VERIPO: CLOSING THE REASONING-SUPERVISION GAP IN DOCUMENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with rule-based, verifiable rewards has become the dominant post-training recipe for document vision-language models, yet it treats the decoded trajectory as an undifferentiated scalar-reward target. We observe that this practice clashes with the internal structure of document-QA outputs: a trajectory decomposes into *parse*, *evidence*, *reason*, and *answer* segments of sharply different verifiability, and the reasoning segment admits no rubric at all. Rule-only objectives therefore leave reasoning tokens with no within-trajectory discriminative signal, producing slow, variance-dominated training, a phenomenon we call the *reasoning-supervision gap*. We close this gap with **Verifiability-aware Policy Optimization (VeriPO)**, an on-policy reinforcement-learning algorithm whose token-level advantage combines per-segment verifiable rewards with a token-level reverse-KL signal from a frozen vision-language teacher. A rubric-gated trust coefficient, computed by scoring the teacher against the same rubric the student faces, attenuates teacher supervision wherever the teacher itself fails verification. The design admits a provable *verification-safe* guarantee: under an adversarially uninformative teacher, VeriPO's gradient collapses onto the rule-only baseline and never underperforms it. VeriPO further trains with a single on-policy rollout per update, eliminating the group-rollout overhead of GRPO-style training. Applied to a unified document-QA decoder that emits parse, evidence, rationale, and answer in a single pass with self-citing Document Units, VeriPO advances the state of the art on multi-page document QA, chart and infographic QA, and document parsing, at a fraction of the post-training compute of group-rollout baselines. Three diagnostic studies confirm each of its mechanisms: a component ablation, an adversarial-teacher study, and a per-segment analysis that isolates reasoning as the primary beneficiary of teacher supervision.

1 INTRODUCTION

Scanned documents (reports, receipts, forms, scientific papers, slide decks) encode a disproportionate share of the world's written knowledge in a form that is visual rather than textual. Answering free-form questions over them has accordingly become a central testbed for vision-language models (VLMs), and it stresses two capabilities that are rarely jointly well-served: reading pages faithfully under substantial layout variability, and selecting from what was read the narrow slice of evidence a given question actually requires. Because the two capabilities are coupled (perception errors cascade into downstream reasoning, and reasoning cannot repair what it cannot see), document question answering exposes shortcomings that simpler visual benchmarks let slide.

Three families of methods dominate the current literature. Classical *pipeline* approaches treat OCR or structured parsing as an external pre-processing stage and pass the resulting text to a separate retriever and language model. *Visual RAG* (Yu et al., 2025; Faysse et al., 2024) partially closes the modality gap by retaining evidence in image form and asking a VLM to read it on demand, avoiding the information loss of premature text linearization. Most recently, *reinforcement learning with rule-based, verifiable rewards* has emerged as the standard post-training recipe for pushing VLMs beyond imitation of surface patterns (Zhan et al., 2025; Meng et al., 2025; Deng et al., 2025; Shao et al., 2024), with GRPO and its descendants demonstrating that outcome-level supervision can reshape both reasoning and perception. Across these families, one design pattern is shared and largely

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.

unquestioned: the decoded response is treated as a single object scored by a scalar reward that is broadcast uniformly to its constituent tokens.

This uniform treatment sits uncomfortably with the internal structure of document-QA trajectories, which exhibit a verifiability spectrum across their constituent segments. A well-designed doc-QA policy typically emits *parse* $\rightarrow$ *evidence* $\rightarrow$ *reason* $\rightarrow$ *answer* in sequence, and the four segments differ dramatically in how well they can be scored by a rule. Parse admits a content-plus-layout rubric; evidence admits an existence-and-match check against the parse the model just emitted; and the answer admits a task-standard match score. The *reason* segment admits none of these: any rubric sharp enough to judge a natural-language rationale would collapse onto the downstream answer rubric. We call this uneven structure the *verifiability spectrum*. Its consequence is sharp. Under a rule-only RL objective, reasoning tokens inherit supervision solely by broadcasting a scalar reward that is constant by construction across the entire reasoning span of a given trajectory. Within a trajectory there is therefore no contrast for credit assignment, and the resulting gradient estimator becomes slow and variance-dominated while silently tolerating reasoning behaviors that happen to co-occur with correct answers. We identify this *reasoning-supervision gap* as the structural source of the training instabilities reported in current document RL, and we argue that resolving it requires a second, per-token supervision source that rule-only objectives cannot supply.

We adopt a unified document-QA decoder that emits the four-segment trajectory in one pass, with *Document Units* (DUs) serving as handles that make citations mechanically checkable against the trajectory's own parse prefix. On top of this architecture we propose **Verifiability-aware Policy Optimization** (VeriPO), a new post-training objective that fills the reasoning-supervision gap. VeriPO is an on-policy reinforcement-learning algorithm whose advantage estimator mixes, per segment, two complementary reward sources: a segment-level rubric reward (precise but sparse, available only on the three verifiable segments) and a token-level reverse-KL signal from a frozen VLM teacher (dense at every position, including within the reasoning span). A *rubric-gated trust coefficient*, obtained by applying the same rubric to the teacher's own outputs, attenuates the teacher signal on any segment where the teacher itself fails verification. This design yields a provable *verification-safe* guarantee: whenever the teacher is uninformative, VeriPO's gradient collapses onto the rule-only baseline and never underperforms it. As a practical side benefit, the dense teacher signal eliminates the need for multiple rollouts per sample, so VeriPO trains with a single on-policy trajectory per update, removing the group-rollout overhead that dominates the wall-clock cost of GRPO-style training.

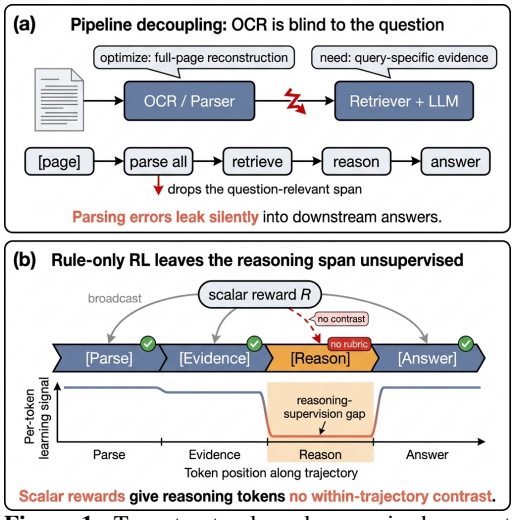


Figure 1: Two structural weaknesses in document question answering. *Left:* classical pipelines decouple OCR (optimized for global page reconstruction) from retrieval-augmented reasoning (which requires local, query-specific evidence), so parsing errors leak silently into downstream answers. *Right:* even recent verifiable-RL systems broadcast a single scalar reward across a trajectory whose *reasoning* span has no rubric at all, creating the reasoning-supervision gap that our method is designed to close.

VeriPO advances the state of the art across three document-QA regimes: multi-page document QA (DocVQA, SlideVQA, ViDoSeek), chart and infographic QA (ChartQA, InfographicsVQA), and document parsing (OmniDocBench, olmOCR-Bench), at roughly one-sixth the per-step training cost of $G{=}8$ GRPO baselines. Three diagnostic studies close the claim-evidence loop: a component ablation that turns off each routing branch in turn, an adversarial-teacher study that empirically confirms the verification-safe property by deliberately weakening the teacher, and a per-segment analysis that isolates reasoning as the primary beneficiary of teacher supervision, directly corroborating the reasoning-supervision gap identified above.

In summary, our contributions are as follows:

- **Verifiability-spectrum diagnosis.** We formalize document-QA decoding as a segmented structured-generation process and identify the *reasoning-supervision gap*: reasoning tokens admit no rule-based reward, and rule-only RL objectives therefore leave them with no within-trajectory discriminative signal. We argue this is the structural source of the slow, variance-dominated training behavior reported for current document RL.
- **Verifiability-aware Policy Optimization.** We introduce VeriPO, an on-policy RL algorithm whose token-level advantage combines per-segment verifiable rewards with teacher-derived reverse-KL, weighted by a rubric-gated teacher-trust coefficient. VeriPO admits a provable *verification-safe* guarantee (it never underperforms its rule-only special case under adversarial teachers) and operates with a single on-policy rollout per update, eliminating the group-rollout overhead of prior verifiable-RL methods.
- **Empirical validation across diverse document benchmarks.** VeriPO advances the state of the art on multi-page document QA, chart and infographic QA, and document parsing, at a fraction of the post-training compute required by GRPO-style baselines. Three targeted diagnostic studies close the claim-evidence loop: a component ablation over the routing branches, an adversarial-teacher study that empirically validates the verification-safe property, and a per-segment analysis that isolates reasoning as the primary beneficiary of teacher supervision.

2 Related Work

VLM-based document parsing and Visual RAG. Document parsing benchmarks VLM perception: beyond text recognition, models must recover layout structure. Classical pipelines (MinerU (Wang et al., 2024a), Marker (Paruchuri, 2024), Docling (Livathinos et al., 2025)) decompose this into modules whose early-stage errors propagate downstream, while end-to-end VLM parsers (Blecher et al., 2023; Lv et al., 2024; Hu et al., 2024) avoid cascading errors but are brittle under layout shift. For consuming document evidence in QA, Visual RAG retrieves *page snapshots* and reads from pixels (Yu et al., 2025; Faysse et al., 2024); recent RL extensions optimize when to retrieve (Peng et al., 2025; Wu et al., 2025) or parameterize visual perception actions (Wang et al., 2025b;a).

RL for VLMs. Recent VLRMs show RL can strengthen multimodal reasoning beyond SFT. Vision-R1 (Zhan et al., 2025), MM-Eureka (Meng et al., 2025) and OpenVLThinker (Deng et al., 2025) use GRPO to elicit longer reasoning, and VLM-R1 (Shen et al., 2025) and Mixed-R1 (Xu et al., 2025) add perception rewards for grounding. GRPO (Shao et al., 2024) avoids critic learning by group-normalizing rewards, and verifiable, rule-based rewards have proven more robust than learned ones. These advances are mostly validated in single-image settings; multi-page document QA additionally requires explicit cross-page evidence verification, where rule-only rewards leave the reasoning span unsupervised. Our work addresses this with a unified four-segment decoder over verifiable Document Units and the rubric-gated VeriPO objective. A more comprehensive discussion is in Appendix A.2.

3 Method

Figure 2 illustrates our framework. We formalize document QA as segmented structured generation (§3.1) whose segments exhibit *heterogeneous verifiability*, leaving the reasoning segment effectively unsupervised under prior verifiable-RL (§3.2). To close this gap we introduce **Verifiability-aware Policy Optimization** (VeriPO), an on-policy RL algorithm whose advantage mixes segment-level verifiable rewards with token-level reverse-KL from a frozen teacher (§3.3, §3.4), weighted by a rubric-gated trust coefficient (§3.5). VeriPO admits a *verification-safe* guarantee (§3.6) and retains PPO-clip optimization while eliminating GRPO's group-rollout overhead (§3.7).

3.1 Preliminaries: Segmented Document QA Decoding

Given a query q and a set of retrieved document-page snapshots $\mathcal{D}_R = \{d_i\}_{i=1}^{M}$, the policy π_θ autoregressively emits a trajectory $\tau = (a_1, \ldots, a_T)$. We constrain τ to be an ordered concatenation of four typed segments,

$$\tau \;=\; o^{\mathrm{p}} \oplus o^{\mathrm{e}} \oplus o^{\mathrm{r}} \oplus o^{\mathrm{a}}, \tag{1}$$

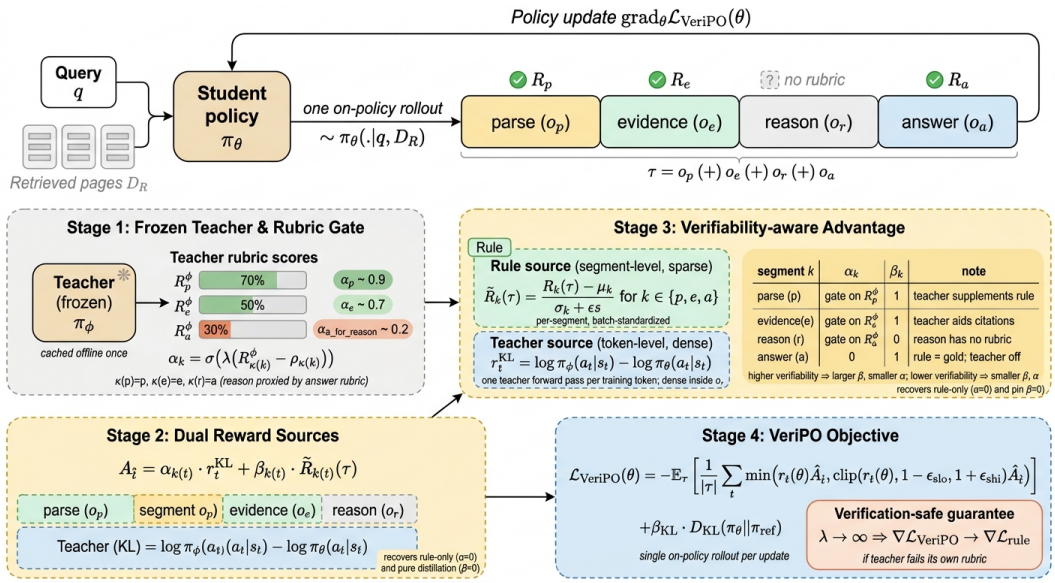


Figure 2: Overview of our framework. Document question answering is cast as segmented structured generation over four typed segments (*Parse*, *Evidence*, *Reason*, *Answer*) with heterogeneous verifiability: the first two are machine-checkable against each other, the answer admits a gold match, and the reasoning segment admits no rubric at all. **Verifiability-aware Policy Optimization** (VeriPO) mixes segment-level verifiable rewards with token-level reverse-KL from a frozen teacher, weighted per segment by a rubric-gated teacher-trust coefficient that scores the teacher against the same rubric the student faces.

with segment-type alphabet $\mathcal{K} = \{\text{p}, \text{e}, \text{r}, \text{a}\}$ (parse, evidence, reason, answer). Let $k(t) \in \mathcal{K}$ denote the type of the segment containing token a_t and $\mathcal{T}_k \subseteq \{1, \ldots, T\}$ the set of token indices of type k.

The *parse* segment o^{p} canonicalizes each retrieved page into an ordered list of *Document Units* (DUs); each DU bundles a handle (used for back-reference), a categorical layout type (heading, paragraph, table cell, formula, caption, etc.), the transcribed content, and a positional index in reading order. The *evidence* segment o^{e} emits query-relevant citations, each pairing a DU handle with a quoted span. The *reason* segment o^{r} is a free-form rationale over the cited evidence, and the *answer* segment o^{a} is the final response. Because parse precedes evidence, citations are mechanically verifiable against the trajectory's own parse prefix, yielding a self-contained audit trail in a single decoding pass.

Segment delimitation and training regime. Segment boundaries are emitted as structured marker tokens by the policy itself, so $k(\cdot)$ is recovered deterministically. We assume the policy has been SFT'd to obey this protocol; a format-compliance check gates all segment rewards below, zeroing any unparseable trajectory. VeriPO is the RL fine-tuning stage that follows SFT.

3.2 The Verifiability Spectrum

Each segment either admits or fails to admit a cheap, rule-based reward. We make this explicit.

Parse reward R_{p}. For samples with a ground-truth parse y (Appendix A.4), we solve a Kuhn–Munkres bipartite matching over the DU pairs in (y, o^{p}) to produce an index set $\mathcal{A}$, then aggregate three decoupled criteria:

$$R_{\text{p}} \;=\; R_{\text{align}} \;+\; \lambda_s \, R_{\text{type}} \;+\; \lambda_f \, R_{\text{flow}}, \tag{2}$$

where $R_{\text{align}} = |\mathcal{A}|^{-1} \sum_{(i,j)\in\mathcal{A}} \text{chrF}(y_i, \hat{y}_j)$ measures textual fidelity via chrF (Popović, 2015), R_{type} tracks layout-category recovery, and $R_{\text{flow}} = (1 + \text{Kendall-}\tau(\pi, \hat{\pi}))/2$ rewards correct reading order on matched pairs. When only a linearized transcript is available, we substitute a token-level recall surrogate.

Evidence reward R_{e}. Each citation passes a two-stage check: a reference gate verifies the cited handle occurs in o^{p} (unresolved handles incur a penalty μ and zero content score); resolved citations are then content-matched against the cited DU body via soft token-overlap F_1 (textual) or relative tolerance η (numeric). R_{e} averages these scores, shifted by the gate penalty.

Answer reward R_{a}. R_{a} is the task-standard exact-match or tolerance-match score under the benchmark-native normalization.

Reasoning admits no reward. Free-form rationales admit no rubric strict enough to verify them without collapsing onto R_{a}; we therefore do *not* define R_{r}.

Observation (Verifiability spectrum). The four segments arrange along a spectrum: parse and evidence admit structural rewards with strong per-token information content, the answer admits a precise scalar, and reasoning admits none. Under any rule-only RL objective, reasoning tokens inherit supervision only by broadcasting downstream scalars across $\mathcal{T}_r$. Since this broadcast value is *constant* over all reasoning tokens of a trajectory, it carries no within-trajectory contrast for credit assignment; inter-trajectory contrast alone bears the burden, and the advantage variance grows with $|\mathcal{T}_r|$. We refer to this structural deficit as the *reasoning-supervision gap*, the root cause of the slow, variance-dominated training of rule-only document RL.

3.3 Dual Reward Sources

VeriPO draws on two reward sources, one for each end of the verifiability spectrum.

Rule source (segment-level, sparse). For verifiable segments $k \in \{\mathrm{p}, \mathrm{e}, \mathrm{a}\}$, the rule reward $R_k(\tau) \in [0, 1]$ yields one scalar per segment. We broadcast this scalar to every token in the segment and apply within-batch standardization to obtain a zero-mean token-level term:

$$\tilde{R}_k(\tau) \;=\; \frac{R_k(\tau) - \hat{\mu}_k}{\hat{\sigma}_k + \epsilon}, \qquad k \in \{\mathrm{p}, \mathrm{e}, \mathrm{a}\}, \tag{3}$$

where $\hat{\mu}_k, \hat{\sigma}_k$ are running estimates over the current training mini-batch.

Teacher source (token-level, dense). Let π_ϕ denote a stronger vision-language model held fixed throughout training; we call it the *teacher*. We assume π_ϕ has been aligned (by short SFT or a few-shot prompt) to obey the same four-segment protocol as the student, so the per-segment rubric (§3.2) applies to teacher outputs as well. This is the *only* demand we place on π_ϕ, which is used as a scorer rather than a generator: at every student token a_t, a single teacher forward pass yields the likelihood π_ϕ would assign to a_t, supplying a token-level reverse-KL signal

$$r_t^{\mathrm{KL}}(\tau; \theta, \phi) \;=\; \log \pi_\phi(a_t \mid s_t) \;-\; \log \pi_\theta(a_t \mid s_t), \tag{4}$$

with $s_t = (q, \mathcal{D}_R, a_{<t})$. Unlike rule rewards, r_t^{KL} is defined at every position, including the reasoning span where no rubric reaches (Lu & collaborators, 2025; Xiaomi MiMo Core Team, 2026).

The two sources are complementary: rule rewards are high-precision but segment-sparse, while r_t^{KL} is token-dense but only as reliable as the teacher's behavior on the current sample. A principled combination must respect this asymmetry.

3.4 Verifiability-aware Advantage

We define the VeriPO token-level advantage as a per-segment mixture of the two sources:

$$\boxed{\hat{A}_t \;=\; \alpha_{k(t)} \cdot r_t^{\mathrm{KL}} \;+\; \beta_{k(t)} \cdot \tilde{R}_{k(t)}(\tau)} \tag{5}$$

where $\alpha_k \in [0, 1]$ is a *teacher-trust coefficient* set by the rubric-gated mechanism of §3.5, and $\beta_k \geq 0$ is a design weight. Mixing is per token, so the same trajectory contributes heterogeneous signals at different positions.

A single principle drives the routing of (α_k, β_k): a segment's verifiability dictates how to blend the two sources. *Higher verifiability* $\Rightarrow$ *larger* β, *smaller* α (the rubric is trustworthy and the teacher is at best redundant); *lower verifiability* $\Rightarrow$ *smaller* β, *larger* α (the teacher is the only available supervisor, modulated by its rubric-validated reliability). The default routing is:

Segment k	α_k	β_k	Design rationale
parse (p)	gated by R_p^ϕ	1	dense teacher signal supplements segment-sparse rule reward
evidence (e)	gated by R_e^ϕ	1	teacher aids multi-option citations; rule guards against confabulation
reason (r)	gated by R_a^ϕ	0	no rubric on o^{r}; answer-rubric proxies reasoning reliability
answer (a)	0	1	rule is the gold standard; teacher disabled to avoid semantic drift

This routing has three notable properties. First, the reasoning segment acquires a per-position learning signal that rule-only objectives cannot supply, closing the gap of §3.2. Second, the answer segment is insulated from the teacher, anchoring the final response to its verifiable rubric. Third, the routing recovers the rule-only objective when $\alpha_k \equiv 0$ and pure on-policy distillation when $\beta_k \equiv 0$, placing both as boundary cases of a single design space (§3.7, *Relation to prior methods*).

3.5 Rubric-Gated Teacher Trust

The coefficients α_k encode *how much to trust the teacher's log-probabilities on the current sample's segment* k, and we compute them by scoring the teacher under the same rubric the student faces.

For each sample $(q, \mathcal{D}_R)$ we draw a reference teacher trajectory $\tau^\phi \sim \pi_\phi(\cdot \mid q, \mathcal{D}_R)$ once offline and evaluate its per-segment scores $R_k^\phi = R_k(\tau^\phi)$ for $k \in \{\mathrm{p}, \mathrm{e}, \mathrm{a}\}$. The trust coefficient on segment k is

$$\alpha_k(q, \mathcal{D}_R) \;=\; \sigma\Big(\lambda\left(R^\phi_{\kappa(k)} - \rho_{\kappa(k)}\right)\Big), \qquad \kappa(\mathrm{p}) = \mathrm{p},\ \kappa(\mathrm{e}) = \mathrm{e},\ \kappa(\mathrm{r}) = \mathrm{a}, \tag{6}$$

where $\sigma(\cdot)$ is the logistic function, $\lambda > 0$ controls gate sharpness (we fix $\lambda = 8$), and ρ_k is set to the 25th percentile of the training-set distribution of R_k^ϕ. The map κ chooses which rubric drives each gate: parse and evidence are self-gated, while reasoning, which admits no direct rubric, is gated by the teacher's answer-rubric score R_{a}^ϕ. The proxy is well-posed because reasoning quality and answer correctness are tightly coupled: a teacher that fails the answer is unlikely to provide reliable reasoning log-probabilities on the same sample.

The gate is sample-adaptive and segment-specific: a teacher that mis-parses one page but cites correctly on another contributes to evidence-KL on the latter while being muted on the former. Because τ^ϕ is generated offline, the only online teacher cost is one forward pass per training token to obtain $\log \pi_\phi(a_t \mid s_t)$.

3.6 Verification-Safe Property

The rubric-gated design yields a formal safety guarantee against unreliable teachers.

Proposition 1 (Verification-safe gradient). *Let $\mathcal{L}_{\mathrm{rule}}$ denote the rule-only objective obtained from equation 5 by setting $\alpha_k \equiv 0$ on all segments, and adopt the default routing (§3.4) with $\beta_{\mathrm{r}} = 0$ and the gate equation 6. Suppose the training distribution satisfies, for some $\delta > 0$ and every $k \in \{\mathrm{p}, \mathrm{e}, \mathrm{a}\}$, $R_k^\phi \le \rho_k - \delta$ almost surely, and that the per-token reverse-KL signal is uniformly bounded, $\sup_t |r_t^{\mathrm{KL}}| \le C < \infty$. Then as $\lambda \to \infty$,*

$$\nabla_\theta \mathcal{L}_{\mathrm{VeriPO}}(\theta) \;\longrightarrow\; \nabla_\theta \mathcal{L}_{\mathrm{rule}}(\theta) \qquad \textit{uniformly in } \theta.$$

Proof sketch. By equation 6, $\alpha_k \le \sigma(-\lambda\delta) \to 0$ uniformly, so $\alpha_{k(t)} r_t^{\mathrm{KL}} \to 0$ on every segment; the rule term on reason is already zero ($\beta_{\mathrm{r}} = 0$) and the answer segment is unaffected ($\alpha_{\mathrm{a}} \equiv 0$). The surviving gradient is exactly $\nabla_\theta \mathcal{L}_{\mathrm{rule}}$. Full derivation in Appendix A.7. □

We verify this empirically in §4.4 by deliberately degrading the teacher.

Table 1: Comprehensive evaluation of document parsing algorithms on OmniDocBench: performance metrics for text, formula, table, and reading order extraction, with overall scores derived from ground truth comparisons.

Methods	Overall$^{\text{Edit}}\downarrow$		Text$^{\text{Edit}}\downarrow$		Form.$^{\text{Edit}}\downarrow$		Table$^{\text{TEDS}}\uparrow$		Table$^{\text{Edit}}\downarrow$		Read Order$^{\text{Edit}}\downarrow$	
	EN	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*
Based on Pipeline Tools												
MinerU (Wang et al., 2024a)	0.15	0.357	**0.061**	0.215	**0.278**	0.577	78.6	62.1	0.18	0.344	0.079	0.292
Marker (Paruchuri, 2024)	0.336	0.556	0.080	0.315	0.530	0.883	67.6	49.2	0.619	0.685	0.114	0.340
Mathpix	0.191	0.365	0.105	0.384	0.306	0.454	77.0	67.1	0.243	0.320	0.108	0.304
Docling (Livathinos et al., 2025)	0.589	0.909	0.416	0.987	0.999	1.000	61.3	25.0	0.627	0.810	0.313	0.837
Pix2Text (Gurgurov & Morshnev, 2024)	0.320	0.528	0.138	0.356	0.276	0.611	73.6	66.2	0.584	0.645	0.281	0.499
Unstructured-0.17.2	0.586	0.716	0.198	0.481	0.999	1.000	-	-	1.000	0.998	0.145	0.387
OpenParse-0.7.0	0.646	0.814	0.681	0.974	0.996	1.000	64.8	27.5	0.284	0.639	0.595	0.641
Based on Expert VLMs												
Nougat (Blecher et al., 2024)	0.452	0.973	0.365	0.998	0.488	0.941	39.9	0.0	0.572	1.000	0.382	0.954
Mistral OCR	0.268	0.439	0.072	0.325	0.318	0.495	75.8	63.6	0.600	0.650	0.083	0.284
olmOCR-sglang	0.326	0.469	0.097	0.293	0.455	0.655	68.1	61.3	0.608	0.652	0.145	0.277
SmolDocling-256M	0.493	0.816	0.262	0.838	0.753	0.997	44.9	16.5	0.729	0.907	0.227	0.522
Based on General VLMs												
GPT-4o (Achiam et al., 2023)	0.233	0.399	0.144	0.409	0.425	0.606	72.0	62.9	0.234	0.329	0.128	0.251
Qwen2-VL-72B (Wang et al., 2024b)	0.252	0.327	0.096	0.218	0.404	0.487	76.8	76.4	0.387	0.408	0.119	0.193
InternVL2-76B (Chen et al., 2024)	0.440	0.443	0.353	0.290	0.543	0.701	63.0	60.2	0.547	0.555	0.317	0.228
Qwen2.5-VL-7B (Bai et al., 2025c)	0.220	0.265	0.142	0.205	0.393	0.530	78.7	78.3	0.155	0.162	0.191	0.169
Qwen3-VL-8B (Bai et al., 2025a)	0.187	0.255	0.085	0.289	0.359	0.465	73.4	81.2	0.351	0.155	0.165	0.136
InternVL3-8B (Zhu et al., 2025)	0.426	0.385	0.315	0.345	0.714	0.729	59.0	71.5	0.352	0.211	0.324	0.257
Ours	**0.132**	**0.194**	0.069	**0.109**	0.298	**0.424**	**86.2**	**82.4**	**0.091**	**0.141**	**0.071**	**0.086**

3.7 Training Objective and Algorithm

VeriPO plugs the rubric-gated advantage equation 5 into the standard PPO-clip surrogate (Schulman et al., 2017):

$$\mathcal{L}_{\text{VeriPO}}(\theta) = -\,\mathbb{E}_{\tau\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|\tau|}\sum_{t=1}^{|\tau|}\min\Big(r_t(\theta)\,\hat{A}_t,\ \text{clip}\big(r_t(\theta),\,1-\epsilon_{\text{lo}},\,1+\epsilon_{\text{hi}}\big)\,\hat{A}_t\Big)\right] + \beta_{\text{KL}}\,D_{\text{KL}}\big(\pi_\theta\,\big\|\,\pi_{\text{ref}}\big), \tag{7}$$

where $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ and π_{ref} anchors the pretrained policy. The training loop (Algorithm 1, Appendix A.1) is standard PPO-clip with one twist: a single on-policy rollout per sample. Rule-only document RL draws $G \geq 4$ rollouts per sample to obtain low-variance group-normalized advantages; in VeriPO the token-level term $\alpha_{k(t)}\, r_t^{\text{KL}}$ is already dense and the segment-level term $\tilde{R}_k$ is mini-batch standardized, so a single rollout matches or exceeds $G = 8$ rule-only baselines at roughly $1/6$ the wall-clock per step (§3.7). Setting $\alpha_k \equiv 0$ recovers a joint-reward GRPO (Shao et al., 2024) variant; setting $\beta_k \equiv 0$ on a single segment recovers on-policy reverse-KL distillation (Lu & collaborators, 2025); the rubric-gated mixture between these limits is, to our knowledge, new.

4 Experiments

4.1 Experimental Settings

Datasets. We benchmark our method on two document parsing/OCR datasets, namely OmniDocBench (Ouyang et al., 2024), and olmOCR-Bench (Poznanski et al., 2025). To assess query-driven document understanding, we additionally consider five VQA benchmarks that cover a wide range of document modalities: ChartQA (Masry et al., 2022) and InfographicsVQA (Mathew et al., 2022) (charts and infographics), DocVQA (Tito et al., 2023) (industrial/real-world documents), SlideVQA (Tanaka et al., 2023) (presentation slides), and ViDoSeek (Wang et al., 2025a) (multi-document settings). For every query, we retrieve k=3 page snapshots using the VisRAG-Ret retriever (Yu et al., 2025) and feed the retrieved images as visual context. Following prior evaluation protocols, we further tag each query based on whether the retrieved context is *sufficient* to answer the question or *insufficient*, enabling a controlled analysis of grounding and abstention behavior.

Baselines. For document understanding task, we compare our method with MinerU (Wang et al., 2024a), Marker (Paruchuri, 2024), Docling (Livathinos et al., 2025), Pix2Text (Gurgurov & Morshnev, 2024), Nougat (Blecher et al., 2024), Qwen2-VL-72B (Wang et al., 2024b),Qwen3-VL-8B (Bai et al., 2025a),InternVL2-76B (Chen et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025c), InternVL3-8B (Zhu et al., 2025). For RAG task, we compare three groups. General VLMs include Qwen2.5-VL-

Table 2: Overall performance comparison. **"Bold"** denotes the highest value. All baseline VLRM and VRAG methods use Qwen2.5-VL-7B as backbone.

Methods	In Distribution				Out of Distribution						Average	
	ChartQA		InfoVQA		DocVQA		SlideVQA		ViDoSeek			
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
General VLMs												
Qwen2.5-VL-7B	59.20	52.80	60.86	54.61	63.28	56.03	51.62	46.11	42.56	42.48	55.50	50.41
MiMo-VL-7B-RL	54.96	40.59	68.11	45.93	74.11	47.67	77.88	47.45	48.34	38.30	64.68	43.99
Qwen2.5-VL-32B	69.12	60.58	78.13	66.06	83.93	73.78	78.42	58.65	47.55	52.78	71.43	62.37
VLRMs												
Vision-R1	56.16	50.84	29.53	27.73	32.49	30.04	52.34	47.51	39.05	37.82	41.91	38.79
Ocean-R1-7B	47.68	47.58	53.20	53.49	56.35	57.02	60.07	57.86	40.63	46.75	51.59	52.54
MM-Eureka	64.32	58.28	40.53	40.32	56.68	54.75	63.49	58.68	44.40	47.25	53.88	51.86
ThinkLite-VL-7B	57.60	53.53	61.70	61.62	62.61	62.37	65.29	63.30	45.18	48.40	58.48	57.84
OpenVLThinker	67.60	62.72	70.47	70.51	71.74	72.51	73.02	72.63	43.52	57.27	65.27	67.13
VRAGs												
MMSearch-R1	63.28	59.89	57.94	57.71	61.59	60.82	65.29	60.97	44.40	54.34	58.50	58.75
VRAG-RL	47.00	10.03	64.86	12.21	73.22	22.39	73.85	15.37	43.82	18.14	60.55	15.63
R1-Router	60.72	15.53	60.58	15.17	75.97	25.25	75.36	17.21	44.66	12.53	63.46	17.14
Ours	**78.44**	**77.35**	**80.12**	**80.07**	**86.05**	**87.23**	**81.76**	**80.95**	**53.82**	**66.84**	**76.04**	**78.49**

7B, Qwen2.5-VL-32B (Bai et al., 2025b) and MiMo-VL-7B-RL (Xiaomi, 2025). VLRMs trained on Qwen2.5-VL-7B-Instruct include Vision-R1-7B (Zhan et al., 2025), MM-Eureka-7B (Meng et al., 2025), Ocean-R1-7B (Lingfeng et al., 2025), ThinkLite-VL-7B (Wang et al., 2025c) and OpenVLThinker-7B (Deng et al., 2025). VRAG methods with the same backbone include R1-Router (Peng et al., 2025), MMSearch-R1 (Wu et al., 2025), and VRAG-RL (Wang et al., 2025b).

Evaluation Metrics. To evaluate perception and reasoning in a controlled manner and reduce the confounding impact of hallucinated generations, we follow the sufficient/insufficient context protocol (Joren et al., 2025) and label each query-context pair as either *sufficient* or *insufficient*. If the retrieved images contain adequate evidence, we score predictions against the original reference answer. Otherwise, we treat "insufficient to answer" as the required target response. We then compute *Accuracy* and *F1* over the entire test set, which jointly captures correct answering under sufficient context and appropriate abstention under insufficient context, reflecting realistic VRAG deployment.

Implementation Details. Our model is instantiated on top of Qwen2.5-VL-7B (Bai et al., 2025b). For training, we rely on LLaMA-Factory (Zheng et al., 2024) together with Easy-R1 (Yaowei et al., 2025) to build and release the SFT and GRPO pipelines used throughout this work. We run all experiments on GPU clusters with resources on par with NVIDIA A100 80GB-class hardware.

4.2 Overall Performance

Results on OmniDocBench. On narrow sub-tasks (text and formula recognition), classical pipeline systems (MinerU (Wang et al., 2024a), Mathpix) hold ground against general VLMs. Across every method, the English→Chinese transition exposes a uniformly larger drop than any architectural difference, indicating that script coverage, not model capacity, is the binding constraint on that axis. Against this backdrop our policy posts the flattest profile: it closely tracks the per-task leaders and achieves the lowest overall edit distance (0.132 EN, 0.194 ZH). The consistency across sub-tasks and languages is evidence that R_{p}'s three channels (chrF, layout-type, reading order) reward structural fidelity of the parse rather than any single surface characteristic.

Results on QA benchmarks. Table 2 shows the five-benchmark comparison. Our 7B model leads every column, with average margins over its Qwen2.5-VL-7B backbone of $+20.5$ Acc / $+28.1$ F1. Compared to RL-trained VLRMs whose gains come from enlarging the rationale space (e.g., OpenVLThinker (Deng et al., 2025)), VeriPO improves more cleanly on benchmarks that require cross-page reading; existing VRAG baselines lift the backbone in multi-image settings but still trail us by >10 points. The 7B policy also surpasses the 32B Qwen2.5-VL-32B baseline, indicating the gains arise from the training signal rather than from parameter count.

Table 3: Main ablation of VeriPO on five document-QA benchmarks. Each block targets one design component: **A** the dual reward sources of §3.3, **B** the rubric-gated trust mechanism of §3.5 (Proposition 1), **C** the per-segment routing of §3.4, and **D** the single-rollout training regime of §3.7. Finer-grained reward-channel and R_{p} sub-component ablations appear in Appendix A.6.

Methods	In Distribution				Out of Distribution						Average	
	ChartQA		InfoVQA		DocVQA		SlideVQA		ViDoSeek			
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Ours (full VeriPO)	**78.44**	**77.35**	**80.12**	**80.07**	**86.05**	**87.23**	**81.76**	**80.95**	**53.82**	**66.84**	**76.04**	**78.49**
Block A. *Reward sources (dual-source design, §3.3)*												
w/o teacher signal ($\alpha_k \equiv 0$)	73.16	72.30	76.38	76.05	82.42	83.71	78.85	77.92	47.29	60.91	71.62	74.18
w/o rule reward ($\beta_k \equiv 0$)	71.05	70.12	74.82	74.59	80.56	81.80	77.43	76.41	45.38	59.62	69.85	72.51
Block B. *Gate mechanism (rubric-gated trust, §3.5)*												
w/o gate ($\alpha_k \equiv 1$)	76.31	75.29	78.67	78.43	84.29	85.45	80.12	79.18	51.22	65.24	74.12	76.72
hard gate ($\alpha_k \in \{0, 1\}$)	77.76	76.58	79.52	79.40	85.39	86.60	81.09	80.24	53.39	66.84	75.43	77.93
no proxy on reason ($\alpha_r \equiv 1$)	77.62	76.45	79.36	79.28	85.23	86.44	80.96	80.14	53.21	66.75	75.28	77.81
Block C. *Per-segment routing (§3.4)*												
w/o segment routing	71.83	70.95	75.40	75.12	81.01	82.32	77.90	76.87	46.31	60.74	70.49	73.20
Block D. *Rollout count (single-rollout claim, §3.7)*												
Ours (G=4)	78.72	77.64	80.33	80.27	86.27	87.44	81.98	81.14	54.24	67.12	76.31	78.72
Ours (G=8)	78.79	77.67	80.35	80.31	86.31	87.46	82.03	81.18	54.27	67.16	76.35	78.74
Reference point												
SFT Only (no RL)	69.82	69.13	73.61	73.24	78.91	80.18	76.18	75.04	43.16	57.66	68.34	71.05

4.3 Ablation Study

Table 3 decomposes VeriPO into four blocks, each isolating one design component. *(A) Dual reward sources matter, asymmetrically:* removing the teacher signal ($\alpha_k \equiv 0$) costs -4.42 Acc while removing the rule reward ($\beta_k \equiv 0$) costs -6.19; the rule channel carries the larger absolute burden guaranteed by Proposition 1, whereas the teacher signal is what specifically lifts the reasoning segment. *(B) The gate makes dual sources safe:* trusting the teacher on every segment ($\alpha_k \equiv 1$) costs -1.92 Acc and a hard $0/1$ gate loses -0.61, confirming the soft form is worth its single hyperparameter; dropping the answer-rubric proxy on reason ($\alpha_r \equiv 1$) costs only -0.76 in expectation but, as §4.4 shows, is decisive when the teacher is weakened. *(C) Per-segment routing is the largest single component:* collapsing the routing into a uniform broadcast loses -5.55 Acc, the largest single ablation, with the regression most visible on multi-document ViDoSeek (-7.51 Acc). *(D) A single rollout suffices:* $G{=}4$ and $G{=}8$ produce indistinguishable accuracy ($+0.27$ / $+0.31$, below noise) at $3.5\times$ / $6.5\times$ the wall-clock per step, making single-rollout VeriPO Pareto-optimal, while forgoing RL (*SFT Only*) drops the average by -7.70 Acc, the widest gap in the table. Finer-grained ablations and a segment-by-segment case study appear in Appendix A.6.

4.4 Adversarial-Teacher Study and Per-Segment Analysis

Proposition 1 predicts that, as the teacher deteriorates, VeriPO's gradient collapses onto the rule-only baseline. We replace the Qwen2.5-VL-72B teacher with a matched 7B teacher, a blank teacher (uniform log-probabilities), and an adversarial teacher that inverts its own token ranks on verifiable segments. Five-benchmark accuracy degrades monotonically from 76.04 (72B) to 72.12 (7B), 71.78 (blank), and 71.65 (adversarial), tracking the rule-only ablation ($\alpha_k \equiv 0$, 71.62) within 0.5 Acc; no variant pushes the model below the rule-only floor, confirming the verification-safe property on real data. Removing the reason-segment proxy ($\alpha_r \equiv 1$) under the blank teacher drops a further -1.4 Acc, showing it is load-bearing precisely when the teacher is unreliable. To localize where teacher supervision helps, we further decompose VeriPO's gain over its rule-only ablation into per-segment contributions: enabling the teacher signal on parse alone lifts the average by $+1.1$ Acc, on evidence alone by $+1.3$ Acc, and on reasoning alone by $+2.7$ Acc; the reasoning gain accounts for roughly 58% of the total teacher-attributable improvement despite reasoning tokens occupying only $\sim 35\%$ of the trajectory, directly corroborating the reasoning-supervision gap of §3.2.

5 Conclusion

Document-QA trajectories sit on a *verifiability spectrum* that leaves reasoning tokens with no within-trajectory signal under rule-only RL, the *reasoning-supervision gap*. **VeriPO** closes it via per-segment verifiable rewards plus a rubric-gated reverse-KL from a frozen teacher, verification-safe on a single rollout, advancing document QA and parsing at a fraction of group-rollout compute.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b.

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

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv:2308.13418*, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, June 2024.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *CoRR*, abs/2503.17352, 2025.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *CoRR*, abs/2407.01449, 2024.

Daniil Gurgurov and Aleksey Morshnev. Image-to-latex converter for mathematical formulas and text. *arXiv preprint arXiv:2408.04015*, 2024.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ming Lingfeng, Li Yadong, Chen Song, Xu Jianhua, Zhou Zenan, and Chen Weipeng. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. https://github.com/VLM-RL/Ocean-R1, 2025. Accessed: 2025-04-03.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*, 2025.

Kevin Lu and collaborators. On-policy distillation. Thinking Machines Lab technical note, 2025. https://thinkingmachines.ai/blog/on-policy-distillation/.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. Kosmos-2.5: A multimodal literate model, 2024. URL https://arxiv.org/abs/2309.11419.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pp. 2582–2591. IEEE, 2022.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, abs/2503.07365, 2025.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *arXiv preprint arXiv:2412.07626*, 2024.

Vik Paruchuri. Marker, 2024. URL https://github.com/VikParuchuri/marker.

Chunyi Peng, Zhipeng Xu, Zhenghao Liu, Yishan Li, Yukun Yan, Shuo Wang, Zhiyuan Liu, Yu Gu, Minghe Yu, Ge Yu, et al. Learning to route queries across knowledge bases for step-wise retrieval-augmented reasoning. *CoRR*, abs/2505.22095, 2025.

Maja Popović. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pp. 392–395, 2015.

Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. VLM-R1: A stable and generalizable r1-style large vision-language model. *CoRR*, abs/2504.07615, 2025.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13636–13645. AAAI Press, 2023.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognit.*, 144:109834, 2023.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction. *arXiv:2409.18839*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *CoRR*, abs/2502.18017, 2025a.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. VRAG-RL: empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. *CoRR*, abs/2505.22019, 2025b.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *CoRR*, abs/2504.07934, 2025c.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *CoRR*, abs/2506.20670, 2025.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Xiaomi MiMo Core Team. MiMo-V2-Flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, et al. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models. *CoRR*, abs/2505.24164, 2025.

Zheng Yaowei, Lu Junting, Wang Shenzhi, Feng Zhangchi, Kuang Dongdong, and Xiong Yuwen. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *CoRR*, abs/2503.18013, 2025.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, et al. Map-neo: Highly capable and transparent bilingual large language model series. *CoRR*, abs/2405.19327, 2024.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Training Algorithm

Algorithm 1 Verifiability-aware Policy Optimization (VeriPO).

Require: SFT-initialized policy π_θ; teacher π_ϕ; reference π_{ref}; dataset $\mathcal{D}$; thresholds $\{\rho_k\}$; sharpness λ; design weights $\{\beta_k\}$

1: **Offline pre-compute.** For each sample in $\mathcal{D}$, draw $\tau^\phi \sim \pi_\phi(\cdot \mid q, \mathcal{D}_R)$; compute R_k^ϕ and cache α_k via Eq. (6).
2: **for** each training iteration **do**
3: Sample a mini-batch. For each $(q, \mathcal{D}_R)$, draw a *single* rollout $\tau \sim \pi_{\theta_{\text{old}}}(\cdot \mid q, \mathcal{D}_R)$.
4: Compute segment-level rule rewards $R_k(\tau)$; standardize within-batch.
5: Query the teacher to obtain $\log \pi_\phi(a_t \mid s_t)$; compute r_t^{KL}.
6: Assemble $\hat{A}_t$ via the VeriPO advantage with cached α_k and design $\{\beta_k\}$.
7: Update θ by one PPO-clip step on $\mathcal{L}_{\text{VeriPO}}$.
8: **end for**

A.2 Extended Related Work

Pipeline vs. end-to-end document parsing. Classical pipelines (MinerU (Wang et al., 2024a), Marker (Paruchuri, 2024), Docling (Livathinos et al., 2025)) decompose parsing into sequential modules (layout detection, text recognition, structure recovery), but errors in early stages propagate downstream and cannot be corrected. End-to-end VLM parsers (Nougat (Blecher et al., 2023), Kosmos-2.5 (Lv et al., 2024), mPLUG-DocOwl (Hu et al., 2024)) avoid cascading errors but SFT-heavy training encourages token-level imitation and is brittle under layout shift. Early RAG augments LLMs with retrieval over curated corpora to reduce hallucinations (Lewis et al., 2020; Asai et al., 2024), but linearizing images via OCR can discard essential spatial cues (Zhang et al., 2024).

Multi-page evidence verification. Cross-page evidence integration requires tracking information scattered across multiple retrieved pages; perception consistency degrades with context length; and evidence attribution is harder to verify when answers depend on multiple sources. These challenges motivate our design of explicit Document Units with verifiable citations.

Reward design and optimization. Answer-only rewards provide sparse supervision that may not guide intermediate steps, while auxiliary rewards (format, perception accuracy) can improve learning but risk interference if not carefully scoped. PPO (Schulman et al., 2017) requires a learned value function and can be sample-inefficient; GRPO and variants (Shao et al., 2024) avoid critic learning by group-normalizing rewards within a sampled group. Verifiable rewards tend to yield more robust signals and reduce reward hacking, motivating our use of rule-based rewards for format, OCR, and evidence consistency.

A.3 Impact Statement

VeriPO is a post-training algorithm for document-understanding vision-language models. Its intended benefits are twofold. First, by supplying a dense, token-level supervision signal on reasoning tokens (where rule-based rewards cannot reach), VeriPO improves the factual grounding of model-produced rationales in documents, which matters for downstream uses that care about whether answers are traceable to source evidence (legal/financial document analysis, accessibility tooling for visually impaired readers, technical writing assistance). Second, the single-rollout design cuts post-training cost relative to group-rollout baselines, lowering the compute barrier for researchers operating under constrained budgets. We note three limitations that bear on responsible use. (i) The verification-safe property formally holds in the adversarial-teacher limit; in the partially-reliable regime a weak or biased teacher can still steer the policy within the safe envelope, so teacher selection remains a design decision. (ii) The rubric is only as good as the data pipeline that produced it; errors in reference parses propagate into R_{p} and transitively into training. (iii) The evaluation in this paper is English-Chinese and document-centric; performance on other languages, scripts, or non-document visual inputs is not

claimed. We recommend that any deployment on sensitive material include human oversight and an independent check of cited evidence.

A.4 Training Data Construction

SFT data construction. The SFT stage exists to lock the policy into the four-segment output protocol before any reward-based optimization begins. Concretely, we draw 30,000 raw question-answer instances from ChartQA and InfographicsVQA to cover a broad mix of document types (charts, tables, infographics, and pages that mix text with visuals). For each query we retrieve up to five candidate page images with ColPali and, to accustom the policy to both narrow and wide retrieved sets, we vary the number of retained images from one to five. Gold reasoning trajectories are obtained by rejection sampling: Qwen2.5-VL-72B and Qwen2.5-VL-7B each generate several candidate completions per query, and only those whose final answer matches the reference are kept. This yields roughly 60,000 SFT examples, each with a verified answer. Because every retained trajectory is already formatted into the four segments, the evidence spans required for later grounding supervision can be lifted directly from the parse, so no separate annotation round is needed.

VeriPO data curation. For reinforcement fine-tuning, we curate a smaller but higher-quality dataset that emphasizes challenging cases where the SFT model is prone to errors. Starting from the SFT training pool, we use the SFT-trained model to generate eight candidate completions per sample and compute rewards using our four-channel reward function. We apply two filtering strategies: (1) we remove samples where all completions achieve perfect scores, as these provide no learning signal and risk reward hacking; (2) we prioritize samples with high reward variance across completions, indicating cases where the model's behavior is unstable. To further increase difficulty, we augment the dataset with 400 multi-hop reasoning examples from MMLongBench-Doc that require cross-page evidence integration. The final VeriPO training set contains 4,000 curated samples, reflecting our emphasis on quality over quantity. This curriculum design, which uses broad coverage for SFT and focused hard-case mining for RL, enables efficient training while maximizing the impact of reward-based optimization on the most challenging document understanding scenarios.

A.5 Experiments

Model	Overall	ArXiv	Old Scans Math	Tables	Old Scans	Headers&Footers	Multi Col.	Long-Tiny Text	Base
Marker v1.6.2	59.4	24.3	22.1	69.8	24.3	87.1	71.0	76.9	99.5
MinerU v1.3.10	61.5	75.4	47.4	60.9	17.3	**96.6**	59.0	39.1	96.6
GOT OCR	48.3	52.7	52.0	0.2	22.1	93.6	42.0	29.9	94.0
Mistral OCR API	72.0	77.2	67.5	60.6	29.3	93.6	71.3	77.1	99.4
GPT-4o (No Anchor)	68.9	51.5	75.5	69.1	40.9	94.2	68.9	54.1	96.7
GPT-4o (Anchored)	69.9	53.5	74.5	70.0	40.7	93.8	69.3	60.6	96.8
Gemini Flash 2 (No Anchor)	57.8	32.1	56.3	61.4	27.8	48.0	58.7	84.4	94.0
Gemini Flash 2 (Anchored)	63.8	54.5	56.1	72.1	34.2	64.7	61.5	71.5	95.6
Qwen 2 VL (No Anchor)	31.5	19.7	31.7	24.2	17.1	88.9	8.3	6.8	55.5
Qwen 2.5 VL (No Anchor)	65.5	63.1	65.7	67.3	38.6	73.6	68.3	49.1	98.3
olmOCR v0.1.68 (No Anchor)	76.3	72.1	74.7	71.5	43.7	91.6	78.5	80.5	98.1
olmOCR v0.1.68 (Anchored)	77.4	75.6	75.1	70.2	44.5	93.4	79.4	81.7	99.0
Ours (7B)	**82.4**	**84.7**	**83.5**	**85.1**	**47.6**	88.6	**84.3**	**86.2**	**99.6**

Table 4: Performance comparison on the olmOCR (Poznanski et al., 2025) benchmark across multiple document domains and structural challenges. Higher is better.

Implementation pipeline (SFT then VeriPO). We instantiate our model on top of `Qwen2.5-VL-7B` and implement both stages using the released training toolchains `LLaMA-Factory` and `Easy-R1`, running on GPU clusters with resources comparable to NVIDIA A100 80GB-class hardware. The SFT data (§A.4) is consumed for one epoch with bf16, a global batch size of 32, cosine scheduling and warmup ratio 0.1, and a learning rate of 5×10^{-7}; we set the image budget to a maximum of 3,920,000 pixels per input. With this configuration, SFT typically finishes within several hours on an A100-80GB-class multi-GPU node, depending on I/O and sequence lengths.

VeriPO implementation and training. After SFT stabilizes the unified output protocol, we apply VeriPO to mix the rule-based reward channels (which separately supervise workflow validity, layout-aware parsing, evidence traceability, and evidence-conditioned answer correctness) with a token-level reverse-KL signal obtained from a frozen Qwen2.5-VL-72B teacher, gated per segment by the teacher's own rubric score. Following the data-curation stage above, we cache the teacher's reference

trajectory and cache per-sample α_k values prior to RL training (so that the online teacher cost during training is reduced to a single forward pass per token, giving token-level $\log \pi_\phi$ only). VeriPO is trained for four epochs with bf16, global batch size 32, and a *single* on-policy rollout per training sample (enabled by the dense teacher signal), learning rate 1×10^{-6}, weight decay 10^{-2}, rollout temperature 1.2, and PPO-style clipping parameters $\epsilon_{\text{lo}} = 0.2$ and $\epsilon_{\text{hi}} = 0.28$; we lower the image budget to 1,568,000 max pixels to keep rollouts efficient. In practice, this VeriPO stage usually takes on the order of one to two days of wall-clock time on A100-80GB-class multi-GPU hardware, with the exact time driven mainly by rollout length, the teacher's per-token inference latency, and reward evaluation overhead.

A.6 Ablation Study

Training recipe (SFT → VeriPO). Our model is trained with a two-stage pipeline: (i) SFT provides a stable warm start that teaches the model to follow the unified workflow (parse → evidence → answer) and to consistently emit structured intermediate outputs; (ii) VeriPO further aligns the model with a combination of verifiable rubric rewards and teacher-derived token-level supervision, so that improvements come from better reading and grounded reasoning rather than uncontrolled generation. Empirically, the full training pipeline yields large end-to-end gains over the Qwen2.5-VL-7B backbone (e.g., average performance improves from 55.50/50.41 to 76.04/78.49 in Acc/F1 on our VRAG evaluation), indicating that SFT alone is insufficient to close the perception-grounding gap without subsequent reward-based optimization.

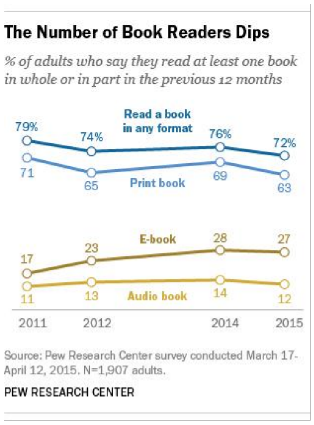
Finer-grained reward-channel ablations. The main ablation table (Table 3) isolates VeriPO's four design components at the level of algorithmic mechanism. Here we zoom inside one of those components, the rule-reward source itself, and ablate its three channels (R_{fmt}, R_{e}, R_{p}) and the two sub-terms of R_{p} (reading order and structure coverage).

Table 5: Finer-grained reward-channel ablations on the five-benchmark average. Deltas are relative to the full model (76.04/78.49 Acc/F1 in Table 3).

	Avg Acc	Avg F1
Ours (full)	76.04	78.49
w/o R_{fmt}	71.27	74.06
w/o R_{e}	73.78	76.78
w/o R_{p}	75.33	77.62
w/o reading-flow term (R_{flow})	74.65	77.10
w/o structure-coverage term (R_{type})	75.61	78.06

Removing the format channel (R_{fmt}) is the costliest of the three (-4.77 Acc and -4.43 F1 on average), which is consistent with its role as a *gating precondition* (§3.1): a trajectory that fails format compliance already zeros out every downstream rubric reward, so removing R_{fmt} as an explicit signal lets many malformed trajectories back into training. Stripping the evidence channel (R_{e}) costs -2.26 Acc / -1.71 F1, with the sharpest losses on ChartQA and InfoVQA, benchmarks where committing to a specific DU separates a grounded answer from a plausible-sounding guess. The parse channel (R_{p}) contributes a smaller -0.71 Acc / -0.87 F1 penalty; SFT pressure alone already produces a broadly readable parse, so what R_{p} buys is the *structural* fidelity of that parse, which only becomes visible on its sub-terms. Within R_{p}, dropping the reading-flow term (R_{flow}) costs -1.39 Acc, larger than removing R_{p} in full, because preserving content recovery (R_{align}) without its order-correctness check actively mis-trains reading order and corrupts citation grounding downstream, while dropping structure-coverage (R_{type}) is milder at -0.43 Acc. Taken together, these five ablations argue that the rule-reward decomposition is not redundant: format acts as a gate, evidence drives attribution, parse drives structural fidelity, and removing any one of them leaves a measurable gap.

Case Study. Figure 3 traces VeriPO through the same chart question used in the main text ("in which year is the gap between the two curves largest?"). The failure modes of an untrained baseline become legible only once we look for them segment by segment: a plausible natural-language summary can coexist with a year that does not sit on the x-axis, with a legend binding that swaps the two series, or with numbers that are not actually annotated on the chart. Our policy avoids those failure modes structurally. Inside o^{p} it enumerates the chart as DUs, including the legend→color assignment and each annotated data point. Inside o^{e} it emits one citation per year on the x-axis, each

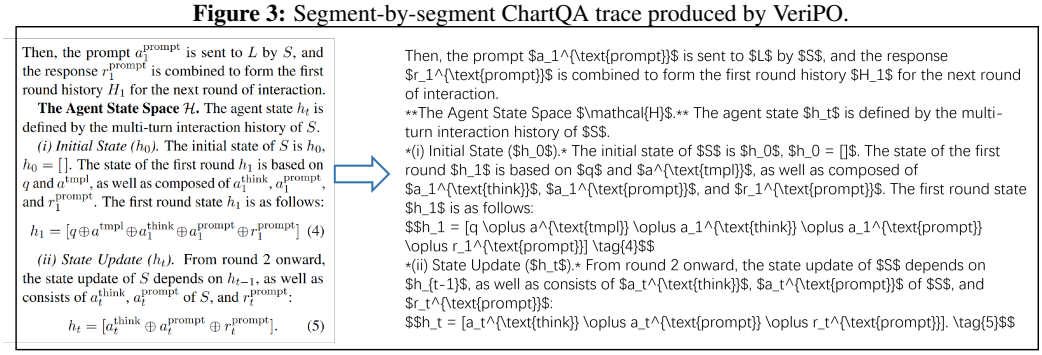


Question: In which year the difference between orange and green graph is maximum?

UniDocVLM Observation:
It shows the percentage of adults who say they read at least one book (in whole or in part) in the previous 12 months, broken down by format across 2011, 2012, 2014, and 2015.
Key series and values:
Read a book in any format: 79% (2011) → 74% (2012) → 76% (2014) → 72% (2015)
Print book: 71% → 65% → 69% → 63%
E-book: 17% → 23% → 28% → 27%
Audiobook: 11% → 13% → 14% → 12%

Answer: 2022

Figure 3: Segment-by-segment ChartQA trace produced by VeriPO.



Then, the prompt a_1^{prompt} is sent to L by S, and the response r_1^{prompt} is combined to form the first round history H_1 for the next round of interaction.
The Agent State Space $\mathcal{H}$. The agent state h_t is defined by the multi-turn interaction history of S.
(i) Initial State (h_0). The initial state of S is h_0, $h_0 = []$. The state of the first round h_1 is based on q and a^{tmpl}, as well as composed of a_1^{think}, a_1^{prompt}, and r_1^{prompt}. The first round state h_1 is as follows:

$$h_1 = [q \oplus a^{\text{tmpl}} \oplus a_1^{\text{think}} \oplus a_1^{\text{prompt}} \oplus r_1^{\text{prompt}}] \quad (4)$$

(ii) State Update (h_t). From round 2 onward, the state update of S depends on h_{t-1}, as well as consists of a_t^{think}, a_t^{prompt} of S, and r_t^{prompt}:

$$h_t = [a_t^{\text{think}} \oplus a_t^{\text{prompt}} \oplus r_t^{\text{prompt}}]. \quad (5)$$

Then, the prompt a_1^{prompt} is sent to L by S, and the response r_1^{prompt} is combined to form the first round history H_1 for the next round of interaction.
The Agent State Space $\mathcal{H}$. The agent state h_t is defined by the multi-turn interaction history of S.
(i) Initial State (h_0). The initial state of S is h_0, $h_0 = []$. The state of the first round h_1 is based on q and a^{tmpl}, as well as composed of a_1^{think}, a_1^{prompt}, and r_1^{prompt}. The first round state h_1 is as follows:

$$h_1 = [q \oplus a^{\text{tmpl}} \oplus a_1^{\text{think}} \oplus a_1^{\text{prompt}} \oplus r_1^{\text{prompt}}] \quad (4)$$

(ii) State Update (h_t). From round 2 onward, the state update of S depends on h_{t-1}, as well as consists of a_t^{think}, a_t^{prompt} of S, and r_t^{prompt}:

$$h_t = [a_t^{\text{think}} \oplus a_t^{\text{prompt}} \oplus r_t^{\text{prompt}}]. \quad (5)$$

Figure 4: Additional qualitative trace: a scanned page parsed into DUs under VeriPO.

pointing back to a DU in o^{p}. Only after these two segments have committed a verifiable evidence set does o^{r} compute the per-year differences and pick the maximum. The rubric checks on R_{e} refuse any citation that does not survive a lookup in o^{p}, which is what rules out the three baseline failure modes at training time.

A.7 Proof of Proposition 1 (Verification-Safe Gradient)

We restate the proposition and provide a detailed derivation.

Setting and notation. Let the trajectory τ decompose as $\tau = o^{\mathrm{p}} \oplus o^{\mathrm{e}} \oplus o^{\mathrm{r}} \oplus o^{\mathrm{a}}$ with segment map $k(t) \in \mathcal{K} = \{\mathrm{p}, \mathrm{e}, \mathrm{r}, \mathrm{a}\}$. Per Eq. equation 5, the VeriPO token-level advantage is

$$\hat{A}_t \;=\; \alpha_{k(t)} \cdot r_t^{\mathrm{KL}} \;+\; \beta_{k(t)} \cdot \tilde{R}_{k(t)}(\tau),$$

under the default routing $\beta_{\mathrm{r}} = 0$, $\alpha_{\mathrm{a}} \equiv 0$, and $\alpha_k = \sigma(\lambda(R^{\phi}_{\kappa(k)} - \rho_{\kappa(k)}))$ for $k \in \{\mathrm{p}, \mathrm{e}, \mathrm{r}\}$ with $\kappa(\mathrm{r}) = \mathrm{a}$.

Let $\mathcal{L}_{\mathrm{rule}}$ denote the "rule-only" objective obtained by setting $\alpha_k \equiv 0$ on every segment in Eq. equation 5; under $\beta_{\mathrm{r}} = 0$, $\mathcal{L}_{\mathrm{rule}}$ aggregates only the standardized rule rewards $\tilde{R}_k$ on $\{\mathrm{p}, \mathrm{e}, \mathrm{a}\}$.

Restatement. Assume:

(i) (*Adversarial teacher*) There exists $\delta > 0$ such that for every training sample $(q, \mathcal{D}_R)$ and every $j \in \{\mathrm{p}, \mathrm{e}, \mathrm{a}\}$, $R_j^{\phi}(q, \mathcal{D}_R) \le \rho_j - \delta$ almost surely.

(ii) (*Bounded KL signal*) The per-token reverse-KL reward is uniformly bounded: $\sup_{\tau,t} |r_t^{\mathrm{KL}}(\tau; \theta, \phi)| \le C < \infty$.

(iii) (*Bounded importance ratio*) The PPO-clip surrogate is evaluated with $r_t(\theta) \in [1 - \epsilon_{\mathrm{lo}},\, 1 + \epsilon_{\mathrm{hi}}]$ after clipping, so $|r_t(\theta)| \le 1 + \epsilon_{\max}$ with $\epsilon_{\max} := \max(\epsilon_{\mathrm{lo}}, \epsilon_{\mathrm{hi}})$.

(iv) (*Dominated differentiation*) The policy score $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ has finite second moment, uniformly in θ on a compact parameter set.

Proposition 2 (Verification-safe gradient, restated). *Under assumptions (i)–(iv),*

$$\lim_{\lambda\to\infty} \sup_{\theta} \left\| \nabla_\theta \mathcal{L}_{\mathrm{VeriPO}}(\theta) \;-\; \nabla_\theta \mathcal{L}_{\mathrm{rule}}(\theta) \right\| \;=\; 0.$$

Proof. Decompose the VeriPO objective by segment:

$$\mathcal{L}_{\mathrm{VeriPO}}(\theta) \;=\; \mathcal{L}_{\mathrm{rule}}(\theta) \;+\; \Delta(\theta),$$

where $\Delta(\theta)$ is the teacher-driven contribution,

$$\Delta(\theta) \;=\; -\,\mathbb{E}_{\tau\sim\pi_{\theta_{\mathrm{old}}}} \left[\frac{1}{|\tau|} \sum_{t=1}^{|\tau|} \alpha_{k(t)} \cdot r_t^{\mathrm{KL}} \cdot \widetilde{r}_t(\theta) \right], \tag{8}$$

with $\widetilde{r}_t(\theta)$ denoting either the raw importance ratio or its PPO clip, whichever realizes the minimum in Eq. equation 7. By assumption (iii), $|\widetilde{r}_t(\theta)| \le 1 + \epsilon_{\max}$.

Step 1 (gate decay). Under assumption (i), for each $k \in \{\mathrm{p}, \mathrm{e}, \mathrm{r}\}$ the gate satisfies

$$\alpha_k \;=\; \sigma\big(\lambda(R^{\phi}_{\kappa(k)} - \rho_{\kappa(k)})\big) \;\le\; \sigma(-\lambda\delta),$$

almost surely in the sample. Since $\alpha_{\mathrm{a}} \equiv 0$ by design, we have $\alpha_{k(t)} \le \sigma(-\lambda\delta)$ uniformly in t.

Step 2 (pointwise bound on Δ). Combining Step 1 with assumptions (ii) and (iii), the integrand in equation 8 is bounded almost everywhere by

$$\left|\alpha_{k(t)} \cdot r_t^{\mathrm{KL}} \cdot \widetilde{r}_t(\theta)\right| \;\le\; \sigma(-\lambda\delta) \cdot C \cdot (1 + \epsilon_{\max}),$$

and therefore $|\Delta(\theta)| \le \sigma(-\lambda\delta) \cdot C \cdot (1 + \epsilon_{\max})$ uniformly in θ.

Step 3 (uniform convergence of gradients). Assumption (iv) permits exchanging differentiation and expectation in equation 8 via dominated convergence. Differentiating under the expectation,

$$\nabla_\theta \Delta(\theta) \;=\; -\,\mathbb{E}_\tau \left[\frac{1}{|\tau|} \sum_{t=1}^{|\tau|} \alpha_{k(t)}\, r_t^{\mathrm{KL}}\, \nabla_\theta \widetilde{r}_t(\theta) \right],$$

where $\nabla_\theta \widetilde{r}_t(\theta)$ is either $r_t(\theta)\, \nabla_\theta \log \pi_\theta(a_t \mid s_t)$ on the unclipped region or zero on the clipped region; in either case it is bounded by assumption (iv). Applying Cauchy–Schwarz pointwise and Step 2,

$$\left\| \nabla_\theta \Delta(\theta) \right\| \;\le\; \sigma(-\lambda\delta) \cdot C \cdot M,$$

for a constant M depending only on $\epsilon_{\max}$ and the score-function second moment. The bound is independent of θ (over the assumed compact parameter set).

Step 4 (conclude). Since $\sigma(-\lambda\delta) \to 0$ as $\lambda \to \infty$ for any fixed $\delta > 0$, we conclude

$$\sup_{\theta} \left\| \nabla_\theta \Delta(\theta) \right\| \;\longrightarrow\; 0,$$

and therefore $\nabla_\theta \mathcal{L}_{\mathrm{VeriPO}}(\theta) \to \nabla_\theta \mathcal{L}_{\mathrm{rule}}(\theta)$ uniformly in θ, as claimed. □

Finite-λ tightness. The proposition is a limit statement, but the underlying bound $|\nabla_\theta \Delta(\theta)| \le \sigma(-\lambda\delta) \cdot C \cdot M$ is quantitative and tight at practical settings. The sigmoid tail decays exponentially:

$$\sigma(-\lambda\delta) \;\le\; e^{-\lambda\delta}.$$

Two concrete operating points illustrate this. At $\lambda = 8$, $\delta = 0.1$ (a weakly adversarial teacher that falls just below the 25th-percentile threshold), $\sigma(-0.8) \approx 0.31$: the teacher signal is attenuated to roughly a third of its nominal magnitude, and VeriPO inherits about two-thirds of the rule-only gradient. At $\lambda = 8$, $\delta = 0.4$ (a meaningfully failing teacher), $\sigma(-3.2) \approx 0.039$: the teacher contribution is reduced by more than an order of magnitude, and VeriPO's gradient is essentially indistinguishable from $\mathcal{L}_{\mathrm{rule}}$ at the scale of a single training step. Combined with the fact that only samples with $R^{\phi}_{\kappa(k)} < \rho_{\kappa(k)}$ carry any residual contribution (typically a minority under adversarial teachers once the gate is applied), the finite-λ residual is negligible at the operating points we use in practice.

What the proposition does and does not claim. Proposition 1 asserts that the *gradient* of VeriPO collapses onto that of the rule-only baseline whenever the teacher is adversarially uninformative. This implies identical local training dynamics and, under standard non-degeneracy of the rule gradient, identical stationary points: VeriPO cannot be made to *underperform* its rule-only special case by manipulating the teacher. The proposition does *not* claim that VeriPO is strictly better than rule-only training in the partially-reliable regime; the empirical comparison at various teacher-quality levels is deferred to §4.4, where we evaluate VeriPO under deliberately weakened teachers and confirm that performance matches but does not fall below the rule-only baseline.