Switching Heads and Softening Tokens: Turnkey Solutions to Visually Grounded Document QA

Ximing Wen[†*] **Wenbo Li**[*] **Sudipta Paul** **Yashas Malur Saidutta**
Kalpa Gunaratna **Srinivas Chappidi**
AI Center-Mountain View, Samsung Electronics
xw384@drexel.edu wenbo.li1@samsung.com

Abstract

Visually Grounded Document Question Answering often lacks robust, end-to-end solutions capable of handling complex, multi-answer queries without reliance on ad-hoc processing. In this work, we propose two turnkey LLM architectures to address this gap. We first introduce a single-head architecture where coordinates are represented as special tokens within the unified vocabulary. While structurally robust, this approach suffers from the limitations of discrete supervision; to address this, we propose a novel "softening token" method that enables differentiable Mean-Squared-Error loss over token probabilities. Although this significantly improves visual grounding, the spatial precision remains bound by discretization. Consequently, we propose a second solution: a dual-head architecture that alternates between text generation and regression-based bounding box prediction. This method offers high spatial precision via a regression head, further stabilized by our introduction of an Intersection-over-Union loss. Finally, by combining the single head model's structural robustness with the high precision of the dual head model, we propose an ensemble method that yields significant performance gains beyond each of individual components.

1 Introduction

Document Question Answering (DQA) addresses the challenge of interpreting natural-language queries directed at visually rich and semi-structured documents, such as invoices, forms, receipts, contracts, scientific papers, and business reports. This task holds significant practical value across various sectors, including finance, enterprise automation, education, and legal or government services. Representative applications range from extracting key fields in receipts and invoices and locating specific clauses in legal documents, to retrieving information from multi-page PDFs. Furthermore, it plays a vital role in document-based customer support and accessibility tools designed to interpret dense or complex layouts.

Current DQA models generally fall into two primary categories: (i) OCR-based multimodal models, such as the LayoutLM series (Xu et al., 2021; Huang et al., 2022), DocFormerv2 (Appalaraju et al., 2024), DocLayLLM (Liao et al., 2025), DocLLM (Wang et al., 2024), LayTextLLM (Lu et al., 2025), and LayTokenLLM (Zhu et al., 2025); and (ii) OCR-free vision–language models, including Donut (Kim et al., 2022), Pix2Struct (Lee et al., 2023), mPLUG-DocOwl2 (Hu et al., 2025), Qwen2.5-VL (Bai et al., 2025), and InternVL3.5 (Wang et al., 2025). Despite substantial progress in the field, accurately visually grounding answers within documents remains a significant hurdle, limiting both the interpretability and real-world applicability of these models.

While recent initiatives (Mohammadshirazi et al., 2025; Chen et al., 2025; Zhou et al., 2025) have attempted to bridge this gap, they fall short of offering a turnkey solution for Visually Grounded DQA (VGDQA). These approaches often rely on oversimplified assumptions regarding user query complexity and depend on auxiliary components external to the LLM or VLM. Specifically, they assume a user's question yields a single, contiguous answer. This contradicts the reality of complex queries that may require multiple distinct answers; for example, "What is the departure date, time, and terminal for my flight to Seattle?" A robust system must identify three separate answers ("departure date," "departure time," and "departure terminal") and visually ground them with three distinct bounding boxes. Furthermore, reliance on ad-hoc pre- or post-processing prevents these solutions from being end-to-end trainable, thereby capping their

[*]Equal contribution. [†]Currently at Drexel University. Work done during an internship at AI Center-Mountain View, Samsung Electronics.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36490–36503

potential performance.

In this work, we explore the adaptation of LLMs into turnkey VGDQA solutions, proposing two distinct architectures, each with unique advantages and drawbacks. Given the maturity of modern OCR technology (Wei et al., 2025), we design our models as OCR-based multimodal models.

Our first proposed architecture is a *single-head solution*. Inspired by (Lu et al., 2025), we introduce 1,001 special tokens (<B0> through <B1000>) to represent bounding box coordinates. This allows the text head to output four special tokens representing a bounding box immediately after generating an answer, eliminating the need for complex architectural changes. However, supervising this architecture is challenging. While Cross-Entropy (CE) loss is effective for classification, it lacks spatial awareness; it penalizes a prediction of <B499> (spatially accurate) just as harshly as <B100> (spatially distant) when the target is <B500>. To overcome this lack of nuance, we propose a *softening token* method. This computes the expected numeric coordinate value based on softmax probabilities, enabling the application of a differentiable Mean-Squared-Error (MSE) loss over the discrete token space. Despite these improvements, the single-head solution remains limited by the discretization of coordinates, which caps its spatial precision.

To address the precision limitations of the single-head approach, we propose a second architecture: a *dual-head solution*. This model features separate decoder heads: a text head for answers and a box head for decoding hidden states directly into continuous bounding box coordinates. Drawing inspiration from (Lai et al., 2024), we utilize a special token, <SWITCH_HEAD>, to govern the transition between these heads. To stabilize training and ensure high precision, specifically for small bounding boxes where standard MSE loss often fails to ensure overlap, we incorporate a scale-invariant Intersection-over-Union (IoU) loss. While the dual-head solution achieves superior spatial accuracy, it suffers from a specific structural drawback: its reliance on the head-switching mechanism. Errors in this mechanism can lead to the generation of a non-negligible number of extraneous answers or "hallucinated" switching events.

To maximize performance, we propose an *ensemble method* that combines the strengths of both architectures. We use the single-head model as a structural anchor—leveraging its stability in determining the set of answers and their initial visual grounding—while using the dual-head model's output to refine the spatial coordinates, thereby injecting regression-based precision into the final result.

2 Related Works

Our work on VGDQA is closely related to DQA, with the key distinction being that VGDQA requires visually grounding the answers, whereas DQA does not. As discussed in §1, existing DQA models generally fall into two categories: OCR-based and OCR-free. Since OCR-based models align more closely with our proposed solution, we focus our review on this area.

Early efforts, such as the LayoutLM series (Xu et al., 2021; Huang et al., 2022), pioneered multimodal pretraining specifically for document understanding, enabling the pretrained models to be finetuned for DQA. Similarly, DocFormerv2 (Appalaraju et al., 2024) also explored multimodal pretraining but concentrated on enhancing local-feature alignment across different modalities through carefully designed unsupervised tasks.

More recent research (Liao et al., 2025; Wang et al., 2024; Lu et al., 2025; Zhu et al., 2025) has focused on extending powerful pretrained LLMs to handle DQA. DocLayLLM (Liao et al., 2025) provides an efficient multimodal extension by integrating visual patch tokens, 2D positional tokens, and OCR information into the LLM's input. It also fully integrates the concept of chain-of-thought into every stage of its training process. DocLLM (Wang et al., 2024) takes a different approach by avoiding the costly image encoder used in DocLayLLM. Instead, it relies exclusively on bounding box information, incorporating spatial layout structure via a redesigned attention mechanism.

To maximize the utility of existing foundational models, other approaches have minimized architectural changes. LayTextLLM (Lu et al., 2025) only incorporates an additional lightweight layout tokenizer and leverages Low-Rank Adaptation (LoRA) (Hu et al., 2022) for finetuning. It fully exploits the autoregressive nature of LLMs by using an input composed of interleaved text and layout tokens. LayTokenLLM (Zhu et al., 2025) builds upon LayTextLLM with two main improvements: introducing a specialized positional encoding to eliminate the need for extra position IDs for layout tokens, and proposing a novel pretraining objective called Next Interleaved Text and Layout Token Prediction (NTLP) to boost cross-modality learning.

The models reviewed above were primarily designed for DQA and do not inherently address the visual grounding requirement of VGDQA. A notable exception is LayTextLLM, which can be adapted for VGDQA. It achieves this by introducing 1,000 special tokens to represent bounding box coordinates and is trained to generate the extracted key information and its corresponding bounding box in an interleaved fashion. However, LayTextLLM uses a standard CE loss to supervise both text and bounding box generation, which as noted in §1, can result in imprecise supervision due to its lack of nuance.

Other efforts to tackle VGDQA often involve enhancing VLMs with specialized pre- or post-processing techniques. DLaVA (Mohammadshirazi et al., 2025) proposes a training-free pipeline for zero-shot visual grounding. It first calls a VLM to generate the initial answer, and cleverly bypasses OCR by creating a single image that contains the detected text regions along with unique bounding box identifiers. This composite image, along with the query, initial answer and bounding box information, is then fed to a VLM to generate the final answer and its corresponding bounding box. DocExplainerV0 (Chen et al., 2025) introduced a plug-and-play, post-processing module specifically for bounding-box prediction, decoupling the visual grounding task from the answer generation process.

Most recently, the Document Grounding and Referring data engine (DOGR-Engine) (Zhou et al., 2025) was introduced to generate synthetic data for pretraining and finetuning VLMs across a spectrum of tasks, including VGDQA. While the resulting DOGR model addresses VGDQA, it stops short of providing a turnkey solution due to its reliance on a rigid post-processing step that maps generated bounding boxes to input word-level coordinates. This dependency limits generalizability, particularly when answers span multiple words or lines, rendering simple word-matching infeasible. In contrast to DOGR, which treats VGDQA as one component of a multi-task framework, BoundingDocs (Giovannini et al., 2025) establishes a specialized dataset dedicated exclusively to this problem. Accordingly, we utilize this focused benchmark to evaluate our proposed methods.

3 Methodology

To address the limitations of existing VGDQA approaches, we propose two distinct architectures for end-to-end VGDQA: a single-head solution (§3.1) and a dual-head solution (§3.2). Both architectures utilize a backbone LLM initialized with pre-trained weights and accept interleaved sequences of text and layout tokens as input. We also propose a simple yet effective ensemble method (§3.3) that combines the best of both solutions. In §3.4, we present the training strategy for our solutions.

3.1 Single-head Solution

We first propose a single-head solution that treats both the answer and bounding box generation as a unified token generation task. This approach avoids complex architectural modifications and allows the model to be trained as a standard autoregressive language model with a single language head.

Architecture. Inspired by (Lu et al., 2025), we discretize the coordinate space into integer bins. We expand the LLM vocabulary with 1,001 special tokens, {<B0>, . . . , <B1000>}, representing normalized coordinates from 0.0 to 1.0. In this setup, a bounding box is represented as a sequence of four tokens appended immediately after the answer text. This setup is structurally robust: the model naturally learns the pattern of following an answer with four coordinate tokens (top left and bottom right coordinates), rarely failing to produce the correct format.

Softening Token for Differentiable MSE. Training an autoregressive language model typically relies on Cross-Entropy loss ($\mathcal{L}_{\text{CE}}$). While $\mathcal{L}_{\text{CE}}$ is effective for text, it lacks a notion of spatial proximity. For example, if the ground truth token is <B500>, $\mathcal{L}_{\text{CE}}$ penalizes a prediction of <B499> (spatially adjacent) just as harshly as <B100> (spatially far-off compared to <B499>). This lack of ordinal awareness makes convergence sluggish and imprecise. Ideally, we would apply Mean-Squared-Error (MSE) loss to enforce spatial proximity. However, standard MSE requires continuous numerical values and is not directly compatible with discrete token classification, nor is a discrete lookup operation differentiable. To solve this, we propose a **Softening Token** mechanism.

Instead of taking the `argmax` of the logits, we compute the *expected coordinate value* from the probability distribution over the coordinate tokens. Let V_{coord} be the set of coordinate tokens {<B0>, . . . , <B1000>} and $v(t_i) \in [0, 1]$ be the normalized value associated with token t_i. The expected

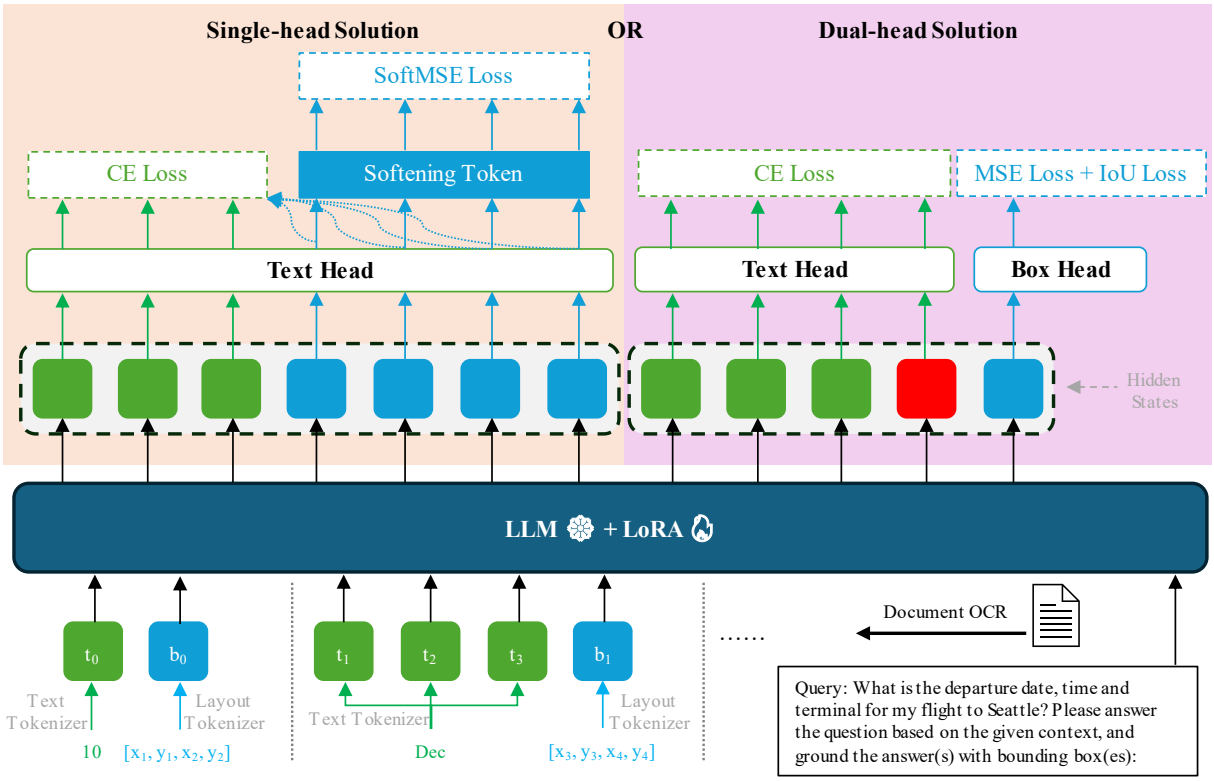

Figure 1: Single-head solution vs. dual-head solution. The green indicates the input, output and supervision for the text, and the blue indicates those for the layout or bounding box. For the single-head solution, after the model generates an answer that corresponds to the three green hidden states, it immediately generates the bounding box represented as four coordinate tokens that corresponds to the four blue hidden states. For the dual-head solution, the red hidden state corresponds to the control token, <SWITCH_HEAD>, after the emission of which the model pauses text decoding and routes the current hidden state (in blue) to the box head for decoding the bounding box.

coordinate prediction $\hat{c}$ is computed as:

$$\hat{c} = \sum_{t_i \in V_{\text{coord}}} \text{softmax}(z)_i \cdot v(t_i) \quad (1)$$

where z represents the logits for all tokens. Because the softmax operation is differentiable, we can now compute a valid MSE loss between the expected value $\hat{c}$ and the ground truth coordinate c:

$$\mathcal{L}_{\text{SoftMSE}} = \|\hat{c} - c\|^2 \quad (2)$$

This allows us to combine the categorical precision of Cross-Entropy with the distance-aware supervision of MSE. The total loss for supervising the bounding box generation of the single-head architecture is defined as:

$$\mathcal{L}_{\text{Single}} = \mathcal{L}_{\text{CE}} + \gamma \mathcal{L}_{\text{SoftMSE}} \quad (3)$$

This hybrid objective ensures that the model not only selects the correct coordinate token but also pushes the probability mass toward the spatial neighborhood of the ground truth, significantly improving grounding accuracy for complex queries.

3.2 Dual-head Solution

While this hybrid objective significantly improves performance, the single-head solution is inherently limited by the discretization granularity. To address the precision limitations of the single-head model, our second architecture decouples the generation of semantic content and spatial coordinates using distinct decoder heads.

Architecture and Head Switching. The dual-head architecture shares a common transformer backbone but bifurcates at the final layer into two distinct decoder heads: a **text head** (a standard LM head) and a **box head** (an MLP regressor). To coordinate these heads, we introduce a control token, <SWITCH_HEAD>. The generation process begins with the text head. When the model generates a textual answer segment that requires visual grounding, it emits the <SWITCH_HEAD> token. This token acts as a rigorous gate: upon its generation, the model pauses text decoding and routes the current hidden state to the box head. The box head then predicts a bounding box tuple $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$.

Once the coordinate prediction is complete, control reverts to the text head to continue generating the next answer or the <END_OF_TEXT> token.

Optimizing Spatial Supervision. A critical challenge in training dual-head architectures is ensuring that generated boxes accurately overlap with ground-truth regions, particularly for small visual elements. Following (Zhu et al., 2025), we initially employed a combined loss function consisting of CE loss for the text head and MSE loss $\mathcal{L}_{\text{MSE}}$ for the box head. However, we observed that $\mathcal{L}_{\text{MSE}}$ is inherently sensitive to the scale of the bounding box. For small objects (*e.g.*, a date), a slight shift in coordinates results in a numerically small MSE value, even if the predicted box fails to overlap with the ground truth entirely (the "non-overlap" issue). This leads to optimization instability where the model converges to low loss values without achieving functional grounding accuracy.

To rectify this, we incorporate the IoU loss:

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{|b_{\text{pred}} \cap b_{\text{gt}}|}{|b_{\text{pred}} \cup b_{\text{gt}}|}, \tag{4}$$

where b_{pred} and b_{gt} are the predicted and ground truth bounding boxes, respectively. Unlike MSE, IoU-based losses are scale-invariant, ensuring that errors in small bounding boxes are penalized as heavily as those in larger ones. This forces the model to prioritize the structural alignment of the predicted box with the ground truth. Our final objective function for the dual-head solution is:

$$\mathcal{L}_{\text{Dual}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}} \tag{5}$$

Empirically, we find that the addition of $\mathcal{L}_{\text{IoU}}$ stabilizes training and significantly reduces the non-overlap rate for fine-grained document elements.

3.3 Ensemble Method

The two architectures exhibit complementary characteristics. The single-head solution is structurally robust—it reliably generates the correct number of answers—but is spatially coarser. The dual-head solution is spatially precise but prone to head-switching artifacts (extra or missing answers and bounding boxes).

To leverage the best of both, we propose an ensemble strategy that treats the single-head output as the structural "anchor". We retain the answers and the count of predictions from the single-head model but attempt to refine the spatial coordinates using the dual-head model's output.

Algorithm 1 Ensemble Strategy for VGDQA

Require: Set of single-head predictions $P_S = \{(a_i^s, b_i^s)\}_{i=1}^{N}$
Require: Set of dual-head predictions $P_D = \{(a_j^d, b_j^d)\}_{j=1}^{M}$
Ensure: Final predictions P_F

$P_F \leftarrow \emptyset$
for each pair (a_i^s, b_i^s) in P_S **do**
  $b_{\text{final}} \leftarrow b_i^s$ ▷ Initialize with single-head box
  Candidates $\leftarrow \emptyset$ ▷ Find all dual-head predictions with matching text
  for each pair (a_j^d, b_j^d) in P_D **do**
    if $a_j^d = a_i^s$ **then**
      Candidates $\leftarrow$ Candidates $\cup \{b_j^d\}$
    end if
  end for
  ▷ Select the spatially closest candidate
  if Candidates $\neq \emptyset$ **then**
    $b_{best} \leftarrow \underset{b \in \text{Candidates}}{\text{argmax}} \ \text{IoU}(b_i^s, b)$
    if $\text{IoU}(b_i^s, b_{\text{best}}) > 0$ **then**
      $b_{\text{final}} \leftarrow b_{\text{best}}$
    end if
  end if
  $P_F \leftarrow P_F \cup \{(a_i^s, b_{\text{final}})\}$
  $P_D \leftarrow P_D \setminus \{(a_i^s, b_{\text{final}})\}$
end for
return P_F

For every answer-box pair generated by the single-head model, we query the dual-head model's output for an identical text answer. If a match is found, we identify all dual-head candidate boxes associated with that answer string and select the one with the highest IoU with the single-head anchor box, provided that at least one candidate achieves an IoU greater than zero; otherwise, the single-head prediction is retained as the final result. This strategy effectively filters out dual-head hallucinations while injecting regression-based spatial precision into the structurally robust single-head predictions. The procedure is outlined in Algorithm 1.

3.4 Training Strategy

Following the curriculum learning paradigm established by Zhu et al. (2025), we train models in two distinct stages: self-supervised pretraining using the Next Interleaved Text and Layout Token Prediction (NTLP) objective, followed by Supervised Fine-Tuning (SFT) on VGDQA.

Stage 1: NTLP Pretraining. In the initial phase, the model is trained to reconstruct document content from OCR data. The NTLP objective requires the model to predict the next text token or layout coordinate given the preceding context. This effectively aligns the textual and spatial modalities, enabling the LLM to process the interleaved input format described in §1. During this stage, we optimize the layout tokenizer, text embeddings, the respective decoder heads (text and box heads for the dual-head model; text head for the single-head model), and the LoRA matrices injected into the frozen base LLM.

Stage 2: SFT. In the second stage, we finetune the models using the VGDQA dataset. Here, the objective shifts from generic document reconstruction to question answering and visual grounding. To ensure training stability and parameter efficiency, we freeze the layout tokenizer, text embeddings, and the decoder heads optimized during pretraining. Gradients are computed exclusively for the LoRA matrices of the base LLM. This focused updating strategy prevents catastrophic forgetting of the structural knowledge acquired during NTLP while allowing the model to adapt to the specific prompt-response dynamics of VGDQA.

Optimization and Loss Consistency. A key feature of our strategy is the consistency of the objective functions across both training stages. For the **single-head solution**, $\mathcal{L}_{\text{Single}}$ (3) is applied throughout the training stages, maintaining the hybrid supervision of Cross-Entropy and SoftMSE to enforce spatial proximity in the token space. For the **dual-head solution**, the model is supervised using $\mathcal{L}_{\text{Dual}}$ (5) in both NTLP and SFT stages, ensuring that the box head continuously refines its regression capabilities under the guidance of IoU and MSE losses.

By leveraging LoRA, the vast majority of parameters—specifically the weights of the backbone LLM—remain frozen and shared. This design allows for a highly efficient training pipeline where the architectural differences between the dual-head and single-head solutions incur minimal computational overhead.

4 Experiments

4.1 Datasets

Pretraining Data. To align the visual and textual modalities, we follow LayTokenLLM (Zhu et al., 2025) to utilize the **Layout-aware SFT** dataset (Luo et al., 2024). This corpus comprises a diverse ensemble of high-quality documents tailored for document understanding and information extraction. To ensure computational efficiency during the pretraining phase, we filter the dataset to exclude documents exceeding a token length of 2K.

SFT Data. For the downstream VGDQA task, we employ **BoundingDocs v2.0**, an extended version of the BoundingDocs dataset (Giovannini et al., 2025). This dataset consolidates data from 11 public sources—spanning invoices, contracts, forms, receipts, and multilingual corpora—into a unified question-answering format. Crucially, it provides normalized bounding boxes for precise answer localization. The dataset consists of 48,151 documents and 249,016 question–answer pairs, serving as a robust benchmark for evaluating both semantic correctness and visual grounding accuracy. We adhere to the standard training, validation, and test splits provided by the dataset.

4.2 Implementation Details

We employ **Qwen3-8B** (Yang et al., 2025) as the backbone LLM and the layout tokenizer from (Zhu et al., 2025) for both our architectures. Other implementation details can be found in §A.1.

4.3 Baselines

To validate the efficacy of our proposed turnkey solutions, we benchmark them against several categories of baselines. First, we compare against state-of-the-art VGDQA-specific methods, including DLaVA (Mohammadshirazi et al., 2025), DocExplainerV0 (Chen et al., 2025), and DOGR (Zhou et al., 2025). While these architectures are primarily designed as OCR-free solutions, DLaVA and DocExplainerV0 also introduce strong OCR-based variants, which we include in our comparison and distinguish with a †in Table 1. Second, we include LayTextLLM (Lu et al., 2025), an OCR-based model that represents the closest prior work to our single-head architecture in its treatment of bounding box coordinates as interleaved tokens. Third, to contextualize our results against general-purpose vision-language models not specifically designed for VGDQA, we evaluate InternVL3.5-8B (Wang et al., 2025) and Qwen3-VL-8B (Bai et al., 2025) in a zero-shot setting. We also attempted to include DeepSeek-OCR (Wei et al., 2025) as a baseline; however, we were unable to prompt the

Table 1: Quantitative results on BoundingDocs v2.0 dataset. Our solutions are highlighted in light green. † marks methods that are OCR-free, and all the other methods are OCR-based. Note that Finetuned DOGR is a VLM baseline. $*$ marks methods that require two LLM calls. ↓ indicates lower is better; ↑ indicates higher is better. The best results are highlighted in **bold**, and the second best are highlighted using underline.

Methods	Overall Metrics		Metrics across Answer Count					Metrics across Bbox Sizes	
	NCE↓	Acc↑	Acc-1	Acc-2	Acc-3	Acc-4	Acc-≥5	Acc-Small	Acc-Large
DLaVA†*	.570	0	.010	0	.068	0	0	.002	.015
DLaVA	.532	.168	.235	.133	.139	.096	.014	.160	.184
DocExplainerV0†	.420	.030	.037	.042	.032	.035	.006	.029	.030
DocExplainerV0	.868	.025	.039	.061	.031	0	0	.004	.068
Finetuned DOGR†	.900	.001	.001	.002	0	0	0	.001	0
Finetuned Qwen3-8B	.101	.655	.739	.647	.672	.608	.444	.593	.788
LayTextLLM	.414	0	0	0	0	0	0	0	.002
Qwen3-VL-8B†	.440	.252	.358	.246	.193	.125	.032	.205	.353
InternVL3.5-8B†	.497	.001	.001	.001	0	0	0	0	.002
Single-head model	**.086**	.796	.866	.816	.827	.775	.618	.758	**.879**
Dual-head model	.113	.854	.883	.838	.861	.838	.793	.847	.871
Ensemble method*	**.086**	**.866**	**.891**	**.845**	**.880**	**.862**	**.813**	**.862**	.875

model to generate the answer and its corresponding bounding box simultaneously.[1] We therefore exclude it from the quantitative comparison.

Regarding the experimental setup, DLaVA operates as a training-free method, and DocExplainerV0 was originally trained on BoundingDocs v2.0; consequently, neither required retraining. In contrast, due to the incompleteness of the publicly released weights for DOGR, we reproduced the model by adhering to its original pre-alignment and pre-training protocols, followed by fine-tuning on BoundingDocs v2.0. For LayTextLLM, we use the publicly released LayTextLLM-Zero checkpoint[2] provided in the official repository, without any further fine-tuning. Finally, we establish a vanilla LLM baseline by pre-training and fine-tuning Qwen3-8B on our datasets. This baseline does not utilize special coordinate tokens but adopts the DOGR output format (*i.e.*, <ocr>answer</ocr><box>x_1, y_1, x_2, y_2</box>).

4.4 Evaluation Metrics

We employ a comprehensive suite of metrics to evaluate performance across answer generation, grounding accuracy, and robustness to query complexity.

Evaluation Metrics. We report **Grounding Accuracy (Acc)**, following the protocol in DOGR (Zhou et al., 2025). A prediction is considered accurate only if: (1) the generated text matches the ground truth, AND (2) the IoU between the predicted bounding box and the ground truth exceeds a threshold of 0.5.

We also define the **Normalized Cardinality Error (NCE)** to assess the model's ability to generate a correct number of answers. This metric accounts for: Extra (the number of hallucinated or extra answers generated), Missing (the number of missing answers), and GT (the number of ground-truth answers), which is defined as:

$$\text{NCE} = \frac{\text{Extra} + \text{Missing}}{\text{GT}} \tag{6}$$

Metrics across Answer Count. To evaluate robustness against multi-answer queries—a key motivation of this work—we stratify Accuracy based on the number of required answers per question: **Acc-1** (single answer), **Acc-2**, **Acc-3**, **Acc-4**, and **Acc-≥5** (five or more answers).

Metrics across Box Sizes. To assess spatial precision, we categorize ground truth bounding boxes into small and large clusters using K-means clustering. We report **Acc-Small** and **Acc-Large** to highlight the model's performance on fine-grained visual elements versus dominant layout features.

[1] We tried the following prompt formats: (1) <image>\n<|grounding|>QUESTION, and what is the bounding box for the answer?; (2) <image>\n<|grounding|>answer the question with OCR boundingbox: QUESTION; (3) <image>\n<|grounding|>Free OCR. Then answer the question with grounding: QUESTION. None of these elicited joint answer and bounding box generation.

[2] https://huggingface.co/LayTextLLM/LayTextLLM-Zero

Table 2: Ablation study. The ablative versions of our solutions are highlighted in light purple, respectively.

Methods	Overall Metrics		Metrics across Answer Count					Metrics across Bbox Sizes	
	NCE↓	Acc↑	Acc-1	Acc-2	Acc-3	Acc-4	Acc-≥5	Acc-Small	Acc-Large
Single-head model w/o $\mathcal{L}_{\text{SoftMSE}}$	.099	.762	.845	.795	.797	.742	.549	.712	.867
Single-head model	**.086**	.796	.866	.816	.827	.775	.618	.758	**.879**
Dual-head model w/o $\mathcal{L}_{\text{IoU}}$	.270	.042	.063	.012	.020	.011	.004	.011	.108
Dual-head model	.113	**.854**	**.883**	**.838**	**.861**	**.838**	**.793**	**.847**	.871

5 Results and Analysis

5.1 Quantitative Results

The quantitative evaluation on the BoundingDocs v2.0 dataset is detailed in Table 1. Our proposed **ensemble method** achieves the highest overall performance (Acc $= 0.866$), establishing a new state-of-the-art by effectively synthesizing the structural robustness of the single-head architecture with the spatial precision of the dual-head regression.

Comparison with Baselines. As evidenced by the experimental results in Table 1, existing baselines exhibit significant performance deficits compared to our turnkey solutions. The pipeline-based approaches, DLaVA and DocExplainerV0, perform poorly (Acc < 0.17). This stems primarily from their lack of end-to-end optimization; they rely on generic, training-free VLMs coupled with ad-hoc pre- or post-processing, which prevents the holistic learning of visual grounding. LayTextLLM, despite sharing a similar token-interleaving design with our single-head model, achieves zero grounding accuracy, suggesting that its training objective does not adequately supervise bounding box prediction for VGDQA. General-purpose VLMs, including Qwen3-VL-8B (0.252) and InternVL3.5-8B (0.001), similarly underperform despite their strong multimodal capabilities, suggesting that general visual understanding does not transfer directly to the precise spatial supervision required by VGDQA. Furthermore, while the fine-tuned DOGR and Qwen3-8B models benefit from training on the target data, they still consistently underperform compared to our proposed methods. This performance gap underscores a fundamental limitation of treating bounding box coordinates as standard textual tokens: such a formulation precludes the integration of geometric loss functions (*e.g.*, IoU or MSE), thereby depriving the model of the inductive bias necessary for fine-grained spatial convergence. Notably, despite specialized finetuning, DOGR struggles more significantly than the vanilla Qwen3-8B baseline in generating the correct set of answers, suggesting that effectively fine-tuning VLMs for the specific demands of VGDQA remains a non-trivial challenge.

Analysis of Proposed Architectures. Our individual solutions demonstrate distinct complementary strengths. The **single-head model** excels in structural consistency, yielding the lowest Extra and Missing Ratio (NCE $= 0.086$), indicating a superior ability to determine the correct cardinality of answers. Conversely, the **dual-head model** dominates in spatial precision, particularly for fine-grained elements, achieving an Acc-Small of 0.847. This confirms that the regression head, guided by continuous loss functions, resolves spatial details better than discretized token classification.

Robustness Analysis. We observe a natural performance drop as query complexity increases. For multi-answer queries where the answer count $N \geq 5$, accuracy drops across all models, reflecting the compounded difficulty of retrieving and grounding multiple distinct entities. Additionally, performance on larger bounding boxes (Acc-Large) consistently exceeds that of smaller ones (Acc-Small). This is attributable to the fact that larger visual elements typically possess more distinct visual features and are less sensitive to minor coordinate regressions than smaller, dense text regions.

Performance Across Document Lengths To assess whether performance degrades on longer documents, we partition BoundingDocs v2.0 into a Long-Context (LC) subset comprising documents exceeding 4K tokens and a Short-Context (SC) subset comprising those falling below this threshold, and report stratified metrics in Table 4. The ensemble method achieves an Acc of 0.873 on SC documents and 0.848 on LC documents, with NCE increasing from 0.075 to 0.118. The degradation in both metrics indicates that longer documents pose a meaningful challenge, likely due to the difficulty

Table 3: Grounding accuracies across different languages on BoundingDocs v2.0 dataset.

Methods	Grounding Accuracies across Languages							
	English	Italian	French	Spanish	Chinese	German	Portuguese	Japanese
Ensemble method	.879	.613	.648	.585	.338	.444	.532	.333

Table 4: Overall performance for Long Context (LC) and Short Context (SC). ↓ indicates lower is better; ↑ indicates higher is better.

Methods	Metrics for LC		Metrics for SC	
	NCE↓	Acc↑	NCE↓	Acc↑
Ensemble method	.118	.848	.075	.873

of maintaining precise attention over extended input sequences. We identify this as a limitation and an area for future work.

Cross-Lingual Grounding Analysis BoundingDocs v2.0 contains documents in eight languages, enabling a stratified analysis of grounding accuracy across linguistic groups (Table 3). The ensemble method achieves strong performance on English (0.879) but exhibits a notable drop for non-Latin-script languages, with Chinese (0.338) and Japanese (0.333) performing substantially below the overall average. As the majority of training documents in BoundingDocs v2.0 are English-based, the reduced performance on lower-resource languages is consistent with data imbalance rather than a fundamental architectural limitation. This result directly quantifies the bias acknowledged in our Ethical Considerations section.

Inference Cost The inference cost analysis is provided in §A.3

5.2 Qualitative Results

The qualitative results are provided in §A.2.

6 Ablation Study

To isolate the contributions of our proposed training objectives, we conducted an ablation study, the results of which are highlighted in Table 2.

Efficacy of Softening Token (Single-Head). We trained a variant of the single-head model using only standard Cross-Entropy loss, removing the $\mathcal{L}_{\text{SoftMSE}}$ component. While the semantic structure remained stable (comparable NCE), the overall grounding accuracy declined from 0.796 to 0.762. The degradation is most pronounced in complex queries (Acc- $\geq$ 5 drops from 0.618 to 0.549). This validates that Softening Token successfully injects ordinal awareness into the discrete vocabulary, enabling the model to learn spatial proximity and significantly improving localization for dense document layouts.

Necessity of IoU Loss (Dual-Head). The impact of the Intersection-over-Union loss ($\mathcal{L}_{\text{IoU}}$) on the dual-head architecture is critical. Removing $\mathcal{L}_{\text{IoU}}$ causes a catastrophic collapse in performance, with grounding accuracy plummeting from 0.854 to 0.042. Without the scale-invariant properties of IoU, the MSE loss fails to penalize non-overlapping predictions for small objects effectively, leading to an Acc-Small of nearly zero (0.011). Furthermore, we observed that training without $\mathcal{L}_{\text{IoU}}$ resulted in severe numerical instability, leading to gradient divergence (NaN loss) after the second epoch. This confirms that $\mathcal{L}_{\text{IoU}}$ is essential not only for enforcing spatial overlap but also for stabilizing the optimization of the regression head.

7 Conclusion

We presented two turnkey LLM architectures for VGDQA that eliminate the need for ad-hoc processing. We first introduced a single-head architecture that bridges the gap between discrete tokens and continuous coordinates via a novel softening token mechanism with differentiable MSE loss. Second, we proposed a dual-head architecture that achieves high spatial precision through a regression head stabilized by IoU loss, though it remains prone to head-switching artifacts. Finally, we demonstrated that a simple ensemble strategy—using the single-head model as a robust anchor and the dual-head model for spatial refinement—significantly outperforms existing baselines, offering a scalable direction for future VGDQA systems.

Limitations

Our work is not without its limitations. Firstly, as noted in §1, our architecture leverages OCR for text detection and recognition. Consequently, the performance of our models are dependent on the quality of the OCR input. Secondly, our method leverages a pretrained LLM for further training. The instruction following capability of our models is a direct consequence of these pretrained LLMs. Hence, our method depends on availability of high quality instruction tuned LLMs. Thirdly, our models exhibit degraded performance on long-context documents, as evidenced by the drop in grounding accuracy and increased NCE on documents exceeding 4K tokens, which we attribute to the difficulty of maintaining precise attention over extended input sequences. Finally, our models exhibit substantially lower grounding accuracy on non-Latin-script languages such as Chinese and Japanese, reflecting the English-dominant distribution of BoundingDocs v2.0 training data.

Ethical Considerations

Bias in Training Data. Our models are trained and evaluated primarily on public datasets which may skew towards Western business document formats and high-resource languages (English, Chinese). Consequently, the model may exhibit degraded performance or biases when processing documents from underrepresented regions or those utilizing non-standard layouts. This could lead to inequitable access to automated document processing tools for specific demographics.

Automation and Employment. As highlighted in the introduction, one of the goals of this work is to facilitate enterprise automation. The deployment of highly accurate, end-to-end VGDQA systems has the potential to displace jobs traditionally focused on manual data entry and document review. While these tools increase efficiency, the socioeconomic impact on the workforce in administrative sectors should be acknowledged.

Reliability in Critical Contexts. Although our ensemble method achieves high accuracy, LLMs are inherently probabilistic and prone to hallucination. In high-stakes environments—such as legal contract analysis or medical record retrieval—a grounding error or a hallucinated date could have serious legal or health consequences. We strongly advise that human-in-the-loop verification remains a component of workflows involving critical decision-making based on these models.

References

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. 2024. Docformerv2: Local features for document understanding. In *AAAI*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*.

Alessio Chen, Simone Giovannini, Andrea Gemelli, Fabio Coppini, and Simone Marinai. 2025. Towards reliable and interpretable document question answering via vlms. *CoRR*.

Simone Giovannini, Fabio Coppini, Andrea Gemelli, and Simone Marinai. 2025. Boundingdocs: a unified dataset for document question answering with spatial annotations. *CoRR*.

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 *ACL*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *ACM MM*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *ECCV*.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. LISA: reasoning segmentation via large language model. In *CVPR*.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*.

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 *CVPR*.

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, Hao Liu, and Can Huang. 2025. A bounding box is worth one token - interleaving layout and text in a large language model for document understanding. In *ACL Findings*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *CVPR*.

Ahmad Mohammadshirazi, Pinaki Prasad Guha Neogi, Ser-Nam Lim, and Rajiv Ramnath. 2025. Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness. In *ICML Workshops*.

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 *ACL*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *CoRR*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*.

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 40 others. 2025. Qwen3 technical report. *CoRR*.

Yinan Zhou, Yuxin Chen, Haokun Lin, Shuyu Yang, Li Zhu, Zhongang Qi, Chen Ma, and Ying Shan. 2025. DOGR: towards versatile visual document grounding and referring. In *CVPR*.

Zhaoqing Zhu, Chuwei Luo, Zirui Shao, Feiyu Gao, Hangdi Xing, Qi Zheng, and Ji Zhang. 2025. A simple yet effective layout token in large language models for document understanding. In *CVPR*.

A Appendix

A.1 Implementation Details

Training Hyperparameters. Both the pretraining and SFT stages are conducted for 10 epochs with a batch size of 16. We utilize the AdamW optimizer with a weight decay of 0.05, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate follows a Linear Warmup Cosine Decay schedule, featuring 300 warmup steps, a warmup start learning rate of 1×10^{-6}, a peak learning rate of 3×10^{-4}, and a minimum learning rate of 1×10^{-5}. The maximum position ID is extended to 16K to accommodate long-context documents.

LoRA Configuration. To maintain training efficiency, we apply LoRA to the backbone LLM. We set the LoRA rank $r = 16$, alpha $\alpha = 32$, and dropout rate $p = 0.1$.

Loss Weights and Hardware. Based on empirical tuning, the loss weighting hyperparameters are set as follows: $\lambda_{\text{MSE}} = 0.3$ and $\lambda_{\text{IoU}} = 0.3$ for the dual-head objective (5), and $\gamma = 0.3$ for the single-head soft-tokening objective (3). All experiments were conducted on 8 AMD MI300X GPUs.

Input Prompts. We maintain a consistent system prompt across both pretraining and fine-tuning stages: "`You are a helpful assistant. Please follow the instructions provided by the user`". For the pretraining phase, we employ the following reconstruction-based prompt: "`{ocr-placeholder}. Please reconstruct this document with grounding:`". Conversely, the Supervised Fine-Tuning (SFT) stage utilizes a QA-centric format to align with the downstream task: "`Please read the document and then answer the question:\n Document: {ocr-placeholder}\n Question: {question-placeholder}`".

Baseline Configuration. We benchmark our approach against four distinct baselines: DLaVA (Mohammadshirazi et al., 2025), DocExplainerV0 (Chen et al., 2025), a reproduced DOGR (Zhou et al., 2025), and a fine-tuned Qwen3-8B (Yang et al., 2025). For DLaVA and DocExplainerV0, we adhere strictly to the evaluation protocols and prompt templates outlined in their respective official codebases. regarding DOGR, due to the unavailability of public weights for the prealigning and pretraining stages, we reproduced the model by following the official pre-alignment and pretraining procedures prior to fine-tuning it on BoundingDocs v2.0. For these three external baselines, we utFilize the input prompts provided in their original repositories. Finally, the Qwen3-8B baseline is trained using the exact same input prompt structure as our proposed models to ensure a fair comparison.

A.2 Qualitative Results

Figures 2 and 3 present a qualitative comparison of the answers and visual groundings generated by five distinct methods: our single-head model, dual-head model, and ensemble method, alongside the fine-tuned DOGR and Qwen3-8B baselines. These visualizations highlight the architectural trade-offs discussed in §3.

Structural Consistency vs. Spatial Precision. By avoiding the complexities of the head-switching mechanism, the **single-head model** demonstrates superior structural integrity. As observed in the figures, it effectively suppresses extraneous predictions, exhibiting a robust ability to determine the correct cardinality of answers. However, its reliance on discretized tokens results in localization limitations; the generated bounding boxes occasionally exhibit minor spatial drifts, particularly when bounding small visual elements.

Conversely, the **dual-head model** leverages regression-based decoding to achieve high spatial precision. It produces significantly tighter and more accurate bounding boxes compared to the classification-based approach. However, this precision comes at the cost of structural stability; the dependency on the head-switching mechanism leads to a higher incidence of hallucinated switching events, resulting in extraneous answer-box pairs (visualized in red).

Comparison with Baselines. The qualitative gap between our turnkey solutions and the baselines is pronounced. Both the fine-tuned **Qwen3-8B** and **DOGR** models exhibit notably coarser visual grounding. This degradation is attributed to their treatment of bounding box coordinates as plain text tokens. This formulation precludes the application of geometric loss functions (*e.g.*, IoU or MSE) during training, thereby depriving the models of the inductive bias necessary for pixel-level spatial convergence.

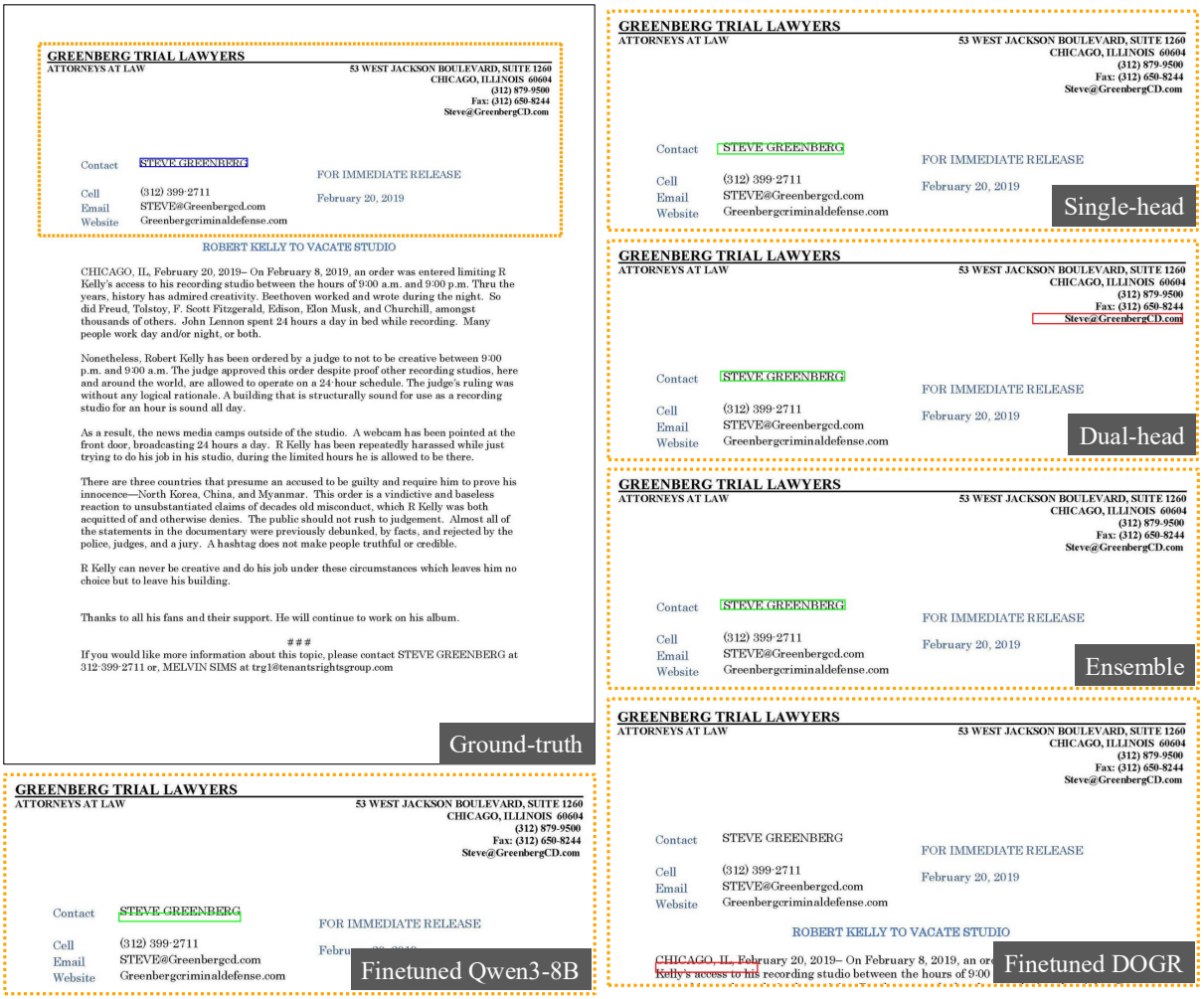

Figure 2: Qualitative results. Question: `Who is listed as the contact?` The ground-truth answer and its visual grounding is visualized using blue bounding boxes. The answer with grounding accuracy as 1 is visualized using green bounding boxes, and the extraneous answers and wrong answers are visualized using red bounding boxes. The answer generated by DOGR is "stephen mitchell".

Ensemble Efficacy. The **ensemble method** anchors the answer set with the single-head model and refines spatial coordinates with the dual-head model, combining structural fidelity with regression-based precision.

A.3 Inference Cost

Table 5 reports runtime and FLOPs for a single forward pass. The dual-head model is faster than the single-head model, as coordinate prediction requires only two forward operations rather than four autoregressively generated tokens. The ensemble method doubles the inference cost as the trade-off for its performance gains, yet our individual models already achieve substantially higher grounding accuracy than the VLM baselines at comparable or lower FLOPs.

Table 5: Overall performance and inference cost. To measure the FLOPs, the input sequence to LLMs is 1K, and the input image size to Qwen3-VL-8B is 1024×1024 which is much smaller than the actual document image sizes in BoundingDocs v2.0. Choosing larger input image size will lead to out-of-memory error when profiling. Our solutions are highlighted in light green. ↓ indicates lower is better; ↑ indicates higher is better. The best results are highlighted in **bold**, and the second best are highlighted using underline.

Methods	Overall Metrics		Inference Cost	
	NCE↓	Acc↑	Avg Runtime↓	FLOPs↓
Qwen3-VL-8B	.440	.252	4.17s	19.42T
InternVL3.5-8B	.497	.001	2.84s	31.59T
LayTextLLM	.414	0	5.82s	13.74T
Single-head model	**.086**	.796	3.58s	13.89T
Dual-head model	.113	.854	2.49s	13.89T
Ensemble method	**.086**	**.866**	6.08s	27.78T

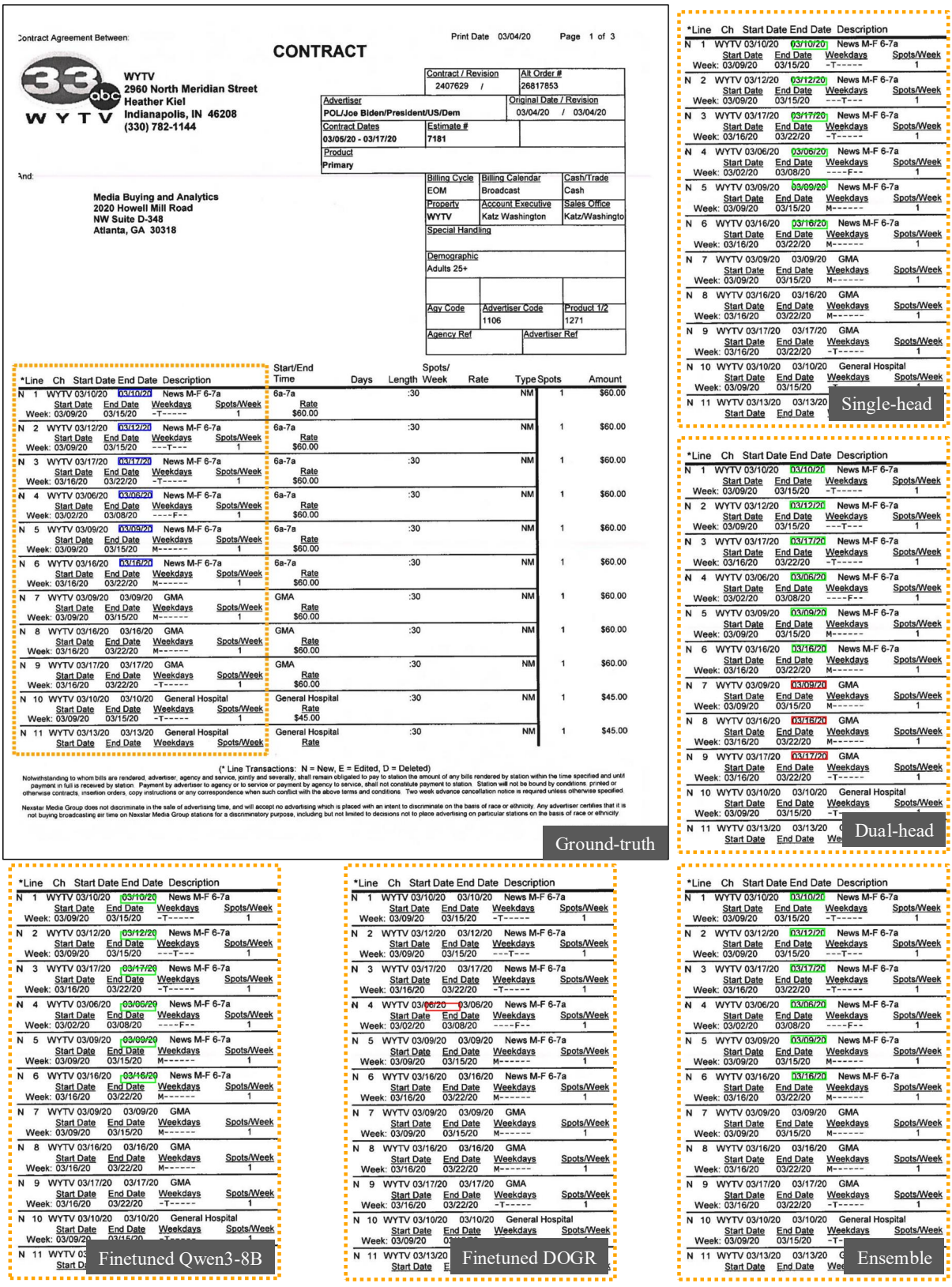

Figure 3: Qualitative results. Question: `What is the end date for the News M-F 6-7a program`? The ground-truth answer and its visual grounding is visualized using blue bounding boxes. The answer with grounding accuracy as 1 is visualized using green bounding boxes, and the extraneous answers and wrong answers are visualized using red bounding boxes. The answer generated by DOGR is "05/04/20".