MessToClean: Evidence-Grounded Structure-Preserving Reconstruction for Real-World Degraded Exam Paper Images

Jiayi Tuo[1,†], Cheng Tang[1,†], Zihan Wang[1], Chenyue Zhou[2], Yao Li[3], Yanbiao Ma[4,5,6,*]
Chao Wang[1], Wei Dai[7], Mingxuan Wang[4], Shitong Qin[8], Ziwei Zhao[9]

[1]University of Science and Technology of China
[2]Nanyang Technological University [3]Kean University
[4]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[5]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[6]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[7]Xidian University [8]Gaotu Techedu Inc. [9]Technical University of Munich

[†] Equal contribution [*] Corresponding author: ybmamail@stu.xidian.edu.cn

Abstract

Intelligent education systems often collect exam sheets as in-the-wild photos. These photos often suffer from distortions and noise caused by handwriting and occlusions, collectively referred to as Real-World Degraded Exam Images (RDEI). Structure-preserving reconstruction is key to converting RDEI into structured assets for downstream educational applications. Existing Multimodal Large Language Models (MLLMs) often fail under RDEI, leading to disrupted structure and evidence-unsupported hallucinations. To tackle these challenges, we propose MessToClean, a backbone-agnostic, evidence-driven pipeline that treats off-the-shelf MLLMs as interchangeable components. By grounding extraction in pixel-aligned evidence and enforcing post-hoc consistency auditing on recovered structures, MessToClean mitigates unsupported hallucinations and enhances both controllability and structural fidelity in question-level reconstruction. We curate RDEI-Exam from our educational platforms and evaluate across 12 state-of-the-art MLLM backbones. Across these, MessToClean improves stem consistency by 1.01-3.18%, figure consistency by 0.50-49.16%, and refusal F1 by 1.06-10.88% across question types.

1 Introduction

Intelligent education platforms have accumulated vast collections of exam sheet resources. However, their utility for automated assessment and learning diagnosis is limited by unstructured storage (Dinh et al., 2024; Zhang et al., 2024b), as they are typically captured as in-the-wild photographs using mobile devices rather than as scanned documents (Das et al., 2021; Yu et al., 2026). Such images often suffer from geometric distortions caused by shooting angles and are further degraded by significant noise from handwritten answers, grading marks, and correction annotations (Zhang et al., 2024c). We collectively refer to these inputs as *Real-World Degraded Exam Images* (RDEI). Moreover, exam papers inherently exhibit complex layouts, hierarchical multi-question structures, and strong coupling between questions and figures, which makes question-level structural recovery especially challenging in real-world settings (Aich et al., 2026).

Under these conditions, restoring the structural integrity of the question, particularly its boundaries, reading order, and correspondence with the diagram, becomes highly unstable and constitutes a critical bottleneck in the scalable exam processing pipeline (Aich et al., 2026; Zhang et al., 2024a; Ramu et al., 2024; Wang et al., 2021). The most straightforward approach is optical character recognition (OCR), which typically follows a *text detection–text recognition* pipeline to transcribe document images into editable text (Li et al., 2023a; Liao et al., 2020). However, when applied to RDEI, the combination of degradations significantly exacerbates character-level errors, text line fragmentation, and missed detections. More importantly, this paradigm lacks explicit mechanisms to recover essential structural relations such as question boundaries, reading order, and question–figure correspondence (Li et al., 2025; Xing et al., 2025).

Although multimodal large language models (MLLMs) can generate structured text from document images in an end-to-end manner (Hu et al., 2025; Liao et al., 2025; Chen et al., 2023), they still suffer from two systematic limitations when processing RDEI, as illustrated in Figure 1. First, handwritten annotations and severe noise often lead MLLMs to over-rely on linguistic priors, resulting in hallucinated content that contradicts pixel-level evidence (Xu et al., 2025a). Second, general-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40304–40322
July 2-7, 2026

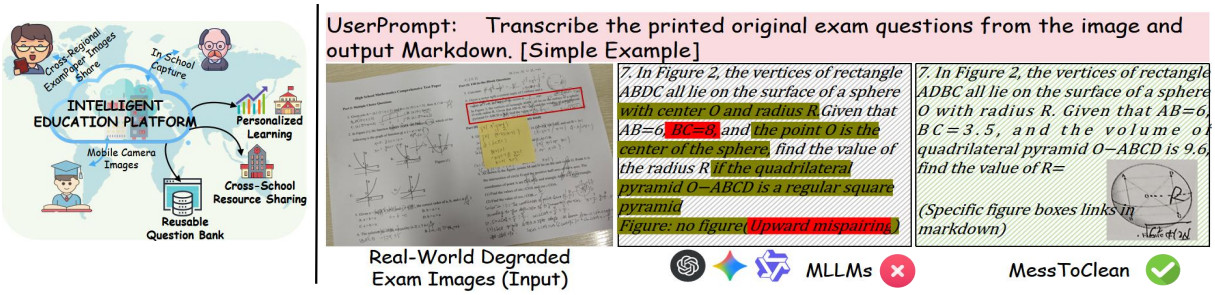

Figure 1: MLLMs Limitations and MessToClean in Action for Structure-Preserving Reconstruction in RDEI.

purpose MLLMs exhibit unstable region-level localization and alignment (Yin et al., 2025; Wang et al., 2025; Xu et al., 2025b), making it difficult to accurately identify question blocks and their associated figures. These issues introduce latent structural errors that may silently corrupt downstream question banks.

Motivated by these observations, we propose **MessToClean**, an evidence-driven pipeline that is decoupled from specific MLLM backbones. MessToClean treats off-the-shelf MLLMs as interchangeable components and suppresses hallucinations through explicit pixel-aligned evidence grounding and post-hoc consistency auditing. Specifically, we first employ a two-stage fine-tuned RF-DETR detector to robustly localize question text blocks and figure blocks under occlusion and large-scale noise, thereby constructing a set of pixel-aligned evidence bounding boxes. Based on these detections, we reconstruct the hierarchical page structure and infer a stable reading order as an integral part of the recovered representation. Finally, cropped question and figure regions are used as pixel-level evidence to drive a *Generator-Verifier-Patcher* pipeline, which performs consistency verification and minimal corrective rewriting, and outputs structured Markdown along with auditable JSONL edit logs. In parallel, we curate RDEI-Exam, a real-world benchmark collected from our intelligent educational platforms across multiple regions and schools, comprising 12,472 Real-World Degraded Exam Images.

In summary, our contributions are as follows:

- We introduce and formalize structure-preserving reconstruction for RDEI, which aims to produce traceable and auditable question-level structured representations.
- We propose MessToClean, an MLLM-backbone-agnostic, evidence-driven pipeline. MessToClean leverages pixel-aligned evidence to recover question structure and enforce globally consistent question-figure pairing, and further applies an evidence-constrained *Generator-Verifier-Patcher* loop to audit and suppress hallucinations.
- Extensive experiments across a broad set of mainstream MLLM backbones demonstrate that MessToClean yields consistent improvements, substantially enhancing stem consistency, question-figure alignment, and refusal reliability under challenging real-world degradations, highlighting its practical potential.

2 Related Work

Structure-preserving reconstruction of exam images aims to recover complete question-centric text structures for intelligent education systems and LLM-based reasoning training, moving beyond fixed-schema field extraction. Early OCR-based methods (Hegghammer, 2022) with layout priors often fail under real-world degraded exam images (RDEI), struggling to distinguish printed text from handwriting. Later approaches introduced RNN (Aggarwal et al., 2020), CNN (Denk and Reisswig, 2019) and Transformer (Majumder et al., 2020; Wang et al., 2023) models to jointly predict layout and reading order, but remain vision-centric and are prone to errors under handwritten corrections or clutter. Recent methods serialize OCR results into text for downstream text-only LLMs (Wang et al., 2024; Do et al., 2025), enhancing semantics but losing layout cues and suffering from OCR noise. The latest paradigm, multimodal large language models (MLLMs) (Wu et al., 2023), integrates vision and language to directly generate structured outputs from document images, bridging pixel-to-text gaps. Therefore, mainstream models such as the Qwen3 series (Yang et al., 2025), GPT

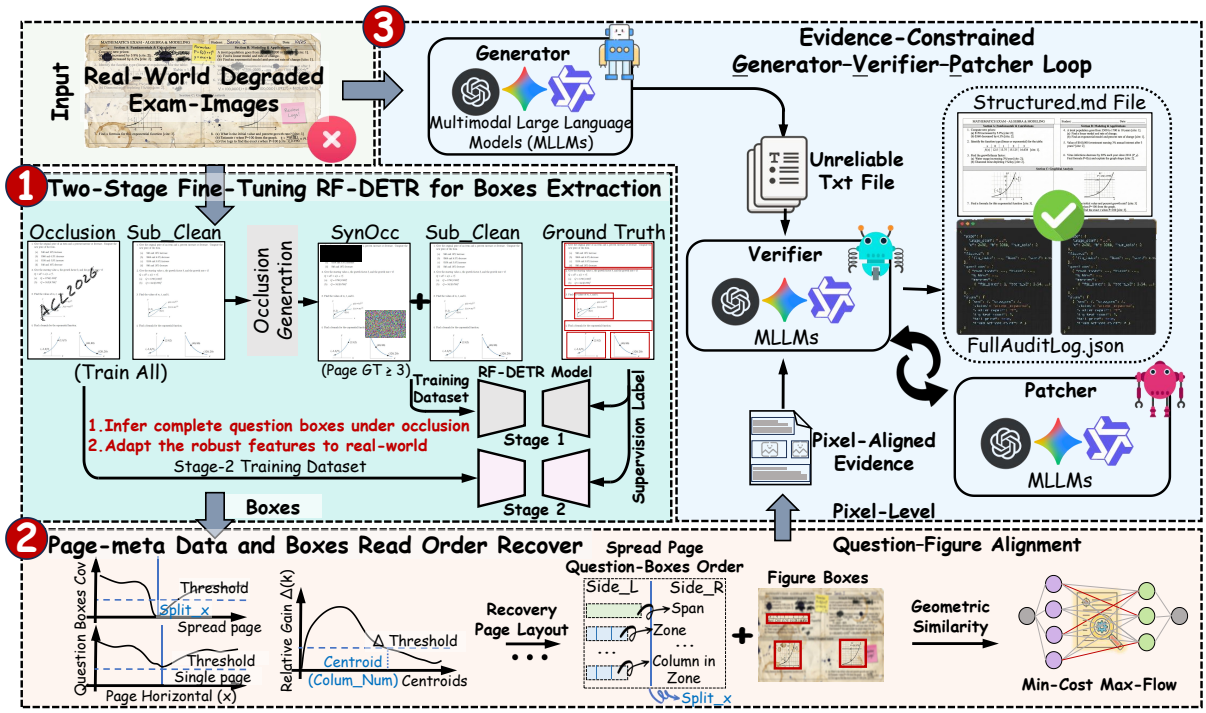

Figure 2: **Overview of MessToClean Pipeline.** ① Detect pixel-aligned text/figure boxes as visual evidence. ② Recover hierarchical layout and a stable global reading order from the box set, providing question-aligned, traceable structure priors for MLLMs. ③ Run an evidence-constrained *Generator-Verifier-Patcher (GVP)* loop to audit and minimally patch local degradations, suppressing evidence-unsupported hallucinations.

series (Sanderson, 2023; OpenAI, 2024, 2025), Gemini series (Comanici et al., 2025), and GLM-V series (Team et al., 2026) serve as our core baselines. However, under handwriting and other visual noise, prompt-driven generation often hallucinates outputs (He et al., 2025; Li et al., 2023b), while occlusions further exacerbate the unstable localization capabilities of general-purpose MLLMs (Tong et al., 2024), limiting effective layout grounding. Fine-tuning MLLMs for explicit layout parsing or font-specific recognition is prohibitively costly in terms of annotation and computation. We therefore incorporate dedicated object detectors to provide pixel-aligned structural evidence.

Object detection aims to predict object locations and categories simultaneously, typically outputting bounding boxes and class labels $(x_1, y_1, x_2, y_2, c_i)$. In the RDEI context, detectors are used to localize question blocks—categorized as Multiple-choice, Short-Answer, and Long-Answer—as well as figure regions, forming the structural backbone for downstream reading order recovery, question–figure alignment, and layout reconstruction. Mainstream detection paradigms include one-stage dense predictors (e.g., YOLO (Redmon et al., 2016), RetinaNet (Lin et al., 2017)) and set-based end-to-end models (e.g., DETR). The former produces numerous overlapping candidates and relies on heuristic post-processing, which degrades under handwriting noise, texture artifacts, or dense layouts (Li et al., 2020; Bolya et al., 2020; Zhang et al., 2020). In contrast, DETR frames detection as a set prediction problem with Hungarian matching, inherently suppressing duplicates (Robinson et al., 2025; Huang et al., 2025; Zhao et al., 2024; Jia et al., 2023; Meng et al., 2021; Carion et al., 2020), and reducing dependence on fragile thresholds (Sun et al., 2021), thus yielding more consistent structural inputs.

3 Proposed Methodology

In this section, we present MessToClean, an evidence-driven structured reconstruction pipeline that treats MLLMs as interchangeable components, constrains their behavior using pixel-aligned box evidence, and enhances structural consistency through an auditable patching loop. Fig. 2 provides an overview of the pipeline. Given a real-world degraded exam image as input, the system first (①) applies a two-stage fine-tuned RF-DETR to detect pixel-aligned bounding boxes for text regions and figure regions, effectively filtering out large-scale

noise and occlusions while reducing localization ambiguity. It then (②) recovers the page-level hierarchy based on these boxes, and infers a stable global reading order for question and text boxes. Finally (③), a *Generator-Verifier-Patcher* agent chain operates under the constraint of the ordered box evidence to further remove local degradations, producing a structure-preserving file.

3.1 Two-Stage Fine-Tuning RF-DETR for Evidence Extraction

The bounding-box set predicted by RF-DETR serves as pixel-level evidence for downstream `Reading Order` recovery and `Question–Figure Pairing`. However, directly using the vanilla RF-DETR, or training it with only a limited number of heavily occluded samples, often leads to severe fragmentation under common RDEI degradations (e.g., large handwritten scribbles and sticker occlusions), where a semantically complete region is split into multiple part-level boxes. Such fragmentation not only disrupts the semantic integrity of originally complete question boxes, but also compromises the reliability of subsequent page-level structure reconstruction.

Therefore, we carefully design a two-stage fine-tuning strategy to improve full-box localization under occlusion and to reduce noise propagation caused by fragmented detections. In Stage-1, we synthesize occlusions on a clean training subset and define the target visibility ratio $\texttt{vis} \in [0, 1]$ as:

$$\texttt{vis} = \frac{\texttt{Area(GT)} - \texttt{Area}(\cup_i(\texttt{GT} \cap \texttt{OCC}_i))}{\texttt{Area(GT)}}$$

where $\texttt{OCC}_i$ denotes the i-th occlusion region. Using two thresholds `low` and `high` (Appendix C.2), we define the visibility-based label mapping as follows: instance with $\texttt{vis} \leq \texttt{low}$ is discarded and if $\texttt{vis} \geq \texttt{high}$, we keep its base class (text or figure). Otherwise, we assign it to partial to explicitly mark *partially visible* hard cases.

A key design in Stage-1 is to regress to full GT boxes rather than visible parts. We retain full-box supervision even under occlusion. This encourages the detector to infer the complete spatial extent from partial observations, thereby suppressing fragmentations. In Stage-2, we fine-tune the detector on the full target-domain training set for effective synthetic-to-real adaptation. This enhances robustness to real-world degradation while maintaining localization consistency. In Exp. 4.3.3, we compare multiple training variants and show that the proposed two-stage fine-tuning yields a clear improvement in detection reliability on RDEI.

3.2 Hierarchical Layout and Read Order Reconstruction

Externalizing layout grounding as ordered, pixel-aligned evidence is **key to enabling MLLMs to reliably exploit document structure, thereby shifting localization uncertainty into an explicit, auditable evidence layer rather than model's implicit guess.** However, three factors make page-level hierarchy recovery and stable reading-order inference from box sets challenging. First, diverse layouts such as multi-column, cross-column, and nested structures violate simple top-to-bottom assumptions, causing naïve sorting to fail in recovering the true relative order. Second, limited fields of view often result in partial page captures rather than complete pages. When only part of one side of a double-column page is visible, the missing regions distort global geometry and interfere with positional ordering and column inference. Third, large occlusions and heavy noise corrupt geometric cues, rendering proximity- or alignment-based heuristics unreliable under severe degradation.

Therefore, we carefully design Stage ② to explicitly address these challenges by converting unordered box set from ① into a coherent, layout-aware representation with stable reading order.

Multi-column detection and gutter estimation. We infer whether a page is single- or two-column by estimating the gutter location `split_x` from a 1D x-axis coverage profile of the detected question boxes. We partition the page width $[0, W]$ into B vertical bins and compute the normalized coverage $\texttt{cov}(b) \in [0, 1]$ as the fraction of eligible boxes whose horizontal span intersects bin I_b. For two-column layouts, $\texttt{cov}(b)$ exhibits a clear valley near the gutter. We therefore search the minimum only within a central interval $S \subset [0, W]$ and apply a relative-saliency test. If the valley is sufficiently salient, we set `split_x` to the minimizer; otherwise we treat the page as single-column.

Zone partition and column recovery. We split the page into one or two sides using `split_x`. Within each side, wide cross-column text blocks are treated as *span barriers* that partition the vertical space into zones. For a zone that may contain multiple columns, we cluster the left-edge coordinates $X = \{x_i\}$ with K-means and choose the

column count via an elbow-style gain test:

$$\mathtt{SSE}(k) = \sum_i \min_{1 \le j \le k} (x_i - \mu_j^{(k)})^2$$

$$\Delta(k) = \frac{\mathtt{SSE}(k-1) - \mathtt{SSE}(k)}{\mathtt{SSE}(k-1)}$$

We increase k until $\Delta(k) \le \tau$ and adopt the resulting k. We then order boxes within each zone by a row-major ("Z") scan and concatenate zones and spans within a side to form an ordered sequence. For two-page spreads, we concatenate sides left-to-right to obtain $\mathtt{Questions}^{\mathtt{ordered}}$ and assign each question box a `read_index`.

Constrained Global Boxes Matching Nearest-neighbor matching based on $y_{\mathtt{mid}}$ or center distance can be brittle under complex layouts. We therefore perform candidate pruning, geometric scoring, and global assignment. For each figure box f, we define a local search window based on its page location and collect candidate question boxes q, forming candidate edges $\mathcal{E} = \{(q, f)\}$. For each candidate pair (q, f), let $A_q = \mathtt{Area}(q)$, $A_f = \mathtt{Area}(f)$, and $A_\cap = \mathtt{Area}(q \cap f)$. We define normalized overlaps

$$\mathtt{IoF}(q,f) = \frac{A_\cap}{\max(\epsilon, A_f)}, \mathtt{IoQ}(q,f) = \frac{A_\cap}{\max(\epsilon, A_q)}$$

where ϵ is a stability constant. Let (W, H) be the page width/height, and let $\mathtt{gap}_x(q,f), \mathtt{gap}_y(q,f)$ be non-negative axis gaps (0 if projections overlap). We compute

$$g_{x(y)} = \frac{\mathtt{gap}_{x(y)}(q,f)}{\max(1, W)}, c_d = \frac{2\|c_q - c_f\|_2}{\sqrt{W^2 + H^2}}$$

where c_q and c_f are box centers, and define

$$\mathtt{inside}(q,f) = \mathbb{I}(c_f \in q)$$

We decompose the score into a reward for reliable evidence and a penalty for weak geometric cues

$$s(q,f) = \boldsymbol{\lambda}^\top \phi^+(q,f) - \boldsymbol{\gamma}^\top \phi^-(q,f)$$

where $\phi^+(q,f) = [\mathtt{IoF}, \mathtt{IoQ}, \mathtt{inside}]$ and $\phi^-(q,f) = [g_x, g_y, c_d]$ are normalized features. $\boldsymbol{\lambda}, \boldsymbol{\gamma} \succ \mathbf{0}$ are fixed constants. We set these weights once on development set and keep them fixed for all experiments. The matching is insensitive to small perturbations within a reasonable range.

Finally, we formulate question–figure association as a capacity-constrained bipartite assignment. Each question can be linked to at most K figures, and each figure has capacity $\mathtt{cap}_{\mathtt{fig}}$. Using $\mathtt{s}_{(q,f)}$ as edge affinity, we solve for a page-level globally consistent matching, producing reading-order-consistent alignments and traceable evidence bindings I_i^q for downstream cropping and generate-verify processing.

3.3 Evidence-Constrained GVP Loop

After ②, we obtain question-level pixel-aligned evidence bindings $\{I_i^q\}$, where each I_i^q denotes the cropped visual evidence associated with question q_i (including its linked regions such as figures when applicable). These bindings provide grounded constraints for this stage. While large-scale degradations are largely removed upstream, residual errors are typically local and can be handled via evidence-constrained auditing and conservative edits.

We thus introduce an Evidence-Constrained GVP (Generator-Verifier-Patcher) Loop, which restricts generation and correction to question-level evidence bindings $\{I_i^q\}$ and forms an auditable closed loop. Generator produces a page-level candidate $M_{\mathtt{page}}$. Verifier checks each question against its cropped evidence I_i^q and outputs a decision v. Patcher is triggered only when needed to apply minimal, whitelist-constrained fixes followed by re-verification. Overall, the loop suppresses free-form completion under insufficient evidence and yields traceable, structure-consistent outputs.

4 Experimental Setup and Results

We systematically evaluate MessToClean on Real-World Degraded Exam Images (RDEI) across 12 state-of-the-art MLLM backbones. Sec. 4.1 details evaluation setup and protocol, and Sec. 4.2 reports the main results. Sec. 4.3.1 shows the gains are not attributable to extra model stacking or increased numbers of model calls, Sec. 4.3.2 studies the effect of capture angles, Sec. 4.3.3 compares alternative detector training recipes against our two-stage fine-tuned RF-DETR. The test set used in our experiments is drawn from the newly curated RDEI dataset, which was collected from our intelligent education platform and contains a total of 12,472 user-uploaded math exam images.

4.1 Experimental Setup

During evaluation, we benchmark MessToClean on **12** mainstream MLLM backbones spanning diverse

Model	Similarity		Hallucination		
	$\text{All}_{\text{ImgSim}} \uparrow$	$\text{All}_{\text{StemSim}} \uparrow$	$\text{All}_{\text{P}} \uparrow$	$\text{All}_{\text{R}} \uparrow$	$\text{All}_{\text{F1}} \uparrow$
GPT-4o-mini	7.25	28.29	22.48	45.91	30.19
GPT-4o-mini w/ MessToClean	**56.41**(+49.16)	**31.47**(+3.18)	**28.61**	**34.44**	**31.25**(+1.06)
GPT-4o	13.77	38.78	24.19	40.48	30.28
GPT-4o w/ MessToClean	**57.85**(+44.08)	**41.52**(+2.74)	**30.37**	**35.13**	**32.52**(+2.24)
GLM-4.6v-flashx	43.39	70.2	47.47	33.98	39.61
GLM-4.6v-flashx w/ MessToClean	**61.48**(+18.09)	**72.45**(+2.25)	**55.34**	**34.61**	**42.59**(+2.98)
Gemini-2.5-flash	51.59	81.16	47.03	39.42	42.89
Gemini-2.5-flash w/ MessToClean	**62.26**(+10.67)	**83.59**(+2.43)	**49.26**	**41.77**	**45.21**(+2.32)
GLM-4.6v-flash	46.97	76.88	53.99	31.01	39.39
GLM-4.6v-flash w/ MessToClean	**62.38**(+15.41)	**79.85**(+2.97)	**58.37**	**33.93**	**42.92**(+3.53)
Gemini-2.5-pro	53.87	83.65	51.94	53.6	52.76
Gemini-2.5-pro w/ MessToClean	**62.49**(+8.62)	**86.57**(+2.92)	**59.15**	**48.46**	**53.27**(+0.51)
GLM-4.5v	54.92	86.19	58.68	26.14	36.17
GLM-4.5v w/ MessToClean	**62.98**(+8.06)	**88.39**(+2.20)	**63.16**	**30.8**	**41.41**(+5.24)
Qwen3-VL-8B	51.75	83.9	51.09	24.13	32.78
Qwen3-VL-8B w/ MessToClean	**63.02**(+11.27)	**86.48**(+2.58)	**59.09**	**32.02**	**41.53**(+8.75)
Qwen3-VL-30B	60.78	86.05	58.05	19.64	29.35
Qwen3-VL-30B w/ MessToClean	**64.30**(+3.52)	**88.13**(+2.08)	**61.97**	**29.79**	**40.23**(+10.88)
GLM-4.6v	57.29	86.74	61.37	25.41	35.94
GLM-4.6v w/ MessToClean	**64.37**(+7.08)	**88.34**(+1.60)	**67.06**	**30.90**	**42.30**(+6.36)
GPT-5	64.89	89.29	62.93	21.19	31.7
GPT-5 w/ MessToClean	**65.39**(+0.50)	**90.30**(+1.01)	**68.05**	**28.52**	**40.20**(+8.50)
Qwen3-VL-235B	64.35	88.81	61.86	21.01	31.37
Qwen3-VL-235B w/ MessToClean	**65.40**(+1.05)	**90.52**(+1.71)	**69.12**	**28.86**	**40.72**(+9.35)

Table 1: Overall results on RDEI across 12 MLLM backbones. "All" denotes an equal-weight macro average over MC/SA/LA; ImgSim/StemSim assess figure binding and question-statement fidelity, and P/R/F1 measure question-level reconstruction accuracy.

scales and inference styles, including open-source models (Qwen3-VL: 8B/30B/235B; GLM-V: Flash 9B, 4.5V/4.6V 106B) and closed-source systems (GPT-5/4o/4o-mini; Gemini-2.5 Pro/Flash). We use a unified protocol without backbone-specific tuning or prompt customization. All experiments run on a single server with 2× Xeon 8488C CPUs and 8× A6000 GPUs.

4.1.1 Evaluation Metrics

Under a unified evaluation setup, we apply consistent metrics and aggregation rules across all MLLM backbones to quantify the quality of structure-preserving reconstruction. Reconstruction performance is assessed separately for each question type $t \in \{\texttt{MC}, \texttt{SA}, \texttt{LA}\}$ (*Multiple-choice*, *Short-Answer*, *Long-Answer*), and the overall average metric (`All`) is reported. The evaluation covers three dimensions to capture system capabilities:

Stem Consistency (StemSim) measures the character-level similarity between the predicted and ground-truth question stems. Let $\mathcal{D}_t$ denote the set of samples for question type t, with S_i^{gt} and S_i^{pred} representing the ground-truth and predicted stems of the i-th sample, respectively. Let $\phi(\cdot)$ be a text normalization function, and $\text{LevRatio}(\cdot,\cdot)$ denote the normalized Levenshtein similarity (higher is better). The final $\texttt{StemSim}_t$ is computed as the average over all samples of type t:

$$\frac{1}{|\mathcal{D}_t|} \sum_{i \in \mathcal{D}_t} \text{LevRatio}\Big(\phi\Big(S_i^{\text{gt}}\Big),\ \phi\Big(S_i^{\text{pred}}\Big)\Big) \in [0,1]$$

Image Consistency (ImgSim) measures the alignment quality between predicted and ground-truth figures, using the ground-truth figure sequence as reference. A similarity score $s(\cdot,\cdot) \in [0,1]$ is computed for each aligned pair, where s is a perceptual similarity function. Missing predictions are penalized with a score of 0. Questions with no ground-truth figures are marked as N/A and excluded from aggregation. Formally, if a question

Model	Setting	$\texttt{All}_{\texttt{StemSim}}$ ↑	$\texttt{All}_{\texttt{F1}}$ ↑
GLM-4.6v-flash	Whole	72.02	29.39
	Ours	**79.85 (+7.83)**	**42.92 (+13.53)**
Qwen3-VL-8B	Whole	80.85	26.88
	Ours	**86.48 (+5.63)**	**41.53 (+14.65)**
Qwen3-VL-30B	Whole	78.6	33.4
	Ours	**88.13 (+9.53)**	**40.23 (+6.83)**
GLM-4.6v	Whole	82.61	32.07
	Ours	**88.34 (+5.73)**	**42.30 (+10.23)**

Table 2: Evidence Granularity Matters: Aligned Crops vs. Whole-Page Inputs to ③.

has m ground-truth figures:

$$\text{ImgSim} = \frac{1}{m}\sum_{i=1}^{m} s(G_i, \hat{G}_i), \text{ with } s(G\_i, \varnothing) = 0$$

where G_i and $\hat{G}_i$ denote the i-th ground-truth and predicted figures, respectively.

Unrecognizable/Refusal Prediction (P/R/F1). Refusals are treated as the positive class, with precision, recall, and F1 computed according to standard definitions.

Details on text normalization $\phi(\cdot)$, refusal labeling, and N/A handling in All are provided in Appendix E.

4.2 Main Results

Table 1 reports the overall (*All*) results of 12 MLLM backbones (in %). Two key findings emerge. First, **consistent gains across backbones:** our method improves $\texttt{All}_{\texttt{F1}}$ for every backbone, increasing the average from 36.04% to 41.18% with a maximum gain of 10.88%, and also boosts $\texttt{All}_{\texttt{StemSim}}$ by 2.31% on average. **These improvements are not tied to a specific model family, indicating robustness across different inference styles.** The gains are mainly driven by a systematic increase in precision ($\texttt{All}_{\texttt{P}}$ +5.71% on average), suggesting that structured evidence and consistency constraints reduce unreliable free-form completion. Fine-grained results are provided in Appendix A.1. Second, **enabling small models to outperform larger ones:** with MessToClean, the 8B Qwen3-VL model surpasses the 30B variant, and is competitive with the 106B GLM-4.6v in stem consistency, figure consistency, and abstention/rejection decisions. **This indicates that the improvements primarily stem from MessToClean rather than backbone scale alone.**

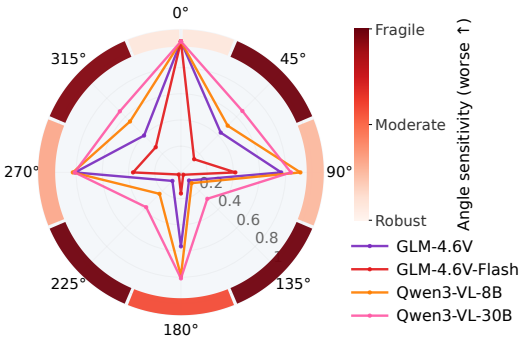

Figure 3: Polar plots of $\texttt{All}_{\texttt{StemSim}}$ under different input orientations. Lighter colors indicate higher $\texttt{All}_{\texttt{StemSim}}$.

4.3 More Analyses

4.3.1 Block-Level Evidence Alignment Drives Stable Gains

This experiment verifies that the gains of `Ours` **stem from question-level pixel evidence and the evidence-constrained pipeline itself, rather than from additional model calls or differences in chain structure.** To this end, we construct a `chain-matched WholePage` baseline that differs from `Ours` only in the granularity of visual evidence: when processing each question, *WholePage* reuses the same full-page image as evidence and does not provide question-level crops or alignment. We evaluate this setting on the full test set.

As shown in Table 2, even with the same backbone and an identical agent chain, replacing question-level evidence with full-page evidence yields consistent degradation: `WholePage` reduces $\texttt{All}_{\texttt{StemSim}}$ by 5.63-9.53% (7.18% on average) and $\texttt{All}_{\texttt{F1}}$ by 6.83-14.65% (11.31% on average) relative to `Ours`. These results indicate that question-level evidence alignment is critical for stable gains; when evidence granularity is insufficient, the same auditing-and-patching chain cannot reliably maintain stem consistency (`StemSim`) or overall quality (`F1`) across backbones.

4.3.2 Effect of Rotation Angles in RDEI on Model Performance

In our experiments, **we find relying solely on the geometric generalization of MLLM backbones is insufficient to stably recover question structure and content under oblique captures.** To quantify the impact of geometric degradations in Real-World Degraded Exam Images (RDEI), we apply controlled rotations to the input while keeping all other settings identical: $\theta \in \{0°, 45°, \ldots, 315°\}$ with a 45° step, and report per-

Exp	Stage1 Train Set	Stage2 Train Set	$\texttt{RDEI}_{\texttt{Target}}$ (%)	
			$\texttt{mAP}_{\texttt{base}}$	$\texttt{AP}_{\texttt{Partial}}$
A	-	$\texttt{All}_{\texttt{Train}}$	71.1	61
B	$\texttt{Sub}_{\texttt{Train}}$	-	73.5	N/A
C	$\texttt{Sub}_{\texttt{Train}}$	$\texttt{All}_{\texttt{Train}}$	79.1	60.7
D	$\texttt{Sub}_{\texttt{Train}}$+Syn	-	73.7	38.1
E	**$\texttt{Sub}_{\texttt{Train}}$+Syn**	**$\texttt{All}_{\texttt{Train}}$**	**79.3**	**67.8**

Table 3: Ablation of Training Recipes in Two-Stage Occlusion-Aware Fine-tuning.

formance at each angle. Figure 3 shows the polar profiles of $\texttt{All}_{\texttt{StemSim}}$ for four representative backbones. Note that, for fair comparison, we apply the same rotation-normalization module to all backbones throughout our experiments, and this study is solely intended to probe the effect of rotation-induced geometric changes.

We observe angular anisotropy: models remain relatively stable under orthogonal rotations ($0°/90°/180°/270°$) but degrade systematically under oblique rotations ($45°/135°/225°/315°$). Across backbones (Table A.4), `StemSim` averages 65.7% at orthogonal angles versus 26.6% at oblique angles, with the worst case at $135°$ (11.3%). This pattern is consistent for all backbones, indicating a geometric sensitivity in the RDEI setting. As rotation alters geometry without changing semantics, the drop in $\texttt{All}_{\texttt{StemSim}}$ highlights limitations of backbones in geometric alignment and reading-order stability (see Appendix A.2 for details).

4.3.3 Effect of 2-Stage Fine-Tune

As stated in Sec. 3.1, we adopt a two-stage fine-tuning scheme: Stage 1 trains the detector to recover complete text boxes under occlusion, while Stage 2 calibrates it on real-domain data. In Table 3, we quantitatively analyze the effects of this two-stage recipe and the Stage 1 `Syn` augmentation on the occlusion class `Partial`, which refers to regions that cannot be reliably classified as text or figures due to occlusion or defects. Here, `RDEI-Target (All)` denotes the full test set. We report two metrics: $\texttt{mAP}_{\texttt{base}}$@50:95 (COCO-style mean Average Precision computed over the base classes, namely text and figure, averaged across IoU thresholds from 0.50 to 0.95), and $\texttt{AP}_{\texttt{Partial}}$@50:95 for the occlusion class. We denote them as $\texttt{mAP}_{\texttt{base}}$ and $\texttt{AP}_{\texttt{Part}}$ for brevity.

Table 3 yields three key findings. (1) Compared to single-stage baseline trained on `Train All` (Exp. A), the full two-stage recipe (Exp. E) improves `mAPbase`@50:95 by 8.2% and $\texttt{AP}_{\texttt{Partial}}$@50:95 by 6.8% on `RDEI-Target (All)`, indicating notable gains under target-domain degradations. (2) Adding Stage 1 `Syn` (Exp. C *to* Exp. E) has negligible impact on base detection (only 0.2-0.4% change in $\texttt{mAP}_{\texttt{base}}$), yet boosts $\texttt{AP}_{\texttt{Partial}}$@50:95 by 7.1%, suggesting the gains primarily come from learning heavy-noise or occlusion boxes rather than a general improvement on base classes. (3) Performing Stage 1 without Stage 2 (Exp. D) results in markedly lower $\texttt{AP}_{\texttt{Partial}}$@50:95; introducing Stage 2 (Exp. D *to* Exp. E) increases $\texttt{AP}_{\texttt{Partial}}$@50:95 by 29.7%, while also improving $\texttt{mAP}_{\texttt{base}}$@50:95 by 5.6%, empirically confirming the importance of Stage 2 calibration for real-world noise.

5 Conclusion

We target Real-World Degraded Exam Images (RDEI), which are pervasive in practical educational deployments, and formalize the task of `structure-preserving reconstruction`. To address text contamination and evidence-unsupported hallucinations caused by unstable localization in off-the-shelf MLLM systems on RDEI, we propose `MessToClean`. `MessToClean` uses a two-stage RF-DETR to provide pixel-aligned evidence, jointly reconstructs hierarchical layout and reading order to enable page-level consistent question-figure binding, and employs an evidence-constrained `GVP` auditing-and-patching loop to perform minimal, interpretable fixes, producing structured Markdown and auditable JSONL logs. Extensive experiments show that `MessToClean` significantly improves stem consistency, figure consistency, and refusal reliability, achieving state-of-the-art performance.

6 Acknowledgements

The authors would like to acknowledge the support of the Public Computing Cloud of Renmin University of China and the fund for building world-class universities (disciplines) of Renmin University of China.

7 Limitations

Since our data come from real-world user-uploaded exam paper images, some raw samples may contain sensitive information such as handwritten names or student identifiers. However, our work does

not aim to extract such information, and the reconstructed outputs are not designed to preserve identity-related content. As a result, privacy risk is limited in our task setting, though appropriate data protection measures are still important when handling the raw images. Beyond these privacy considerations, we acknowledge the limitations of `MessToClean` to encourage further progress in this direction. While our system achieves strong *structure-preserving reconstruction* of degraded exam sheets, it does not fully leverage the *handwritten* solution traces in these images. Enabling reliable extraction and structuring of handwritten reasoning could support large-scale, automated construction of step-by-step math datasets for LLMs. We plan to pursue this direction by integrating robust handwriting understanding into the reconstruction pipeline.

References

Milan Aggarwal, Hiresh Gupta, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. Form2Seq : A framework for higher-order form structure extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3830–3840, Online. Association for Computational Linguistics.

Utathya Aich, Shinjini Chakraborty, Deepan Sadhukhan, Swarnendu Ghosh, and Tulika Saha. 2026. Hilex: Image-based hierarchical layout extraction from question papers. In *Document Analysis and Recognition – ICDAR 2025*, pages 485–505, Cham. Springer Nature Switzerland.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. 2020. Tide: A general toolbox for identifying object detection errors. In *Computer Vision – ECCV 2020*, pages 558–573, Cham. Springer International Publishing.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham. Springer International Publishing.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, and 10 others. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Sagnik Das, Kunwar Yashraj Singh, Jon Wu, Erhan Bas, Vijay Mahadevan, Rahul Bhotika, and Dimitris Samaras. 2021. End-to-end piece-wise unwarping of document images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4268–4277.

Timo I. Denk and Christian Reisswig. 2019. {BERT}grid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. 2024. SciEx: Benchmarking large language models on scientific exams with human expert grading and automatic grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610, Miami, Florida, USA. Association for Computational Linguistics.

Thao Do, Dinh Phu Tran, An Vo, and Daeyoung Kim. 2025. Reference-based post-ocr processing with llm for precise diacritic text in historical document recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27951–27959.

Zhentao He, Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, and Minghui Qiu. 2025. Seeing is believing? mitigating OCR hallucinations in multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Thomas Hegghammer. 2022. Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, 5(1):861–882.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834, Vienna, Austria. Association for Computational Linguistics.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. 2025. Deim: Detr

with improved matching for fast convergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15162–15171.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. 2023. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19702–19712.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023a. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21002–21012. Curran Associates, Inc.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, Shanshan Jiang, Bin Dong, and Le Sun. 2025. READoc: A unified benchmark for realistic document structured extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21889–21905, Vienna, Austria. Association for Computational Linguistics.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. 2025. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4038–4049.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6504, Online. Association for Computational Linguistics.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3651–3660.

OpenAI. 2024. Hello gpt-4o. OpenAI Blog. Accessed: 2026-01-05.

OpenAI. 2025. Gpt-5 is here. OpenAI Product Page. Accessed: 2026-01-05.

Pritika Ramu, Sijia Wang, Lalla Mouatadid, Joy Rimchala, and Lifu Huang. 2024. re^2: Region-aware relation extraction from visually rich documents. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8731–8747, Mexico City, Mexico. Association for Computational Linguistics.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. 2025. Rf-detr: Neural architecture search for real-time detection transformers. *Preprint*, arXiv:2511.09554.

Katharine Sanderson. 2023. Gpt-4 is here: what scientists think. *Nature*, 615(7954):773.

Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, and Ping Luo. 2021. What makes for end-to-end object detection? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9934–9944. PMLR.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Zifeng Wang, Zizhao Zhang, Jacob Devlin, Chen-Yu Lee, Guolong Su, Hao Zhang, Jennifer Dy, Vincent Perot, and Tomas Pfister. 2023. QueryForm: A simple zero-shot form entity query framework. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4146–4159, Toronto, Canada. Association for Computational Linguistics.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zining Wang, Tongkun Guan, Pei Fu, Chen Duan, Qianyi Jiang, Zhentao Guo, Shan Guo, Junfeng Luo, Wei Shen, and Xiaokang Yang. 2025. Marten: Visual question answering with mask generation for multi-modal document understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14460–14471.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

Hangdi Xing, Feiyu Gao, Qi Zheng, Zhaoqing Zhu, Zirui Shao, and Ming Yan. 2025. Intelligent document parsing: Towards end-to-end document parsing via decoupled content parsing and layout grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19987–19998.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. 2025a. Mitigating hallucinations in multimodal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1571–1590, Albuquerque, New Mexico. Association for Computational Linguistics.

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2025b. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17675–17687.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. 2025. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14358–14368.

Wenhao Yu, Tianrui Zong, He Zhang, Lin Yang, He Yang, Guanhao Wu, Ruohua Xu, Qinqin Yan, and Liangcai Gao. 2026. Icdar 2025 competition on understanding chinese college entrance exam papers. In *Document Analysis and Recognition – ICDAR 2025*, pages 523–536, Cham. Springer Nature Switzerland.

Chong Zhang, Yi Tu, Yixi Zhao, Chenshu Yuan, Huan Chen, Yue Zhang, Mingxu Chai, Ya Guo, Huijia Zhu, Qi Zhang, and Tao Gui. 2024a. Modeling layout reading order as ordering relations for visually-rich document understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9658–9678, Miami, Florida, USA. Association for Computational Linguistics.

Dacao Zhang, Kun Zhang, Le Wu, Mi Tian, Richang Hong, and Meng Wang. 2024b. Path-specific causal reasoning for fairness-aware cognitive diagnosis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4143–4154, New York, NY, USA. Association for Computing Machinery.

Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, and Lianwen Jin. 2024c. Docres: A generalist model toward unifying document image restoration tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15654–15664.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16965–16974.

Appendices A–F provide detailed supplementary information for the experiments:

- **Appendix A** reports type-wise results across question types (MC, SA, LA) and MLLM backbones, including ImgSim, StemSim, Precision, Recall, and F1.
- **Appendix B** summarizes datasets, splits, and descriptive statistics, covering a clean 3-class subset and synthetic noise augmentation.
- **Appendix C** documents the synthetic occlusion procedure and the label-mapping strategies for 3-class vs. 4-class training.
- **Appendix D** outlines the fine-tuning setup and initialization details for the RF-DETR-Medium detector.
- **Appendix E** explains the computation of evaluation metrics (e.g., StemSim and ImgSim), with an emphasis on the precision/recall decomposition and the overall macro-averaged F1.
- **Appendix F** provides implementation details on spread detection, zone decomposition, geometric terms, and the min-cost flow construction for figure–text alignment (For Sec. 3.2).

A Detailed Results

A.1 Type-wise results (MC/SA/LA)

This section reports type-wise results that complement the overall (All) scores in Table 1. We evaluate three question types: multiple-choice (MC), short-answer (SA), and long-answer (LA). For each MLLM backbone, we compare the *Direct* (one-shot) baseline with `MessToClean` (ours), and report type-specific metrics: ImgSim ↑, StemSim ↑, and hallucination-related P ↑, R ↑, $F1$ ↑. All numbers are percentages (%). For ImgSim/StemSim/$F1$, we additionally report Δ in parentheses, defined as `MessToClean`−Direct (pp); we omit Δ for P and R. Detailed results for MC, SA, and LA are provided in Tables A.1–A.3, respectively.

A.2 Rotation robustness under controlled in-plane rotations

Rotation sensitivity is strongly angle-dependent: oblique rotations (e.g., 135°/225°) cause the largest degradation, whereas 90°/270° are comparatively less harmful on average (Table A.4). We rotate each test page by $\{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$ and report per-backbone $\mathrm{All}_{\mathrm{StemSim}}$ for each angle; the last row shows the mean across backbones.

A.3 Precision/Recall decomposition for key ablations

We report All_P and All_R alongside $\mathrm{All}_{\mathrm{StemSim}}$. Tables A.5–A.6 should be read by comparing *Ours* against the ablated setting within each backbone.

A.3.1 No-Patcher: detailed breakdown

Table A.5 reports *Ours* vs. *w/o Patcher* for each backbone.

A.3.2 WholePage (chain-matched) vs. Ours: detailed breakdown

Table A.6 compares *Ours* with *WholePage* (chain-matched). Across backbones, WholePage (chain-matched) consistently reduces $\mathrm{All}_{\mathrm{StemSim}}$ and $\mathrm{All}_{\mathrm{F1}}$, typically accompanied by *simultaneous* drops in All_P and All_R. This supports that full-page evidence provides weaker region-level grounding than question-level crops, leading to both more mismatches and more missing content, and thus lower overall reconstruction quality.

B Datasets, Splits, and Statistics

This appendix summarizes the datasets, split protocols, and descriptive statistics used in this work. Unless otherwise stated, all random splits use a fixed seed (**seed=42**) and the train/valid/test partitions are mutually exclusive. We report both page-level counts (Pages) and instance-level counts (Anns); all class ratios (e.g., MC%) are computed over **instance counts** within each split.

B.1 Clean 3-Class Subset (No Occlusion)

To enable controlled ablations and avoid distributional confounding introduced by occlusions, we derive an occlusion-free three-class subset from the source-domain data, retaining only Multiple-choice (MC), Short-answer (SA), and Long-answer (LA). We additionally verify that *Partial* annotations are absent in all three splits (Table B.1).

B.2 Synthetic Noise Augmentation (Syn)

To simulate severe interferences commonly observed in real exam sheets—such as sticky-note occlusions and heavy smearing—we apply Synthetic Noise Augmentation (Syn) to the training set

Model	Similarity		Hallucination		
	MC_{ImgSim} ↑	$MC_{StemSim}$ ↑	MC_{P} ↑	MC_{R} ↑	MC_{F1} ↑
GPT-4o-mini	11.57	26.99	22.68	45.70	30.32
GPT-4o-mini w/ MessToClean	**56.00(+44.43)**	**29.55(+2.56)**	**30.47**	**33.31**	**31.83(+1.51)**
GPT-4o	11.23	37.64	19.52	36.23	25.37
GPT-4o w/ MessToClean	**57.21(+45.98)**	**39.49(+1.85)**	**26.11**	**33.70**	**29.42(+4.05)**
GLM-4.6v-flashx	42.17	72.02	43.15	35.15	38.74
GLM-4.6v-flashx w/ MessToClean	**58.35(+16.18)**	**74.48(+2.46)**	**55.19**	**33.48**	**41.68(+2.94)**
Gemini-2.5-flash	50.61	81.81	45.05	40.93	42.89
Gemini-2.5-flash w/ MessToClean	**59.70(+9.09)**	**83.00(+1.19)**	**49.73**	**43.11**	**46.18(+3.29)**
GLM-4.6v-flash	46.02	76.99	52.13	28.13	36.54
GLM-4.6v-flash w/ MessToClean	**58.81(+12.79)**	**80.73(+3.74)**	**55.53**	**33.83**	**42.05(+5.51)**
Gemini-2.5-pro	51.87	83.98	51.41	50.57	50.99
Gemini-2.5-pro w/ MessToClean	**58.73(+6.86)**	**86.22(+2.24)**	**57.71**	**48.01**	**52.42(+1.43)**
GLM-4.5v	50.35	85.80	56.29	21.74	31.37
GLM-4.5v w/ MessToClean	**58.23(+7.88)**	**88.12(+2.32)**	**61.05**	**31.00**	**41.12(+9.75)**
Qwen3-VL-8B	44.43	83.98	49.47	23.41	31.78
Qwen3-VL-8B w/ MessToClean	**58.09(+13.66)**	**86.06(+2.08)**	**55.92**	**33.67**	**42.03(+10.25)**
Qwen3-VL-30B	58.52	85.83	55.59	19.76	29.16
Qwen3-VL-30B w/ MessToClean	**59.53(+1.01)**	**88.99(+3.16)**	**60.28**	**31.07**	**41.00(+11.84)**
GLM-4.6v	52.61	87.44	60.11	23.42	33.71
GLM-4.6v w/ MessToClean	**59.92(+7.31)**	**88.41(+0.97)**	**65.16**	**30.77**	**41.80(+8.09)**
GPT-5	59.44	90.22	62.37	21.45	31.92
GPT-5 w/ MessToClean	**60.88(+1.44)**	**91.50(+1.28)**	**68.41**	**30.32**	**42.02(+10.10)**
Qwen3-VL-235B	59.03	90.62	60.80	22.96	33.33
Qwen3-VL-235B w/ MessToClean	**60.79(+1.76)**	**91.81(+1.19)**	**68.63**	**32.23**	**43.86(+10.53)**

Table A.1: MC results across 12 MLLM backbones (%).

only, following the protocol below to prevent data leakage:

(i) Only training images and their VOC-format annotations are used; the augmentation process never accesses validation or test sets.

(ii) Augmented variants are generated only for training pages that contain at least three ground-truth bounding boxes ($bbox \geq 3$). Pages with $bbox < 3$ are skipped.

(iii) For each eligible page, we generate at most one augmented full-page image. Keeping the augmented image alongside its clean counterpart depends on the training recipe (Sec. B.3).

Under the split with seed = 42, the training set contains 5,572 eligible pages with $bbox \geq 3$, among which 5,245 full-page augmented samples are successfully generated (coverage: 94.1%). Remaining attempts are discarded due to missing valid occluders satisfying sampling constraints.

B.3 Stage-1 training recipe statistics (Ablation Control)

In the Stage-1 ablation on the clean three-class subset (Sec. B.1), we vary only the training-data recipe while keeping all other training settings unchanged (e.g., model architecture, optimizer, maximum epochs, and early-stopping policy). Table B.2 summarizes the resulting training-set composition under different recipes. The key differences across recipes are:

(i) Whether the **Syn** variant is included as an additional training sample alongside the original clean page;

(ii) For pages with bbox ≥ 3, whether both the clean and Syn versions are retained, or only the Syn version is kept.

B.4 RDEI-Target (All): 4-Class splits and statistics

The target-domain dataset **RDEI-Target (All)** adopts a unified 4-class taxonomy: *MC*, *SA*, *LA*,

Model	Similarity		Hallucination		
	SA_{ImgSim} ↑	$SA_{StemSim}$ ↑	SA_{P} ↑	SA_{R} ↑	SA_{F1} ↑
GPT-4o-mini	5.40	24.53	18.78	51.30	27.49
GPT-4o-mini w/ MessToClean	**56.22(+50.82)**	**26.14(+1.61)**	**24.22**	**32.98**	**27.93(+0.44)**
GPT-4o	10.78	33.80	20.13	42.55	27.33
GPT-4o w/ MessToClean	**57.36(+46.58)**	**35.55(+1.75)**	**26.10**	**31.73**	**28.64(+1.31)**
Qwen3-VL-8B	49.61	82.65	45.11	20.27	27.97
Qwen3-VL-8B w/ MessToClean	**61.01(+11.40)**	**84.91(+2.26)**	**51.26**	**28.56**	**36.68(+8.71)**
GLM-4.5v	51.91	86.06	54.52	30.86	39.41
GLM-4.5v w/ MessToClean	**61.70(+9.79)**	**89.29(+3.23)**	**60.52**	**31.27**	**41.23(+1.82)**
GLM-4.6v-flashx	44.57	71.64	44.10	32.00	37.09
GLM-4.6v-flashx w/ MessToClean	**59.94(+15.37)**	**74.37(+2.73)**	**51.55**	**31.57**	**39.16(+2.07)**
Gemini-2.5-pro	51.51	83.29	43.11	53.48	47.74
Gemini-2.5-pro w/ MessToClean	**60.74(+9.23)**	**87.27(+3.98)**	**50.74**	**46.32**	**48.43(+0.69)**
GLM-4.6v-flash	45.91	78.21	50.41	30.73	38.18
GLM-4.6v-flash w/ MessToClean	**60.89(+14.98)**	**81.71(+3.50)**	**56.16**	**31.11**	**40.04(+1.86)**
Qwen3-VL-30B	55.92	84.42	50.77	15.76	24.05
Qwen3-VL-30B w/ MessToClean	**62.23(+6.31)**	**86.09(+1.67)**	**54.25**	**26.56**	**35.66(+11.61)**
Gemini-2.5-flash	48.03	78.89	32.65	29.36	30.92
Gemini-2.5-flash w/ MessToClean	**59.77(+11.74)**	**82.19(+3.30)**	**32.65**	**32.21**	**32.43(+1.51)**
GLM-4.6v	53.62	85.59	56.90	28.68	38.14
GLM-4.6v w/ MessToClean	**62.11(+8.49)**	**88.07(+2.48)**	**62.35**	**31.04**	**41.45(+3.31)**
Qwen3-VL-235B	61.51	87.09	53.92	18.43	27.47
Qwen3-VL-235B w/ MessToClean	**62.97(+1.46)**	**89.66(+2.57)**	**60.81**	**29.09**	**39.35(+11.88)**
GPT-5	62.22	88.64	54.00	18.01	27.01
GPT-5 w/ MessToClean	**62.70(+0.48)**	**89.25(+0.61)**	**60.04**	**29.14**	**39.24(+12.23)**

Table A.2: SA results across 12 MLLM backbones (%).

and *Partial* (class ids = 0/1/2/3). This split is used for training and main evaluation unless otherwise specified (Table B.3).

B.5 RDEI-Target (Partial+): a hard-distribution test subset

To characterize performance on a harder distribution where *Partial* instances are present, we construct **Partial+** by filtering the **test** split of **RDEI-Target (All)** to pages that contain at least one Partial ground-truth box. Statistics of this filtered subset are reported in Table B.4, which avoids diluting the analysis when Partial instances are relatively rare in the full test set. Importantly, **Partial+** is a *filtered view* of the same test split and is used only for supplementary analysis; it is *not* used for training, early stopping, or model selection.

C Augmentation Strategies and Label Mapping

This appendix describes (i) the synthetic occlusion procedure used to augment training pages and (ii) the visibility-aware label mapping rules applied to align supervision under 3-class vs. 4-class training.

C.1 Synthetic Occlusion

We synthesize occlusions to mimic strong local corruption commonly observed in real exam sheets (e.g., sticky notes and heavy smearing). The augmentation protocol is as follows: **Inputs.** We apply augmentation only to the training split, using training images and their VOC-format ground-truth bounding boxes under the 3-class taxonomy (MC/SA/LA). The pipeline never accesses the validation or test sets, preventing data leakage.

Outputs. For each eligible training page, we generate one occlusion-augmented page image. Whether the augmented image is retained alongside the clean page is determined by the Stage-1 training recipe (Sec. B.3).

Number of occluders. The number of occlusion blocks per augmented page is sampled as $num\_occ \in \{1, 2, 3\}$.

Occlusion patterns. To approximate realistic cor-

Model	Similarity		Hallucination		
	$\mathrm{LA}_{\mathrm{ImgSim}} \uparrow$	$\mathrm{LA}_{\mathrm{StemSim}} \uparrow$	$\mathrm{LA}_{\mathrm{P}} \uparrow$	$\mathrm{LA}_{\mathrm{R}} \uparrow$	$\mathrm{LA}_{\mathrm{F1}} \uparrow$
GPT-4o-mini	4.78	33.35	25.99	40.74	31.73
GPT-4o-mini w/ MessToClean	**57.00(+52.22)**	**38.71(+5.36)**	**31.13**	**37.02**	**33.82(+2.09)**
GPT-4o	19.29	44.91	32.92	42.67	37.17
GPT-4o w/ MessToClean	**58.98(+39.69)**	**49.52(+4.61)**	**38.91**	**39.57**	**39.24(+2.07)**
Gemini-2.5-flash	56.13	82.78	63.39	47.97	54.61
Gemini-2.5-flash w/ MessToClean	**67.30(+11.17)**	**85.59(+2.81)**	**65.39**	**50.00**	**56.67(+2.06)**
Gemini-2.5-pro	58.22	83.67	61.31	56.76	58.95
Gemini-2.5-pro w/ MessToClean	**68.00(+9.78)**	**86.21(+2.54)**	**68.99**	**51.04**	**58.67(-0.28)**
GLM-4.6v	65.63	87.20	67.11	24.14	35.51
GLM-4.6v w/ MessToClean	**71.09(+5.46)**	**88.55(+1.35)**	**73.68**	**30.88**	**43.52(+8.01)**
GLM-4.6v-flashx	43.44	66.93	55.17	34.80	42.68
GLM-4.6v-flashx w/ MessToClean	**66.16(+22.72)**	**68.50(+1.57)**	**59.29**	**38.79**	**46.90(+4.22)**
GLM-4.6v-flash	48.99	75.44	59.42	34.17	43.39
GLM-4.6v-flash w/ MessToClean	**67.45(+18.46)**	**77.10(+1.66)**	**63.41**	**36.86**	**46.62(+3.23)**
GLM-4.5v	62.49	86.71	65.22	25.82	36.99
GLM-4.5v w/ MessToClean	**69.00(+6.51)**	**87.77(+1.06)**	**67.91**	**30.13**	**41.74(+4.75)**
Qwen3-VL-8B	61.22	85.07	58.69	28.72	38.57
Qwen3-VL-8B w/ MessToClean	**69.95(+8.73)**	**88.47(+3.40)**	**70.10**	**33.82**	**45.63(+7.06)**
Qwen3-VL-30B	67.90	87.91	67.79	23.39	34.78
Qwen3-VL-30B w/ MessToClean	**71.14(+3.24)**	**89.30(+1.39)**	**71.38**	**31.73**	**43.93(+9.15)**
Qwen3-VL-235B	72.51	88.72	70.87	21.65	33.17
Qwen3-VL-235B w/ MessToClean	**72.45(-0.06)**	**90.09(+1.37)**	**77.93**	**25.26**	**38.15(+4.98)**
GPT-5	73.01	89.01	72.42	24.10	36.16
GPT-5 w/ MessToClean	**72.58(-0.43)**	**90.14(+1.13)**	**75.70**	**26.11**	**38.83(+2.67)**

Table A.3: LA results across 12 MLLM backbones (%).

ruption, we occlude a single question region in most cases (90%), and additionally sample small cross-question blocks (10%) to model incidental spill-over occlusions.

Reproducibility. The augmentation pipeline is deterministic given the fixed seed (seed=42).

C.2 Visibility-based Label Mapping

Let **vis** $\in [0, 1]$ denote the post-occlusion visibility ratio of an instance. Using two thresholds **low** $= 0.2$ and **high** $= 0.8$, we define label-mapping rules for two training settings: **3-class training** (without Partial) and **4-class training** (with an explicit Partial class). In both cases, instances with extremely low visibility are removed to avoid unreliable supervision.

(i) **2S Mapping (3-class training).**

- If vis $\leq$ low, drop the annotation.
- If vis $>$ low, keep the original base-class label (id $\in \{0, 1, 2\}$).

(ii) **3S Mapping (4-class training with Partial).**

- If vis $\geq$ high, keep the original base-class label (id $\in \{0, 1, 2\}$).
- If vis $\leq$ low, drop the annotation.
- If low $<$ vis $<$ high, map the instance to the Partial class (id = 3).

D Implementation Details

This appendix provides the key implementation details used in our experiments.

D.1 Fine-tuning Setup

We fine-tune an **RF-DETR-Medium** detector with the following configuration:

- **Batch size:** 4
- **Gradient accumulation:** 4 steps (effective batch size = 16)
- **Learning rate:** 1×10^{-4}
- **Max epochs:** 70

Backbone	0°	45°	90°	135°	180°	225°	270°	315°
GLM-4.6v	89.2	38.3	68.3	7.8	50.3	8.2	72.2	35.5
GLM-4.6v-flash	82.5	11.7	34.2	2.1	13.3	1.9	30.2	22.4
Qwen3-VL-30B	90.2	59.8	75.7	25.7	73.0	33.9	73.2	59.5
Qwen3-VL-8B	84.7	42.7	77.2	9.8	66.9	19.6	69.9	46.6
Mean (over backbones)	86.6	38.1	63.9	11.3	50.9	15.9	61.4	41.0

Table A.4: Per-angle $\mathrm{All}_{\mathrm{StemSim}}$ (%) under in-plane rotations.

Backbone	Setting	$\mathrm{All}_{\mathrm{StemSim}}\uparrow$	$\mathrm{All}_{P}\uparrow$	$\mathrm{All}_{R}\uparrow$	$\mathrm{All}_{\mathrm{F1}}\uparrow$
GLM-4.6v	Ours	88.34	67.06	30.90	42.30
GLM-4.6v	w/o Patcher	80.30	32.35	48.88	38.93
GLM-4.6v-flash	Ours	79.85	58.37	33.93	42.92
GLM-4.6v-flash	w/o Patcher	68.93	31.72	51.87	39.37
Qwen3-VL-8B	Ours	86.48	59.09	32.02	41.53
Qwen3-VL-8B	w/o Patcher	77.23	36.07	39.94	37.91
Qwen3-VL-30B	Ours	88.13	61.97	29.79	40.23
Qwen3-VL-30B	w/o Patcher	79.56	30.08	43.28	35.49

Table A.5: No-Patcher ablation (%).

- **Early stopping and checkpoint selection:** Early stopping is enabled based on validation performance, and the best checkpoint on the validation set is selected for final evaluation.

D.2 Initialization

We initialize the detector from the official **RF-DETR COCO-pretrained weights**, which improves convergence speed and training stability.

E Metric Computation Details

This appendix describes how we compute the metrics used to evaluate (i) structured extraction quality and (ii) refusal reliability when visual evidence is insufficient. Unless otherwise noted, we compute all metrics per question type $t \in \{\mathrm{MC}, \mathrm{SA}, \mathrm{LA}\}$ and report All scores via equal-weight macro averaging over the three types. For F1, we do not average $F1_t$ directly; instead, $\mathrm{All}_{\mathrm{F1}}$ is computed from macro-averaged precision/recall as described in Sec. E.4.

Similarity metrics (StemSim and ImgSim) are computed in $[0, 1]$ and reported as percentages by multiplying by 100 in all tables.

E.1 StemSim Normalization $\phi(\cdot)$

Before computing stem-text similarity (StemSim), we apply a shared normalization function $\phi(\cdot)$ to both the ground-truth (GT) stem and the predicted stem, removing evaluation-irrelevant artifacts and improving backbone comparability. The normalization consists of:

(i) **Audit-only removal.** We remove markers and metadata that are only used for auditing/visualization and should not contribute to stem matching (e.g., "hallucination fixed" /"suspected hallucination" tags, or action traces such as $actions = ...$).

(ii) **Canonicalization.** We canonicalize casing, punctuation, and whitespace. In particular, when computing Levenshtein similarity, we ignore all whitespace characters so that spaces and line breaks do not affect the score. We also remove weakly semantic template noise using deterministic rules, including:

- generic figure-referencing prompts ;
- choice-format hints;
- long underscore placeholders that do not carry semantic content.

(iii) **Penalty for missing alignment.** If a prediction cannot be aligned to its GT counterpart or the aligned predicted stem is empty, we set StemSim of that question to 0, explicitly penalizing failed/missing extraction.

Backbone	Setting	$\text{All}_{\text{StemSim}}\uparrow$	$\text{All}_{P}\uparrow$	$\text{All}_{R}\uparrow$	$\text{All}_{F1}\uparrow$
GLM-4.6v	Ours	88.34	67.06	30.90	42.30
GLM-4.6v	WholePage	82.61	58.01	22.16	32.07
GLM-4.6v-flash	Ours	79.85	58.37	33.93	42.92
GLM-4.6v-flash	WholePage	72.02	52.15	20.46	29.39
Qwen3-VL-8B	Ours	86.48	59.09	32.02	41.53
Qwen3-VL-8B	WholePage	80.85	35.18	21.75	26.88
Qwen3-VL-30B	Ours	88.13	61.97	29.79	40.23
Qwen3-VL-30B	WholePage	78.60	49.97	25.08	33.40

Table A.6: Ours vs. WholePage (chain-matched) (%).

Split	Pages	MC	SA	LA	Partial	MC%	SA%	LA%
train	8256	21820	6384	11737	0	54.63	15.98	29.39
valid	1032	2683	815	1438	0	54.36	16.51	29.13
test	1032	2694	815	1482	0	53.98	16.33	29.69

Table B.1: Split statistics of the clean 3-class subset (seed=42).

E.2 ImgSim via Pairwise Scoring $s(\cdot, \cdot)$

For any GT figure g and predicted figure p, we define a pairwise similarity score:

$$s(g, p) \in [0, 1]$$

produced by a fixed visual similarity model, where higher values indicate greater visual similarity.

To compute question-level **ImgSim**, we align predicted figures to the GT sequence (using GT order as reference) and take the average over aligned pairs. If a GT position has no matched prediction, its score is set to 0 to explicitly penalize missed or incorrect detection. Predicted figures not aligned to any GT position are ignored.

E.3 Refusal / Unrecognizable Binarization and P/R/F1

To evaluate whether the system refuses appropriately when evidence is insufficient (thereby avoiding hallucinated or fabricated outputs), we treat refusal/unrecognizable cases as the positive class and compute Precision/Recall/F1 under a binary labeling scheme.

Ground-truth label. For question i, we define: if the GT segment contains any marker from a curated list of *unrecognizable* patterns; otherwise $y_i = 0$.

Prediction label. We define: $\hat{y}_i = 1$ if the system explicitly rejects the question or fails to find a valid alignment to the GT; otherwise $\hat{y}_i = 0$. Concretely, $\hat{y}_i = 1$ if `is_rejected = 1` or `alignment_found = 0`.

We treat failed alignment as refusal-equivalent because it indicates the system did not produce verifiable, groundable content for the corresponding GT instance under our protocol. **Counts.** For question type t with sample set $\mathcal{D}_t$, we compute:

$$\begin{aligned} TP_t &= \sum_{i \in D_t} \mathbf{1}(y_i = 1 \wedge \hat{y}_i = 1), \\ FP_t &= \sum_{i \in D_t} \mathbf{1}(y_i = 0 \wedge \hat{y}_i = 1), \\ FN_t &= \sum_{i \in D_t} \mathbf{1}(y_i = 1 \wedge \hat{y}_i = 0). \end{aligned} \tag{E.1}$$

Metrics. We compute:

$$\begin{aligned} P_t &= \frac{TP_t}{TP_t + FP_t}, \\ R_t &= \frac{TP_t}{TP_t + FN_t}, \\ F1_t &= \frac{2P_tR_t}{P_t + R_t}. \end{aligned} \tag{E.2}$$

Denominator-zero convention. If a denominator is zero (e.g., $TP_t + FP_t = 0$, $TP_t + FN_t = 0$, $P_t + R_t = 0$), we set the corresponding metric to 0 by convention.

E.4 Overall Macro Averaging and All_{F1}

To avoid biases introduced by imbalanced question-type distributions, we report overall metrics using

Exp	Cls	Train Pages	Train Anns	Occ Pages	Clean Pages	Clean (<3)	Clean (≥ 3)	Partial Pages	Partial Anns
Train Sub	3	8256	39941	0	8256	2684	5572	0	0
Train Sub+Syn	3	13501	73426	5245	8256	2684	5572	0	0
Train Syn	3	7929	37683	5245	2684	2684	0	0	0
Train Sub+Syn	4	13501	73426	5245	8256	2684	5572	4205	5881

Table B.2: Stage-1 training recipe statistics (Ablation Control).

Split	Pages	MC	SA	LA	Partial	MC%	SA%	LA%	Partial%
train	9763	28097	8293	12995	1768	54.93	16.21	25.40	3.46
valid	1247	3608	1058	1617	232	55.38	16.24	24.82	3.56
test	1462	4540	1354	1803	486	55.47	16.54	22.03	5.94

Table B.3: Split statistics of RDEI-Target (All) under seed=42 (4 classes).

equal-weight macro averaging across MC/SA/LA. For $X \in \{ImgSim, StemSim, P, R\}$, we define:

$$All_X = \frac{X_{MC} + X_{SA} + X_{LA}}{3}. \tag{E.3}$$

Importantly, All_{F1} is not computed as the arithmetic mean of $F1_t$.Instead, we first macro-average precision and recall and then compute:

$$All_{F1} = \frac{2 \cdot All_P \cdot All_R}{All_P + All_R}. \tag{E.4}$$

This ensures the reported overall F1 is consistent with the reported macro P/R, avoiding inconsistencies caused by the non-linearity of $F1$ across question types.

F Additional Implementation Notes for Sec. 3.2

This appendix records auxiliary implementation conventions omitted in Sec. 3.2.

F.1 Spread detection: B_{elig} and relative saliency

The coverage profile is computed from an **eligible set** B_{elig} consisting of layout-stable carriers. We construct B_{elig} by category filtering and robust scale filtering to suppress tiny fragments and abnormally large boxes that distort the gutter valley, yielding $N_{elig} = |B_{elig}|$.

Valley search is restricted to a **central horizontal band** (an interior range of the page height) to avoid spurious minima caused by margins, bindings, and partial-view artifacts, and $\mathsf{cov}(\cdot)$ may be lightly smoothed along the x-axis before locating the minimum.

We accept **`split_x`** only when the minimum is sufficiently low relative to a robust reference statistic (e.g., median and MAD computed within the band); otherwise, we revert to single-column. When N_{elig} is below a minimal reliability threshold, we conservatively disable spread detection.

If multiple minima satisfy the criterion, we select the one **closest to the band center**, and fix the page-level order as $\langle side_L, side_R \rangle$.

F.2 Side-to-zone/column decomposition and read_index writeback

Within each side, horizontally spanning question/text carriers are treated as ***span barriers***. Barriers that exhibit substantial vertical overlap or negligible vertical separation are merged to prevent empty or excessively thin zones. Zones are then defined by the vertical intervals between consecutive merged barrier groups, under a **no-drop convention**.

Elements are assigned to zones primarily by **vertical center**. Boundary cases are resolved by maximum vertical overlap with candidate zones; remaining ties are broken deterministically by lexicographic order on $\langle y_1, x_1 \rangle$.

Intra-zone columns are estimated from the left-boundary set X built from morphologically stable carriers, with a small cap $k_{\max}$ and a single-column fallback when $|X|$ is insufficient; **1D k-means** uses deterministic initialization or a fixed seed, and smaller k is preferred when gains are marginal (e.g., small reduction in 1D within-cluster SSE).

Split	Pages	MC	SA	LA	Partial	MC%	SA%	LA%	Partial%
test (Partial+)	324	1372	397	229	486	**55.23**	**15.98**	**9.22**	**19.57**

Table B.4: Statistics of the RDEI-Target (Partial+) test subset.

During **writeback**, elements are ordered within each column by increasing top coordinate y, with deterministic tie-breaks by x and then area. Span/barrier nodes are placed strictly between their adjacent zones to avoid cross-level interleaving.

F.3 Weak geometric terms: gap_x, gap_y and *inside*

Let $q = \langle x_1^q, y_1^q, x_2^q, y_2^q \rangle$ and $f = \langle x_1^f, y_1^f, x_2^f, y_2^f \rangle$, with centers

$$c_q = \left\langle \frac{x_1^q + x_2^q}{2}, \frac{y_1^q + y_2^q}{2} \right\rangle, \quad c_f = \left\langle \frac{x_1^f + x_2^f}{2}, \frac{y_1^f + y_2^f}{2} \right\rangle. \tag{F.1}$$

We define non-negative axis separations as

$$\mathrm{gap}_x(q,f) = \max\Big(0, x_1^q - x_2^f, x_1^f - x_2^q\Big), \quad \mathrm{gap}_y(q,f) = \max\Big(0, y_1^q - y_2^f, y_1^f - y_2^q\Big). \tag{F.2}$$

We further define the indicator term as $\mathit{inside}(q,f) = \mathbf{1}(c_f \in q)$. The normalized terms g_x, g_y and the center-distance term c_d follow Sec. 3.2.

F.4 Min-cost flow construction: idle capacity and determinism

We construct a directed flow network from $\mathcal{E}$ with source s and sink t. Edges are defined as follows: (i) $s \to f$ has capacity $cap_{fig}(f)$ and zero cost; (ii) for each $(q,f) \in \mathcal{E}$, $f \to q$ has unit capacity and cost $-score_{base}(q,f)$; and (iii) $q \to t$ has capacity K and zero cost. We then solve a **min-cost flow** that routes the total supply $\sum_f cap_{fig}(f)$ from s to t. To avoid forcing unreliable matches, we allow **idle capacity** by adding a **skip edge** $f \to t$ with capacity $cap_{fig}(f)$ and a **small non-negative cost** when necessary, so that unused figure capacity can be sent directly to t.

When discrete costs permit multiple optima, we enforce **determinism** via a **fixed tie-break** aligned to the recovered read order. Specifically, edges are ranked by **(figure read order, then question read order)**, and we apply a **tiny integer perturbation** to costs to impose **lexicographic preference** without altering the primary objective. Finally, **positive-flow edges** on $f \to q$ are mapped back to $\mathcal{P}$, and the evidence binding I_i^q is written for downstream use.