IMPROVING OCR FAITHFULNESS VIA GATED AND ATTENUATED ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models may rewrite anomalous text in images into linguistically plausible expressions, compromising OCR transcription faithfulness. Sequence-level task rewards and local teacher guidance are complementary, but guidance from the same teacher may not remain equally effective as the student improves. Offline analysis shows that supervision from a fixed teacher becomes progressively less favorable as the student improves, both across training checkpoints and across response groups with different task rewards. Motivated by this observation, we introduce **GAD-RL**, which adaptively regulates teacher supervision during joint post-training according to the student's current task performance and local distributions. A frozen teacher conditions on reference transcriptions and student-generated prefixes. GAD-RL disables distillation for response groups containing an output with task reward at least 0.95 and continuously attenuates distillation strength as group-mean reward increases. It also weights forward KL by the student's probability of the teacher's Top-1 token, moderating local auxiliary updates when student support for that candidate is low. On Qwen3.5-2B, GAD-RL achieves 59.92% Micro Recall on CHAOS-Bench, surpassing GRPO and GRPO+OPD (fixed-weight) by 8.45 and 4.43 percentage points, respectively, while achieving an Overall score of 91.18 on OmniDocBench v1.6.

1 INTRODUCTION

Document parsing aims to faithfully convert the text and structure in page images into machine-readable representations. Vision-language models (VLMs) integrate text recognition, layout understanding, and structured generation, advancing complex document parsing (Bai et al., 2025; Poznanski et al., 2025). Yet strong parsing ability does not guarantee faithful transcription: misspellings, visually confusable characters, or unexpected words may be replaced with linguistically plausible expressions that contradict the image. This linguistic-prior hallucination turns reading into implicit rewriting (Yao et al., 2026; Lee et al., 2026). We focus on OCR transcription faithfulness: when visual evidence conflicts with linguistic priors, models should preserve the content actually visible in the image, even when the source itself contains errors, while maintaining reliable parsing of ordinary text and document structure.

To optimize anomalous-content retention and transcription quality, we adopt Group Relative Policy Optimization (GRPO; Shao et al., 2024), which has been applied to document parser post-training (Poznanski et al., 2025; Wang et al., 2026). Unlike supervised fine-tuning (SFT), which learns under reference prefixes, GRPO uses relative rewards within groups of student-generated responses. However, identical group rewards yield zero group-relative advantages (Yu et al., 2025), while rarely sampled faithful outputs receive limited direct reinforcement. We therefore complement GRPO with on-policy distillation (OPD), which provides token-level supervision at student-generated prefixes (Agarwal et al., 2024).

Recent work has explored regulating on-policy teacher supervision through student-aware target reformulation (Jang et al., 2026) or trajectory-level selection (Akhondzadeh et al., 2026). Our focus is complementary: we study how the utility of a fixed teacher changes with the student's current task performance, and use this observation to regulate when and how strongly the teacher intervenes during joint GRPO post-training. Offline analysis of GRPO-only checkpoints shows rising

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.

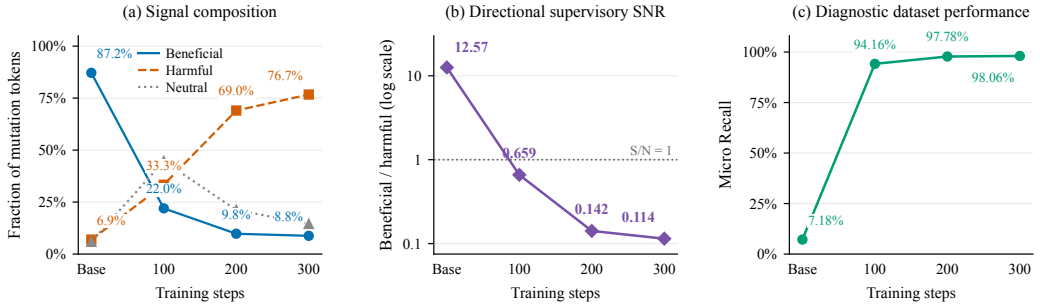


Figure 1: **Student performance improves while directional supervisory SNR declines.** The same frozen teacher scores GRPO checkpoints offline on a separate analysis set of 1,000 synthetic pages. (a) Signal composition on perturbed-word-associated tokens: relative to student probabilities at the same prefixes, teacher signals are beneficial when favoring correct emitted tokens or disfavoring errors, and potentially harmful in the reverse direction; small differences are neutral (Appendix B.4). (b) Beneficial-to-harmful token-count ratio on a log scale; the dotted line marks equal counts. (c) Student perturbed-word Micro Recall.

perturbed-word Micro Recall (7.18% to 98.06%) and declining directional supervisory signal-to-noise ratio (SNR; 12.57 to 0.114) on a fixed analysis set (Figure 1).

Motivated by these observations, we propose GAD-RL, a policy-state-aware framework that regulates teacher intervention according to the student's current task performance. Group-level gating determines when teacher supervision remains active, while reward-aware attenuation continuously reduces its strength as group performance improves. Separately, to avoid overly strong local FKL updates under substantial teacher–student disagreement, we use a simple student-probability weight to moderate the distillation gradient without altering the teacher target.

We evaluate GAD-RL on GlitchText (Yao et al., 2026) and CHAOS-Bench (Li et al., 2026), covering controlled text anomalies and character perturbations in realistic document layouts. On CHAOS-Bench, GAD-RL improves perturbed-word Micro Recall from 51.47% for GRPO to 59.92%, a gain of **8.45 percentage points**, while maintaining comparable general document-parsing performance (Table 1). Comparisons with fixed-weight distillation, step-based decay, and low-score response selection examine state-aware regulation against simpler controls (Section 5.2).

Our main contributions are:

- **Revealing changes in teacher supervision with student state.** Through token-level analysis, we identify a decline in the same frozen teacher's directional supervisory signal-to-noise ratio as the student improves, providing empirical motivation for dynamically regulating teacher intervention.
- **Introducing state-aware teacher intervention for OCR faithfulness.** Based on the observed reward-dependent variation in supervision quality, GAD-RL combines group-level mastery gating with continuous reward-aware attenuation to regulate when and how strongly a frozen teacher intervenes during joint GRPO post-training. We additionally use student-probability-weighted FKL to moderate large local distillation updates under teacher–student disagreement.
- **Improving transcription faithfulness across student backbones.** GAD-RL improves CHAOS-Bench Recall over GRPO by 8.45 and 3.82 percentage points on Qwen3.5-2B and Qwen3-VL-2B, respectively, while maintaining comparable OmniDocBench performance. Ablations and controlled comparisons further support the effectiveness of regulating teacher intervention according to the student's current task performance.

2 Related Work

Document parsing. LayoutLM and LayoutLMv3 integrate text, image, and layout features (Xu et al., 2020); Donut, Pix2Struct, and Nougat recover structured content from images (Kim et al., 2022; Lee et al., 2023; Blecher et al., 2024). RL-based parsing uses verifiable rewards in

olmOCR 2 (Poznanski et al., 2025) and layout-aware rewards in Infinity-Parser (Wang et al., 2026). OvisOCR2 combines real-document annotations with synthetic HTML-derived pages, then uses SFT, RL on a larger branch, OPD into a compact parser, and model fusion (Lu et al., 2026). OmniDocBench measures general parsing quality (Ouyang et al., 2025). However, current VLM-based parsers remain vulnerable to text perturbations: FaithC4 reveals unfaithful rewriting and error amplification on unperturbed content (Lee et al., 2026), while CHAOS-Bench, introduced with HunyuanOCR-1.5, reports low perturbed-word recall across evaluated OCR models (Li et al., 2026).

Multimodal hallucination mitigation. VCD, OPERA, and VISTA modify decoding or visual representations (Leng et al., 2024; Huang et al., 2024; Li et al., 2025b). PAR targets OCR overcorrection through positional perturbation and attention recycling at inference time, without additional training (Yao et al., 2026). Training methods use factual feedback (Sun et al., 2024), preference optimization (Zhao et al., 2023; Yu et al., 2024), or phrase-level alignment (Sarkar et al., 2025). Our work focuses on OCR transcription faithfulness: mitigating overcorrection driven by linguistic priors while preserving the text actually present in document images.

On-policy learning and distillation. GKD supports on-policy distillation jointly with RL (Agarwal et al., 2024). KDRL combines GRPO and reverse KL, with linear coefficient decay and reward-guided response/group masks; group distillation stops when any response succeeds (Xu et al., 2025). OPSD uses privileged-context self-distillation (Zhao et al., 2026); TGPO supplies teacher-preferred next tokens at student prefixes under large policy divergence (Liu et al., 2026). Veto constructs student-dependent geometric targets (Jang et al., 2026).

Teacher supervision reliability and control. RG-OPD filters trajectories by agreement between verifier advantages and teacher–student log-likelihood gaps (Akhondzadeh et al., 2026). RSTG (*Distill Where You Fail*) combines all-incorrect zero-variance group selection, teacher-confidence weighting, token selection, and SFT on correct teacher trajectories (Han et al., 2026). I-SDPO routes all-incorrect groups to privileged self-distillation and any-success groups to GRPO (Zhang et al., 2026a). PACED favors intermediate student pass rates based on cross-problem gradient SNR (Xu et al., 2026). Analyzing teacher noise, Ding & Zhang (2026) find that fixed negative advantages on low-probability sampled tokens can match OPD in their reverse-KL reasoning-distillation settings. We track a fixed teacher's correctness-aware beneficial-to-harmful token-count ratio across OCR student checkpoints as transcription improves. GAD-RL combines group gating and continuous group-mean-reward attenuation with GRPO; student-probability weighting separately moderates local FKL updates.

3 METHODS

GAD-RL combines GRPO with state-aware on-policy distillation from a frozen teacher. Group-level gating and reward-aware attenuation weaken or disable teacher guidance as task performance improves, while student-adaptive forward KL scales token-level updates. GRPO remains active throughout training.

Let $(x, z) \sim \mathcal{D}$ be a document example, where x contains the page image and parsing instruction, and z is the ground truth (GT). The student is π_θ. The frozen teacher $\pi_{\theta_{\mathrm{T}}}$ receives x_{T}, a text prompt combining z with a content-preserving rewrite instruction, without the page image. For each input, the rollout policy $\pi_{\theta_{\mathrm{old}}}$ generates a group $\mathcal{Y} = \{y^{(i)}\}_{i=1}^{G}$, with $y^{(i)} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x)$. Here, i indexes responses, t indexes tokens, and $T_i = |y^{(i)}|$ is the valid response length. Student and teacher score tokens under the same generated prefix $y^{(i)}_{<t}$. We write $\mathbb{E}_{\mathrm{roll}}$ for expectation over $(x, z) \sim \mathcal{D}$ and these sampled groups.

GRPO objective. Let $R_i = R(y^{(i)}, z)$ be the task reward, defined in Section 3.2, and let $\bar{R}$ and σ_R be its group mean and standard deviation. The token-level policy ratio and sequence advantage are

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})}{\pi_{\theta_{\mathrm{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})}, \qquad A_i = \frac{R_i - \bar{R}}{\sigma_R}, \qquad A_{i,t} = A_i. \tag{1}$$

When $\sigma_R = 0$, we set all group advantages to zero. GRPO (Shao et al., 2024) maximizes

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{\mathrm{roll}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i} \min(\rho_{i,t}(\theta)A_i, \mathrm{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\, A_i)\right], \tag{2}$$

where ε is the clipping parameter. Tokens are averaged within each response, and responses are averaged within each group.

3.1 Mastery-Gated Distillation

When to distill. When the student can already generate a sequence that receives the maximum task reward for an input, that sequence requires no further correction under the current reward criterion. Continuing to match the teacher distribution may then introduce unnecessary supervision noise and conflict with task-reward optimization (Han et al., 2026). We therefore disable group-level distillation when any sampled response reaches the success threshold. The gate is

$$g = \mathbb{I}\left[\max_{i=1,\ldots,G} R_i < \tau\right]. \tag{3}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function, and τ is chosen to tolerate minor formatting differences such as whitespace and heading levels. GRPO always uses all responses and reinforces those with above-average rewards when group rewards differ.

3.2 Reward-Aware Attenuation

How strongly to distill. A binary gate assigns the same weight to low-performing and nearly solved groups retained for distillation. Motivated by the lower directional supervisory SNR observed in higher-reward groups within individual checkpoints (Figure 4), we use group-average task reward to reduce guidance as performance improves, while the gate determines whether distillation remains active.

For an output y and its GT z, we combine normalized edit similarity R_{edit} with exact perturbed-word recall R_{recall}:

$$R(y,z) = \begin{cases} R_{\mathrm{edit}}(y,z), & \text{regular documents,} \\ \eta R_{\mathrm{edit}}(y,z) + (1-\eta)R_{\mathrm{recall}}(y,z), & \text{text-perturbed documents.} \end{cases}$$

Here, $\eta \in [0,1]$ controls the balance between edit similarity and perturbed-word recall. Both components lie in $[0,1]$; their definitions are given in Appendix D.

The group-mean reward is $\bar{R} = G^{-1}\sum_{i=1}^{G} R_i$. For groups with $g = 1$, we use

$$f_\kappa(\bar{R}) = \frac{e^{-\kappa\bar{R}} - e^{-\kappa}}{1 - e^{-\kappa}}. \tag{4}$$

The parameter $\kappa > 0$ controls the decay rate. With $f_\kappa(0) = 1$ and $f_\kappa(1) = 0$, this factor continuously weakens teacher guidance as group-mean reward increases. Gating and attenuation together give the group weight $g f_\kappa(\bar{R})$.

3.3 Student-Adaptive Forward KL

For the local gradient comparison at a fixed generated prefix, write $\mathbf{p}$ and $\mathbf{q}$ for the student and teacher probability vectors. Sampled-token estimators of reverse KL, $D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q})$, naturally use student rollouts to provide on-policy token-level credit. Their direct signals act on sampled actions: they can penalize a student-preferred error without explicitly specifying which unsampled alternative to promote (Liu et al., 2026; Han et al., 2026). FKL instead supplies teacher-distribution targets at student-generated prefixes (Agarwal et al., 2024), including teacher-supported candidates absent from the sampled response.

This direct correction can be strong under teacher–student disagreement. Let $\mathbf{h}$ be the student logit vector, so $\mathbf{p} = \mathrm{softmax}(\mathbf{h})$. Let a denote the sampled token and A its fixed group-relative advantage. The negative unclipped token-level GRPO surrogate, denoted ℓ_{GRPO}, at $\theta = \theta_{\mathrm{old}}$ and the full-vocabulary FKL loss $\ell_{\mathrm{FKL}} = D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{p})$ have logit gradients

$$\begin{aligned} \nabla_{\mathbf{h}} \ell_{\mathrm{GRPO}} &= A(\mathbf{p} - \mathbf{e}_a), \\ \nabla_{\mathbf{h}} \ell_{\mathrm{FKL}} &= \mathbf{p} - \mathbf{q}, \end{aligned} \tag{5}$$

where $\mathbf{e}_a$ is the one-hot vector for a. A teacher-supported candidate with very low student probability can receive a much stronger FKL signal than its policy-gradient signal, particularly when the student is highly confident in its sampled token. The FKL logit gradient is bounded, but its relative scale can be large. Since GRPO clipping does not constrain an added FKL term, excessive auxiliary contributions may destabilize policy updates.

We use Student-Adaptive Forward KL (SA-FKL) to moderate local updates when the student assigns low probability to the teacher's preferred candidate. Let $b_{i,t}$ denote the teacher's Top-1 token at the student-generated prefix. The probability weight and forward-KL term are

$$\begin{aligned} b_{i,t} &= \arg\max_{v \in \mathcal{V}} \pi_{\theta_{\mathrm{T}}}(v \mid x_{\mathrm{T}}, y^{(i)}_{<t}), \\ w_{i,t}(\theta) &= \mathrm{sg}\left[\pi_\theta(b_{i,t} \mid x, y^{(i)}_{<t})\right], \\ d_{i,t}(\theta) &= D_{\mathrm{KL}}\left(\pi_{\theta_{\mathrm{T}}}(\cdot \mid x_{\mathrm{T}}, y^{(i)}_{<t}) \,\middle\|\, \pi_\theta(\cdot \mid x, y^{(i)}_{<t})\right). \end{aligned} \tag{6}$$

Here, $\mathcal{V}$ is the vocabulary and sg stops gradient propagation through the weight. The token contribution is $w_{i,t}(\theta)d_{i,t}(\theta)$, so the student's probability of the teacher's Top-1 token scales the full local KL gradient without changing its direction. When the teacher favors a candidate with little student support, this weight attenuates the potentially strong FKL update. The gradient is derived in Appendix C.2.

3.4 Joint optimization.

We average the weighted KL terms over valid tokens in each group and apply the group controls inside the rollout expectation:

$$\begin{aligned} \mathcal{J}_{\mathrm{OPD}}(\theta) &= \mathbb{E}_{\mathrm{roll}}\left[g f_\kappa(\bar{R}) \frac{\sum_{i=1}^{G} \sum_{t=1}^{T_i} w_{i,t}(\theta) d_{i,t}(\theta)}{\sum_{i=1}^{G} T_i} \right], \\ \mathcal{J}(\theta) &= \mathcal{J}_{\mathrm{GRPO}}(\theta) - \lambda \mathcal{J}_{\mathrm{OPD}}(\theta). \end{aligned} \tag{7}$$

GAD-RL maximizes $\mathcal{J}$. Since $\mathcal{J}_{\mathrm{OPD}}$ is a distillation cost, it enters with a minus sign; λ controls its strength. The gate and attenuation are computed separately for each sampled group and held fixed during the update. The vocabulary-truncated approximation is specified in Appendix D.

4 Experiments

Experimental setup. For each backbone, the teacher and student start from the same base model. The teacher is fine-tuned on 200,000 transcription-task examples. Document-parsing training uses a mixed dataset of 14,400 samples, with a 3:2 ratio of general to text-perturbed documents. General documents are sampled from Infinity-Doc2-5M (Huang et al., 2026) and MonkeyDoc (Li et al., 2025a). We cross-check their annotations against parsing outputs from PaddleOCR-VL-1.6 (Zhang et al., 2026b) and discard samples with text similarity below 0.9. All student training strategies start from the base checkpoint and run for 300 steps with a learning rate of 10^{-6}. Unless otherwise specified, we report benchmark results from the checkpoint after 300 training steps. Detailed training settings are provided in Appendix D.

Baselines. We evaluate SFT, GRPO, and GRPO+OPD (fixed-weight) on both Qwen3-VL-2B and Qwen3.5-2B. **GRPO+OPD (fixed-weight)** adds a fixed-weight distillation loss directly to the GRPO objective on all responses. The default distillation coefficient is $\lambda = 0.005$, used by both GAD-RL and GRPO+OPD (fixed-weight) in the main experiments.

4.1 Benchmarks and Metrics

GlitchText (Yao et al., 2026) introduces controlled errors into familiar passages rendered on a plain background. We report Identification Rate (Ident) and Correction Rate (Cor), averaged equally across Chinese and English. *CHAOS-Bench* (Li et al., 2026) introduces character corruptions into 500 academic-paper page images with realistic layouts; we report Micro Recall. For general document parsing, we use *OmniDocBench v1.6* (Ouyang et al., 2025) and report its Overall score. Metric definitions, matching rules, and aggregation details are provided in Appendix E.

4.2 Main Results

Table 1 compares the original model (Baseline), the post-training baselines, and GAD-RL for each student model.

Table 1: Transcription faithfulness and general document parsing, grouped by student model. Micro Recall is a ratio; other scores are percentages. Bold marks the best result within each model block.

Method	CHAOS-Bench	GlitchText		OmniDocBench v1.6
	Micro Recall↑	Ident↑	Cor↓	Overall↑
Student: Qwen3.5-2B				
Baseline	0.0402	76.48	14.94	80.06
SFT	0.3763	82.20	10.71	90.29
GRPO	0.5147	91.30	4.21	90.80
GRPO+OPD (fixed-weight)	0.5549	91.31	**3.14**	90.90
GAD-RL	**0.5992**	**93.39**	3.40	**91.18**
Student: Qwen3-VL-2B				
Baseline	0.0235	83.91	8.28	46.39
SFT	0.3804	87.75	7.00	87.97
GRPO	0.5255	88.33	4.39	89.86
GRPO+OPD (fixed-weight)	0.4167	89.26	4.69	87.67
GAD-RL	**0.5637**	**92.15**	**3.59**	**90.00**

On CHAOS-Bench, GAD-RL improves Micro Recall over GRPO by 8.45 and 3.82 percentage points on Qwen3.5-2B and Qwen3-VL-2B, respectively, and outperforms fixed-weight GRPO+OPD on both backbones.

On GlitchText, GAD-RL improves anomaly identification and reduces overcorrection relative to SFT and GRPO on both backbones; fixed-weight OPD retains a lower correction rate on Qwen3.5-2B. GAD-RL also maintains comparable general document-parsing performance on OmniDocBench.

Training dynamics. GRPO leads early in training, but GAD-RL overtakes it and finishes with higher mean reward and perturbed-word Pass@8 (Figure 2a,b).

The two controls play complementary roles over training. The gate disables an increasing fraction of groups, while the mean attenuation factor among active groups falls from about 0.12 to 0.02 (Figure 2c,d). Teacher guidance therefore weakens even on groups that remain active.

4.3 Ablation Studies

KL formulation in on-policy distillation. Following prior work on divergence choices in on-policy distillation (Agarwal et al., 2024), we compare K1 PG, reverse KL (RKL), Jensen–Shannon divergence (JSD), FKL, and SA-FKL on Qwen3.5-2B within the GRPO+OPD (fixed-weight) setup, replacing only the distillation loss while keeping the coefficient at λ and all other settings fixed.

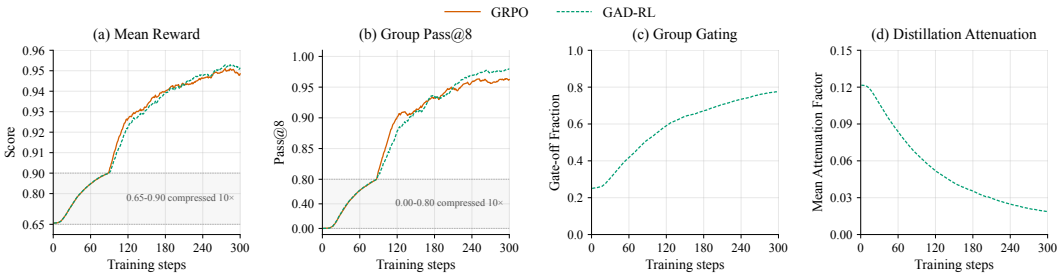


Figure 2: **GRPO and GAD-RL training dynamics.** (a) Mean reward. (b) Fraction of groups with at least one of eight responses achieving perturbed-word Recall of 1. For GAD-RL, (c) the fraction of groups with distillation disabled by the gate and (d) the mean attenuation factor among the remaining active groups.

SA-FKL achieves the highest CHAOS-Bench Recall after 100 steps (Table 2). FKL learns faster initially but shows a reward dip around steps 50–60; SA-FKL is smoother over this interval and achieves higher Recall from step 40 onward (Figure 3).

Distillation coefficient. Among the four nonzero coefficients in Table 3, only $\lambda = 0.005$ outperforms GRPO. Neither halving nor increasing this value helps, showing that fixed-weight OPD is sensitive to coefficient selection.

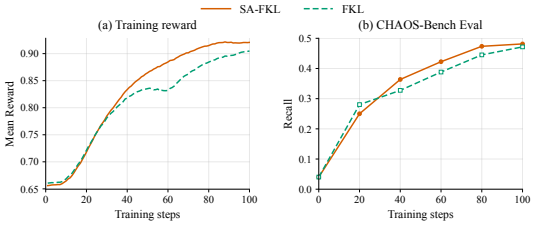


Figure 3: **FKL and SA-FKL training dynamics.** With otherwise identical settings: (a) mean training reward (EMA 0.95); (b) CHAOS-Bench Micro Recall, with a step-0 reference of 0.0402.

Table 2: Distillation objectives on CHAOS-Bench at step 100.

Objective	**Recall↑**
K1 PG	0.4500
RKL	0.4441
JSD	0.4578
FKL	0.4716
SA-FKL	**0.4814**

Table 3: Fixed-weight distillation at different coefficients.

Coefficient	**Recall↑**
0 (GRPO)	0.5147
0.5λ	0.5010
λ	**0.5549**
5λ	0.4873
10λ	0.4637

Gating and attenuation. Table 4 removes student-probability weighting (SW), mastery-gated distillation (MGD), or reward-aware attenuation (RAA) at λ. Removing SW or RAA causes larger Recall losses than removing MGD. The smaller incremental benefit of gating is consistent with RAA already suppressing high-reward groups: at $\bar{R} = 0.9$, it retains only 1.83% of the base coefficient before token weighting.

Table 4: Component ablations on CHAOS-Bench at λ.

Variant	**Recall↑**
GRPO	0.5147
GAD-RL (full)	**0.5992**
w/o SW	0.5598
w/o MGD	0.5931
w/o RAA	0.5414

5 ANALYSIS

Building on Figure 1, we examine whether the teacher prefers incorrect candidates at correct student outputs and how supervision quality varies with group reward within each checkpoint. We then compare teacher-intervention strategies through training experiments.

5.1 TEACHER SUPERVISION AS THE STUDENT IMPROVES

Lowering a correct token's probability does not necessarily imply an incorrect teacher Top-1 prediction. We therefore inspect teacher Top-1 predictions at correct student outputs, reusing Figure 1's 1,000 pages excluded from training, GRPO-only student checkpoints, and frozen teacher.

At the same student-generated prefix, we count a teacher error when its Top-1 token conflicts with the GT continuation. Among correctly transcribed perturbed-word-associated tokens, this rate rises from 9.52% at step 100 to 42.88% at step 300 (Table 5), showing an increasing preference for incorrect candidates at positions the student already transcribes correctly.

Table 5: Teacher Top-1 errors at correct student outputs. Counts refer to perturbed-word-associated tokens. The final columns give teacher errors as percentages of all eligible tokens and of correct student tokens.

Checkpoint	All tokens	Correct tokens	Teacher errors	% of all	% of correct
Base	4,281	1,029	131	3.06	12.73
Step 100	10,325	10,159	967	9.37	9.52
Step 200	12,903	12,865	4,234	32.81	32.91
Step 300	13,391	13,366	5,731	42.80	42.88

Token counts vary across checkpoints because the generated responses differ, resulting in different numbers of tokens aligned to perturbed-word spans; detailed statistics and matching rules are provided in Appendix B.4.

Reward dependence within each checkpoint. Across-checkpoint trends combine changes in training stage and task performance. We therefore examine reward dependence within GRPO+OPD checkpoints at steps 50 and 200, scoring current-policy rollouts with the same frozen teacher. Figure 4 partitions 995 matched diagnostic page groups by mean task reward; scoring and bootstrap details are in Appendix B.5.

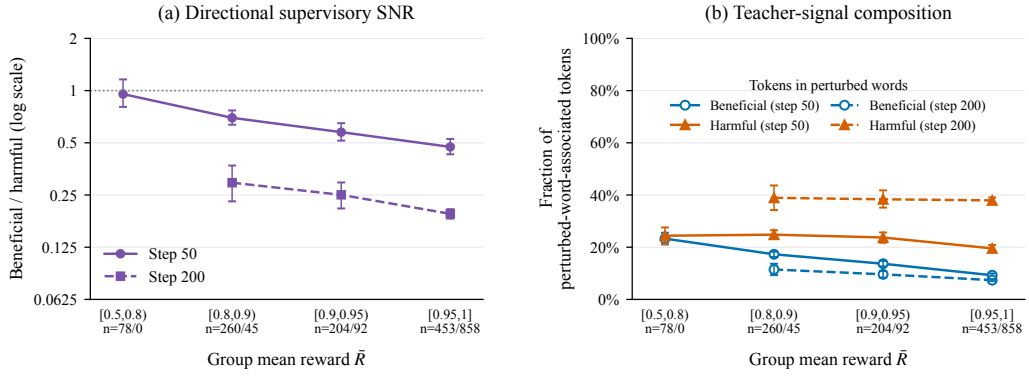


Figure 4: **Teacher supervision becomes less favorable as group reward increases.** GRPO+OPD groups at steps 50 and 200 are binned by mean reward $\bar{R}$. For perturbed-word-associated tokens: (a) beneficial-to-harmful count ratio (log scale; dotted line: equal counts); (b) beneficial and harmful fractions, with the remainder neutral. The absolute probability-change margin is 0.1. n gives group counts at step 50 / 200; error bars show 95% input-group-bootstrap confidence intervals.

Within both checkpoints, beneficial signals become less frequent as group reward increases, while harmful signals do not decline proportionally. Consequently, useful teacher corrections become scarcer relative to conflicting supervision as the student performs better on an input. This within-checkpoint pattern supports using group reward to regulate teacher intervention: a training-step schedule assigns the same coefficient to groups with different relative corrective opportunities. Reward-aware attenuation can instead reduce guidance on groups that already achieve high task reward.

5.2 Distillation Control Strategies

How should teacher intervention be controlled? Section 5.1 motivates comparing uniform changes in strength, response selection, and decay over training steps with state-aware regulation. Building on the coefficient ablation in Table 3, Table 6 compares filtering and decay strategies. These Qwen3.5-2B comparisons retain the GRPO task loss and use a base coefficient of $\lambda = 0.005$. All distillation-control comparisons use SA-FKL; fixed-weight OPD applies it to all responses. Recall is evaluated on CHAOS-Bench. Low-score OPD selects individual responses with total reward $R_i < 0.5$; GAD-RL gates whole groups and attenuates their weight using group-mean reward.

Table 6: Distillation decay and filtering strategies. The best Recall is bold.

Method	Coefficient[1]	Recall↑
GRPO	0	0.5147
+ Linear-decay OPD	$\lambda\, s(u)$	0.5693
+ Low-score OPD	λ	0.5775
GAD-RL	$\lambda g f_\kappa(\bar{R})$	**0.5992**
	$10\lambda g f_\kappa(\bar{R})$	0.5882

[1]The base coefficient is $\lambda = 0.005$; token-level weights are omitted. The gate g and attenuation $f_\kappa(\bar{R})$ are defined in Sections 3.1 and 3.2, respectively. The factor $s(u)$ decreases linearly with training step u.

Response selection and temporal decay. Compared with GRPO+OPD (fixed-weight), low-score OPD and linear-decay OPD improve Recall by 2.26 and 1.44 percentage points, respectively. These gains are consistent with the changing utility of teacher supervision: response selection concentrates guidance on outputs needing correction, while temporal decay relaxes teacher constraints as training progresses.

GAD-RL coefficient. Increasing GAD-RL's base coefficient tenfold lowers Recall by 1.10 percentage points; both tested settings outperform GRPO. The strong attenuation on active groups shown in Figure 2d may help explain this modest change: teacher guidance weakens as group rewards improve, even when the base coefficient is larger.

6 Conclusions

We presented GAD-RL, which combines GRPO with group-gated, reward-attenuated, and student-weighted on-policy distillation. Offline analysis shows that teacher–student disagreement increasingly affects correctly transcribed perturbed-word-associated tokens as the student improves, while favorable suppression of residual errors remains common. GAD-RL improves transcription faithfulness over SFT and GRPO on two student models while maintaining comparable OmniDocBench performance. On Qwen3.5-2B, GAD-RL outperforms GRPO at both tested distillation coefficients. These findings support adapting teacher intervention to the student's evolving task performance and local distributions.

Limitations and future work. Our evaluation is limited to OCR transcription faithfulness in document parsing. Future work will explore GAD-RL on other tasks to assess whether policy-state-aware distillation control generalizes beyond OCR.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search and summarization, the design of mathematical notation, and formula verification. AI tools also generated most of the code used for synthetic data generation, figure visualization, and method implementation. The authors manually reviewed and validated the code and take responsibility for the final manuscript, mathematical claims, implementation, and reported results.

REPRODUCIBILITY STATEMENT

Appendix A describes data synthesis, Appendix B details teacher-signal analysis, and Appendix C provides the gradient derivations. Training configurations and objectives are specified in Appendix D, and evaluation metrics in Appendix E. Upon acceptance, we will release the training, evaluation, and analysis code, the data-synthesis pipeline, teacher-training data, and the 1,000-page analysis set. Data will be distributed as files where their licenses permit, or through source manifests and reconstruction instructions where redistribution is restricted.

ETHICS STATEMENT

This work studies faithful transcription using synthetic text perturbations and existing document benchmarks. Synthetic documents are derived from arXiv LaTeX sources, whose reuse conditions depend on each source's license. Third-party resources, including GlitchText, CHAOS-Bench, and OmniDocBench, remain subject to their original licenses and terms. Releases of derived data will preserve source attribution and respect redistribution restrictions. Perturbed text is intended to test transcription fidelity and should not be treated as factual content. Improved transcription faithfulness does not establish the truth of the underlying documents.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation. *arXiv preprint arXiv:2607.04037*, 2026. URL https://arxiv.org/abs/2607.04037.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*, 2024.

Yi Ding and Ruqi Zhang. Does on-policy distillation really distill? From noisy teacher to self-improvement. *arXiv preprint arXiv:2608.31046*, 2026. URL https://arxiv.org/abs/2608.31046.

Zhuowen Han, Jinwei Xiao, Zhengxi Lu, Renren Jin, Zhiyuan Yao, Yuxin Liu, Hongyan Hao, Yueqing Sun, Yu Yang, Qi Gu, Xunliang Cai, and Deyi Xiong. Distill where you fail: Recovering learning signals of negative RL-groups from adaptive teacher guidance. *arXiv preprint arXiv:2608.00782*, 2026. URL https://arxiv.org/abs/2608.00782.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

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.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 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. 498–517, 2022.

Gwang Gook Lee, Kenan Emir Ak, Jay Mohta, Yan Xu, and Dimitrios Dimitriadis. Do VLMs read or rewrite? On transcription faithfulness in vision-language models. *arXiv preprint arXiv:2607.21617*, 2026. URL https://arxiv.org/abs/2607.21617.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Gengluo Li, Xingyu Wan, Shangpin Peng, Weinong Wang, Hao Feng, Yongkun Du, Binghong Wu, Zheng Ruan, Zhiqiong Lu, Liang Wu, et al. HunyuanOCR-1.5: Making lightweight OCR VLMs faster and better. *arXiv preprint arXiv:2607.04884*, 2026. URL https://arxiv.org/abs/2607.04884.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Biao Yang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. MonkeyOCR: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*, 2025a. URL https://arxiv.org/abs/2506.05218.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *International Conference on Machine Learning*, pp. 35799–35819, 2025b.

Xinyu Liu, Kechen Jiao, Chunyang Xiao, Runsong Zhao, Junhao Ruan, Bei Li, Jiahao Liu, Qifan Wang, Xin Chen, Jingang Wang, Chenglong Wang, Tong Xiao, and Jingbo Zhu. Teacher-guided policy optimization for on-policy reasoning distillation under large policy divergence. *arXiv preprint arXiv:2605.13230*, 2026. URL https://arxiv.org/abs/2605.13230.

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.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24838–24848, 2025.

Jake Poznanski, Luca Soldaini, and Kyle Lo. olmOCR 2: Unit test rewards for document OCR. *arXiv preprint arXiv:2510.19817*, 2025. URL https://arxiv.org/abs/2510.19817.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, 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.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Baode Wang, Biao Wu, Weizhen Li, Meng Fang, Zuming Huang, Jun Huang, Yanjie Liang, Haozhe Wang, Ling Chen, Wei Chu, and Yuan Qi. Infinity-Parser: Layout-aware reinforcement learning with high-quality document parsing dataset. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1647–1667, 2026.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. KDRL: Post-training reasoning LLMs via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025. URL https://arxiv.org/abs/2506.02208.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. pp. 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. PACED: Distillation and on-policy self-distillation at the frontier of student competence. *arXiv preprint arXiv:2603.11178*, 2026. URL https://arxiv.org/abs/2603.11178.

Yao Yao, Manwen Liao, Weitian Zhang, Zuchao Li, and Hai Zhao. PAR: Training-free positional perturbation and attention recycling for faithful OCR. In *Annual Meeting of the Association for Computational Linguistics*, pp. 23258–23273, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Yubo Zhang, Xinhong Ma, Zezhong Tan, and Ziqiang Dong. I-SDPO: Instance-level adaptive self-distillation policy optimization. *arXiv preprint arXiv:2608.12957*, 2026a. URL https://arxiv.org/abs/2608.12957.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, Tingquan Gao, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *arXiv preprint arXiv:2606.03264*, 2026b. URL https://arxiv.org/abs/2606.03264.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. URL https://arxiv.org/abs/2601.18734.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. URL https://arxiv.org/abs/2311.16839.

A Data Synthesis

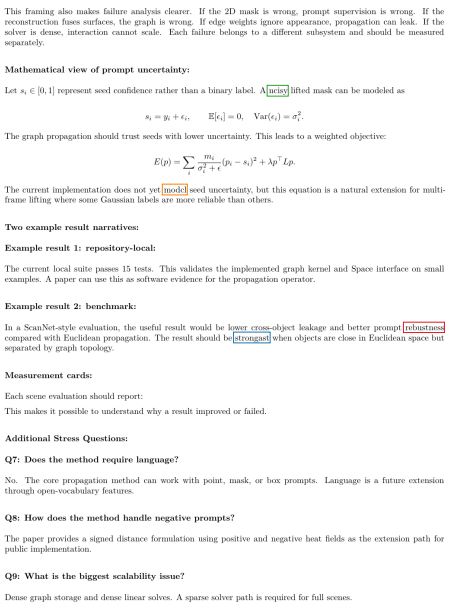
A.1 Text-perturbed documents

We construct text-perturbed documents from arXiv LaTeX projects. Text, headings, formulas, and tables are reflowed into academic layouts to produce page images and matching Markdown targets. The target preserves the anomalous text printed in the image.

Page construction. Source extraction removes citations, cross-references, bibliography entries, and figure content. Tables and formulas retain their source-derived content; tables use HTML in the target. Pages preserve the association of tables with captions and headings with following content. Unsupported structures and overflowing pages are rejected.

Controlled perturbations. Candidate words contain at least four letters, occur uniquely on the page, and have unambiguous source positions. Each edit replaces one lowercase letter without changing word length, using a shared weighted character-pair table. We request three edits per page with probability 0.4 and four with probability 0.6. Headings, captions, tables, formulas, code, and numerical content are protected; replacements colliding with existing words are skipped. Figure 5 shows an example.

Verification. Substitutions are applied consistently to the rendering source and target. Extracted PDF text and positions verify consistency and detect overflow, without generating or repairing the target. A separate analysis set contains 1,000 pages from 147 papers, including 140 table pages and 3,564 perturbed words.



This framing also makes failure analysis clearer. If the 2D mask is wrong, prompt supervision is wrong. If the reconstruction fuses surfaces, the graph is wrong. If edge weights ignore appearance, propagation can leak. If the solver is dense, interaction cannot scale. Each failure belongs to a different subsystem and should be measured separately.

Mathematical view of prompt uncertainty:

Let $s_i \in [0, 1]$ represent seed confidence rather than a binary label. A noisy lifted mask can be modeled as

$$s_i = y_i + \epsilon_i, \qquad \mathbb{E}[\epsilon_i] = 0, \quad \mathrm{Var}(\epsilon_i) = \sigma_i^2.$$

The graph propagation should trust seeds with lower uncertainty. This leads to a weighted objective:

$$E(p) = \sum_i \frac{m_i}{\sigma_i^2 + \epsilon}(p_i - s_i)^2 + \lambda p^\top L p.$$

The current implementation does not yet model seed uncertainty, but this equation is a natural extension for multi-frame lifting where some Gaussian labels are more reliable than others.

Two example result narratives:

Example result 1: repository-local:

The current local suite passes 15 tests. This validates the implemented graph kernel and Space interface on small examples. A paper can use this as software evidence for the propagation operator.

Example result 2: benchmark:

In a ScanNet-style evaluation, the useful result would be lower cross-object leakage and better prompt robustness compared with Euclidean propagation. The result should be strongest when objects are close in Euclidean space but separated by graph topology.

Measurement cards:

Each scene evaluation should report:

This makes it possible to understand why a result improved or failed.

Additional Stress Questions:

Q7: Does the method require language?

No. The core propagation method can work with point, mask, or box prompts. Language is a future extension through open-vocabulary features.

Q8: How does the method handle negative prompts?

The paper provides a signed distance formulation using positive and negative heat fields as the extension path for public implementation.

Q9: What is the biggest scalability issue?

Dense graph storage and dense linear solves. A sparse solver path is required for full scenes.

Figure 5: Example of a synthetically perturbed document page. Boxes highlight character-level substitutions in “noisy”, “model”, “robustness”, and “strongest”.

A.2 Teacher rewrite data

The teacher is fine-tuned on 200,000 transcription-task examples. We build copying and controlled-format rewrite pairs from source-derived transcriptions, without OCR or LLM content rewriting. Samples contain 1,000–7,800 tokens with tables kept intact, and train/validation splits are made by source paper. Approximately 10% of eligible word occurrences are perturbed, with at least three edits, using the same character-pair table.

Input and target preserve identical content, including anomalous words, numbers, formulas, and tables. Only existing heading prefixes and an optional outer Markdown fence may change; heading-level changes need not preserve hierarchy. Validation requires exact target reconstruction using only these permitted edits.

B Token Classification and Teacher-Signal Analysis

B.1 Fixed-trajectory scoring

For each page, we generate a response using the Qwen3.5-2B-based student with thinking disabled and a maximum generation length of 8,192 tokens. Student and teacher score the same emitted token ID at the same response prefix: the student is conditioned on the image, and the teacher on a heading-rewrite prompt containing the reference text. All probabilities are computed over the full vocabulary.

B.2 Token alignment, categories, and correctness

Decoded responses are aligned to the reference after Unicode NFKC normalization and removal of recognized markup and whitespace. Annotated perturbed words must match unambiguous whole-word reference spans. Ambiguous or invalid alignments are excluded from correctness statistics.

Each response token is assigned one category, in priority order: *perturbed-word-associated* if it overlaps the aligned full span of a perturbed word; *formatting* if it contains only whitespace or markup; and *body* otherwise. Perturbed-word association includes correct copies and overcorrections, not just the edited character. Visible numbers, punctuation, formulas, table-cell text, and code remain content; tokens mixing markup and content are body unless perturbed-word overlap takes priority.

Categories do not imply correctness. A content token is correct only when all its aligned normalized content characters match the reference; a correct subtoken within an erroneous word can therefore remain correct. Statistics count eligible response tokens equally, excluding prompts, trailing special tokens, and invalid alignments.

B.3 Teacher information reliability

We compare the original Qwen3.5-2B self-teacher and the SFT teacher specified in Appendix D on the same 1,000 student responses. Both receive identical reference text and the training-time teacher prompt. Correctness denominators exclude 5,179 records with inconsistent decoded-token offsets.

Error suppression. With $\Delta p = p^{\mathrm{T}} - p^{\mathrm{S}}$, the SFT teacher suppresses 69.58% of the 37,631 erroneous content tokens by more than 0.1, compared with 20.54% for the self-teacher. Harmful reinforcement falls from 5.88% to 1.48% (Figure 6). On perturbed-word errors, beneficial suppression reaches 99.85%. These directions concern the emitted token; suppression alone does not verify a faithful alternative.

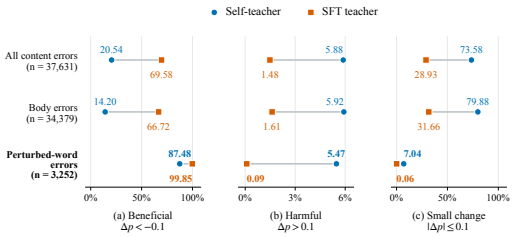


Figure 6: Teacher signals on erroneous tokens. For the same emitted token, $\Delta p < -0.1$ is beneficial suppression and $\Delta p > 0.1$ is harmful reinforcement. Panel (b) uses an enlarged horizontal scale.

Teacher-choice ablation. Table 7 uses the same Qwen3.5-2B GRPO and default SFT-teacher results as Table 1. The SFT teacher improves Recall over the self-teacher by 5.51 percentage points, supporting the use of a teacher trained for transcription.

B.4 Frozen-teacher supervision across GRPO checkpoints

Figure 1 evaluates each GRPO-only checkpoint on the same analysis set described in Section 5.1, with teacher scoring performed offline. Initialization, training data, and the teacher prompt match the main experiments (Appendix D). Base is the RL initial checkpoint (step 0); students receive no OPD updates.

Table 7: Teacher-choice ablation for GAD-RL on CHAOS-Bench.

Method	Teacher	Recall↑
GRPO	None	0.5147
GAD-RL	Self-teacher (2B)	0.5441
GAD-RL	SFT Qwen3.5-2B	0.5992

Signal definition. For an eligible perturbed-word-associated token, let $c \in \{0, 1\}$ indicate correctness, define $\Delta p = p_{\mathrm{T}} - p_{\mathrm{S}}$, and set $h = (2c - 1)\Delta p$. A signal is beneficial for $h > 0.1$, harmful for $h < -0.1$, and neutral otherwise. Invalid alignments and formatting-only tokens are excluded. With counts B, H, and U, the plotted fractions use denominator $N = B + H + U$, and directional SNR is B/H.

Table 8: Fixed-teacher signal counts and perturbed-word Recall on the analysis set. Signal statistics use eligible emitted tokens; Recall uses all 3,564 annotated words, including omissions.

Student	N	Beneficial	Harmful	Neutral	B/H	Recall (%)
Base	4,281	3,732	297	252	12.566	7.18
Step 100	10,325	2,270	3,443	4,612	0.659	94.16
Step 200	12,903	1,261	8,908	2,734	0.142	97.78
Step 300	13,391	1,175	10,273	1,943	0.114	98.06

Token-count variation. These counts follow the tokens each checkpoint generates. Across the same 3,493 perturbed words represented at every checkpoint, the average number of tokens per word rises from 1.22 at Base to 2.94, 3.65, and 3.79 at steps 100, 200, and 300. Finer output tokenization accounts for most of the increase, while the 1,000 input pages remain fixed.

Preservation as the student improves. Recall uses normalized reference-span matching. From Base through steps 100, 200, and 300, the teacher assigns probabilities more than 0.1 below the student's to 28.57/33.88/69.24/76.86% of correctly transcribed perturbed-word-associated tokens. Favorable suppression signals on remaining erroneous tokens occur at rates of 99.85/97.59/100/100%, with denominators 3,252/166/38/25. At step 300 all potentially harmful signals occur on correct tokens. The aggregate decline thus accompanies greater directional disagreement on currently correct output, while favorable signals on residual errors remain prevalent.

Teacher Top-1 at suppressed correct tokens. We split correct perturbed-word-associated tokens with $\Delta p < -0.1$ by teacher Top-1 compatibility (Table 9). Using Appendix B's normalization, we require consecutive exact-match GT anchors for the emitted token. Nonempty Top-1 content matching the GT continuation's prefix is compatible, including alternative tokenizations; nonmatching content is incompatible. Unavailable alignments and formatting-only or undecodable candidates remain unresolved. This tests local compatibility, not whole-word correctness.

Table 9: Teacher Top-1 at correct perturbed-word-associated tokens with $\Delta p < -0.1$. Entries are counts (percent of N), including unresolved cases in the denominator.

Student	N	GT-compatible	Incompatible	Unresolved
Base	294	136 (46.26)	131 (44.56)	27 (9.18)
Step 100	3,442	2,413 (70.10)	967 (28.09)	62 (1.80)
Step 200	8,908	4,615 (51.81)	4,234 (47.53)	59 (0.66)
Step 300	10,273	4,458 (43.40)	5,731 (55.79)	84 (0.82)

These diagnostics count tokens on the analysis set. Checkpoints emit different token sets, and small remaining error subsets limit rate comparisons.

B.5 Reward-stratified teacher supervision

Figure 4 uses Qwen3.5-2B GRPO+OPD checkpoints at steps 50 and 200 and the same frozen SFT teacher. The two checkpoints share 995 matched diagnostic page groups. Of the original 1,000 diagnostic pages, five are excluded because model inference repeatedly failed for these inputs despite retries;

We use the correctness-aware signal definition in Appendix B.4 with an absolute probability-change margin of 0.1. Within each bin, beneficial and harmful fractions divide the respective pooled token counts by all evaluable perturbed-word-associated tokens, including neutral tokens. Directional SNR is the ratio of the pooled beneficial and harmful counts. Error bars are 95% bootstrap confidence intervals from 2,500 resamples of input groups within each reward bin, keeping responses from the same input clustered. Empty bins have no estimate.

C Gradient Scale of Controlled Distillation

C.1 Policy gradients and unweighted FKL

All gradients below are with respect to the student logits $\mathbf{h}$ at unit temperature. Using Equation 5, let $p(a) = 1 - \xi$ and $q(a) \le 1 - \delta$ for $0 < \xi < \delta \le 1$: ξ is the student's probability mass away from sampled token a, and δ lower-bounds the corresponding teacher mass. At $\theta = \theta_{\text{old}}$, before clipping and group controls,

$$\|\nabla \ell_{\text{GRPO}}\|_1 = 2|A|\xi, \qquad 2(\delta - \xi) \le \|\nabla \ell_{\text{FKL}}\|_1 \le 2. \tag{8}$$

For fixed disagreement and bounded nonzero advantages, FKL can be much stronger than PG at a high-confidence error, although its logit gradient remains bounded. This comparison holds at fixed prefixes and does not differentiate through the rollout distribution.

C.2 Teacher-preferred-token probability weighting

At a fixed prefix, let $b = \arg\max_v q(v)$ be the teacher's Top-1 token, $w = \text{sg}[p(b)]$, and $d = D_{\text{KL}}(\mathbf{q}\|\mathbf{p})$. The weighted token loss is $\ell_{\text{SA-FKL}} = wd$. With the probability weight detached,

$$\nabla_{\mathbf{h}} \ell_{\text{SA-FKL}} = p(b)(\mathbf{p} - \mathbf{q}). \tag{9}$$

This scales the FKL gradient without changing its direction. If the teacher prefers a candidate to which the student assigns little probability, the local update is attenuated even when the student is highly confident in its sampled token a. In the high-confidence setting of Appendix C.1, $b \ne a$ implies $p(b) \le \xi$, hence

$$\|\nabla_{\mathbf{h}} \ell_{\text{SA-FKL}}\|_1 \le 2p(b) \le 2\xi. \tag{10}$$

This trades correction strength for update moderation when the student assigns low probability to the teacher's preferred candidate.

For the Top-K approximation, let $\mathbf{q}^{\mathcal{K}}$ retain the original teacher probabilities on $\mathcal{K}$ and be zero elsewhere, and set $s = \sum_{v \in \mathcal{K}} q(v)$. Writing $d^{\mathcal{K}}$ for the truncated KL term gives

$$\begin{aligned} \nabla_{\mathbf{h}}\big(wd^{\mathcal{K}}\big) &= p(b)(s\mathbf{p} - \mathbf{q}^{\mathcal{K}}), \\ \big\|\nabla_{\mathbf{h}}\big(wd^{\mathcal{K}}\big)\big\|_1 &\le 2sp(b). \end{aligned} \tag{11}$$

The scalar gate g, attenuation $f_\kappa(\bar{R})$, and coefficient λ further scale these local contributions. Actual parameter updates also depend on model Jacobians, batch aggregation, and the optimizer.

D Training Configuration and Objectives

Objective averaging. Equation 2 first averages valid tokens within each response, then averages responses within a group. The OPD term in Equation 7 instead divides the sum of weighted token KL terms by $\sum_i T_i$ for each group, before taking the rollout expectation. Groups with $g = 0$ contribute zero to OPD and remain in the GRPO objective. Rewards, advantages, and group controls are held fixed during policy updates. The equivalent minimized training loss is $-\mathcal{J}_{\text{GRPO}} + \lambda \mathcal{J}_{\text{OPD}}$.

Table 10: Shared GAD-RL training configuration for Qwen3.5-2B and Qwen3-VL-2B. Each student uses a teacher trained from the same backbone.

Setting	Value
Student / teacher backbone	Qwen3.5-2B or Qwen3-VL-2B; matched within each pair
Initialization	Student: base checkpoint; teacher: transcription SFT from the same base
Teacher SFT data	200,000 transcription-task examples
Mixed dataset	14,400 samples (perturbed + general)
Mixture (general:perturbed)	3:2
Learning rate / duration	10^{-6} / 300 steps
Input batch / PPO minibatch	48 / 48; microbatch 1 per GPU
Responses per input	8 (384 per rollout batch)
Rollout temperature / top-p	1 / 1; top-k filtering disabled
Teacher Top-K / default coefficient	32 / $\lambda = 0.005$
Main-result coefficient	λ
Group controls	Disable distillation if $\max_i R_i \geq 0.95$; attenuation $\kappa = 3$
Teacher / student distillation temperature	1 / 1
Student prompt / response limits	8,192 / 8,192 tokens
Maximum image pixels	4,194,304
Hardware	8 H100 GPUs

Teacher input. The teacher receives reference text in a standalone heading-rewrite prompt, with no image. It changes existing heading prefixes, preserves all other characters, and outputs no outer Markdown fence. The same teacher prompt is used for offline signal analysis.

Training reward components. For text-perturbed documents, we use $\eta = 0.5$ to give equal weight to edit similarity and perturbed-word recall. For output y and GT z, the reward components in Section 3.2 are

$$\begin{aligned} R_{\text{edit}}(y, z) &= 1 - \frac{D_{\text{lev}}(y, z)}{\max(|y|, |z|)}, \\ R_{\text{recall}}(y, z) &= \frac{\sum_w \min\big(c_{M(z)}(w), c_y(w)\big)}{|M(z)|}. \end{aligned} \tag{12}$$

Here, D_{lev} is the Levenshtein distance, and $|y|$ and $|z|$ are text lengths. $M(z)$ is the multiset of annotated perturbed words as printed; $c_{M(z)}(w)$ and $c_y(w)$ count exact occurrences of word w in that multiset and the output. The sum ranges over distinct annotated perturbed words, and $|M(z)|$ counts all annotated occurrences. These training rewards are distinct from the benchmark metrics in Appendix E.

Truncated FKL. For the vocabulary-truncated approximation, let $\mathcal{K}_{i,t}$ contain the teacher's Top-K candidates, with $K = 32$. We replace $d_{i,t}$ in Equation 7 with

$$\begin{aligned} d^{\mathcal{K}}_{i,t}(\theta) = \sum_{v \in \mathcal{K}_{i,t}} & \pi_{\theta_{\text{T}}}(v \mid x_{\text{T}}, y^{(i)}_{<t}) \\ & \times \log \frac{\pi_{\theta_{\text{T}}}(v \mid x_{\text{T}}, y^{(i)}_{<t})}{\pi_\theta(v \mid x, y^{(i)}_{<t})}. \end{aligned} \tag{13}$$

The teacher probabilities retain their original full-vocabulary mass. Without renormalization, this truncated surrogate can be negative. The corresponding token loss is

$$\ell^{\mathcal{K}}_{i,t}(\theta) = w_{i,t}(\theta) d^{\mathcal{K}}_{i,t}(\theta), \tag{14}$$

where $w_{i,t}(\theta) = \text{sg}[\pi_\theta(b_{i,t} \mid x, y^{(i)}_{<t})]$ and $b_{i,t}$ is the teacher's Top-1 token. The weight uses the student's full-vocabulary probability of this teacher-preferred token and is detached; gradients pass

only through the KL term. No additional rollout-ratio weight is applied to OPD. Teacher-choice comparisons use the same objective and group controls; component comparisons remove the indicated control in Table 4.

E Evaluation Metrics

We use benchmark normalization and matching rules for the metrics in Section 4.1.

GlitchText. For K annotated anomalies, K_{vis} counts predictions matching the visible perturbed form and K_{orig} counts restorations to the original form under official span alignment:

$$\text{Ident} = K_{\text{vis}}/K, \qquad \text{Cor} = K_{\text{orig}}/K. \tag{15}$$

Omissions and other errors count toward neither numerator. Table 1 reports equal-weight means of the separate GlitchText-ZH and GlitchText-EN-Padding200 scores, expressed as percentages.

CHAOS-Bench. For perturbed words $\mathcal{P}_i$ and prediction $\hat{y}_i$ on page i, with h indicating a case-insensitive whole-word match,

$$\text{MicroRecall}_{\text{CHAOS}} = \frac{\sum_i \sum_{w \in \mathcal{P}_i} h(w, \hat{y}_i)}{\sum_i |\mathcal{P}_i|}. \tag{16}$$

Each annotated target has equal weight across pages; unperturbed words are excluded.

General parsing. We report the official OmniDocBench v1.6 Overall score as a percentage.

F Case Study

Figure 7 illustrates both useful and conflicting guidance. In (a), the teacher suppresses an incorrect restoration to ordinary spelling; in (b), it reinforces a faithful perturbed-word token. In (c) and (d), it suppresses correctly transcribed perturbed-word-associated tokens. Correctness follows the printed GT, including intentional spelling perturbations. Thus privileged conditioning can produce both directionally beneficial and potentially conflicting signals, depending on token correctness and the direction of probability change.

(a) Beneficial signal

Suppressing an incorrect token



Image detail

ned classifier, it is important to verify

GT `inportant`

Response `important`

Original word: important

Original input page

Pos.	Token ID	Response token	Student P	Teacher P
751	369	`·is`	1.00000	0.99772
752	**2894**	**`·important`**	**0.99998**	**4.00e-08**
753	310	`·to`	1.00000	0.07977

(b) Beneficial signal

Reinforcing a correct token



Image detail

deteriorates with growimg model complexity

GT `growimg`

Response `growimg`

Original word: growing

Original input page

Pos.	Token ID	Response token	Student P	Teacher P
382	440	`·with`	1.00000	0.99842
383	2965	`·grow`	0.97030	0.99859
384	**1827**	**`img`**	**0.32344**	**0.99899**
385	1558	`·model`	1.00000	0.99817

(c) Harmful signal

Suppressing a correct token



Image detail

e severe accuracy ccllapses (42.9%–

GT `ccllapses`

Response `ccllapses`

Original word: collapses

Original input page

Pos.	Token ID	Response token	Student P	Teacher P
95	13016	`·accuracy`	0.99960	0.99779
96	272	`·c`	0.93244	0.99814
97	549	`cl`	0.99829	0.99841
98	75	`l`	0.89470	0.99895
99	2538	`aps`	0.49705	0.00026
100	**287**	**`es`**	**1.00000**	**1.11e-06**
101	318	`·(`	1.00000	0.99853

(d) Harmful signal

Suppressing a correct token



Image detail

aid the crisis will wcrsen next year"). Ten

GT `wcrsen`

Response `wcrsen`

Original word: worsen

Original input page

Pos.	Token ID	Response token	Student P	Teacher P
482	668	`·will`	1.00000	0.99855
483	288	`·w`	0.99840	0.99647
484	66	`c`	0.79300	9.76e-06
485	5259	`rs`	0.99223	0.00050
486	**268**	**`en`**	**1.00000**	**2.09e-06**
487	1727	`·next`	1.00000	0.99858

Figure 7: Beneficial and harmful teacher signals. Panels show the page, enlarged perturbed-word region, GT, student response, and token probabilities. Yellow rows mark focal tokens; · denotes a space. Student and teacher score the same emitted token at the same prefix, conditioned on image and privileged text, respectively.