LAYOUT-AWARE LATENT EVIDENCE REASONING FOR OCR-FREE KEY INFORMATION EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Key information extraction (KIE) requires associating textual values with their spatial context and requested fields. OCR-free models typically learn these associations implicitly from answer supervision, while OCR-augmented methods expose explicit evidence at the cost of external OCR dependence and variable-length context at inference. We propose LAYER, a **LAY**out-aware latent **E**vidence **R**easoning framework that uses OCR as training supervision to learn a compact internal representation of document evidence. Given a document image and a field query, LAYER predicts a fixed-length sequence of continuous visual latent tokens in parallel. The same language model then conditions on these latents and the original image-query context to autoregressively generate structured answers. Two complementary objectives guide learning: Layout-aware Feature Alignment aligns the predicted latents with visual features of OCR text rendered in its page layout, while Confidence-Gated Contrastive Learning associates answer-value representations with supporting OCR blocks through reliable field-region matches. A two-stage training scheme initializes the latent interface and then jointly optimizes structured generation with both objectives as auxiliary supervision. Inference requires only the document image and query. Experiments on public KIE benchmarks show pooled micro-F1 gains of 1.66%–3.45% over answer-only fine-tuning across three backbone sizes. A controlled timing comparison further shows a $2.03\times$ end-to-end speedup over OCR as Chain-of-Thought with answer length fixed.

1 INTRODUCTION

Key information extraction (KIE) converts document images into structured key-value pairs, supporting the processing of invoices, receipts, forms, and other business documents (Kim et al., 2022; Wan et al., 2024). Its central challenge is to associate textual content with the role it plays in a document. For example, an invoice contains several monetary amounts whose relevance to a requested field depends on nearby labels and spatial organization. Therefore, accurate extraction requires both recognizing candidate values and interpreting their layout context (Xu et al., 2020).

Existing KIE systems can be broadly divided into two categories: OCR-free and OCR-augmented. The former methods perform end-to-end extraction directly from document images, typically by fine-tuning document understanding models or vision-language models (VLMs) to generate structured answers (Kim et al., 2022; Wan et al., 2024; Yu et al., 2025), as shown in Figure 1(a). This formulation removes the need for an external OCR pipeline at inference time, and supervision is usually applied only to the final answer. As a result, the correspondence between a requested field, its textual evidence, and its source region must be learned implicitly. In contrast, OCR-augmented methods (Fujitake, 2024; Liao et al., 2025; Lu et al., 2025) explicitly provide recognized text and layout information to the model, making document evidence directly accessible during extraction, as shown in Figure 1(b). Such explicit evidence facilitates key-value association, but also introduces dependence on an external OCR system and requires processing variable-length OCR sequences at inference time. Consequently, the extraction pipeline becomes dependent on OCR quality and incurs additional cost for constructing and processing textual context.

This motivates using OCR-derived textual and spatial information as training supervision for internal evidence representations. Recent studies on visual latent reasoning suggest that explicit textual inter-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

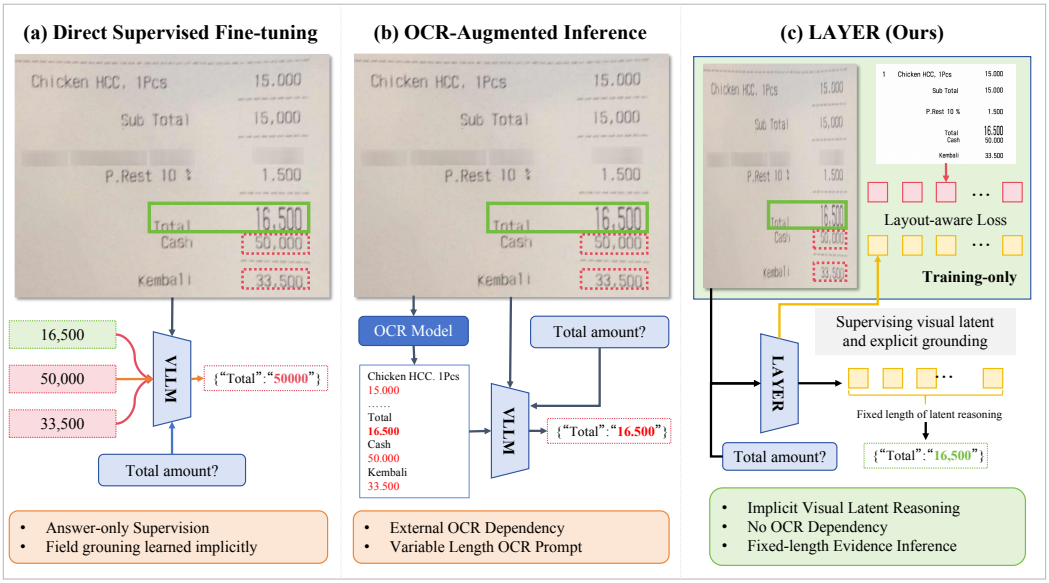


Figure 1: **Comparison of KIE paradigms.** (a) OCR-free answer-only fine-tuning provides no explicit supervision for field–region associations. (b) OCR-augmented methods provide explicit textual and layout evidence, but require an external OCR system and variable-length OCR context at inference time. (c) LAYER uses OCR only during training to supervise layout-aware latent reasoning and field-level grounding, enabling fixed-length latent reasoning without external OCR at inference time.

mediate information can be internalized into continuous visual representations (Wang et al., 2026b; Xing et al., 2025; Li et al., 2025), providing a compact alternative to supplying such information as textual tokens during inference. However, it is non-trivial to directly transfer this idea to KIE. Unlike general reasoning traces, document evidence is inherently two-dimensional: the interpretation of a candidate value often depends on its spatial relationship with labels, neighboring values, and other layout elements. Moreover, KIE requires not only modeling page-level content, but also establishing field-specific correspondences between requested values and their source regions. A linearized textual representation could lose such spatial structure, whereas page-level representation alignment alone does not guarantee correct field grounding. This raises a natural question: *can OCR evidence be internalized during training into a compact, layout-aware latent representation that preserves field-level grounding, while keeping KIE fully OCR-free at inference time?*

To address this question, we propose **LAYER**, a **LAY**out-aware latent **E**vidence **R**easoning framework that learns a compact evidence representation from OCR-derived layout and field-level supervision, while remaining fully OCR-free at inference time. Given a document image, LAYER predicts a fixed-length sequence of continuous visual tokens in parallel. The language model conditions on these latent tokens together with the image-query context to generate the structured answer. This latent sequence provides an intermediate evidence representation whose length is independent of the number of OCR tokens.

LAYER learns this latent evidence representation through two complementary forms of OCR-derived supervision. At the page level, inspired by Render-of-Thought (RoT) (Wang et al., 2026b), we render OCR strings at their normalized page coordinates and encode the rendered page into spatially structured visual features. **Layout-aware Feature Alignment** (LFA) aligns these OCR-derived layout features with the predicted visual latent tokens, encouraging the latent representation to retain the two-dimensional organization of document evidence. At the field level, **Confidence-Gated Contrastive Learning** (CGCL) explicitly links answer-value representations to OCR blocks that jointly encode textual content and spatial location. A confidence-based gating mechanism restricts this supervision to reliable correspondences, reducing interference from ambiguous or noisy matches. Training proceeds in two stages: we first initialize the latent reasoning module with Layout-aware Feature Alignment, and then jointly optimize structured answer generation while re-

taining Layout-aware Feature Alignment and Confidence-gated Contrastive Learning as auxiliary supervision.

Experiments on public KIE benchmarks demonstrate the effectiveness and efficiency of LAYER. Across three backbone sizes, LAYER improves pooled micro-F1 over answer-only fine-tuning by 1.66%–3.45%. Its 4B variant achieves F1 scores of 91.7% on CORD and 91.8% on SROIE, without external OCR at inference. In a controlled timing comparison with answer length fixed, LAYER achieves a $2.03\times$ end-to-end speedup over OCR-CoT. These results show that OCR-derived supervision can improve structured extraction while supporting efficient latent evidence construction and a fully OCR-free inference pipeline.

Our contributions are summarized as follows:

- We propose LAYER, a layout-aware latent evidence reasoning framework that internalizes OCR-derived textual and spatial evidence into fixed-length visual latent tokens during training, while requiring no external OCR at inference time.
- We introduce two complementary OCR-derived supervision mechanisms. Layout-aware Feature Alignment injects the two-dimensional organization of document evidence into visual latent representations while Confidence-gated Contrastive Learning establishes reliable field-level associations between answer values and their source regions.
- Experiments on public KIE benchmarks demonstrate consistent gains over answer-only fine-tuning across three backbone sizes and a $2.03\times$ end-to-end speedup over OCR-CoT under a fixed answer-length setting.

2 Related Work

2.1 Key Information Extraction

Key Information Extraction (KIE) has traditionally been formulated as structured prediction over OCR-derived representations, with datasets such as FUNSD, SROIE, and CORD establishing standard benchmarks for entity and key-value extraction. Representative methods, including LayoutLM (Xu et al., 2020), LayoutLMv2 (Xu et al., 2021), LayoutLMv3 (Huang et al., 2022), DocFormer (Appalaraju et al., 2021), and StrucTexT (Li et al., 2021), jointly model OCR text, layout, and visual information, but remain sensitive to recognition errors, missing regions, and inaccurate spatial annotations. Generative document understanding explores several input representations: Donut (Kim et al., 2022) directly maps document images to serialized outputs without external OCR, while Pix2Struct (Lee et al., 2023), UDOP (Tang et al., 2023), DocLLM (Wang et al., 2024a), and Layout-LLM (Fujitake, 2024) investigate vision-to-text generation, multimodal pretraining, and instruction tuning. Within this landscape, LAYER focuses on transferring OCR-derived textual and spatial supervision into latent evidence for structured extraction. It requires neither an explicit OCR prompt nor full-page transcription at inference time.

2.2 Visual Latent Reasoning

Recent studies explore continuous representations for intermediate reasoning. Coconut (Hao et al., 2024) feeds hidden states back as continuous reasoning steps, while LVR (Li et al., 2026) extends this approach to visual latent reasoning. Closely related to our work, Render-of-Thought (RoT) (Wang et al., 2026b) uses rendered textual reasoning as visual supervision and autoregressively generates projected visual embeddings. Parallel alternatives have also emerged: SCOLAR (Wang et al., 2026a) generates auxiliary visual latents through a dedicated transformer branch, while DeepLatent (Lu et al., 2026) uses learnable spatial queries and a LatentFormer for parallel latent construction. Separately, Text or Pixels (Li et al., 2025) and DeepSeek-OCR (Wei et al., 2025) investigate visual representations for compact text transmission and reconstruction. Building on RoT, LAYER predicts fixed-length visual latents in parallel rather than autoregressively, using OCR-derived layout and field-level supervision for KIE without external OCR at inference time.

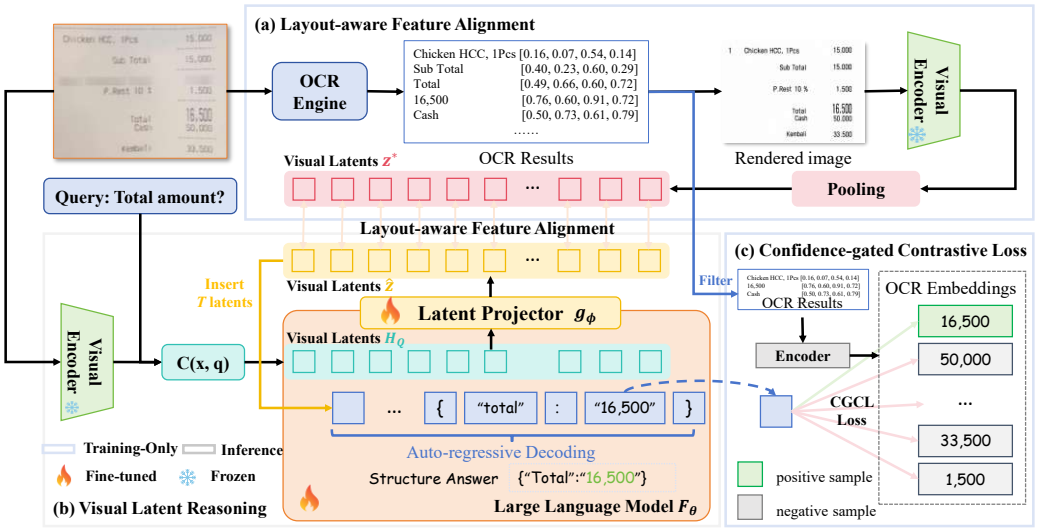


Figure 2: **Overview of LAYER.** (a) Layout-aware Feature Alignment transfers OCR-derived layout evidence into visual latents. (b) Visual Latent Reasoning predicts a fixed-length latent sequence in parallel for autoregressive structured answer generation. (c) Confidence-Gated Contrastive Learning aligns answer-value representations with their supporting OCR blocks through reliable field–region matches. OCR is used only during training; inference requires only the document image and query.

3 Methodology

3.1 Problem Formulation and Overview

Given a document image x and a query q that specifies a set of fields $\mathcal{K} = \{k_1, \ldots, k_m\}$, KIE produces a serialized structured answer $y = \{k_i : v_i\}_{i=1}^{m}$. Identifying each value requires relating its textual content to its position and context on the page. LAYER learns a fixed-length sequence of continuous visual latent tokens $\hat{Z} \in \mathbb{R}^{T \times d}$, where d is the language-model hidden dimension and $T = 64$ in our main model. Unlike Render-of-Thought (Wang et al., 2026b), LAYER predicts all T latents in parallel from the image-query context, then supplies them to the same language model for autoregressive answer generation. We refer to this intermediate evidence modeling as *visual latent reasoning*. In detail, two complementary objectives supervise this representation. *Layout-aware Feature Alignment* (LFA) provides document-level supervision from spatially rendered OCR text, while *Confidence-Gated Contrastive Learning* (CGCL) associates answer-value representations with reliably matched OCR regions. Both modules use OCR results only during training, and inference requires only (x, q).

3.2 Layout-Aware Feature Alignment

Layout-preserving visual targets. During training, OCR provides blocks $\mathcal{B} = \{(s_b, r_b, c_b)\}_{b=1}^{B}$ containing recognized text, normalized bounding boxes, and confidence scores. A deterministic renderer R places the strings within their corresponding boxes on a blank canvas. The rendered page provides the teacher target $Z^\star = A_T(V(R(\mathcal{B})))$. Here, V is the same frozen visual pathway used by the VLM, including its native projection into the language-model hidden space. A_T adaptively averages the complete visual feature sequence into T vectors. Rendering retains their page layout, whereas pooling provides a fixed token budget without truncating the resulting visual sequence. The pooled target supplies layout-informed supervision without imposing a one-to-one correspondence between latent slots and OCR blocks.

Parallel visual latent prediction. We predict visual latents from a fixed set of query slots conditioned on the document image and field query. Let $C(x, q)$ denote the multimodal input prefix and

F_θ the hidden-state mapping of the language backbone. We append T copies of a shared learnable query embedding e_{r}, bracketed by the begin and end embeddings e_{b} and e_{e}:

$$\begin{aligned} H_Q &= \left[F_\theta\big([C(x,q); e_{\mathrm{b}}; e_{\mathrm{r}}^{\times T}; e_{\mathrm{e}}]\big)\right]_Q, \\ \hat{Z} &= g_\phi(H_Q), \end{aligned} \tag{1}$$

where $[\cdot]_Q$ extracts the hidden states at the query positions, and g_ϕ projects them into the visual embedding space defined by $Z^\star$. The three special embeddings correspond to $\langle \mathrm{img_begin} \rangle$, $\langle \mathrm{img} \rangle$, and $\langle \mathrm{img_end} \rangle$, respectively. The query slots are distinguished by positional encoding and processed together under the causal attention mask of the language backbone. No vocabulary decoding or feedback of predicted latents is used within this step. Thus, all T visual latents are obtained in one query forward pass. The projected sequence $\hat{Z}$ is aligned with $Z^\star$ during training and subsequently supplied as continuous input embeddings for answer generation.

Answer Generation. We restore the original image-query prefix cache and replace the query-slot sequence with $[e_{\mathrm{b}}; \hat{Z}; e_{\mathrm{e}}]$. The same language model then generates the answer using standard next-token prediction. Training uses teacher-forced answer tokens; inference decodes them autoregressively. Both use predicted latents rather than teacher features. The projection head therefore remains part of the inference path. Parallel prediction applies to the latent sequence, not to the subsequent answer tokens.

3.3 Confidence-Gated Contrastive Learning

LFA supervises document-level features but does not directly associate individual fields with their source regions. CGCL adds a training-only contrastive objective between answer-value states and OCR block representations.

Reliable field-region matches. OCR errors and repeated values can introduce incorrect or ambiguous positive pairs for contrastive grounding. We therefore restrict supervision to reliable correspondences: a value must have exactly one candidate OCR block, with a string alignment score of at least T_{align} and OCR confidence of at least T_{conf}. We further require the value to occur exactly once in the serialized answer so that its answer span can be identified unambiguously. The alignment score measures string similarity between the normalized ground-truth value and OCR text; candidate construction and scoring are detailed in Section C. At most two eligible values are retained per document, forming the set $\mathcal{I}$.

Contrastive grounding. For each non-empty OCR block k, we mean-pool its input token embeddings, apply a learned projection, and add an MLP encoding of its normalized bounding box. The resulting vector b_k is ℓ_2-normalized. For each retained answer value i, tokenizer offsets identify the overlapping answer tokens. We mean-pool their hidden states from the teacher-forced answer pass, project them into the same space, and normalize the result to obtain a_i.

Let k_i^+ denote the matched block and $\mathcal{B}_{\mathrm{valid}}$ the indices of non-empty blocks. We contrast the positive block with every other valid block on the same page:

$$\mathcal{L}_{\mathrm{CGCL}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \log \frac{\exp(a_i^\top b_{k_i^+}/\tau)}{\sum_{k \in \mathcal{B}_{\mathrm{valid}}} \exp(a_i^\top b_k/\tau)}, \tag{2}$$

where $\tau = 0.07$; the loss is zero when $\mathcal{I}$ is empty. This objective encourages associations between answer values and their source blocks using lexical and spatial cues. These associations provide auxiliary supervision; no region prediction is required at inference.

3.4 Two-Stage Optimization and OCR-Free Inference

Training. In Stage I, we freeze the backbone and initialize the latent interface through LFA, updating the projection head and the begin/query embeddings. In Stage II, the visual pathway remains

frozen, while we optimize language-model LoRA parameters, the projection head, the special embeddings, and the CGCL modules. The objectives are

$$\begin{aligned} \mathcal{L}_{\mathrm{I}} &= \mathcal{L}_{\mathrm{MSE}} + \lambda_{\cos}\,\mathcal{L}_{\cos}, \\ \mathcal{L}_{\mathrm{II}} &= \mathcal{L}_{\mathrm{ans}} + \lambda_{\mathrm{MSE}}\,\mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{CGCL}}\,\mathcal{L}_{\mathrm{CGCL}}. \end{aligned} \tag{3}$$

Here, $\mathcal{L}_{\mathrm{MSE}}$ measures mean squared error between $\hat{Z}$ and $Z^{\star}$, $\mathcal{L}_{\cos}$ is their mean per-slot cosine distance, and $\mathcal{L}_{\mathrm{ans}}$ is the token-averaged answer negative log-likelihood. Stage II retains the MSE component of LFA alongside CGCL as auxiliary supervision.

During Stage II, the model predicts $\hat{Z}$ from the document image and query, and generates the structured answer conditioned on both $\hat{Z}$ and the original image-query context. The same conditioning scheme is used at inference.

OCR-free inference. Given only (x, q), LAYER first computes the image–query prefix and predicts $\hat{Z}$ in one forward pass. It then restores the same prefix cache, inserts the predicted latent sequence, and autoregressively generates the JSON format answer. The OCR system, renderer, teacher encoding, reliability gate, and copy-grounding modules are absent from this inference procedure. Thus, OCR contributes to learning the evidence interface without requiring OCR execution, OCR prompt tokens, or auxiliary copy prediction at deployment.

4 Experiments

We organize the evaluation around the role of the latent evidence interface: extraction performance, the contributions of the three components in Figure 2, robustness of gains across backbone sizes, the marginal effect of inference-time OCR, and the cost of intermediate evidence construction.

4.1 Experimental Setup

Our unified public KIE test set contains 1,394 documents and 11,952 fields from CORD (Park et al., 2019), FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), POIE(Kuang et al., 2023). We parse predictions into key-value dictionaries and report field-level micro-F1 for exact matches, average normalized Levenshtein similarity (ANLS) for partial textual correctness, and strict record-level accuracy (Record Acc). All experiments are conducted on an NVIDIA H20 GPU. More detailed settings are shown in the Appendix.

For controlled fine-tuning comparisons, we use the same 4,013 training examples and evaluate with greedy decoding. LAYER uses a fixed latent budget of $T = 64$, with adaptive average pooling of teacher features. In Stage II, the answer loss has unit weight, while $\lambda_{\mathrm{MSE}} = 0.05$ and $\lambda_{\mathrm{CGCL}} = 0.005$. For CGCL, we set the contrastive temperature to $\tau = 0.07$ and the reliability thresholds to $T_{\mathrm{align}} = 0.98$ and $T_{\mathrm{conf}} = 0.9$, retaining at most two eligible values per document. The visual pathway remains frozen throughout training. All experiments are conducted on an NVIDIA H20 GPU. Additional implementation details are provided in the appendix.

4.2 Benchmark Comparisons

Competitive extraction at small scale. Table 1 shows that LAYER-4B exceeds all listed published F1 scores, while LAYER-0.8B exceeds the listed specialists on CORD and approaches STNet on SROIE. This suggests that OCR-derived supervision can support competitive extraction with a compact backbone and no inference-time OCR. Cross-source protocol differences qualify these comparisons, which provide context rather than a controlled ranking.

Adaptation complements model capacity. LAYER-0.8B surpasses the 4B zero-shot model in F1 on both datasets, showing that task adaptation can outweigh a substantial increase in backbone size. Gains in both F1 and TED accuracy over each same-size zero-shot model further indicate improvements in exact extraction and structural similarity. These gains combine fine-tuning with evidence supervision; the SFT controls below assess the latter's additional value. FUNSD and POIE results appear in Appendix B.

Table 1: **Benchmark comparison on CORD and SROIE.** Scores are reported in %; Acc denotes TED accuracy. Parentheses give absolute gains over the same-size zero-shot Qwen3.5 model, computed before rounding. Published results retain their source protocols; score sources and evaluation details are provided in Appendix A.

	CORD		SROIE	
Method	F1 ↑	Acc ↑	F1 ↑	Acc ↑
End-to-end specialist models				
Donut (Kim et al., 2022)	84.1	90.9	83.2	92.8
Dessurt (Davis et al., 2022)	82.5	-	84.9	-
DocParser (Dhouib et al., 2023)	84.5	-	87.3	-
$\text{SeRum}_{\text{total}}$ (Cao et al., 2023)	80.5	85.8	85.6	92.8
OmniParser (Wan et al., 2024)	84.8	88.0	85.6	93.6
OmniParser V2 (Yu et al., 2025)	85.0	88.7	85.8	94.0
CREPE (Okamoto et al., 2024)	85.0	-	-	-
STNet (Liu et al., 2024)	88.8	93.5	88.3	97.4
MLLMs: CC-OCR evaluation (Yang et al., 2024)				
GPT-4o (OpenAI, 2024)	81.3	-	72.6	-
Claude-3.5-Sonnet (Anthropic, 2024)	71.5	-	69.8	-
Gemini-1.5-Pro (Gemini Team, 2024)	78.7	-	75.0	-
Qwen2-VL-72B (Wang et al., 2024b)	69.0	-	75.5	-
InternVL2-76B (OpenGVLab Team, 2024)	66.1	-	57.1	-
MLLMs: zero-shot				
Qwen3.5-0.8B (Qwen Team, 2026)	64.5	83.7	61.5	77.3
Qwen3.5-2B (Qwen Team, 2026)	82.7	92.9	76.1	84.6
Qwen3.5-4B (Qwen Team, 2026)	88.4	96.3	73.9	80.1
Our methods				
LAYER (Qwen3.5-0.8B)	89.4 (+24.9)	96.1 (+12.4)	87.8 (+26.3)	97.1 (+19.8)
LAYER (Qwen3.5-2B)	90.2 (+7.5)	96.0 (+3.2)	90.8 (+14.7)	98.0 (+13.4)
LAYER (Qwen3.5-4B)	91.7 (+3.3)	97.3 (+1.0)	91.8 (+17.9)	98.2 (+18.1)

Table 2: **Component ablations with Qwen3.5-4B.** Components follow Figure 2. (a) LFA: Layout-aware Feature Alignment; (b) VLR: Visual Latent Reasoning; (c) CGCL: Confidence-gated Contrastive Loss ($\mathcal{L}_{\text{CL}}$). LAYER w/o CGCL is the Layout-RoT variant. Components (a) and (c) are used only during training.

	Components			Extraction scores ↑		
Variant	(a) LFA	(b) VLR	(c) CGCL	micro-F1	ANLS	Record Acc.
SFT	–	–	–	88.32	95.35	63.27
RoT	–	✓	–	89.70	96.52	66.28
LAYER w/o CGCL	✓	✓	–	89.55	96.18	66.42
LAYER (full)	✓	✓	✓	**89.98**	**96.89**	**66.78**

4.3 Component Analysis

Table 2 connects the three components in Figure 2 to controlled Qwen3.5-4B comparisons. SFT uses answer supervision alone; RoT introduces visual latent reasoning with its standard teacher; LAYER w/o CGCL replaces that teacher with layout-aware supervision; full LAYER adds confidence-gated field-level grounding.

Benefit of the latent evidence interface. All latent variants outperform SFT across the three metrics, with standard RoT accounting for much of the improvement. Full LAYER achieves the best scores, including a 1.66% micro-F1 gain over SFT. The accompanying gains in ANLS and strict Record Acc. show that improvement extends from partial textual correctness to exact field and document recovery. Thus, the benefit of the training approach is not confined to a single extraction metric.

Table 3: **Gains over SFT across backbone sizes.** All LAYER models include layout-aware feature alignment, visual latent reasoning, and CGCL. Scores and absolute gains Δ are reported in %. The 4B results match Table 2.

	micro-F1 ↑			ANLS ↑		
Backbone	SFT	LAYER	Δ	SFT	LAYER	Δ
Qwen3.5-0.8B	81.70	**85.15**	+3.45	91.30	**94.27**	+2.97
Qwen3.5-2B	85.87	**87.81**	+1.94	94.36	**95.92**	+1.56
Qwen3.5-4B	88.32	**89.98**	+1.66	95.35	**96.89**	+1.54

Contribution of field-level grounding. Replacing the standard teacher with layout-aware supervision alone slightly reduces micro-F1 and ANLS, despite improving Record Acc. Page-level alignment therefore does not consistently improve extraction in this comparison. Adding CGCL to the layout-aware variant improves all three metrics, including a 0.43% gain in micro-F1. This pattern is consistent with the value of explicitly supervising answer–region associations in addition to the page representation. The ablation evaluates CGCL as a whole, rather than isolating its gating rule. Relative to standard RoT, full LAYER yields modest gains across all three metrics; the larger improvement over SFT indicates that the latent evidence interface accounts for much of the overall benefit.

4.4 Effect of Backbone Size

Table 3 compares LAYER with SFT across three Qwen3.5 backbone sizes, using the full LAYER configuration throughout. LAYER improves both micro-F1 and ANLS at every tested size, with the largest absolute gains at 0.8B. Notably, LAYER-0.8B reaches 85.15 micro-F1 and 94.27 ANLS, trailing SFT-2B by only 0.72% and 0.09%, respectively. These results show that evidence supervision can substantially narrow the performance gap to a larger answer-only baseline.

The improvements persist at 2B and 4B. LAYER-2B exceeds SFT-4B by 0.57% in ANLS but remains 0.51% below it in micro-F1, indicating that improved textual similarity does not necessarily imply equivalent exact extraction performance. Meanwhile, both metrics improve as LAYER's backbone grows, supporting the complementary benefits of evidence supervision and model scaling within the evaluated range.

4.5 Marginal Effect of Inference-Time OCR

To assess whether LAYER internalizes OCR-derived evidence, we append OCR text at inference and measure the resulting F1 changes. Figure 3 compares zero-shot and LAYER models across three backbone sizes on CORD and SROIE. This intervention is separate from our standard OCR-free evaluation.

LAYER consistently obtains smaller marginal gains from explicit OCR. On CORD, OCR improves LAYER by 1.93%–3.15%, compared with 2.35%–6.95% for zero-shot models. On SROIE, OCR improves zero-shot models by 0.79%–3.31% points but reduces LAYER's F1 by 0.22%–1.30% points. Together with LAYER's higher OCR-free extraction scores, these results support the effectiveness of learning from OCR-derived supervision during training: additional transcription offers less benefit once document evidence is learned internally. The remaining gains on CORD indicate that this internalization is partial, with explicit OCR still providing complementary evidence.

4.6 Inference Efficiency

Table 4 tests the computational consequence of predicting latent evidence in parallel. The intermediate-stage speedup is much larger than the reduction in sequence length, consistent with removing serial token generation in addition to shortening the intermediate state. The resulting end-to-end speedup is $2.03\times$ with answer length fixed, so the improvement does not come from generating a shorter answer.

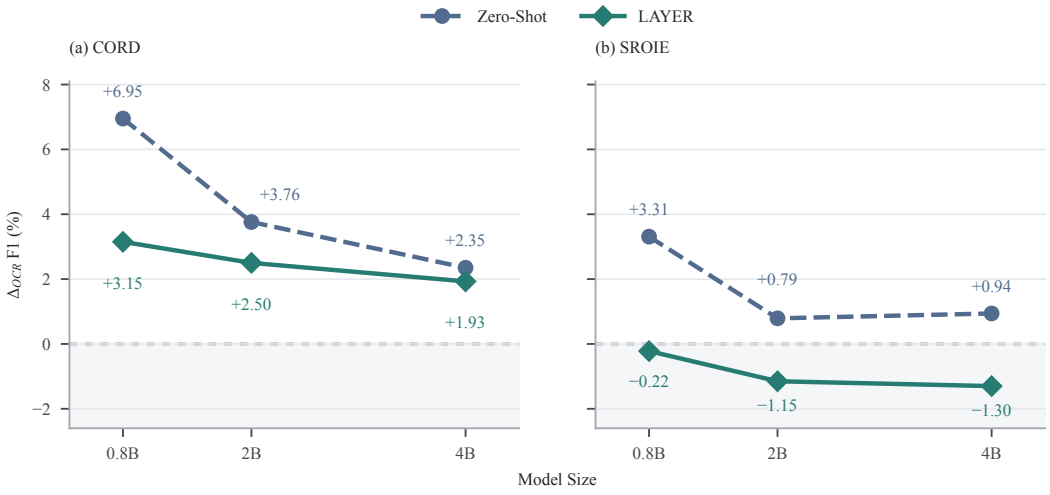


Figure 3: **Marginal effect of explicit OCR.** Each point shows $\Delta_{\text{OCR}} = \text{F1}_{\text{with OCR}} - \text{F1}_{\text{without OCR}}$ (%) for the same model. Positive values indicate improvement after appending OCR text; negative values indicate degradation. Lines connect backbone sizes, and the dashed horizontal line marks no change.

Table 4: **Inference efficiency relative to OCR-CoT.** Controlled end-to-end timing includes a 265-token answer for both methods. The final row reports the reduction in intermediate sequence length and latency speedups relative to OCR-CoT.

			Latency (s) ↓	
Method	Token budget	Generation	Intermediate	End-to-end
OCR-CoT	291	Autoregressive	6.976	13.397
LAYER	64	One parallel pass	**0.112**	**6.599**
Reduction / speedup	4.55×	–	62.1×	2.03×

The smaller end-to-end gain also identifies the remaining bottleneck: once latent construction is inexpensive, answer decoding and other shared costs dominate runtime. LAYER therefore accelerates evidence construction while retaining autoregressive structured generation; it does not eliminate the cost of producing the final answer. This controlled comparison measures latency under the stated output length, rather than an accuracy–latency trade-off between OCR-CoT and LAYER.

5 Conclusion

We introduced LAYER, a framework that uses OCR-derived supervision to learn compact, layout-aware latent evidence for key information extraction. LAYER predicts a fixed-length visual latent sequence in parallel and uses it alongside the image-query context for autoregressive structured generation. Layout-aware Feature Alignment and Confidence-Gated Contrastive Learning provide complementary document-level and field-level supervision, while inference requires only the image and query. Experiments demonstrate improvements over answer-only fine-tuning across three backbone sizes and lower inference latency than OCR-CoT under a fixed answer-length setting. These findings support using OCR as a source of training supervision for accurate and efficient OCR-free extraction. Future work will investigate query-adaptive latent allocation, learning from noisy OCR, and generalization to unseen document layouts and languages.

ETHICS STATEMENT

This work studies key information extraction using publicly available datasets and does not involve new human-subject data collection. We are not aware of additional ethical concerns specific to the experiments reported here. The broader use of document extraction systems nevertheless requires attention to privacy, since documents may contain personal or confidential information. Deployment and reuse should respect applicable dataset licenses, privacy requirements, and authorization to process the underlying documents. Extracted values should be verified before use in consequential decisions, as recognition and field-assignment errors may propagate to downstream applications.

AI USE STATEMENT

AI tools were used solely for manuscript polishing, including improving grammar, wording, clarity, and organization. Their use was limited to the presentation of this paper and did not include generating research ideas, training supervision, experimental data, or results. The authors reviewed and verified the AI-assisted revisions to ensure that they accurately reflect the methodology and findings, and take full responsibility for the research and the final manuscript.

REFERENCES

Anthropic. Introducing Claude 3.5 Sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 973–983. IEEE, 2021.

Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. Attention where it matters: Rethinking visual document understanding with selective region concentration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19517–19527, 2023. URL https://arxiv.org/abs/2309.01131.

Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. End-to-end Document Recognition and Understanding with Dessurt. *arXiv preprint arXiv:2203.16618*, 2022. URL https://arxiv.org/abs/2203.16618.

Mohamed Dhouib, Ghassen Bettaieb, and Aymen Shabou. DocParser: End-to-end OCR-free Information Extraction from Visually Rich Documents. *arXiv preprint arXiv:2304.12484*, 2023. URL https://arxiv.org/abs/2304.12484.

Masato Fujitake. Layoutllm: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10219–10224, 2024.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 4083–4091, 2022.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 international conference on document analysis and recognition workshops (ICDARW)*, volume 2, pp. 1–6. IEEE, 2019.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pp. 36–53. Springer, 2023.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026.

Yanhong Li, Zixuan Lan, and Jiawei Zhou. Text or pixels? evaluating efficiency and understanding of llms with visual text inputs. In *EMNLP (Findings)*, pp. 10564–10578, 2025.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1912–1920, 2021.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4038–4049. IEEE, 2025.

Shuhang Liu, Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Qing Wang, Jianshu Zhang, and Chenyu Liu. See then tell: Enhancing key information extraction with vision grounding. *arXiv preprint arXiv:2409.19573*, 2024. URL https://arxiv.org/abs/2409.19573v2.

Dongchen Lu, Zhimo Li, Mao Shu, and Huo Cao. Deeplatent: Think with images via parallel latent visual reasoning. *arXiv preprint arXiv:2606.00562*, 2026.

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, et al. A bounding box is worth one token-interleaving layout and text in a large language model for document understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7252–7273, 2025.

Yamato Okamoto, Youngmin Baek, Geewook Kim, Ryota Nakao, DongHyun Kim, Moon Bin Yim, Seunghyun Park, and Bado Lee. CREPE: Coordinate-aware end-to-end document parser. *arXiv preprint arXiv:2405.00260*, 2024. URL https://arxiv.org/abs/2405.00260.

OpenAI. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on document intelligence at NeurIPS*, volume 2019, pp. 5, 2019.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19254–19264. IEEE, 2023.

Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. OmniParser: A Unified Framework for Text Spotting, Key Information Extraction and Table Recognition. *arXiv preprint arXiv:2403.19128*, 2024. URL https://arxiv.org/abs/2403.19128.

Chenfeng Wang, Wei He, Xuhan Zhu, Chunpeng Zhou, Qizhen Li, Song Yan, Yufei Zheng, Chengjun Yu, Fan Lu, Wei Zhai, et al. Self-consistent latent reasoning: Long latent sequence reasoning for vision-language model. *arXiv preprint arXiv:2605.12163*, 2026a.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8529–8548, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024b. URL https://arxiv.org/abs/2409.12191.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Zheng Wei, and Yang Tang. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45236–45253, 2026b.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Ling Xing, Rui Yan, Alex Jinpeng Wang, Zechao Li, and Jinhui Tang. See the text: From tokenization to visual reading. *arXiv preprint arXiv:2510.18840*, 2025.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 2579–2591, 2021.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1192–1200, 2020.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, Lianwen Jin, and Junyang Lin. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*, 2024. URL https://arxiv.org/abs/2412.02210v3.

Wenwen Yu, Zhibo Yang, Jianqiang Wan, Sibo Song, Jun Tang, Wenqing Cheng, Yuliang Liu, and Xiang Bai. OmniParser V2: Structured-points-of-thought for unified visual text parsing and its generality to multimodal large language models. *arXiv preprint arXiv:2502.16161*, 2025. URL https://arxiv.org/abs/2502.16161v2.

The appendix provides:

- A: Structured extraction protocols.
- B: Results on other benchmarks.
- C: OCR matching and reliability filtering.
- D: Implementation and training details.

A Structured Extraction Protocols

Table 1 preserves each source's evaluation protocol. Specialist scores, except CREPE and STNet, follow the compilation in OmniParser V2 (Yu et al., 2025); CREPE and STNet follow their respective papers (Okamoto et al., 2024; Liu et al., 2024). The listed STNet scores correspond to its coordinate-supervised STNet* variant. Published MLLM results follow CC-OCR (Yang et al., 2024). Differences in evaluation and aggregation prevent treating these published scores as a fully controlled comparison. Acc in that table denotes tree-edit-distance (TED) accuracy, rather than the strict record-level accuracy used in the component analysis. Parenthetical gains are computed from unrounded same-backbone results before display rounding.

B Results on Other Benchmarks

Table 5 compares zero-shot Qwen3.5, SFT, and LAYER on FUNSD and POIE. LAYER improves both F1 and ANLS over SFT at every tested backbone size on both datasets.

On POIE, the largest gains occur at 0.8B: F1 increases from 81.71 to 87.75, and ANLS from 89.11 to 93.90, corresponding to gains of 6.04 and 4.79 percentage points. The improvements remain positive at 2B and 4B, consistent with the benefit of evidence supervision across backbone sizes.

On FUNSD, LAYER-0.8B improves F1 from 53.51 to 54.71 and ANLS from 78.59 to 81.46. Both metrics also improve at 2B and 4B, although the gains over SFT are smaller than those on POIE. These results show that OCR-derived supervision benefits both exact field matching and textual similarity, while the magnitude of improvement varies across datasets.

Table 5: **Extraction performance on FUNSD and POIE.** All scores are reported in %. Zero-shot denotes the original Qwen3.5 backbone without task fine-tuning. Bold marks the best score within each backbone size.

		FUNSD		POIE	
Backbone	Method	F1 ↑	ANLS ↑	F1 ↑	ANLS ↑
Qwen3.5-0.8B	Zero-shot	45.01	58.92	50.01	55.21
	SFT	53.51	78.59	81.71	89.11
	LAYER	**54.71**	**81.46**	**87.75**	**93.90**
Qwen3.5-2B	Zero-shot	56.97	78.86	68.06	75.52
	SFT	57.17	81.26	87.24	93.46
	LAYER	**57.46**	**81.46**	**91.07**	**96.29**
Qwen3.5-4B	Zero-shot	61.05	83.29	78.86	84.99
	SFT	65.36	88.48	89.03	93.86
	LAYER	**66.55**	**88.87**	**91.79**	**95.61**

C OCR Matching and Reliability Filtering

String alignment score. We apply Unicode NFKC normalization, remove all whitespace, and lowercase both the ground-truth field value and OCR block text. Punctuation, including decimal separators, currency symbols, and minus signs, is retained. For the resulting strings, the alignment score is computed as follows:

- If either string is empty, the score is zero.

- If the two nonempty strings are identical, the score is one.
- If either string is a substring of the other, the score is the shorter string length divided by the longer string length.
- Otherwise, the score is returned by `difflib.SequenceMatcher(None, v, o).ratio()`, where `v` and `o` are the normalized value and OCR text.

The final case uses Python's sequence-matching ratio, rather than normalized Levenshtein similarity. This string score is used only to construct supervision targets; it is distinct from the learned cosine similarity in CGCL and does not participate in backpropagation.

Candidate construction and filtering. OCR recognition errors and repeated values can produce incorrect or ambiguous positive correspondences. We therefore first collect all OCR blocks with an alignment score of at least 0.82 and store their scores rounded to four decimal places. A value is eligible for CGCL only if this candidate set contains exactly one block, its stored alignment score is at least T_{align}, and its OCR confidence is at least T_{conf}. We use $T_{\text{align}} = 0.98$ and $T_{\text{conf}} = 0.9$. The value must also occur exactly once in the serialized answer, allowing its answer span to be identified unambiguously. At most two eligible values are retained per document.

Uniqueness is checked on the initial candidate set, before applying the stricter alignment threshold. Thus, a value with multiple candidates is rejected even if only one candidate reaches T_{align}. For example, candidates with scores 1.0000 and 0.9231 both pass the initial threshold, making the correspondence ineligible.

Contrastive candidates and weighting. The reliability gate selects positive field–block correspondences; it does not restrict the candidate blocks used in the contrastive denominator. For each retained answer value, the uniquely matched block is the positive, and all other nonempty OCR blocks on the same page serve as negatives. The alignment score and OCR confidence act as hard gates rather than continuous loss weights. Retained values contribute equally to the per-document CGCL loss, which is zero when no value passes the gate.

D Implementation and Training Details

Data and evaluation. The controlled fine-tuning comparisons use 4,013 training examples. The unified test set contains 1,394 documents and 11,952 fields from CORD, FUNSD, SROIE, and POIE. Predictions are decoded greedily and parsed into key–value dictionaries for evaluation. Experiments are conducted on an NVIDIA H20 GPU.

Visual targets and latent representations. During training, recognized OCR strings are rendered within their normalized bounding boxes on a blank canvas. The rendered image is encoded through the same frozen visual pathway used for the original document image, including the VLM's native projection into the language-model hidden space. Adaptive average pooling reduces the complete teacher feature sequence to $T = 64$ vectors.

The query-slot states H_Q are decoder hidden states, whereas the predicted visual latents are the projection outputs $\hat{Z} = g_\phi(H_Q)$. The latent projection head g_ϕ is separate from the VLM's native visual projection. All query slots are processed together under causal attention, without feeding a predicted latent back to generate the next one. The projected sequence is subsequently inserted into the image-query context for autoregressive answer generation.

Special embeddings. The begin, query, and end embeddings are initialized from the backbone's native vision-start, image-pad, and vision-end token embeddings, respectively. They are implemented as three separate trainable parameters, without unfreezing the complete input embedding matrix. The shared query embedding is repeated at all T query positions; positional encoding distinguishes the slots.

Two-stage optimization. Stage I freezes the language and visual backbones and trains the latent projection head and begin/query embeddings using feature alignment. The objective combines mean squared error with per-slot cosine distance, as defined in the main text. Stage II keeps the visual

Table 6: **Latent and auxiliary-supervision settings.**

Setting	Value
Latent budget T	64
Teacher feature resampling	Adaptive average pooling
Stage II answer-loss weight	1.0
Stage II MSE-loss weight λ_{MSE}	0.05
Stage II CGCL weight λ_{CGCL}	0.005
Contrastive temperature τ	0.07
Initial string-match threshold	0.82
Alignment threshold T_{align}	0.98
OCR confidence threshold T_{conf}	0.9

pathway frozen and updates the language-model LoRA parameters, latent projection head, special embeddings, and CGCL modules. The Stage II objective weights are 1.0 for answer generation, 0.05 for latent MSE alignment, and 0.005 for CGCL. Both training and inference condition answer generation on predicted latents together with the original image-query context.

CGCL implementation. For each nonempty OCR block, we tokenize its text with a maximum length of 64 tokens and mean-pool the language model's input token embeddings. A learned text projection and a bounding-box encoder map textual and spatial information into a shared space. Their outputs are added and ℓ_2-normalized to obtain the block representation.

For each retained value, tokenizer offsets identify the overlapping answer tokens. Their decoder hidden states from the teacher-forced answer pass are mean-pooled, projected, and ℓ_2-normalized. CGCL applies a temperature-0.07 softmax contrastive loss over the nonempty OCR blocks on the same page. These block representations and the associated matching machinery are used only during training.