± 0.68	82.88 ± 0.54	83.38 ± 0.89	82.44 ± 0.61	**84.23** ± 0.13
	350	82.55 ± 1.24	81.26 ± 1.32	83.18 ± 1.13	83.21 ± 0.92	**83.94** ± 0.48
	400	83.33 ± 0.51	81.81 ± 1.17	82.94 ± 0.47	82.27 ± 0.68	**84.14** ± 0.31
	450	82.19 ± 1.19	82.51 ± 1.33	82.37 ± 1.23	82.32 ± 0.65	**83.83** ± 0.41
	500	82.70 ± 0.64	82.09 ± 1.01	83.16 ± 1.03	82.20 ± 0.79	**84.36** ± 0.65

D SUPPLEMENTARY THEORETICAL ANALYSIS

We establish three properties of the credit construction: the information lost by aggregation, the feedback restored by UAP, and the conditions under which SCR increases allocation to effective targets. Alignments, token supports, and masks are held fixed. The aggregation model uses common matched rollouts and a fixed additive utility; the UAP and SCR results concern the coefficients defined in Section 2.

D.1 WHAT IS LOST IN AGGREGATION?

Consider $G \geq 2$ rollouts in which all K reference blocks are matched, and let $\mathbf{Q} \in [0,1]^{G\times K}$ contain their scores. Partition the blocks into C nonempty element groups $\mathcal{J}_c$, with $1 < C < K$. Here a group can contain several blocks of the same element type; it is coarser than an individual reference element. Let $M_{k,c} > 0$ for $k \in \mathcal{J}_c$ and $M_{k,c} = 0$ otherwise, with $\sum_{k\in\mathcal{J}_c} M_{k,c} = 1$. Let $\alpha_c > 0$ and $\sum_c \alpha_c = 1$. Group scores are $\mathbf{R} = \mathbf{QM}$ and the additive page utility is $\mathbf{r} = \mathbf{R}\boldsymbol{\alpha}$. Define $\mathbf{H} = \mathbf{I}_G - G^{-1}\mathbf{1}_G\mathbf{1}_G^\top$ and the unstandardized block contrasts $\mathbf{B} = \mathbf{HQ}$.

Proposition D.1 (Strict hierarchy of aggregation losses)**.** *Block contrasts determine group contrasts* $\mathbf{HR} = \mathbf{BM}$*, which determine page contrasts* $\mathbf{Hr} = \mathbf{BM}\boldsymbol{\alpha}$*. Each reverse determination is impossible in general. More precisely, on the space of centered matrices* $\mathcal{B} = \{\mathbf{B} \in \mathbb{R}^{G\times K} : \mathbf{HB} = \mathbf{B}\}$*,*

$$\{\mathbf{0}\} \subsetneq \{\mathbf{B} \in \mathcal{B} : \mathbf{BM} = \mathbf{0}\} \subsetneq \{\mathbf{B} \in \mathcal{B} : \mathbf{BM}\boldsymbol{\alpha} = \mathbf{0}\}. \tag{11}$$

Both strict losses occur for scores in $[0,1]$*. Consequently, deterministic processing of pooled rewards alone cannot recover all distinctions available at the preceding resolution.*

Proof. The contrast identities follow by linearity. For the first strict inclusion, choose two blocks k, ℓ in the same group c, and a vector $\mathbf{x}$ with $x_k = M_{\ell,c}$, $x_\ell = -M_{k,c}$, and all other entries zero. Then $\mathbf{x} \neq 0$ but $\mathbf{x}^\top\mathbf{M} = 0$. For the second, choose two distinct groups c, d and a vector $\mathbf{v} \in \mathbb{R}^C$ with $v_c = \alpha_d$, $v_d = -\alpha_c$, and other entries zero. Set $x_k = v_j$ for $k \in \mathcal{J}_j$. Column normalization gives $\mathbf{x}^\top\mathbf{M} = \mathbf{v}^\top \neq 0$ but $\mathbf{v}^\top\boldsymbol{\alpha} = 0$. In either construction, take a nonzero centered vector $\mathbf{u} \in \mathbb{R}^G$ and set $\mathbf{B} = \mathbf{u}\mathbf{x}^\top$. For sufficiently small $\delta > 0$, the two score matrices $\mathbf{Q}_0 = \frac{1}{2}\mathbf{1}_G\mathbf{1}_K^\top$ and $\mathbf{Q}_1 = \mathbf{Q}_0 + \delta\mathbf{B}$ lie in $[0,1]^{G\times K}$. In the first construction they have identical group scores but different block contrasts; in the second they have identical page scores but different group contrasts. A deterministic function of these identical pooled inputs cannot distinguish the respective pair. □

This hierarchy concerns scores before standardization. Separate block normalization retains within-block contrasts when their dispersion exceeds the tolerance, but does not preserve their original scales. No information ordering is asserted after independent standardization. The result applies to a page reward only when it has the stated fixed additive form; it does not assume that every evaluation metric or training reward has this form.

Auxiliary routing lemma. Fix one rollout with disjoint nonempty block supports of lengths L_k, and let b_k be its normalized P2BPO coefficients. Their broadcast token vector on the union of these matched-block supports, before page/local composition, is the reconstruction target. Let D_{B}, D_{E}, and D_{P} be the minimum squared token errors using unrestricted coefficients per block, per element group, and per page, respectively. Thus

$$D_{\mathrm{E}} = \min_{\boldsymbol{\gamma}} \sum_c \sum_{k\in\mathcal{J}_c} L_k(b_k - \gamma_c)^2, \qquad D_{\mathrm{P}} = \min_v \sum_k L_k(b_k - v)^2.$$

Write $N_c = \sum_{k\in\mathcal{J}_c} L_k$, $\bar{b}_c = N_c^{-1}\sum_{k\in\mathcal{J}_c} L_k b_k$, and $\bar{b} = (\sum_k L_k)^{-1}\sum_k L_k b_k$.

Lemma D.2 (Resolution of block, element-group, and page routing)**.** *The optimal reconstruction errors satisfy*

$$\begin{aligned} D_{\mathrm{B}} &= 0, \qquad D_{\mathrm{E}} = \sum_c \sum_{k\in\mathcal{J}_c} L_k(b_k - \bar{b}_c)^2, \\ D_{\mathrm{P}} &= D_{\mathrm{E}} + \sum_c N_c(\bar{b}_c - \bar{b})^2. \end{aligned} \tag{12}$$

Hence $D_{\mathrm{B}} \leq D_{\mathrm{E}} \leq D_{\mathrm{P}}$*. The first inequality is strict exactly when some group contains unequal block coefficients; the second is strict exactly when the group means are not all equal.*

Proof. Independent block coefficients reproduce the target. Minimizing each group sum gives $\gamma_c = \bar{b}_c$, and minimizing the page sum gives $v = \bar{b}$. Expanding $b_k - \bar{b} = (b_k - \bar{b}_c) + (\bar{b}_c - \bar{b})$, the cross terms vanish because $\sum_{k \in \mathcal{J}_c} L_k(b_k - \bar{b}_c) = 0$. The decomposition and strictness conditions follow from positivity of the lengths. □

Both inequalities can be strict for actual P2BPO coefficients. With equal block lengths of one, groups $\{1, 2\}$ and $\{3\}$, and rollout scores $(0.9, 0.2, 0.9)$ and $(0.2, 0.9, 0.2)$, population normalization gives $\mathbf{b} = (1, -1, 1)$ in the first rollout when $\epsilon < 0.35$. The errors are $0 < 2 < 8/3$. A global coefficient cannot distinguish blocks within a shared group, while routing indexed by individual reference blocks matches P2BPO's resolution. The minima allow free coefficients: a particular reward-normalization rule need not attain them, and its actual reconstruction errors need not follow this ordering. The lemma compares coefficient expressivity, not policy performance or the optimality of the chosen target coefficients.

D.2 How does UAP restore missing feedback?

For one reference block, consider $n \geq 1$ valid rollouts with completed outcomes $Q_i \in [0, c]$, where $c > 0$ is fixed and missing outcomes equal zero. Let μ and σ be their mean and population standard deviation. The algorithm uses the unit-ceiling specialization $c = 1$. Outcome completion makes omissions available to the comparison, while the ceiling gap supplies feedback when dispersion is too small for relative normalization.

Proposition D.3 (Feedback activation and delivery by UAP)**.** *Let $\lambda > 0$ and $\epsilon \geq 0$. UAP has the following properties.*

1. *If $0 < m < n$ matched rollouts all have score $s > 0$ and the other $n - m$ outcomes are missing, then $\mu = ms/n$ and $\sigma = s\sqrt{m(n-m)}/n$. When $\sigma > \epsilon$,*

$$A^{\mathrm{U}}_{\mathrm{matched}} = \sqrt{\frac{n-m}{m}}, \qquad A^{\mathrm{U}}_{\mathrm{missing}} = -\sqrt{\frac{m}{n-m}}. \tag{13}$$

   Matched-only centering instead gives zero feedback to all matched rollouts and no coefficient for omissions.

2. *Whenever $\sigma \leq \epsilon$, every valid rollout receives $A^{\mathrm{U}}_i = -\lambda(c - \mu)$. This coefficient is negative exactly when $\mu < c$ and vanishes at ceiling quality. In particular, an imperfect tie at $r < c$ receives $-\lambda(c - r)$ and an all-missing group receives $-\lambda c$.*

3. *A missing outcome with coefficient a and routing weights $\nu_t \geq 0$, $\sum_t \nu_t = 1$, contributes $a\nu_t$ at generated tokens, preserving its signed and absolute mass:*

$$\sum_t a\nu_t = a, \qquad \sum_t |a\nu_t| = |a|. \tag{14}$$

Proof. For the mixed group, direct evaluation gives $\sigma^2 = m(n-m)s^2/n^2$. Substitution of the matched score s and missing score zero into $(Q_i - \mu)/\sigma$ yields Eq. (13); the matched-only scores are constant. For the low-dispersion branch, $0 \leq \mu \leq c$ implies $h(\mu, c) = c - \mu$, proving the signs and special cases. Since every $Q_i \leq c$, $\mu = c$ holds only when all outcomes equal c. More generally, centered outcomes alone cannot distinguish $r\mathbf{1}_n$ from $c\mathbf{1}_n$, as both center to zero; UAP retains their distinct uncentered means. Finally, nonnegative normalized routing weights give Eq. (14) by summation. □

With unequal matched scores, omissions receive negative coefficients in the relative branch when $\mu > 0$, while a matched coefficient is positive only if its score exceeds μ. In the low-dispersion branch, all coefficients share the ceiling-gap value, even if an individual outcome reaches the ceiling. The routing identities hold before loss masking and composition; mass assigned to inactive tokens does not enter the masked objective. Section 6 discusses the distinction between coefficient delivery and downstream recovery.

D.3 When does SCR direct feedback to effective targets?

Assume SCR is enabled. Fix an eligible matched block of length $L \geq 2$ with coefficient $A \neq 0$, redistribution strength $\beta \in [0, 1]$, and pre-rebate cap $D > 1$. Its selected mask S targets errors when $A < 0$ and correct positions when $A > 0$. Assume $p = |S|/L \in (0, 1)$. Let $\mathcal{E}$ be an externally specified target set (no statistical independence from S is assumed), $q = |\mathcal{E}|/L$ its prevalence, and $\varrho = |S \cap \mathcal{E}|/|S|$ the mask precision. Thus effectiveness is evaluated against the intended target, rather than defined by membership in S.

Proposition D.4 (Validity and target allocation of SCR)**.** *Algorithm 2's cap-and-rebate operation yields*

$$d_t = 1 - \eta + \frac{\eta}{p}\mathbf{1}[t \in S], \qquad \eta = \min\{\beta, p(D - 1 + \beta)\} \in [0, \beta]. \tag{15}$$

The weights are nonnegative and preserve the signed and absolute source-block budgets:

$$\sum_{t=1}^{L} Ad_t = LA, \qquad \sum_{t=1}^{L} |Ad_t| = L|A|. \tag{16}$$

The target's share of absolute coefficient mass satisfies

$$F_{\mathcal{E}} = \frac{\sum_{t\in\mathcal{E}} |Ad_t|}{L|A|} = q + \eta(\varrho - q). \tag{17}$$

It exceeds uniform allocation exactly when $\eta > 0$ and $\varrho > q$.

Proof. Before capping, selected and unselected weights are $1 - \beta + \beta/p$ and $1 - \beta$. Since $D > 1$, only selected weights can be capped. Set $a = \min\{1 - \beta + \beta/p, D\}$ and $b = 1 - \beta$. Adding the uniform rebate $1 - pa - (1 - p)b$ gives selected weight $1 + (1 - p)(a - b)$ and unselected weight $1 - p(a - b)$. With $\eta = p(a - b) = \min\{\beta, p(D - 1 + \beta)\}$, these equal Eq. (15). Because $0 \leq \eta \leq \beta \leq 1$, weights are nonnegative, and their sum is $L(1 - \eta) + |S|\eta/p = L$. Multiplication by A or $|A|$ proves conservation. Finally,

$$\frac{1}{L}\sum_{t\in\mathcal{E}} d_t = q(1 - \eta) + \frac{|S \cap \mathcal{E}|}{L}\frac{\eta}{p} = q + \eta(\varrho - q).$$

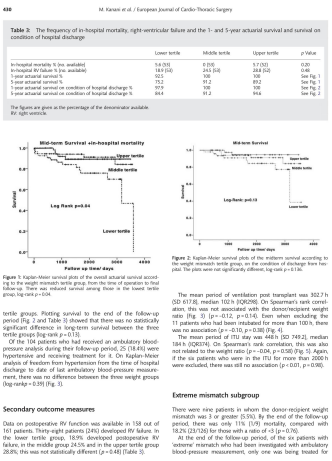
Subtracting the uniform share q proves the strict-improvement condition. □

Disabling SCR, empty or full masks, ineligible blocks, and zero coefficients leave allocation uniform; $\beta = 0$ also gives uniform weights. The cap applies before the rebate, so final weights can exceed D; they satisfy $d_t \leq D + \beta$. With the target set, mask size, and redistribution parameters fixed, a uniformly random mask has expected precision q and expected target allocation q. SCR strictly improves on this expectation when $\eta > 0$ and its mask is more precise than the target's prevalence.

For an eligible reference block matched in every valid rollout with identical scores $0 \leq r < c$, let $\lambda > 0$. UAP supplies $A = -\lambda(c - r)$ while matched-only normalization supplies zero. Substituting this coefficient into Eq. (17), SCR changes the target's absolute mass relative to uniform UAP by $L\lambda(c-r)\eta(\varrho-q)$. This change is strictly positive when $\eta > 0$ and $\varrho > q$, linking feedback activation to localization under an informative mask. All conservation and allocation statements concern the source block before composition with page or omission contributions.

E Qualitative Case Studies

Figures 7–9 supplement Section 4.3 with selected OmniDocBench v1.6 outputs from separate GRPO and P2TPO runs after 100 training updates. Each figure places the source page on the left and reference excerpts, generated excerpts, and per-page edit distances (ED; lower is better) on the right. References come from the evaluator's ground-truth records, and generated excerpts come from saved model outputs. Ellipses indicate abridgment; italic annotations identify omitted content and are not generated text. These examples illustrate output differences without isolating individual credit-assignment components.



Clinical article: omission of a heading and paragraph

Reference excerpt

```
Secondary outcome measures
Data on postoperative RV function was
available in 158 out of 161 patients.
Thirty-eight patients (24%) developed RV
failure.  ...
```

GRPO output excerpt

```
Figure 1:  Kaplan-Meier survival plots ...
```

The intervening text, including the heading and paragraph shown above, is absent.

```
Figure 2:  Kaplan-Meier survival plots of
the midterm survival ...
```

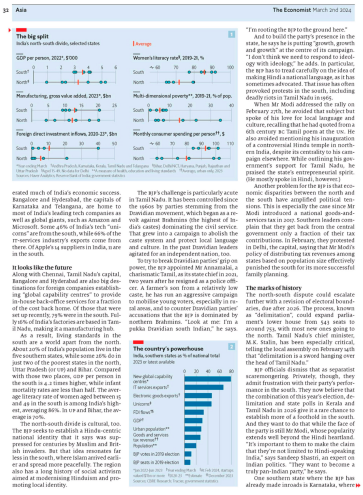
P2TPO output excerpt

```
Secondary outcome measures
Data on postoperative RV function was
available in 158 out of 161 patients.
Thirty-eight patients (24%) developed RV
failure.  ...
```

Comparison. GRPO skips the displayed heading and paragraph between two figure captions; P2TPO transcribes both. The table edit distance is unchanged.

Metric ↓	GRPO	P2TPO
Text ED	0.4780	0.0000
Reading-order ED	0.5000	0.0000
Table ED	0.0663	0.0663

Figure 7: Qualitative comparison of heading and paragraph transcription in a clinical article from OmniDocBench v1.6 (scihub_j.ejcts.2011.04.042.pdf_3).



Magazine article: omission within a multi-column page

Reference excerpt

```
It looks like the future
Along with Chennai, Tamil Nadu's capital,
Bangalore and Hyderabad are also big
destinations for foreign companies
establishing ``global capability centres''
...
```

GRPO output excerpt

```
Monthly consumer spending per person...
```

The subheading and accompanying text shown above are absent.

```
The BJP's challenge is particularly acute
in Tamil Nadu.  ...
```

P2TPO output excerpt

```
It looks like the future
Along with Chennai, Tamil Nadu's capital,
Bangalore and Hyderabad are also big
destinations for foreign companies
establishing ``global capability centres''
...
```

Comparison. GRPO omits the displayed subheading and accompanying text, resuming with the paragraph beginning "The BJP's challenge." P2TPO retains the passage.

Metric ↓	GRPO	P2TPO
Text ED	0.2721	0.0153
Reading-order ED	0.4286	0.0000

Figure 8: Qualitative comparison of text coverage in a multi-column magazine page from OmniDocBench v1.6 (magazine_TheEconomist.2024.03.02_page_032).



Chemistry article: omissions and mathematical notation

Reference excerpt

```
Measurements
Elemental analysis was performed with a
Elementar-CHNO-Rapiol to determine the
carbon, nitrogen, and hydrogen content.
...
```

GRPO output excerpt

```
2,855 cm -1 .  1 H NMR(CDCl 3 ):  ...
```

The "Measurements" heading and elemental-analysis paragraph are absent.

P2TPO output excerpts

```
2,855 cm$^{-1}$.  ...
Measurements
Elemental analysis was performed with a
Elementar-CHNO-Rapiol to determine the
carbon, nitrogen, and hydrogen content.
...
```

Comparison. GRPO omits the displayed "Measurements" heading and paragraph and renders the superscript as plain text. P2TPO retains the passage and uses inline LaTeX notation for the superscript. P2TPO has lower text and reading-order edit distances, whereas GRPO has lower table edit distance.

Metric ↓	GRPO	P2TPO
Text ED	0.4625	0.0427
Reading-order ED	0.5833	0.0000
Table ED	0.0734	0.1152

Figure 9: Qualitative comparison of text coverage and superscript notation in a chemistry article from OmniDocBench v1.6 (docstructbench_llm-raw-scihub-o.O-s00396-003-0935-9.pdf_2).