MeDocVL: A Visual Language Model for Medical Document Understanding and Parsing

Visual Computing Group, Ping An Property & Casualty Insurance Company

[1] Ping An Group, Shenzhen, Guangdong 518001, China
[2] Ping An Property & Casualty Insurance, Shenzhen, China info@pingan.com
https://www.pingan.com

Abstract. Medical document OCR is challenging due to complex layouts, domain-specific terminology, and noisy annotations, while requiring strict field-level exact matching. Existing OCR systems and general-purpose vision–language models often fail to reliably parse such documents. We propose **MeDocVL**, a post-trained vision–language model for query-driven medical document parsing. Our framework combines *Training-driven Label Refinement* to construct high-quality supervision from noisy annotations, with a *Noise-aware Hybrid Post-training* strategy that integrates reinforcement learning and supervised fine-tuning to achieve robust and precise extraction. Experiments on medical invoice benchmarks show that MeDocVL consistently outperforms conventional OCR systems and strong VLM baselines, achieving state-of-the-art performance under noisy supervision.

Keywords: OCR · Vision–Language Model · Medical Document Parsing · Post-training

1 Introduction

Modern Optical Character Recognition (OCR) has long been a cornerstone of information technology (IT) and continues to play an important role in pro-
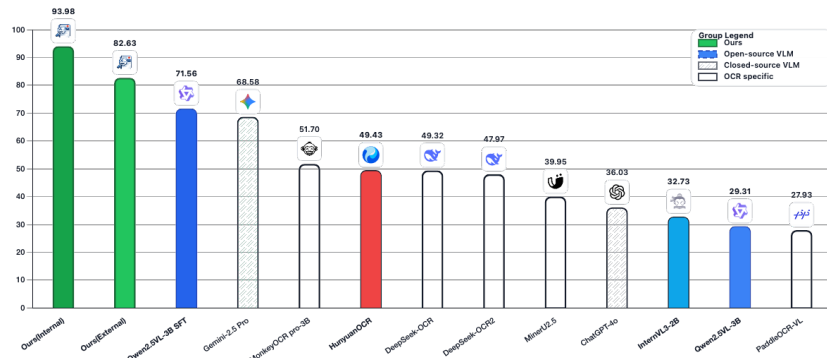

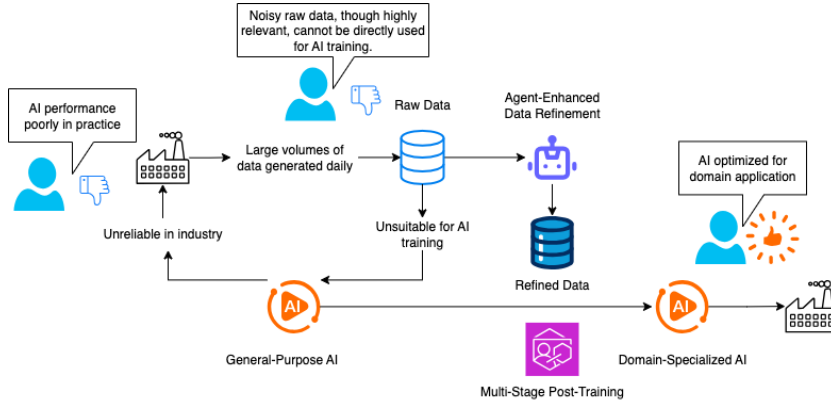**Fig. 1:** Overall metrics of our work.

Fig. 2: Overall pipeline of the proposed MeDocVL framework. Large-scale raw industrial data are highly relevant but unsuitable for direct training due to noisy or missing annotations. MeDocVL employs Training-driven Label Refinement to convert raw data into reliable domain supervision, followed by Noise-aware Hybrid Post-training to adapt a general-purpose VLM into a high-precision, domain-specialized model.

moting document digitization and information automation. Recent advances in artificial intelligence (AI) have significantly improved OCR performance and expanded its scope to document parsing, information extraction, and visual question answering. However, high-stakes, domain-specific document parsing remains challenging. In medical documents, insurance claims, and financial invoices, the objective is not merely transcription but accurate extraction of fields and their values (e.g., dates, amounts, identifiers), where a single-character error may invalidate the entire field and break downstream usage.

Meanwhile, practical document parsing systems increasingly require flexible, query-driven access to document content, allowing users to retrieve information via natural-language queries rather than a predefined schema. This motivates query-driven document content parsing, which demands both precise visual recognition and alignment with user intent.

Existing approaches rarely satisfy both requirements. Pipeline-based OCR systems convert images into intermediate representations (e.g., text/markup) and then rely on downstream components (often LLMs) for field extraction, which introduces error propagation, system complexity, and limited domain adaptability. In contrast, MLLM-based OCR/parsing models offer unified vision-language reasoning and natural-language interaction, but they remain unreliable for high-precision extraction, frequently producing token-level recognition errors (and occasional hallucinations) under dense layouts and strict formatting constraints.

These limitations reveal a fundamental gap, as high-precision visual recognition and flexible, query-driven parsing are rarely achieved simultaneously. We therefore study high-precision, end-to-end, query-driven document content pars-

ing, focusing on medical documents and invoices. Our goal is to transform a general-purpose MLLM into a queryable document parsing expert that achieves OCR-level recognition fidelity while preserving the flexibility of natural-language querying. Crucially, we target realistic training conditions, where large-scale document data are available but annotations are incomplete or noisy—an unavoidable setting in industrial applications.

To this end, we propose a backbone-agnostic post-training framework for query-driven document content parsing under noisy supervision, built upon three tightly coupled technical components. (1) **Training-driven Label Refinement (TLR)** formulates annotation correction as a learnable process. Starting from a small Clean Data with expert annotations, we generate structured pseudo-label prompts using OCR and MLLM systems, whose outputs constitute Noisy annotations. An Annotation Refinement Model is then trained to map these noisy annotations to corrected predictions, using expert labels only to supervise the refinement behavior. Importantly, TLR does not aim to eliminate all noise. Instead, it distills the error characteristics of OCR/MLLMs and suppresses their systematic biases, producing Refined Data that provides more stable and reliable supervision for downstream optimization. (2) Building upon the refined supervision, we introduce **Noise-aware Hybrid Post-training (NHP)**, which integrates supervised fine-tuning (SFT) and reinforcement learning (RL) in a noise-aware manner. Rather than directly fitting potentially unreliable annotations, RL is first employed on Refined Data with task-specific rewards to improve robustness under residual noise. Subsequently, SFT on Clean Data is used to consolidate extraction behavior, output formatting, and domain specialization. This staged design balances robustness and stability when adapting general-purpose MLLMs to domain-specific document parsing. (3) Finally, we design a **token-wise GRPO objective for noise-robust document extraction**, which aligns RL with the exact-match requirements of structured field parsing. Instead of sequence-level optimization, we derive fine-grained token-level rewards from refined annotations, combined with token masking, optional confidence weighting, and reference regularization. This formulation focuses learning on reliable tokens and mitigates the adverse impact of noisy or ambiguous supervision.

The proposed framework is model-agnostic and applicable to a wide range of multimodal backbones. **MeDocVL** represents one concrete instantiation, enabling end-to-end, query-driven document parsing with recognition accuracy comparable to strong OCR baselines. Through systematic experiments, we analyze the impact of annotation noise, validate each component of the framework, and demonstrate improved robustness under noisy supervision.

Our contributions can be summarized as follows:

- We formulate query-driven, high-precision document content parsing as a distinct problem setting that bridges OCR-level recognition accuracy with the flexibility of MLLMs.
- We propose a backbone-agnostic post-training framework for adapting general MLLMs to document parsing under noisy supervision, built upon Training-

driven Label Refinement, Noise-aware Hybrid Post-training, and token-wise alignment.
– We conduct systematic experiments, including controlled noise injection, comparative evaluations, and ablation studies, to validate the role of each technical component.
– In addition, we release a curated medical and invoice document content parsing benchmark with refined annotations and evaluation protocols to facilitate reproducible research.

2 Related Work

Traditional document parsing approaches [4, 18, 19] typically decompose document understanding into a sequence of specialized sub-tasks, including layout analysis [8, 26, 32], reading order prediction [28], optical character recognition (OCR) [10,12], formula recognition [27], and table structure recognition [6,21,30], each handled by an independent model. Such modular pipelines are effective across diverse document formats and allow fine-grained optimization for individual components. However, errors introduced at early stages tend to propagate downstream, and the lack of holistic semantic modeling often necessitates extensive post-processing to obtain coherent, human-readable outputs, limiting robustness in complex real-world scenarios.

The emergence of large language models (LLMs) [9,15,24] has partially alleviated these issues by providing strong text understanding and reasoning capabilities. Several works [3,11,17,22,25,29] incorporate LLMs to aggregate OCR outputs, interpret document content, or perform structured information extraction. While these methods improve semantic consistency and flexibility, LLMs operate purely on textual inputs and lack access to spatial and visual context, which makes them sensitive to layout errors and limits their reliability on documents with complex structures.

Recent vision–language models (VLMs) [2,20,31,33] address this limitation by jointly modeling visual and textual information through a unified architecture. By coupling a vision encoder with a language model, VLMs can directly process document images and reason over layout, typography, and semantics in an end-to-end manner, achieving strong performance on general OCR and document understanding benchmarks. Representative models, including the Monkey series [13,14,16], mPLUG-DocOwl2 [7], and olmOCR [23], demonstrate the effectiveness of large-scale VLM fine-tuning on document-centric corpora. Despite these advances, most existing VLM-based approaches prioritize general-purpose versatility and benchmark coverage, rather than the strict precision, reliability, and robustness required in domain-specific, high-stakes applications.

In settings such as medical document recognition, where field-level exactness, interpretability, and error tolerance are critical, the above limitations become particularly pronounced. This motivates our proposed **MeDocVL** framework, which integrates training-driven data refinement with noise-aware post-training into a unified adaptation pipeline. By explicitly addressing annotation noise, systematic OCR errors, and preference alignment during post-training, MeDocVL

provides a scalable and robust approach for adapting general-purpose VLMs to specialized industrial domains.

3   Training-driven Label Refinement

In this section, we introduce Training-driven Label Refinement (TLR), a principled, data-centric post-training component designed to stabilize learning for query-driven document content parsing under noisy supervision. Unlike conventional data cleaning pipelines that aim to produce perfectly accurate annotations, TLR explicitly targets the dominant sources of systematic annotation noise arising from OCR systems and MLLMs, and seeks to reduce their bias while preserving realistic supervision variability.

In real-world document understanding scenarios, large-scale training data are typically obtained through a combination of human annotation and automated labeling systems. While extensive filtering is often applied to remove severely corrupted samples, residual noise in annotations remains inevitable due to human mistakes and consistent failure patterns of OCR and MLLM-based extractors. Such noise is particularly harmful for document content parsing, where supervision is defined at the field level, and even minor token-level errors can invalidate an entire key-value pair. Directly applying SFT on noisy annotations therefore leads to unstable optimization and brittle extraction behavior.

TLR addresses this challenge by formulating label refinement itself as a learnable task. Starting from a small Expert Annotations Dataset with reliable ground truth, we construct structured pseudo labels by jointly leveraging OCR systems and MLLMs. These pseudo labels, although imperfect, encode rich domain knowledge and characteristic error patterns. An Annotation Refinement Model is then trained to revise these pseudo labels by explicitly learning how OCRMLLM predictions deviate from expert annotations. Importantly, this process does not aim to eliminate all noise; instead, it distills correction behavior and suppresses systematic biases, producing a Refined Data distribution that is more suitable for downstream optimization under noisy supervision.

The TLR pipeline consists of three stages:

1. **Pseudo-label construction and prompt synthesis.** We generate structured pseudo labels from expert-annotated samples using OCR and MLLM predictions, and convert them into instruction-style prompts that expose typical annotation errors to the refinement model.
2. **Correction distillation.** The Annotation Refinement Model is trained by comparing its corrected outputs against expert annotations, learning to revise noisy key–value pairs while preserving valid information from pseudo labels.
3. **Large-scale refinement and filtering.** The trained refiner is applied to the Noisy Annotations Dataset to produce Refined Data, which exhibits reduced systematic noise and improved consistency, and serves as the primary supervision source for subsequent noise-aware post-training.
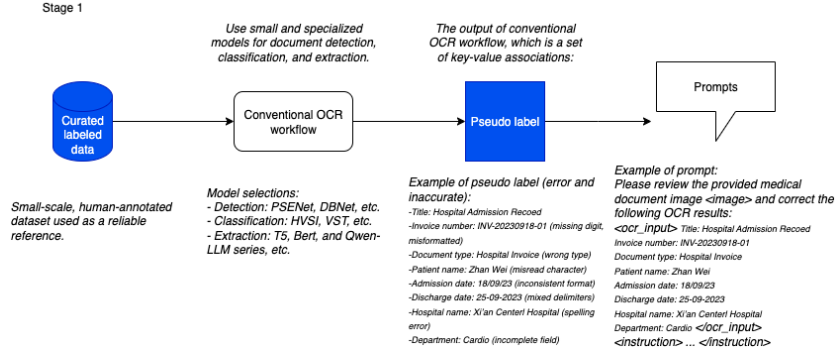
Fig. 3: Stage 1 — Pseudo-label construction and prompt synthesis. Expert-annotated documents are processed by OCR systems and MLLMs to produce structured but imperfect key–value predictions. These predictions are converted into instruction-style prompts that explicitly expose annotation errors, enabling the refinement model to learn correction behaviors rather than generating labels from scratch.

By transforming noisy annotations into a more stable and structured supervision signal, TLR establishes a robust foundation for the Noise-aware Hybrid Post-training strategy and token-wise reinforcement learning objectives introduced in later sections.

3.1   Stage 1: Pseudo-label Construction and Prompt Synthesis

The first stage of TLR focuses on constructing informative pseudo labels and corresponding instruction prompts that expose typical annotation errors to the refinement model. Starting from a small Expert Annotations Dataset with reliable ground truth, we generate pseudo labels by jointly leveraging OCR systems and multimodal large language models (MLLMs). These pseudo labels consist of structured key-value pairs that reflect realistic extraction outputs, including both correct fields and systematic errors commonly observed in medical documents.

Unlike conventional data augmentation or masking-based self-regression strategies, we intentionally preserve the full key-value structure of pseudo labels, even when values are incorrect. This design choice is motivated by two observations. First, OCR and MLLM predictions encode substantial domain knowledge and layout priors, which would be discarded if values were masked or regenerated from scratch. Second, the dominant failure modes in document parsing arise from consistent, structured errors (e.g., misaligned fields, truncated spans, or semantic confusion between similar medical terms), which must be explicitly observed in order to be corrected.

To this end, pseudo labels are transformed into instruction-style prompts that present the model with both the document image and the predicted key-value pairs, and request the correction of erroneous fields. The refinement model
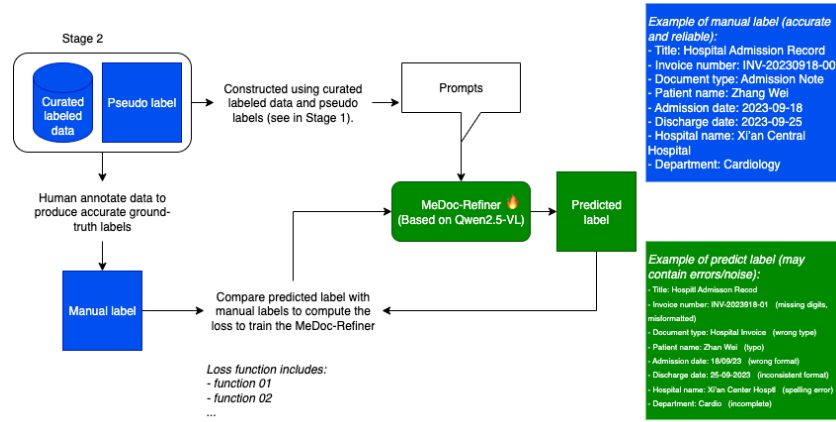
Fig. 4: Stage 2 — Correction distillation training. The Annotation Refinement Model is trained to revise pseudo labels by comparing its predictions against expert annotations. The training objective emphasizes learning systematic correction behaviors while preserving the general multimodal representations of the base model.

is therefore trained to revise existing predictions rather than hallucinating new ones. This prompt formulation shifts the learning objective from unconditional generation to conditional correction, enabling the model to focus on identifying and fixing systematic deviations between pseudo labels and expert annotations.

Crucially, this process should not be interpreted as an attempt to construct noise-free supervision. Instead, pseudo labels serve as a structured intermediate representation that captures the characteristic error distribution of upstream OCR and MLLM systems. By conditioning on these imperfect predictions, the refinement model learns how annotation noise manifests at the field and token levels, which forms the foundation for subsequent correction distillation and large-scale refinement.

Through this design, Stage 1 establishes a supervision interface that maximizes the reuse of existing extraction signals, exposes realistic error patterns, and prepares the refinement model for learning robust correction behaviors in the presence of noisy annotations.

3.2   Stage 2: Correction Distillation Training

Given the pseudo-label prompts constructed in Stage 1, the second stage focuses on training the Annotation Refinement Model to acquire robust correction capabilities. The model is initialized from a pre-trained multimodal backbone (Qwen2.5-VL) and adapted using parameter-efficient fine-tuning. Rather than performing broad re-training, this stage explicitly targets the ability to identify and correct systematic annotation errors introduced by OCR systems and MLLMs.

The key objective of correction distillation is not to directly predict ground-truth labels from scratch, but to learn how pseudo labels deviate from expert annotations. For each curated sample, the refinement model is presented with the document image and the corresponding pseudo-label prompt, and is trained to produce a corrected version that aligns with expert-provided annotations. This formulation encourages the model to focus on error localization and revision, instead of unconditional generation, which is known to be unstable under noisy supervision.

We formulate this training process as a structured distillation problem. Pseudo labels serve as a teacher signal that encodes domain-specific priors and layout awareness, while expert annotations provide the target correction direction. By optimizing the refinement model to bridge the gap between the two, we effectively distill the correction behavior that maps noisy but informative supervision to more reliable annotations. Importantly, this process suppresses systematic bias without collapsing supervision diversity, resulting in refined labels that remain realistic for downstream training.

To support heterogeneous document structures, the correction objective integrates complementary supervision signals at different granularities, including field-level classification, sequence correction for text spans, and spatial alignment for tabular content. In addition, regularization terms are introduced to preserve the general multimodal reasoning ability of the base model and prevent overfitting to the small expert-annotated set. The full mathematical formulation of the optimization objective is provided in the Appendix.

Through iterative correction distillation, the Annotation Refinement Model learns stable and transferable error-correction strategies. As demonstrated in our experiments, this stage plays a critical role in reducing annotation noise and establishing a reliable supervision foundation for large-scale refinement and subsequent noise-aware post-training.

The full mathematical formulation of the correction distillation objective, including all loss terms, notation, and optimization details, is provided in Appendix A.

3.3   Stage 3: Large-scale Refinement for Noise-aware Post-training

In the final stage of TLR, the trained Annotation Refinement Model is deployed to process large volumes of domain-specific documents and generate Refined Data for downstream optimization. Following the same prompt construction strategy as in Stage 1, the refiner revises OCR/MLLM-derived annotations by applying the correction behaviors learned during distillation. The resulting annotations exhibit substantially reduced systematic errors compared to raw OCR outputs, while still retaining residual noise that reflects realistic extraction uncertainty.

Importantly, the objective of this stage is not to construct a perfectly clean dataset. In practical industrial settings, large-scale document corpora inevitably contain annotation noise due to human mistakes and consistent failure patterns of automated labeling systems. Over-aggressive filtering or manual correction
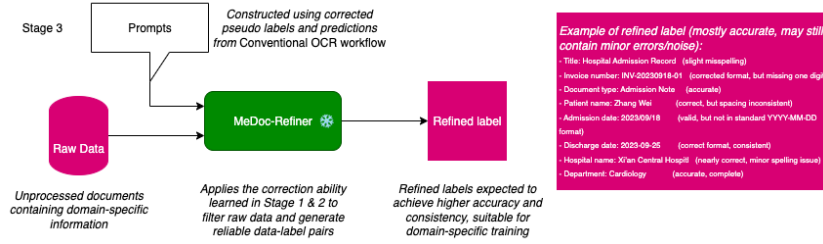
Fig. 5: Stage 3 — Large-scale refinement for noise-aware post-training. The trained Annotation Refinement Model is applied to large-scale raw documents to transform noisy annotations into refined supervision. Rather than eliminating all errors, the resulting Refined Data reduces systematic bias while preserving realistic annotation variability, serving as a stable supervision source for noise-aware post-training.

may reduce noise but often leads to limited data scale and distorted supervision distributions. Instead, TLR deliberately produces Refined Data that strikes a balance between consistency and diversity: systematic biases are suppressed, but controlled noise is preserved.

This property is crucial for subsequent Noise-aware Hybrid Post-training. As shown in our experiments, supervised fine-tuning is highly sensitive to annotation noise and tends to overfit to corrupted labels, whereas reinforcement learning objectives exhibit greater tolerance to residual noise. By reshaping the supervision distribution rather than fully denoising it, Refined Data provides a stable foundation on which noise-aware optimization strategies can effectively operate.

Figure 6 provides an intuitive illustration of this effect. Raw OCR supervision yields a highly irregular optimization landscape, making learning unstable. Expert annotations constrain optimization to a narrow region, which accelerates convergence but increases the risk of overfitting under limited data. In contrast, Refined Data reshapes the supervision distribution into a smoother yet sufficiently diverse space, enabling robust optimization under noisy supervision.

Through this large-scale refinement process, TLR produces a supervision source that is cost-effective, scalable, and explicitly designed to support noise-aware post-training. This stage bridges data refinement and optimization, and directly motivates the hybrid training strategy introduced in the next section.

4 Noise-aware Hybrid Post-training

In this section, we introduce Noise-aware Hybrid Post-training (NHP), a post-training strategy designed to adapt multimodal large language models to high-precision document content parsing under noisy supervision. Unlike conventional fine-tuning pipelines that rely solely on supervised learning, NHP explicitly accounts for the intrinsic noise and bias present in large-scale document anno-
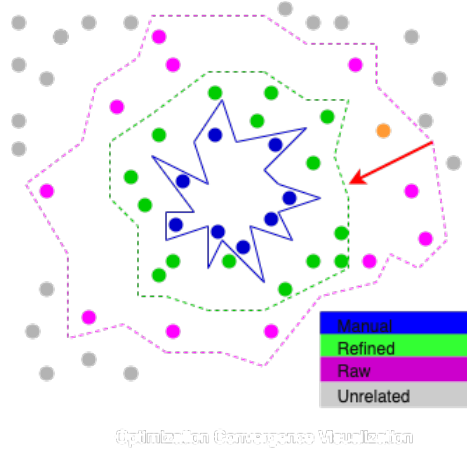
Fig. 6: Optimization landscape under different supervision sources. Raw OCR annotations induce a broad and irregular optimization space that hampers convergence. Expert annotations define a narrow but potentially overfitting-prone region. Refined Data occupies an intermediate regime, reducing irregularity while maintaining sufficient coverage for stable and robust optimization.

tations, and leverages complementary optimization paradigms to achieve both robustness and precision.

As established in Section 3, even after training-driven label refinement, large-scale supervision inevitably contains residual noise. For document parsing tasks, this noise is particularly harmful, as supervision is defined at the field and token levels, and small annotation errors can lead to incorrect extraction outcomes. Directly applying supervised fine-tuning (SFT) under such conditions often results in overfitting to corrupted labels and unstable optimization, especially when exact matching is required.

To address this challenge, NHP adopts a hybrid post-training paradigm that combines reinforcement learning (RL) and supervised fine-tuning in a noise-aware manner. The core insight is that these two optimization regimes play fundamentally different roles under noisy supervision. Reinforcement learning introduces soft preference constraints, allowing the model to improve alignment while tolerating imperfect annotations, whereas supervised fine-tuning enforces hard token-level targets, which is effective only once the model behavior has been sufficiently stabilized.

As illustrated in Figure 7, NHP decomposes post-training into two stages. In the first stage, we apply reinforcement learning on Refined Data produced by TLR, using preference-based objectives to guide the model away from systematic errors without strictly enforcing noisy labels. This stage emphasizes robustness and correction behavior under uncertainty. In the second stage, we apply supervised fine-tuning on curated high-quality annotations to consolidate extraction accuracy and strengthen domain-specific behavior. Importantly, the
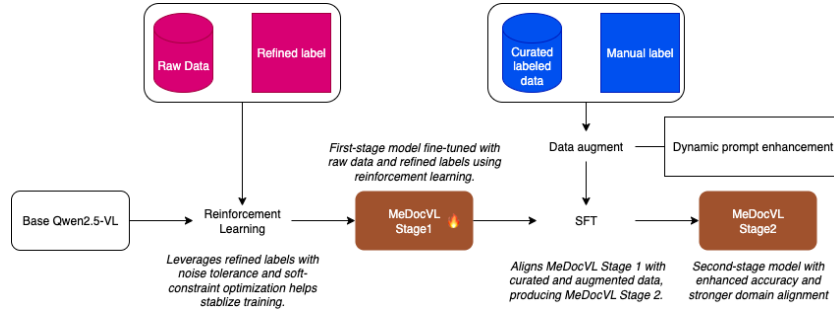
Fig. 7: Training recipe of MeDocVL. In Stage 1, raw data and initial labels are corrected by a VLM (*MeDocVL-Stage1*) to generate refined data and labels. In Stage 2, the refined data and labels are used to further train a correction model (*MeDocVL-Stage2*), enhancing the quality of data–label pairs and improving model robustness.

order of these stages is non-interchangeable: RL serves to stabilize learning under noise, while SFT refines and solidifies the aligned behavior.

Throughout NHP, the vision encoder of the multimodal backbone is kept frozen to preserve strong visual representations learned during pre-training. By decoupling robustness-oriented alignment from precision-oriented refinement, NHP provides a principled and efficient post-training strategy for high-precision, query-driven document parsing in realistic noisy supervision settings.

4.1   Reinforcement Learning under Noisy Supervision

After Training-driven Label Refinement (TLR), the resulting supervision is substantially cleaner than raw OCR outputs, yet still inevitably contains residual noise. For document content parsing, such noise is particularly harmful, as training signals are defined at the field and token levels, and even minor annotation errors can lead to incorrect extraction behavior. Directly applying supervised fine-tuning (SFT) on this data risks overfitting to noisy labels and may push the model toward brittle solutions.

Motivated by recent findings that naïve SFT on multimodal reasoning data can degrade alignment performance [5], we adopt reinforcement learning (RL) as the first post-training stage. Instead of enforcing strict one-to-one correspondence between predictions and labels, RL introduces *preference-based soft constraints*, allowing the model to improve correction behavior while remaining tolerant to imperfect supervision. This property makes RL particularly suitable for learning from Refiner-labeled data, where annotation noise is reduced but not eliminated.

Why reinforcement learning before SFT. Applying SFT prior to preference optimization has been shown to induce pseudo reasoning paths, i.e., superficial patterns that mimic correct outputs without learning robust correction strategies. Once such patterns are reinforced, subsequent optimization becomes unsta-

ble. In contrast, RL optimizes relative preferences between candidate outputs, guiding the model toward more reliable behaviors without over-committing to noisy targets. This stage therefore serves as a behavioral stabilization step before high-precision supervision is enforced in later SFT.

From sequence-level to token-level preference optimization. We implement a customized variant of Generalized Reinforcement Preference Optimization (GRPO). Standard GRPO computes token-level rewards but aggregates them into sequence-level advantages, while GSPO further averages these rewards to improve stability for long-chain reasoning tasks. However, document recognition fundamentally differs from reasoning-centric tasks: correctness depends on precise token predictions (e.g., digits, dates, medical codes), rather than global sequence coherence. Aggregating rewards at the sequence level may obscure localized errors that are critical for document parsing.

To address this mismatch, we adopt token-wise preference optimization, where preference signals are applied directly at the token level. This design provides fine-grained corrective supervision, ensuring that small but critical extraction errors are explicitly penalized. The resulting optimization objective emphasizes local accuracy while retaining the robustness advantages of RL.

Token-wise GRPO Objective Given an input document x with a preferred response y^+ and a dispreferred response y^-, we define the token-wise preference loss as:

$$\mathcal{L}_{\text{tok-GRPO}} = \mathbb{E}_{(x,y^+,y^-)} \left[-\frac{1}{T} \sum_{t=1}^{T} w_t \log \sigma\Big(\beta\big[\Delta_t(\theta) - \kappa\,\Delta_t^{\text{ref}}\big]\Big) \right], \qquad (1)$$

where $\Delta_t(\theta)$ denotes the token-level log-probability gap between preferred and dispreferred responses, w_t is a token confidence weight, and Δ_t^{ref} constrains policy drift relative to the frozen backbone. Full definitions and derivations are provided in Appendix B.

Stabilization via KL regularization. To preserve the strong general representations of the pre-trained backbone, we include a KL regularization term that limits deviation from the reference policy. The detailed formulation is provided in Appendix C. The overall RL objective is:

$$\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{tok-GRPO}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}}. \qquad (2)$$

Discussion. Compared to sequence-level GRPO or GSPO, token-wise GRPO better aligns with the error characteristics of document recognition. By weighting tokens according to confidence and avoiding unnecessary length penalties, the optimization remains both stable and precise. This RL stage equips the model with robust correction behavior under noisy supervision, forming a reliable foundation for subsequent supervised fine-tuning on clean data.

4.2   Supervised Fine-tuning on Clean Data (Precision Consolidation Stage)

Following reinforcement learning under noisy supervision, we perform a final stage of Supervised Fine-Tuning (SFT) on clean, high-confidence annotations. Unlike the preceding RL phase, which primarily focuses on robustness and preference alignment under imperfect supervision, this stage serves as a precision consolidation step. Its goal is to sharpen token-level accuracy, reduce residual variance, and calibrate the model's predictions to satisfy the strict exact-match requirements of document content parsing.

Document recognition differs fundamentally from open-ended reasoning tasks in that correctness is defined by fine-grained, field-level precision rather than long-chain inference. After RL has guided the model toward robust and noise-tolerant behaviors, SFT provides strong deterministic supervision to consolidate these behaviors into stable and highly accurate extraction patterns. Importantly, this stage operates exclusively on clean or refined annotations produced by the Training-driven Label Refinement (TLR) pipeline, avoiding the risk of reintroducing noisy supervision.

To mitigate catastrophic forgetting and maintain balanced performance across heterogeneous document formats, we adopt a mixed-stage SFT strategy. Instead of sequentially fine-tuning the model on individual domains, all data sources are jointly used in a single SFT phase. This design preserves cross-domain generalization while allowing the model to specialize in high-precision document parsing.

Instruction Data Construction The SFT dataset is constructed from curated and refined instruction–response pairs, covering approximately 2 million samples. The data are evenly distributed across three representative document categories: free-form documents, tabular documents, and mixed-format documents. Each sample follows a unified instruction-driven format, explicitly specifying target fields to be extracted from the input document.

Both text-only and multimodal samples are included to maintain alignment with the pre-trained language capabilities of the backbone. However, multimodal samples are prioritized for domain-specific consolidation, as they expose the model to realistic visual–textual interactions involving layout, tables, stamps, and handwritten content. This balanced construction enables MeDocVL to consolidate high-precision extraction skills without sacrificing general instruction-following behavior.

Dynamic Prompt-based Precision Augmentation To further strengthen precision and robustness, we introduce dynamic prompt-based augmentation during SFT. Instead of perturbing the document images, controlled noise is injected into the instruction prompts. Specifically, individual fields within the prompt are randomly masked or replaced with plausible but incorrect values, forcing the model to verify information against the visual content rather than relying on prompt priors.

Fig. 8: Representative samples from the medical invoice dataset. Although all documents belong to a single category (medical invoices), they exhibit substantial visual and semantic complexity, including domain-specific medical terminology, mixed layout structures, printed text, tabular regions, handwritten annotations, and official seals, posing significant challenges for accurate recognition and field-level extraction.

Each field is assigned a perturbation probability in the range $[0, 1]$, and fields with probabilities exceeding a threshold of 0.3 are modified. This threshold is selected empirically and validated in Section 5. The perturbation process is lightweight yet effective, as it encourages systematic cross-checking between textual instructions and visual evidence.

This strategy directly complements the preceding RL phase. While RL improves tolerance to annotation noise and preference misalignment, dynamic prompt augmentation during SFT tightens decision boundaries at the token level. As a result, the model becomes less prone to hallucinated field values and more reliable in exact-match extraction scenarios.

Overall, this precision consolidation stage transforms the robustness gained from reinforcement learning into stable, high-accuracy document parsing performance, completing the Noise-aware Hybrid Post-training (NHP) pipeline.

5   Experiments

We conduct extensive experiments to evaluate the effectiveness of our proposed method for query-driven document content parsing. Specifically, we compare MeDocVL with state-of-the-art (SOTA) OCR and MLLM-based approaches for medical document parsing, analyze the impact of annotation noise on model fine-tuning, assess the effectiveness of training-driven label refinement, and perform ablation studies to verify the contribution of each component in our framework.

5.1   Experiment Settings

Dataset preparation We evaluate our method on a publicly available medical invoice dataset released by Alibaba Cloud [1]. The dataset contains 800 real-world medical billing documents collected from diverse scenarios, with 700 images used for training and 100 for testing.

To better reflect practical application requirements, we re-annotated the dataset by refining and expanding field definitions while preserving the original document content and fixed data split. Semantically redundant fields are consolidated, and additional fine-grained fields commonly required in real-world

medical invoice processing are introduced. The same re-annotated labels and experimental settings are applied to all baselines and our method, ensuring a fair and controlled comparison with full reproducibility.

Although the dataset contains a single document category, medical invoices exhibit substantial intra-class variability and domain-specific complexity. As illustrated in Figure 8, each document integrates heterogeneous information sources, including invoice metadata, medical terminology, itemized billing records, and patient-related fields, and often combines printed text, tabular layouts, handwritten content, and official stamps or seals. These characteristics result in dense layouts and complex visual–textual alignments, posing significant challenges for accurate recognition and query-driven field extraction.

Metrics Our experiments aim to assess practical field-level recognition performance under realistic document parsing scenarios. Conventional sequence-level metrics, such as exact match accuracy, are often insufficient to capture extraction quality across heterogeneous fields and may lack discriminative power when documents contain multiple keys.

We therefore adopt Field Match Ratio (FMR) as the primary evaluation metric. Since document annotations are represented as key–value pairs, FMR measures the proportion of correctly extracted field values across documents, directly reflecting field-level correctness.

Formally, let D denote the set of documents, K_d the set of keys in document d, $v_{d,k}$ the ground-truth value for key k, and $\hat{v}_{d,k}$ the predicted value. Since labels are represented as key–value associations, the model must accurately extract the correct value for each key. We define the micro-averaged FMR as:

$$\text{FMR}_{\text{micro}} = \frac{\sum_{d \in D} \sum_{k \in K_d} [\phi(\hat{v}_{d,k}) = \phi(v_{d,k})]}{\sum_{d \in D} |K_d|}, \tag{3}$$

where $\phi(\cdot)$ denotes optional normalization (e.g., case folding or date reformatting). For completeness, we also report the macro-averaged FMR across keys:

$$\text{FMR}_{\text{macro}} = \frac{1}{|K|} \sum_{k \in K} \left(\frac{\sum_{d \in D_k} [\phi(\hat{v}_{d,k}) = \phi(v_{d,k})]}{|D_k|} \right), \tag{4}$$

where $K = \bigcup_{d \in D} K_d$ and $D_k = \{d \in D \mid k \in K_d\}$. Unless otherwise specified, we report micro-averaged FMR in all experiments.

5.2   Comparison with the SOTA Models

Table 1 reports the Field Match Ratio (FMR) of different methods on the medical invoice dataset. We compare MeDoc-VL with representative OCR systems and MLLMs for document understanding.

The compared methods include general-purpose MLLMs with native visual–textual reasoning capabilities, such as Qwen2.5-VL (3B), InternalVL3 (2B), ChatGPT-4o, and Gemini-2.5 Pro, as well as OCR-oriented or document-focused models,

Table 1: FMR ↑ on medical invoice

Method	Medical Invoice
Qwen2.5-VL 3B	0.2931
InternalVL3 2B	0.5995
DeepSeek-OCR	0.4932
PaddleOCR-VL	0.2793
MonkeyOCR Pro 3B	0.517
MinerU2.5	0.3995
ChatGPT-4o	0.3603
Gemini-2.5 Pro	0.6858
MeDoc-VL (Internal)	**0.9398**
MeDoc-VL (External)	**0.8244**

Note: FMR = Field Match Ratio, best results per column are in **bold**.

including DeepSeek-OCR, PaddleOCR-VL, MonkeyOCR Pro (3B), and MinerU2.5. Notably, most OCR-based models do not natively support query-driven document parsing, i.e., they cannot directly retrieve fields or values conditioned on user-provided natural-language queries.

To enable a fair comparison under a unified query-driven setting, we adopt a two-stage evaluation protocol for such methods. Specifically, OCR models are first used to extract document content into structured JSON representations. A fixed Qwen3-8B model is then employed as an intermediate parser to extract fields and values from the OCR outputs based on user queries. The same extraction prompt and Qwen3 configuration are used across all OCR-based baselines, and the prompt is provided in the appendix to ensure reproducibility. All methods are ultimately evaluated under the same query-driven document parsing protocol and measured using FMR.

As shown in Table 1, MeDoc-VL achieves the highest FMR on medical invoices, substantially outperforming both OCR-based pipelines and general-purpose MLLMs. While MLLMs such as Qwen2.5-VL and Gemini-2.5 Pro demonstrate reasonable performance, they still exhibit noticeable field-level errors under complex medical invoice layouts. OCR-centric approaches, even when combined with a strong downstream parser, remain limited by recognition errors and imperfect semantic alignment, resulting in lower FMR.

We report two configurations of MeDoc-VL, denoted as Internal and External, which differ only in the source of supervision used during post-training. The Internal model is trained with additional large-scale proprietary data and is included solely to demonstrate the scalability and upper-bound performance of the proposed framework. Due to privacy and compliance constraints, its weights are not publicly released. The External model is trained exclusively on the publicly available Alibaba Cloud dataset and is fully reproducible. Both configurations achieve identical performance on this benchmark, indicating that the proposed framework is robust to supervision sources and does not rely on proprietary data.

Table 2: Impact of annotation noise on supervised fine-tuning and the effectiveness of training-driven label refinement.

Training Data	FMR
20% Noisy Data	0.7812
30% Noisy Data	0.7296
50% Noisy Data	0.5820
70% Noisy Data	0.2601
Refined Data (from 30% noise)	**0.8133**

5.3 Noisy Effectiveness Analysis

This experiment investigates the impact of annotation noise on SFT and evaluates the effectiveness of Training-driven Label Refinement (TLR) in mitigating noise-induced performance degradation. Starting from the same foundation model, we synthetically inject noise into training annotations by randomly corrupting a controlled proportion of key-value labels, with noise ratios set to 20%, 30%, 50%, and 70%.

We emphasize that noise is injected into annotations rather than raw document images to faithfully simulate practical post-training scenarios. In real-world document processing pipelines, input data are typically pre-filtered to remove severely corrupted or low-quality samples, ensuring the visual content itself is relatively reliable. However, annotations often contain non-negligible noise due to human labeling errors or systematic biases introduced by automatic annotation systems (e.g., OCR or MLLM-based labeling). Accordingly, we focus on modeling annotation noise rather than data corruption.

To reflect realistic annotation errors, corrupted labels are not replaced with random characters. Instead, values are substituted with semantically plausible but incorrect alternatives (e.g., valid-looking dates, amounts, or medical terms), preserving surface-level realism while violating field-level correctness. All models are then fine-tuned under identical settings and evaluated on the clean test set.

As shown in Table 2, model performance degrades rapidly as the noise ratio increases. Even moderate noise (30%) leads to a noticeable drop in FMR, while higher noise levels (50% and 70%) cause a severe performance collapse. These results indicate that SFT for document content parsing is highly sensitive to annotation noise, primarily due to the strict field-level exact-match requirements.

We further evaluate the effectiveness of label refinement by applying the Annotation Refinement Model to the 30% Noisy Data, producing a Refined Data set. Fine-tuning on Refined Data yields an FMR of 0.8133, which not only recovers the performance loss caused by noisy supervision but also outperforms all SFT results trained on unrefined noisy annotations. This demonstrates that TLR effectively suppresses systematic annotation noise and provides more reliable supervision for downstream optimization.

Overall, this analysis confirms that annotation noise constitutes a critical bottleneck for SFT in query-driven document parsing, and that training-driven

Table 3: Ablation study results under different model configurations.

Pipeline Configuration		Invoice
Stage-1	**Stage-2**	FMR
(A) Baseline		
baseline		0.2931
(B) Single-Stage Training		
SFT on clean data	—	0.753
SFT on corrected data	—	0.8133
Vanilla RL	—	0.8197
Token-wise GRPO	—	0.8214
(C) Two-Stage Training		
SFT	**SFT**	0.8152
SFT	**Vanilla RL**	0.8198
Vanilla RL	**SFT**	0.8225
Token-wise GRPO	**SFT**	0.8244
(D) Full Pipeline		
Token-wise GRPO	**SFT + DynPrompt**	**0.8263**

Best results per column are in **bold**.

label refinement serves as an effective mechanism for stabilizing learning under realistic noisy supervision.

5.4   Ablation Study

We conduct comprehensive ablation studies to analyze the contribution of each component in the proposed framework and to compare the behavior of SFT and RL under noisy supervision. All variants are evaluated on the medical invoice dataset using the same evaluation protocol, and results are summarized in Table 3.

Baseline and Single-Stage Training. The baseline model achieves an FMR of 0.2931, indicating that direct foundation model under noisy supervision yields limited performance. As shown in Block (B), single-stage SFT on clean data provides only marginal improvement, while SFT on corrected data leads to a substantial gain, demonstrating the importance of refined supervision. Single-stage RL further improves performance, with token-wise GRPO outperforming vanilla RL, highlighting the benefit of fine-grained reward design even without multi-stage training.

RL vs. SFT under Noisy Supervision. Block (C) compares different two-stage training orders. Notably, configurations that apply RL before SFT consistently outperform those that apply SFT before RL. This observation suggests that SFT is more prone to overfitting noisy or partially corrected annotations, whereas RL better tolerates residual noise by optimizing task-level rewards. Applying SFT after RL helps consolidate extraction behavior and formatting, resulting in improved overall performance. These results empirically confirm that RL exhibits higher robustness to noisy supervision than SFT in query-driven document parsing.

Effect of Token-wise GRPO. Across both single-stage and two-stage settings, token-wise GRPO consistently outperforms vanilla RL. This improvement can be attributed to its token-level optimization objective, which directly targets field-level correctness and avoids diluting localized extraction errors that are common in sequence-level reinforcement learning.

Full Pipeline. The full pipeline, which integrates token-wise GRPO followed by SFT with dynamic prompt augmentation, achieves the best performance with an FMR of 0.8534. Compared with the baseline, this represents a significant improvement, validating the complementary roles of training-driven label refinement, noise-aware hybrid post-training. Overall, the ablation results demonstrate that each component contributes to performance gains and that their combination is essential for robust, high-precision document content parsing under realistic noisy supervision.

6   Conclusion

We presented **MeDocVL**, a noise-aware post-training framework for high-precision medical document parsing with vision–language models. By integrating Training-driven Label Refinement and Noise-aware Hybrid Post-training, our approach effectively addresses annotation noise, complex layouts, and strict field-level accuracy requirements. Experimental results demonstrate that MeDocVL consistently outperforms conventional OCR systems and strong general-purpose VLMs.

Beyond medical documents, the proposed framework offers a scalable and cost-effective pathway for adapting VLMs to other high-stakes vertical domains. Moreover, the refined supervision and specialist models produced by our pipeline can serve as reliable data sources for downstream domain specialization and post-training distillation in general-purpose multimodal models.

Contributors

- **Project Leaders:** Zengjian Fan, Wenfeng Xie, Ziling Lin, Ling Huang, Kaihe Xu, Hong Li, De Shi
- **Implementation Contributors:** Wenjie Wang, Wei Wu, Ying Liu, Yuan Zhao, Xiaole Lv, Liang Diao

A   Correction Distillation Objective

A.1   Notation

We define the notation used in the correction distillation stage as follows:

- x: model input, consisting of a document image and a structured instruction prompt.
- $y = (y_1, \ldots, y_T)$: expert-annotated target token sequence.
- $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$: predicted token sequence produced by the refinement model.
- $p_\theta(\cdot \mid x)$: token distribution of the refinement model with parameters θ.
- $p_T(\cdot \mid x)$: teacher distribution induced by pseudo labels (e.g., OCR/MLLM predictions).
- θ_0: frozen parameters of the pre-trained base model.
- $\mathcal{D}_{\exp}$: expert-annotated dataset.
- $\mathcal{D}_{\text{gen}}$: optional replay buffer for representation preservation.
- τ: temperature parameter for distillation.

A.2   Token-level Correction Distillation

To align the refinement model with informative pseudo-label distributions while correcting their deviations, we adopt a token-level knowledge distillation objective. Let $p_T^{(\tau)}$ and $p_\theta^{(\tau)}$ denote the softened teacher and student distributions with temperature τ. The distillation loss is defined as:

$$\mathcal{L}_{\text{KD}} = \frac{1}{|\mathcal{D}_{\exp}|} \sum_{(x,y) \in \mathcal{D}_{\exp}} \frac{1}{T} \sum_{t=1}^{T} \tau^2 \, \text{KL}\Big(p_T^{(\tau)}(\cdot \mid x, y_{<t}) \, \big\| \, p_\theta^{(\tau)}(\cdot \mid x, y_{<t}) \Big). \quad (5)$$

To avoid propagating uncertain supervision, distillation is applied only to tokens where the teacher confidence exceeds a predefined threshold.

A.3   Task-Specific Supervision

To support heterogeneous document structures, we integrate complementary task-specific objectives.

Field-level Classification. For document or field type supervision, we use a cross-entropy loss:

$$\mathcal{L}_{\text{CLS}} = -\mathbb{E}_{(x,c) \in \mathcal{D}_{\exp}} \log p_\theta(c \mid x). \quad (6)$$

Sequence Correction. For text spans extracted from documents, we apply a sequence-level negative log-likelihood loss:

$$\mathcal{L}_{\text{SEQ}} = -\mathbb{E}_{(x,y) \in \mathcal{D}_{\exp}} \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(y_t \mid x, y_{<t}). \quad (7)$$

Spatial Alignment. For structured elements such as tables, we incorporate a localization-aware alignment objective combining L1 distance, IoU loss, and classification loss over matched cells.

A.4   Representation Preservation

To prevent overfitting to the small expert-annotated set and preserve the general multimodal capabilities of the base model, we introduce regularization terms.

Parameter Regularization.

$$\mathcal{L}_{\text{SP}} = \sum_p w_p \left\| \theta_p - (\theta_0)_p \right\|_2^2. \tag{8}$$

Logit Preservation. Optionally, a replay buffer $\mathcal{D}_{\text{gen}}$ is used to enforce logit-level consistency:

$$\mathcal{L}_{\text{KLP}} = \mathbb{E}_{(x,y) \in \mathcal{D}_{\text{gen}}} \frac{1}{T} \sum_{t=1}^{T} \text{KL}\big(p_{\theta_0}(\cdot \mid x, y_{<t}) \,\big\|\, p_\theta(\cdot \mid x, y_{<t})\big). \tag{9}$$

A.5   Overall Objective

The full correction distillation objective is a weighted combination of the above terms:

$$\mathcal{L}_{\text{TOTAL}} = \alpha \mathcal{L}_{\text{KD}} + \beta \mathcal{L}_{\text{CLS}} + \gamma \mathcal{L}_{\text{SEQ}} + \delta \mathcal{L}_{\text{ALIGN}} + \varepsilon \left(\mathcal{L}_{\text{SP}} + \mathcal{L}_{\text{KLP}} \right), \tag{10}$$

where $\alpha, \beta, \gamma, \delta, \varepsilon$ are hyperparameters.

B   Details of Token-wise GRPO

This appendix provides the full formulation and technical details of the proposed token-wise Generalized Reinforcement Preference Optimization (GRPO), which is briefly introduced in Section 4.

B.1   Preference Data Construction

We construct a preference dataset $\mathcal{D} = \{(x, y^+, y^-)\}$, where x denotes the input document (image and instruction prompt), y^+ is a preferred response, and y^- is a dispreferred response.

For document parsing, preferences are derived from: (1) refined annotations produced by the MeDoc-Refiner, (2) alternative candidate outputs generated by the model, and (3) heuristic or rule-based validators (e.g., field consistency, format validity). Unlike reasoning-centric tasks, preferences are typically localized to specific fields or spans rather than the entire sequence.

B.2   Token-level Log-probability Gaps

Let $y^+ = (y_1^+, \ldots, y_T^+)$ and $y^- = (y_1^-, \ldots, y_T^-)$ denote preferred and dispreferred sequences. Given model parameters θ, we define the token-level log-probability gap as:

$$\Delta_t(\theta) = \log p_\theta\big(y_t^+ \mid x, y_{<t}^+\big) - \log p_\theta\big(y_t^- \mid x, y_{<t}^-\big). \tag{11}$$

To prevent excessive deviation from the pre-trained backbone, we additionally introduce a frozen reference policy p_{ref} (Qwen2.5-VL) and compute:

$$\Delta_t^{\text{ref}} = \log p_{\text{ref}}\big(y_t^+ \mid x, y_{<t}^+\big) - \log p_{\text{ref}}\big(y_t^- \mid x, y_{<t}^-\big). \tag{12}$$

These quantities capture relative preference strength at each token position.

B.3   Token-wise GRPO Objective

The token-wise GRPO loss is defined as:

$$\mathcal{L}_{\text{tok-GRPO}} = \mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \left[-\frac{1}{T} \sum_{t=1}^{T} w_t \log \sigma\Big(\beta\big[\Delta_t(\theta) - \kappa\,\Delta_t^{\text{ref}}\big]\Big) \right], \tag{13}$$

where:

- $\sigma(\cdot)$ is the sigmoid function;
- $\beta > 0$ controls preference sharpness;
- $\kappa \geq 0$ regulates reference-policy correction;
- $w_t \in [0,1]$ is a token-level confidence weight.

The confidence weight w_t allows selective learning: tokens corresponding to unreliable OCR spans or ambiguous fields can be down-weighted or masked ($w_t = 0$), while high-confidence tokens receive stronger supervision.

B.4   Relation to GRPO and GSPO

Standard GRPO aggregates token-level rewards into sequence-level advantages, while GSPO further averages rewards to improve stability in long-chain reasoning. For document recognition, this aggregation is suboptimal, as correctness depends on localized token accuracy rather than holistic sequence coherence.

Token-wise GRPO removes this aggregation step and applies preference optimization directly at the token level. This design ensures that fine-grained extraction errors (e.g., single-digit mistakes) are explicitly penalized, which is critical for high-precision document parsing. latex Copy code

C   KL Stabilization and Reference Policy

To stabilize reinforcement learning and preserve the strong general representations of the pre-trained backbone, we incorporate KL regularization with respect to a frozen reference policy.

C.1   Token-level KL Divergence

Given a preferred response y^+ and a dispreferred response y^-, we define the token-averaged KL divergence as:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{KL}} = {} & \mathbb{E}_{(x,y^+)} \left[\frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}\Big(p_\theta(\cdot \mid x, y^+_{<t}) \,\|\, p_{\mathrm{ref}}(\cdot \mid x, y^+_{<t}) \Big) \right] \\
& + \mathbb{E}_{(x,y^-)} \left[\frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}\Big(p_\theta(\cdot \mid x, y^-_{<t}) \,\|\, p_{\mathrm{ref}}(\cdot \mid x, y^-_{<t}) \Big) \right].
\end{aligned}
\tag{14}
$$

This symmetric formulation constrains both preferred and dispreferred trajectories, preventing policy collapse or uncontrolled drift.

C.2   Total Reinforcement Learning Objective

The final RL objective combines token-wise preference optimization with KL regularization:

$$
\mathcal{L}_{\mathrm{RL}} = \mathcal{L}_{\text{tok-GRPO}} + \lambda_{\mathrm{KL}}\, \mathcal{L}_{\mathrm{KL}}, \qquad \lambda_{\mathrm{KL}} \geq 0.
\tag{15}
$$

The hyperparameter λ_{KL} balances stability and adaptability. Larger values enforce conservative updates, while smaller values accelerate domain adaptation. In all experiments, λ_{KL} is selected to ensure stable convergence without sacrificing correction accuracy.

C.3   Discussion

KL stabilization plays a critical role in adapting large multimodal models to domain-specific document parsing tasks. By anchoring optimization to the pre-trained backbone, the model retains general visual–textual representations while gradually acquiring robust correction behaviors under noisy supervision.

References

1. Alibaba Tianchi: Medical invoice ocr dataset. https://tianchi.aliyun.com/dataset/131815 (2023), accessed: 2026-01-29 14
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 4
3. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418 (2023) 4
4. breezedeus: Pix2text: An open-source tool for recognizing layouts, tables, math formulas, and text in images. https://github.com/breezedeus/Pix2Text (2025), accessed: 2025-10-11 4
5. Chen, H., Tu, H., Wang, F., Liu, H., Tang, X., Du, X., Zhou, Y., Xie, C.: Sft or rl? an early investigation into training r1-like reasoning large vision-language models. arXiv preprint arXiv:2504.11468 (2025) 11

6. He, Y., Qi, X., Ye, J., Gao, P., Chen, Y., Li, B., Tang, X., Xiao, R.: Pingan-vcgroup's solution for icdar 2021 competition on scientific table image recognition to latex. arXiv preprint arXiv:2105.01846 (2021) 4

7. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. arXiv preprint arXiv:2409.03420 (2024) 4

8. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia (MM '22). pp. 4083–4091 (2022). https://doi.org/10.1145/3503161.3548112, https://arxiv.org/abs/2204.08387, arXiv preprint arXiv:2204.08387 4

9. Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., et al.: Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186 (2024) 4

10. JaidedAI: Easyocr: A ready-to-use ocr library supporting 80+ languages. https://github.com/JaidedAI/EasyOCR (2025), accessed: 2025-10-11 4

11. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision. pp. 498–517. Springer (2022) 4

12. Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., et al.: Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. arXiv preprint arXiv:2206.03001 (2022) 4

13. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025) 4

14. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26763–26773 (2024) 4

15. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024) 4

16. Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., Bai, X.: Textmonkey: An ocr-free large multimodal model for understanding document. arXiv preprint arXiv:2403.04473 (2024) 4

17. Liu, Y., Zhang, J., Peng, D., Huang, M., Wang, X., Tang, J., Huang, C., Lin, D., Shen, C., Bai, X., et al.: Spts v2: single-point scene text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(12), 15665–15679 (2023) 4

18. Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., Vagenas, P., Ramis, C.B., Omenetti, M., Dinkla, K., Kim, Y., Gupta, S., de Lima, R.T., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., Staar, P.W.J.: Docling: An efficient open-source toolkit for ai-driven document conversion (2025), https://arxiv.org/abs/2501.17887 4

19. Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., Jin, Z., Liang, G., Zhang, R., Zhang, W., Qu, Y., Ren, Z., Sun, Y., Zheng, Y., Ma, D., Tang, Z., Niu, B., Miao, Z., Dong, H., Qian, S., Zhang, J., Chen, J., Wang, F., Zhao, X., Wei, L., Li, W., Wang, S., Xu, R., Cao, Y., Chen, L., Wu, Q., Gu, H., Lu, L., Wang, K., Lin, D., Shen, G., Zhou, X., Zhang, L., Zang, Y., Dong, X., Wang, J., Zhang, B., Bai, L., Chu, P., Li, W., Wu, J., Wu, L., Li, Z., Wang, G., Tu, Z., Xu, C., Chen, K., Qiao, Y., Zhou, B., Lin, D., Zhang, W.,

He, C.: Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing (2025), https://arxiv.org/abs/2509.22186 4

20. OpenAI: Hello GPT-4o. https://openai.com/index/hello-gpt-4o (2024), accessed: YYYY-MM-DD 4

21. Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., Altman, E.: Tabular transformers for modeling multivariate time series. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3565–3569. IEEE (2021) 4

22. Peng, D., Wang, X., Liu, Y., Zhang, J., Huang, M., Lai, S., Li, J., Zhu, S., Lin, D., Shen, C., et al.: Spts: single-point text spotting. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4272–4281 (2022) 4

23. Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Wilhelm, C., Lo, K., Soldaini, L.: olmocr: Unlocking trillions of tokens in pdfs with vision language models. arXiv preprint arXiv:2502.18443 (2025) 4

24. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4

25. Wan, J., Song, S., Yu, W., Liu, Y., Cheng, W., Huang, F., Bai, X., Yao, C., Yang, Z.: Omniparser: A unified framework for text spotting key information extraction and table recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15641–15653 (2024) 4

26. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems **37**, 107984–108011 (2024) 4

27. Wang, B., Gu, Z., Liang, G., Xu, C., Zhang, B., Shi, B., He, C.: Unimernet: A universal network for real-world mathematical expression recognition. arXiv preprint arXiv:2404.15254 (2024) 4

28. Wang, Z., Xu, Y., Cui, L., Shang, J., Wei, F.: Layoutreader: Pre-training of text and layout for reading order detection. arXiv preprint arXiv:2108.11591 (2021) 4

29. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024) 4

30. Xia, R., Mao, S., Yan, X., Zhou, H., Zhang, B., Peng, H., Pi, J., Fu, D., Wu, W., Ye, H., et al.: Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. arXiv preprint arXiv:2406.11633 (2024) 4

31. Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., et al.: Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. arXiv preprint arXiv:2508.06471 (2025) 4

32. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024). https://doi.org/10.1109/CVPRXXXX.XXXXXXX, arXiv preprint arXiv:2304.08069, version v3 4

33. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 4