Is Cognition Consistent with Perception? Assessing and Mitigating Multimodal Knowledge Conflicts in Document Understanding

Zirui Shao[1*], **Feiyu Gao**[2*], **Zhaoqing Zhu**[2*], **Chuwei Luo**[2†], **Hangdi Xing**[1], **Zhi Yu**[1,3†], **Qi Zheng**[2], **Ming Yan**[2], **Jiajun Bu**[1]

[1] Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, Zhejiang University

[2]Alibaba Group,

[3] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and DataSecurity

{shaozirui, yuzhirenzhe}@zju.edu.cn, feiyu.gfy@alibaba-inc.com

{zzhaoqing.z, luochuwei}@gmail.com

Abstract

Multimodal large language models (MLLMs) have shown impressive capabilities in document understanding, a rapidly growing research area with significant industrial demand. As a multimodal task, document understanding requires models to possess both perceptual and cognitive abilities. However, due to different types of annotation noise in training, current MLLMs often face conflicts between perception and cognition. Taking a document VQA task (cognition) as an example, an MLLM might generate answers that do not match the corresponding visual content identified by its OCR (perception). This conflict suggests that the MLLM might struggle to establish an intrinsic connection between the information it "sees" and what it "understands". Such conflicts challenge the intuitive notion that cognition is consistent with perception, hindering the performance and explainability of MLLMs. In this paper, we define the conflicts between cognition and perception as *Cognition and Perception (C&P) knowledge conflicts*, a form of multimodal knowledge conflicts, and systematically assess them with a focus on document understanding. Our analysis reveals that even GPT-4o, a leading MLLM, achieves only 75.26% C&P consistency. To mitigate the C&P knowledge conflicts, we propose a novel method called *Multimodal Knowledge Consistency Fine-tuning*. Our method reduces C&P knowledge conflicts across all tested MLLMs and enhances their performance in both cognitive and perceptual tasks.

1 Introduction

In recent years, multimodal large language models (MLLMs) (OpenAI, 2023; Team et al., 2023; OpenAI, 2024; Chen et al., 2024b; Bai et al., 2025; Ye et al., 2024; Li et al., 2024a) have witnessed rapid development and have demonstrated remarkable capabilities across a wide range of multimodal tasks (Antol et al., 2015; Mathew et al., 2021; Hossain et al., 2019). Of particular note is their application in document understanding (Cui et al., 2021; Xu et al., 2020, 2021; Huang et al., 2022; Luo et al., 2023), an area of high academic and industrial value, where significant progress has been made (Zhang et al., 2023a; Ye et al., 2023a,b; Luo et al., 2024; Wang et al., 2023; Hu et al., 2024).

As a multimodal task, document understanding requires models to accurately perceive visual content (perception, e.g., OCR) and then generate coherent responses (cognition, e.g., VQA) based on that perception. However, current MLLMs train perception and cognition using different sources of annotation (Bai et al., 2024; Hu et al., 2024). Perception typically relies on external OCR engines, while cognition often depends on human-annotated or LLM-generated data (Mathew et al., 2021; Van Landeghem et al., 2023). This discrepancy leads to different noise profiles, creating conflicts between perception and cognition. As shown in Figure 1, GPT-4o (OpenAI, 2024) recognizes the text in a certain region of an image as "Doral" but responds to a related VQA question with the text "Doraf". This conflict suggests that GPT-4o struggles to establish a consistent connection between what it "sees" and what it "understands". Statistical analysis further underscores this issue, as Figure 2 demonstrates that leading MLLMs like GPT-4o achieve only 75.26% consistency between perception and cognition (Section 3).

In this paper, we define intrinsic conflicts between cognitive knowledge and perceptual knowledge within MLLMs, which result in inconsistencies in responses related to cognition and percep-

[*] Equal contribution.

[†] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30923–30944

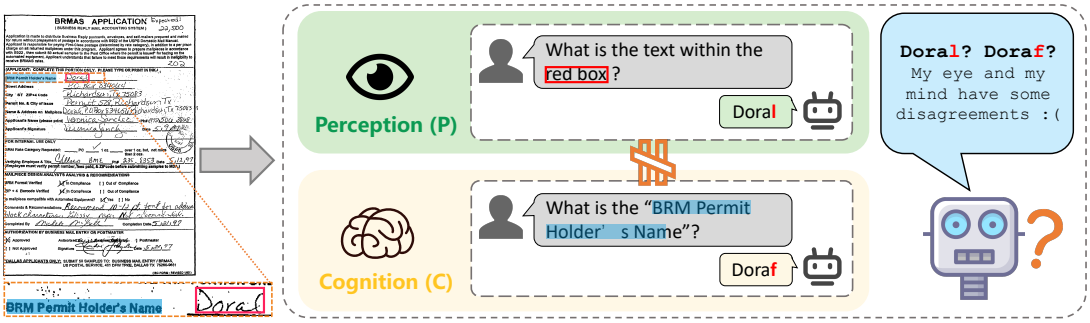

Figure 1: GPT-4o generates a VQA (cognition) answer that conflicts with the corresponding visual content identified by its OCR (perception). We refer to these multimodal knowledge conflicts in MLLMs as *Cognition and Perception (C&P) knowledge conflicts*.

tion, as *Cognition and Perception (C&P) knowledge conflicts* (Section 2.1). These conflicts undermine the explainability of MLLMs, as they challenge the intuitive notion that cognition is consistent with perception. Unlike previous research on multimodal knowledge conflicts (e.g., hallucination) (Zhai et al., 2024; Li et al., 2023; Guan et al., 2024; Liu et al., 2023), which focuses solely on conflicts within either cognition or perception, we highlight, for the first time, the conflicts that arise between the two.

We systematically evaluate C&P knowledge conflicts in the five current MLLMs (Section 3), focusing on document understanding. For documents, the primary perception task is the recognition of optical characters, while the primary cognitive task is the comprehension of text content. Therefore, we select OCR as the perceptual task and document-related VQA as the cognitive task. To ensure the validity of our evaluation, we eliminate potential confounding factors, such as model failures in following instructions. The experimental results reveal substantial C&P knowledge conflict in current MLLMs, highlighting the need to resolve these conflicts. To address this, we introduce a novel method called *Multimodal Knowledge Consistency Fine-tuning*. This method aims to strengthen the connection between cognitive and perceptual tasks through two key components (Section 4). First, a special token called *C&P Link Token* is introduced as a prompt prefix and suffix to connect cognitive and perceptual knowledge. Second, we design a *C&P Connector* that guides the model to cross-verify cognitive knowledge using perceptual knowledge.

Comprehensive experiments are conducted on three open-source MLLMs across two series and two parameter sizes. The results indicate that multimodal knowledge consistency fine-tuning improves C&P consistency (Section 5.2). Notably, our method also enhances MLLM performance in both cognitive and perceptual tasks (Section 5.3). This suggests that reducing C&P knowledge conflicts allows the model to better integrate perceptual and cognitive knowledge, thereby improving its overall capabilities.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to identify and introduce the concept of *Cognition and Perception knowledge conflicts*, a form of multimodal knowledge conflicts, in MLLMs.
- A systematic evaluation is conducted on current MLLMs to assess the Cognition and Perception knowledge conflicts in document understanding, showing that such conflicts are commonly present in current MLLMs.
- A novel method called *Multimodal Knowledge Consistency Fine-tuning* is introduced to mitigate the C&P knowledge conflicts in current MLLMs. Extensive experiments on five public document understanding benchmarks in three MLLMs demonstrate the effectiveness of the proposed method.

2 Problem Statement

2.1 The Definition of Cognition and Perception Knowledge Conflicts

For a given MLLM $f(\cdot)$, an image x_I, and a pair of queries consisting of a cognitive query x_C and a perceptual query x_P, we denote the ground truth for this pair as GT. The MLLM's responses for cognitive and perceptual tasks are represented as

$y_C = f(x_C, x_I)$ and $y_P = f(x_P, x_I)$, respectively.

In the training process of current MLLMs, annotations for perceptual tasks (e.g., OCR) and cognitive tasks (e.g., VQA) are often derived from different sources. For example, in the widely used DocVQA dataset (Mathew et al., 2021), OCR annotations are generated by commercial OCR solutions, while VQA annotations are crowd-sourced. Differences in annotation origins introduce discrepancies in noise and content, resulting in inconsistent bias that creates conflicts between cognitive and perceptual knowledge, referred to as *Cognition and Perception (C&P) knowledge conflicts*. Such conflicts manifest when y_C and y_P are inconsistent, i.e., $\delta(y_C, y_P) = 0$. It is important to note that C&P knowledge conflicts do not consider whether $y_C = GT$ or $y_P = GT$. To quantify the severity of these conflicts, we introduce *C&P consistency*. Let N denote the number of query pairs, with the C&P consistency calculated as follows:

$$\text{C\&P Consistency} = \frac{\sum_{i=1}^{N} \delta(y_{C_i}, y_{P_i})}{N}. \quad (1)$$

In this paper, we focus on document understanding and follow common practice (Fu et al., 2024; Chen et al., 2024a) by using OCR as a representative perceptual task and VQA as a representative cognitive task. Specifically, given a text GT within x_I bounded by Box, x_C is a VQA query using GT as the answer, and x_P is an OCR query operating solely within Box. In practice, Box may contain additional text besides GT. Consequently, C&P knowledge conflicts occur when y_P does not fully contain y_C. The $\delta(y_C, y_P)$ can be specifically defined as follows:

$$\delta(y_C, y_P) = \begin{cases} 1, & \text{if } y_C \subseteq y_P \\ 0, & \text{if } y_C \nsubseteq y_P \end{cases}. \quad (2)$$

Furthermore, performance gaps may cause models to exhibit C&P inconsistency. For example, MLLMs may fail to comprehend VQA questions. Therefore, we introduce an auxiliary metric, called "*Idealized C&P Consistency*," which evaluates inconsistencies only when both $ANLS(y_C, GT)$ and $ANLS(y_P, GT)$ are at least 0.5. The ANLS metric (Biten et al., 2019) is widely used in document understanding to measure text similarity on a scale from 0 to 1. Generally, cases with ANLS below 0.5 are considered complete failures of the model's response to a query. By filtering out these poor cases caused by model performance, this metric provides additional insight into the C&P consistency under ideal conditions.

2.2 The Construction of Evaluation Samples

To calculate C&P consistency, we construct pairs of cognitive (VQA) query and perceptual (OCR) query, i.e., (x_C, x_P), with each pair using the same ground truth GT from the image x_I. The process is as follows:

Given an image x_I with its QA annotation (Q, A), we assign $GT = A$ and $x_C = Q$. We construct x_P using visual prompting (Wu et al., 2024b; Yang et al., 2023). x_P is a simple question: "`What is the text within the red box?`" The corresponding image x_I^B is obtained by drawing a red box in x_I at the location of Box, denoted as $x_I^B = \text{VisP}(x_I, Box)$, where $\text{VisP}(\cdot)$ represents the visual prompting process and Box is the bounding box containing GT. In practice, responses for cognitive and perceptual tasks are obtained as $y_C = f(x_C, x_I)$ and $y_P = f(x_P, x_I^B)$, respectively.

Additionally, constructing (x_C, x_P) pairs involves several preprocessing steps. According to the definition in Section 2.1, the questions must pertain to the text in the image. However, certain questions, such as those related to comparisons or yes/no answers, do not directly reference the text. Moreover, since the current document datasets do not provide Box annotations, we also need to locate Box based on the OCR annotations of x_I. We employ GPT-4o to perform these preprocessing steps. Specific details are provided in Section A.2.

In particular, we consider five document understanding datasets to construct evaluation samples, which are categorized into three tasks: Document Question Answering (DocVQA (Mathew et al., 2021) and DUDE (Van Landeghem et al., 2023)), Document Information Extraction (DeepForm (Svetlichnaya, 2020) and FUNSD (Jaume et al., 2019)), and Chart Question Answering (ChartQA (Masry et al., 2022)). The evaluation samples are constructed from the test sets of these datasets. Section A.1 and A.2 provides additional details, including dataset descriptions, an example evaluation sample, and comprehensive statistics.

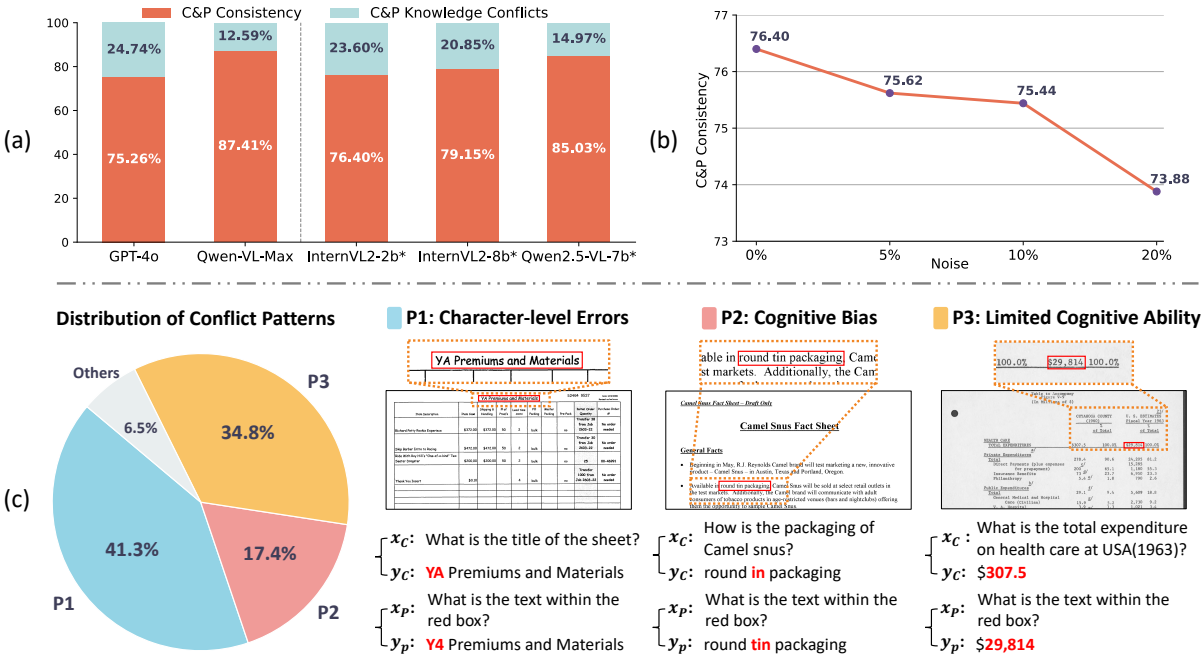

Figure 2: *a*: C&P knowledge conflicts in current MLLMs. "*" denotes the "SFT-baseline" (see Section 3). Additional quantitative results are provided in Section A.4 and Table 1. *b*: Results of the synthetic noise experiment, with additional details provided in Section A.5. *c*: The distribution of conflict patterns, including character-level errors (P1), cognitive bias (P2), and limited cognitive ability (P3), with one illustrative example for each.

3 The Cognition and Perception Knowledge Conflicts in Current MLLMs

Two closed-source and three open-source MLLMs are evaluated. The closed-source models, GPT-4o (OpenAI, 2024) and Qwen-VL-Max (Bai et al., 2024, 2025), are well-regarded in the community. We evaluate these models using their publicly available APIs, disabling all randomness-inducing hyperparameters. Additionally, to ensure that MLLMs follow instructions, we carefully adjust the prompts based on the characteristics of each dataset. Details are provided in Section A.4.

The open-source models include InternVL2-2b (Chen et al., 2024b), InternVL2-8b (Chen et al., 2024b), and Qwen2.5-VL-7b (Bai et al., 2025), which differ in size and architecture. We perform the evaluation by disabling all randomness-inducing hyperparameters on an Nvidia A100 GPU. Furthermore, we observe that using the original weights for inference leads to issues with instruction following (see Section A.4 for details), and thus we construct SFT data using the training sets from all datasets following the procedure outlined in Section 2.2 to train baseline models for each MLLM, referred to as "SFT-baseline". Training details are provided in Section 5.1.

Figure 2 (a) presents the evaluation results, with more quantitative results provided in Section A.4 and Table 1. Overall, C&P knowledge conflicts are common in current MLLMs, with inconsistencies observed in 12%–25% of cases. Furthermore, the severity of these conflicts appears comparable between open-source and closed-source models.

To further investigate the potential cause of C&P knowledge conflicts, we train InternVL2-2b with varying levels of synthetic noise (OCR: shape mix-ups, missing or extra letters; VQA: typos, omitted details). Synthetic noise is injected into 5%, 10%, and 20% of the training data. Figure 2 (b) shows the results, with additional quantitative analysis provided in Section A.5. Overall, as the level of noise increases, the C&P consistency declines.

We also randomly sample 10% of all inconsistent cases generated by InternVL2-2b and manually inspect them. Three main types of conflicts are identified, as shown in Figure 2 (c). The majority of conflicts (41.3%, P1) stem from character-level errors when the model responds to either the OCR or VQA query. P2 (17.4%) arises from cognitive bias. Although the model "sees" the correct text (its OCR output is accurate), it prefers a linguistically more plausible answer (e.g., substituting "round tin packaging" with "round in packaging"). The synthetic noise experiment, together with P1 and P2,

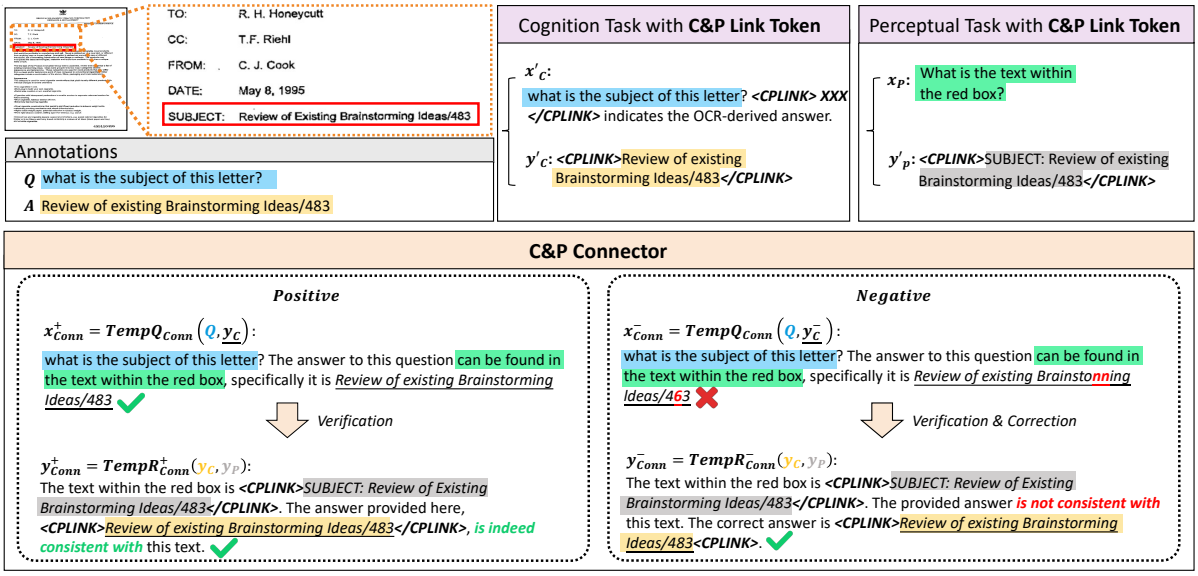

Figure 3: An example illustrates the source data and its corresponding *Multimodal Knowledge Consistency Fine-tuning* sample. All mathematical symbols in the figure are consistent with those in Section 4. Corresponding relationships use the same colors for clarity.

supports our hypothesis that heterogeneous VQA and OCR annotations are a primary source of C&P knowledge conflicts. Specifically, P1 reflects perception noise introduced by external OCR engine annotators, while P2 reflects semantic bias introduced by human or LLM annotators. P3 (34.8%) reveals a limitation in the model's cognitive ability, where the VQA response is hallucinated despite an accurate OCR output. To focus on purer conflict conditions, we exclude P3 from the idealized C&P consistency (Section 2.1).

4 Multimodal Knowledge Consistency Fine-tuning

Section 3 demonstrates that even state-of-the-art MLLMs exhibit C&P knowledge conflicts. To resolve these conflicts, we propose *Multimodal Knowledge Consistency Fine-tuning*, illustrated in Figure 3, which comprises two components: *C&P Link Tokens* and the *C&P Connector*. As heterogeneous VQA and OCR annotations are a primary source of C&P conflicts (Section 3), this method aims to reinforce the connection between cognitive and perceptual tasks, thereby mitigating C&P knowledge conflicts.

4.1 C&P Link Tokens

Previous research (Wu et al., 2024a) indicates that special tokens can effectively connect knowledge across different tasks. Therefore, we define a pair of *C&P Link Tokens* to connect cognitive and perceptual tasks, namely `<CPLINK>` and `</CPLINK>`, and add them to the original MLLM vocabulary. When the MLLM responds to a query using text extracted from an image, it encloses that text with the two C&P link tokens, for example, "`<CPLINK>XXX</CPLINK>`." Given an image x_I with QA annotation (Q, A), the cognitive task's query and response are (x_C, y_C) and the perceptual task's are (x_P, y_P). According to Section 2.2, both y_C and y_P are texts derived from x_I, i.e. A. Therefore, the C&P link tokens can be applied to the responses of both tasks, denoted as y'_C and y'_P, thereby strengthening their connection. Additionally, for guiding linked responses, we design x'_C to more explicitly prompt the model by adding a special instruction: "`<CPLINK>XXX</CPLINK> indicates the OCR-derived answer.`"

4.2 C&P Connector

The second component is the *C&P Connector*, which uses the question Q as an intermediary to link y_P and y_C, thereby bridging cognitive and perceptual tasks. The C&P Connector consists of positive and negative samples, denoted as (x^+_{Conn}, y^+_{Conn}) and (x^-_{Conn}, y^-_{Conn}), respectively. In terms of input images, the connector takes images with bounding boxes, x^B_I, as input (see Section 2.2 for details).

Positive samples aim to guide the model to use perceptual knowledge to verify cognitive knowledge. Specifically, as shown in Figure 3,

	DocVQA	DUDE	DeepForm	FUNSD	ChartQA	Average
InternVL2-2b*	80.59 90.62	64.69 83.00	72.05 77.40	80.84 87.95	83.80 91.27	76.40 86.05
InternVL2-2b (Ours)	**83.39** 91.32	**69.49** 84.75	**78.56** 82.20	**81.50** 89.60	**87.64** 93.17	**80.12** 88.21
InternVL2-8b*	84.28 91.32	67.82 83.14	74.19 77.70	82.60 91.82	86.88 91.86	79.15 87.17
InternVL2-8b (Ours)	**87.32** 93.03	**73.26** 84.70	**79.17** 82.22	**83.48** 90.13	**90.53** 94.22	**82.75** 88.86
Qwen2.5-VL-7b*	93.79 96.44	79.30 91.87	75.20 85.36	84.80 90.38	92.06 95.37	85.03 91.88
Qwen2.5-VL-7b (Ours)	**94.95** 97.10	**84.04** 94.22	**79.57** 86.73	**90.31** 94.07	**93.09** 95.74	**88.39** 93.57

Table 1: Performance comparison between the original MLLM and the MLLM after multimodal knowledge consistency fine-tuning (ours) across all datasets. All values are percentages (%). The main number is C&P Consistency, and the smaller number is Idealized C&P Consistency. Bolded numbers indicate superior performance. The average results are the macro-averages of all datasets. "*" denotes the "SFT-baseline" (see Section 3).

(x^+_{Conn}, y^+_{Conn}) is constructed as follows:

$$\begin{cases} x^+_{Conn} = \text{TempQ}_{Conn}(Q, y_C) \\ y^+_{Conn} = \text{TempR}^+_{Conn}(y_C, y_P) \end{cases}. \quad (3)$$

Here, $\text{TempQ}_{Conn}(\cdot)$ is the template for constructing C&P connector queries, and $\text{TempR}^+_{Conn}(\cdot)$ is the template for constructing positive sample responses. The model is required to first answer y_P, and then y_C, thus using perceptual knowledge to verify cognitive knowledge.

In addition to verification, negative samples further guide the model to use perceptual knowledge to correct erroneous cognitive results. Specifically, as shown in Figure 3, (x^-_{Conn}, y^-_{Conn}) is constructed as follows:

$$\begin{cases} x^-_{Conn} = \text{TempQ}_{Conn}(Q, y^-_C) \\ y^-_{Conn} = \text{TempR}^-_{Conn}(y_C, y_P) \end{cases}. \quad (4)$$

Here, the template for constructing queries is the same as that used for positive samples. y^-_C is an OCR-error version of y_C, generated using GPT-4o (refer to the Section A.7 for the specific prompt). $\text{TempR}^-_{Conn}(\cdot)$ is the template for generating negative sample responses, which require the model to first answer y_P, then indicate that y^-_C is incorrect, and finally provide the correct y_C.

The final training data, given N pairs of (Q, A), is represented as follows:

$$\mathcal{X} = \{(x'_{C_i}, y'_{C_i}), (x_{P_i}, y'_{P_i}), (x^+_{Conn_i}, y^+_{Conn_i}), (x^-_{Conn_i}, y^-_{Conn_i})\}^N_{i=0}. \quad (5)$$

5 Experiment

5.1 Implementation

We construct the training data using the training sets from the five datasets mentioned in Section 2.2. For the multimodal knowledge consistency fine-tuning experiment, we focus on three open-source MLLMs (Section 3): InternVL2-2b, InternVL2-8b, and Qwen2.5-VL-7b. We train all models using the original weights from Huggingface with a learning rate of 1e-5 and a batch size of 128, while keeping other hyperparameters at their default settings. We freeze the visual encoder and optimize only the language model. Each model trains for 1 epoch using 8 Nvidia A100 GPUs. We disable all randomness-inducing hyperparameters during inference.

5.2 C&P Consistency Results

The evaluation is conducted on the dataset constructed in Section 2.2. The experimental results, presented in Table 1, demonstrate that our multimodal knowledge consistency fine-tuning method enhances C&P consistency across all five datasets. Specifically, InternVL2-2b and InternVL2-8b show improvements of 3.72% and 3.60% in C&P consistency, respectively, while Qwen2.5-VL-7b exhibits

	Doc VQA		DUDE		Deep Form		FUNSD		Chart QA	
	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.
InternVL2-2b*	83.44	91.71	60.29	86.64	72.42	91.70	73.87	87.39	72.76	96.39
InternVL2-2b (Ours)	**85.37**	**93.24**	**62.44**	**88.78**	**75.50**	**94.09**	**76.34**	**88.69**	**75.84**	**97.28**
InternVL2-8b*	88.54	92.27	65.09	88.88	76.58	92.70	78.01	87.33	78.52	96.95
InternVL2-8b (Ours)	**89.47**	**94.01**	**67.18**	**90.41**	**77.08**	**94.58**	**78.16**	**89.77**	**82.80**	**97.55**
Qwen2.5-VL-7b*	94.79	90.67	70.11	87.56	50.17	95.64	79.75	89.39	87.76	95.29
Qwen2.5-VL-7b (Ours)	**95.40**	**91.85**	**71.10**	**88.66**	**57.58**	**96.90**	**80.52**	**91.29**	**88.32**	**95.74**

Table 2: The performance of cognitive and perceptual tasks. "C.T." and "P.T." stand for cognitive task (VQA) and perceptual task (OCR), respectively. Metrics are detailed in Section 5.3; all values are percentages (%), with bold indicating superior performance. "*" denotes the "SFT-baseline" (see Section 3).

#	Link.	Conn.	Doc VQA	Deep Form	Chart QA	Average
1			80.59 90.62	72.05 77.40	83.80 91.27	76.40 86.05
2		✓	82.97 91.52	77.85 80.97	87.45 93.66	79.10 87.77
3	✓		82.71 91.14	77.24 80.72	87.45 93.26	79.24 87.51
4	✓	✓	**83.39** 91.32	**78.56** 82.20	**87.64** 93.17	**80.12** 88.21

Table 3: Ablation study based on InternVL2-2b. All values are percentages (%), with the primary number representing C&P Consistency and the smaller representing Idealized C&P Consistency. The best results are in bold. "Link." and "Conn." denote C&P link token and C&P connector, respectively (see Section 4).

			Doc VQA		Deep Form		Chart QA	
#	Link.	Conn.	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.
1			83.4	91.7	72.4	91.7	72.8	96.4
2		✓	85.0	92.9	75.3	93.5	75.6	96.8
3	✓		85.1	93.1	75.2	94.0	75.4	97.1
4	✓	✓	**85.4**	**93.2**	**75.5**	**94.1**	**75.8**	**97.3**

Table 4: Ablation study based on InternVL2-2b. "C.T." and "P.T." denote cognitive (VQA) and perceptual (OCR) tasks. Metrics are in Section 5.3; values are percentages (%), with bold numbers indicating best performance. "Link." and "Conn." denote C&P link token and C&P connector, respectively (see Section 4).

a 3.36% increase. Under ideal conditions, consistency also improves. These findings indicate that our method effectively reduces C&P knowledge conflicts by linking perceptual and cognitive tasks. The comparison between Qwen2.5-VL-7b and the InternVL2 models highlights the general applicability of our approach across different MLLM architectures. Additionally, we perform two-sided paired t-tests using InternVL2-2b in Section A.9, showing that all gains in Table 1 are statistically significant.

5.3 The Performance of Cognitive and Perceptual Tasks

To assess the impact of C&P consistency on model performance, we evaluate the model's effectiveness on cognitive and perceptual tasks. For the cognitive task, following previous works (Borchmann et al., 2021; Lee et al., 2023; Luo et al., 2024), we evaluate DocVQA and FUNSD using ANLS (Biten et al., 2019), DeepForm using the F1 score, and ChartQA using relaxed accuracy (Methani et al., 2020). For the perceptual task, all datasets are evaluated using ANLS.

As shown in Table 2, the three MLLMs show improved performance on both cognitive and perceptual tasks across all datasets after the multimodal knowledge consistency fine-tuning. We attribute this improvement to our fine-tuning approach, which reduces the conflict between perceptual and cognitive knowledge, thereby promoting their integration. We believe that the results suggest that enhancing C&P consistency can strengthen the capabilities of MLLMs. Similar to Section 5.2, the t-tests in Section A.9 show that the performance gains are statistically significant.

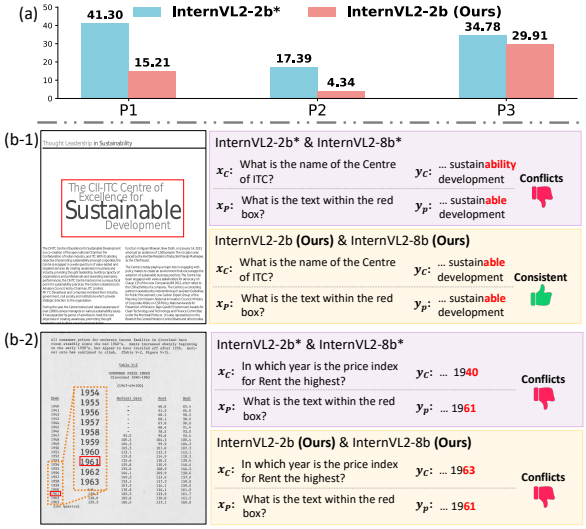

Figure 4: *a*: Comparison of the distribution of conflict patterns between InternVL2-2b* and InternVL2-2b (Ours). *b*: Two cases: *b-1* demonstrates the effectiveness of our method, while *b-2* reveals a limitation.

5.4 Ablation Study

To evaluate the contribution of each component in multimodal knowledge consistency fine-tuning, we conduct a series of ablation experiments using InternVL2-2b, as shown in Table 3 and Table 4. Due to space limits, we show three datasets here and provide the rest in Section A.6. Each experiment, with different fine-tuning tasks, is trained according to the settings outlined in Section 5.1. #2 removes all C&P link tokens from the training data, including those in the C&P connector. The results in Table 3 validate our hypothesis that both components in multimodal knowledge consistency fine-tuning are crucial for enhancing C&P consistency. For instance, on average, the C&P link token improves by 2.70%, and the C&P connector improves by 2.84%. Furthermore, Table 4 shows that our method achieves the best performance on cognitive and perceptual tasks.

5.5 Analysis of Conflict Patterns and Case Evidence

To further evaluate the effectiveness of multimodal knowledge consistency fine-tuning, we reuse the procedure described in Section 3 to analyze conflict patterns. Figure 4 (a) shows that, after fine-tuning, character-level errors (P1) and cognitive bias (P2) decrease significantly, making limited cognitive ability (P3) the dominant pattern. This shift supports our claim that heterogeneous VQA and OCR annotations are the primary sources of C&P knowledge conflicts and confirms that our method mitigates them effectively. The qualitative evidence in Figure 4 (b) illustrates these statistics. In case (b-1), categorized as P2, both InternVL2-2b and InternVL2-8b recognize "sustainable development" but incorrectly respond with "sustainability development" in the VQA task due to cognitive bias. The conflicts disappear after fine-tuning, as the models better integrate cognitive and perceptual knowledge. Notably, a similar P1 case with the same conclusion is provided in the Section A.8. In case (b-2), categorized as P3, the result indicates that our method cannot fundamentally extend the model's cognitive boundaries. In Figure 4, the responses of InternVL2-2b and InternVL2-8b are identical, reflecting the representativeness of these cases, though they differ in most other cases.

6 Related Work

MLLMs for Document Understanding Document understanding (Cui et al., 2021; Xu et al., 2021; Huang et al., 2022; Luo et al., 2023, 2024; Shao et al., 2023; Wang et al., 2023; Zhu et al., 2025; Mo et al., 2025b) is a rapidly growing research area driven by increasing industrial demand. Its main objective is to comprehend complex typeset images that contain rich textual information, such as scanned document pages (Mathew et al., 2021; Svetlichnaya, 2020; Stanisławek et al., 2021), charts (Masry et al., 2022; Kafle et al., 2018; Methani et al., 2020), tables (Pasupat and Liang, 2015; Chen et al., 2019; Mo et al., 2025a), and other formats (Tanaka et al., 2021; Mathew et al., 2022; Xing et al., 2024; Shao et al., 2024). As a multimodal task, document understanding involves automated processes for understanding, classifying, and extracting information, requiring models to possess both perceptual and cognitive capabilities (Cui et al., 2021). Recent studies (Chen et al., 2024b; Hong et al., 2024; Dong et al., 2024; Bai et al., 2025) for general MLLMs improve the encoding resolution of document images, significantly boosting performance in document understanding tasks. Several MLLMs are developed to focus on addressing document understanding problems, such as mPLUG-DocOwl (Ye et al., 2023a; Hu et al., 2024) and UReader (Ye et al., 2023b).

Knowledge Conflicts in LLMs LLMs are distinguished for encapsulating an extensive repository of world knowledge, known as the memory. Simultaneously, LLMs continue to engage with external

contextual knowledge post-deployment (Pan et al., 2023). The discrepancies between the contexts and the model's memory knowledge, i.e. context-memory conflicts, are being intensively studied recently (Xie et al., 2023). Another notable challenge arises with intra-memory conflict—a condition where LLMs exhibit unpredictable behaviors to inputs that are semantically equivalent but syntactically distinct (Chang and Bergen, 2023; Bartsch et al., 2023; Li et al., 2024b; Zhu et al., 2024; Zhang et al., 2024). This variance can be attributed to the conflicting knowledge embedded within the LLM's memory, which stems from the inconsistencies present in the complex and diverse pre-training datasets.

Hallucination issues in MLLMs MLLMs provide powerful tools for content generation across a wide range of tasks. However, they are susceptible to hallucinations (Bang et al., 2023; Zhang et al., 2023b; Liu et al., 2024b), where the generated outputs contain information not present in the visual input. These hallucinations typically arise when the models overly rely on the strong priors of their language modules. Such conflicts between MLLMs' language and visual perception raise concerns about their reliability and limit their applications (Ji et al., 2023; Kaddour et al., 2023). Current research primarily focuses on detecting and evaluating hallucinations (Li et al., 2023; Zhang et al., 2023b), as well as methods to reduce them (Liu et al., 2024a; Wang et al., 2024). To mitigate hallucinations, efforts have been directed toward enhancing data collection and training procedures (Liu et al., 2024a; Wang et al., 2024). Nevertheless, research on how MLLMs integrate perception and cognition knowledge, which is also vital for interpreting and debugging these models, has not progressed at the same pace.

7 Conclusion

In this paper, we identify that current MLLMs often face conflicts between cognition and perception, referred to as *Cognition and Perception (C&P) knowledge conflicts*. The severity of these conflicts is systematically assessed across five document understanding datasets, revealing that even leading MLLMs still struggle with these multimodal knowledge conflicts. To address this problem, a novel method called *Multimodal Knowledge Consistency Fine-tuning* is introduced. Comprehensive experiments demonstrate the effectiveness of our method in reducing C&P knowledge conflicts. Additionally, our method improves the performance of MLLMs in both cognitive and perceptual tasks.

Limitations

Despite contributing to the identification and mitigation of C&P knowledge conflicts, several limitations remain. This work simplifies cognition and perception to VQA and OCR tasks, potentially overlooking other cognitive abilities (e.g., multi-step reasoning, layout-aware inference) and perceptual channels (e.g., color, shape). We address these omissions in future work. Moreover, the current focus is on document understanding. We plan to extend our research to broader multimodal domains, such as general open-world images and video streams, to further explore C&P knowledge conflicts.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62372408).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. 2023. Self-consistency of large language models under ambiguity. *EMNLP 2023*, page 89.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. 2021. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey. *Preprint*, arXiv:2303.11504.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. 2024a. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1086–1104.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, and 1 others. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and 1 others. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and 1 others. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

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 *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei

Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2024b. Benchmarking and improving generator-validator consistency of language models. In *The Twelfth International Conference on Learning Representations*.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jentse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. 2024b. Insight over sight? exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7092–7101.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Ye Mo, Gang Huang, liangcheng li, Dazhen Deng, Zhi Yu, Yilun Xu, Kai Ye, Sheng Zhou, and Jiajun Bu. 2025a. Tablenarrator: Making image tables accessible to blind and low vision people. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Ye Mo, Zirui Shao, Kai Ye, Xianwei Mao, Bo Zhang, Hangdi Xing, Peng Ye, Gang Huang, Kehan Chen, Zhou Huan, and 1 others. 2025b. Doc-cob: Enhancing multi-modal document understanding with visual chain-of-boxes reasoning. *arXiv preprint arXiv:2505.18603*.

OpenAI. 2023. Gpt-4v(ision) system card.

OpenAI. 2024. Gpt-4o system card.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2023. Knowledge-in-context: Towards knowledgeable semi-parametric language models. *Preprint*, arXiv:2210.16433.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480.

Zirui Shao, Feiyu Gao, Zhongda Qi, Hangdi Xing, Jiajun Bu, Zhi Yu, Qi Zheng, and Xiaozhong Liu. 2023. Gem: Gestalt enhanced markup language model for web understanding via render tree. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6132–6145.

Zirui Shao, Feiyu Gao, Hangdi Xing, Zepeng Zhu, Zhi Yu, Jiajun Bu, Qi Zheng, and Cong Yao. 2024. Webrpg: Automatic web rendering parameters generation for visual presentation. In *European Conference on Computer Vision*, pages 56–74. Springer.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer.

S Svetlichnaya. 2020. Deepform: Understand structured documents at scale.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. 2024. Vigc: Visual instruction generation and correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5309–5317.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, and 1 others. 2024a. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, and 1 others. 2024b. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Hangdi Xing, Changxu Cheng, Feiyu Gao, Zirui Shao, Zhi Yu, Jiajun Bu, Qi Zheng, and Cong Yao. 2024. Dochienet: A large and diverse dataset for document hierarchy parsing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1129–1142.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2021. 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)*, pages 2579–2591.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. 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*, pages 1192–1200.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, and 1 others. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, and 1 others. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2024. Halle-switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023a. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14472–14482.

A Additional Details

A.1 Details of Selected Datasets

We consider five document understanding datasets to assess C&P knowledge conflicts, categorized into the following three tasks:

Document QA. DocVQA (Mathew et al., 2021) contains 50k question-answer pairs from 12k document images in the UCSF Industry Documents Library. DUDE (Van Landeghem et al., 2023) covers diverse domains, including medical, legal, technical, and financial, providing 41k question-answer pairs from 5k documents. We exclude all multi-page VQA annotations from DUDE, retaining only single-page annotations.

Document IE. DeepForm (Svetlichnaya, 2020) and FUNSD (Jaume et al., 2019) are two Information Extraction datasets. DeepForm consists of 1.1k documents related to election spending. FUNSD contains 0.2k document images from the RVL-CDIP dataset (Harley et al., 2015). The annotations for DeepForm and FUNSD are transformed into a question-answer format, with DeepForm following Hu et al. (2024), and FUNSD following Luo et al. (2024). The annotations in Hu et al. (2024) for DeepForm incorrectly assume that all key values are on the first page, ignoring that DeepForm documents are multi-page. We correct this issue (see Section A.3 for details), ensuring information extraction occurs on the correct pages.

Chart QA. ChartQA (Masry et al., 2022) compiles a diverse range of topics and chart types from four primary sources: Statista (statista.com), The Pew Research Center (pewresearch.org), Our World in Data (ourworldindata.org), and the OECD (oecd.org). In total, the dataset includes 21k chart images and 32k question-answer pairs.

Notably, OCR annotations are required in Section 2.2. For DocVQA and DUDE, the official OCR annotations are utilized, whereas the other datasets employ OCR annotations generated by a commercial OCR solution.

A.2 Details of Evaluation Sample Construction

As described in Section 2.2, the construction of (x_C, x_P) pairs involves several preprocessing steps. According to the definition in Section 2.1, the questions must pertain to the text in the image. However, certain questions, such as those related to comparisons or yes/no answers, do not directly reference the text. To address this, we filter out such QA pairs using GPT-4o with the prompt detailed in Table 12. Moreover, since the Box annotations are not provided, we employ GPT-4o to locate Box based on the OCR annotations of x_I with the prompt detailed in Table 13. We use GPT-4o to find Box because a single image may contain multiple occurrences of the text A in different locations. Therefore, identifying the correct Box requires semantic understanding, which GPT-4o excels at. QA pairs for which GPT-4o cannot find a Box, or the Box found does not contain A, are also excluded. Additionally, an example of an evaluation sample is provided in Section 5. Table 5 provides the statistics of evaluation data, including the number of (x_C, x_P) pairs and their corresponding images.

	Doc VQA	DUDE	Deep Form	FUNSD	Chart QA
# (x_C, x_P)	4575	1855	984	454	1562
# Images	1268	1101	248	46	1278

Table 5: Data statistics for C&P knowledge conflicts evaluation. The number of evaluation samples, i.e., cognitive (VQA) query and perceptual (OCR) query (x_C, x_P) pairs, along with the corresponding images for each dataset.

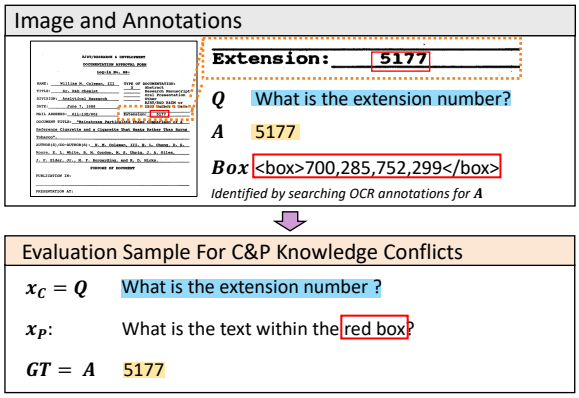

Figure 5: A specific example illustrates the evaluation sample. All mathematical symbols in the figure are consistent with those in Section 2.2. Corresponding relationships are represented using the same colors for clarity.

A.3 Details of DeepForm Single-page QA Annotations

As described in Section A.1, Hu et al. (2024) provide incorrect annotations for DeepForm because they assume all key values are on the first page, overlooking that DeepForm documents are multi-page. To address this, we use GPT-4o to identify

	DocVQA	DUDE	DeepForm	FUNSD	ChartQA	Average
GPT-4o	85.58 93.35	67.84 87.30	62.70 71.20	78.76 90.52	81.41 92.38	75.26 86.95
Qwen-VL-Max	95.66 97.15	82.54 91.35	83.23 86.69	83.19 90.52	92.44 95.68	87.41 92.28

Table 6: C&P knowledge conflicts in current MLLMs. All values are percentages (%), where the primary number represents C&P Consistency and the smaller number represents Idealized C&P Consistency.

#	Link.	Conn.	DUDE	FUNSD
1			64.69 83.00	80.84 87.95
2		✓	67.49 84.14	79.74 88.56
3	✓		68.84 84.46	79.96 87.97
4	✓	✓	**69.49** 84.75	**81.50** 89.60

Table 7: Ablation study based on InternVL2-2b. All values are percentages (%), with the primary number representing C&P consistency and the smaller representing idealized C&P consistency. Best results are in bold. "Link." and "Conn." denote C&P link token and C&P connector, respectively, as detailed in Section 4.

			DUDE		FUNSD	
#	Link.	Conn.	C.T.	P.T.	C.T.	P.T.
1			60.29	86.64	73.87	87.39
2		✓	62.32	87.56	75.89	88.02
3	✓		61.68	88.43	76.20	88.51
4	✓	✓	**62.44**	**88.78**	**76.34**	**88.69**

Table 8: Ablation study based on InternVL2-2b. "C.T." and "P.T." denote cognitive (VQA) and perceptual (OCR) tasks. Metrics are in Section 5.3; values are percentages (%), with bold numbers indicating best performance. "Link." and "Conn." denote C&P link token and C&P connector, respectively (see Section 4).

the correct page for information extraction using the prompt detailed in Table 14, ensuring that all single-page QA annotations in DeepForm are correct.

A.4 Additional Details of C&P Knowledge Conflicts Evaluation

As described in Section 3, to ensure that closed-source MLLMs follow instructions, we carefully adjust the prompts based on the characteristics of each dataset. For cognitive tasks, the prompts for DocVQA and DUDE are detailed in Table 15, DeepForm in Table 16, FUNSD in Table 17, and ChartQA in Table 18. For perceptual tasks, the prompts are detailed in Table 19. Table 6 presents the additional evaluation results of C&P knowledge conflicts in closed-source MLLMs.

Additionally, Table 9 presents the performance of closed-source MLLMs on cognitive and perceptual tasks. The results demonstrate that closed-source MLLMs perform well on both tasks, indicating that they effectively follow instructions and validating the results reported in Section 3.

We also report in Table 9 the performance of open-source MLLMs with original weights on cognitive and perceptual tasks. The results show that open-source MLLMs perform exceptionally poorly on some datasets, highlighting the necessity of using the "SFT-baseline" in Section 3.

We evaluate the closed-source MLLMs via their publicly available APIs. Specifically, we use the snapshots of GPT-4o[1] from 2024-11-20 and Qwen-VL-Max[2] from 2025-04-08. For open-source MLLMs, we use the model weights available on Hugging Face, including InternVL2-2B[3], InternVL2-8B[4], and Qwen2.5-VL-7B[5].

A.5 Additional Results of the Synthetic Noise Experiment

The additional results of the synthetic noise experiment based on InternVL2-2b (Section 3) show in

[1] https://platform.openai.com/docs/models/gpt-4o

[2] https://www.alibabacloud.com/help/en/model-studio

[3] https://huggingface.co/OpenGVLab/InternVL2-2B

[4] https://huggingface.co/OpenGVLab/InternVL2-8B

[5] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

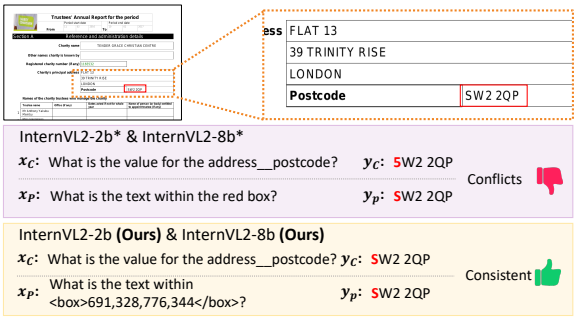

Figure 6: An additional case demonstrating the effectiveness of our method.

Table 10.

A.6 Additional Results of the Ablation Study

Due to space constraints, we report the results of only three datasets in Section 5.4. The results of the remaining two datasets are presented in Table 7 and Table 8.

A.7 Additional Details of C&P Connector

As described in Section 4, the negative samples for the C&P Connector are required to use the OCR-error version of y_C, denoted as $y_{\bar{C}}$, which is generated using GPT-4o with the prompt detailed in Table 20.

A.8 Additional Case Study

We present a case in Figure 6, categorized as P1 (Section 3), which provides evidence that multimodal knowledge consistency fine-tuning mitigates C&P knowledge conflicts.

A.9 Additional Results of the T-test

We perform two-sided paired t-tests using InternVL2-2b, and the results are shown in Table 11.

	Doc VQA		DUDE		Deep Form		FUNSD		Chart QA	
	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.	C.T.	P.T.
GPT-4o	89.14	86.55	62.70	74.58	37.17	85.51	75.95	87.39	68.04	95.22
Qwen-VL-Max	95.88	92.25	70.94	87.91	45.33	97.00	83.18	91.06	87.48	94.28
InternVL2-2b	87.03	66.66	59.96	52.25	19.37	34.28	74.02	59.08	76.40	65.96
InternVL2-8b	91.73	74.18	65.96	59.80	21.63	63.67	75.84	70.22	83.12	73.05
Qwen2.5-VL-7b	95.55	87.95	69.79	82.60	37.98	94.79	78.37	82.09	87.60	92.66

Table 9: The performance of cognitive and perceptual tasks, consisting of two groups: the results of closed-source models and the results of open-source models with original weights. "C.T." and "P.T." stand for cognitive task (VQA) and perceptual task (OCR), respectively. Metrics are detailed in Section 5.3, and all values are reported as percentages (%).

	DocVQA	DUDE	DeepForm	FUNSD	ChartQA	Average
0%	**80.59** 90.62	**64.69** 83.00	**72.05** 77.40	**80.84** 87.95	**83.80** 91.27	**76.40** 86.05
5%	79.12 90.97	64.09 82.70	71.36 76.31	79.74 87.67	83.81 90.12	75.62 85.55
10%	79.01 90.08	62.58 81.80	72.17 76.85	80.18 87.34	83.24 90.39	75.44 85.29
20%	77.92 89.93	61.61 81.17	70.24 75.38	77.53 85.45	82.08 90.52	73.88 84.49

Table 10: The synthetic noise experiment based on InternVL2-2b. All values are percentages (%), where the primary number represents C&P Consistency and the smaller number represents Idealized C&P Consistency.

	DocVQA	**DUDE**	**DeepForm**	**FUNSD**	**ChartQA**
C&P Consistency	5.22 (1.84×10^{-7})	4.75 (2.18×10^{-6})	5.41 (7.94×10^{-8})	2.36 (7.22×10^{-3})	4.72 (2.61×10^{-6})
Cognitive Task	4.69 (2.76×10^{-6})	3.28 (1.05×10^{-3})	3.32 (9.17×10^{-4})	2.10 (3.60×10^{-2})	4.55 (5.70×10^{-6})
Perceptual Task	6.52 (7.99×10^{-11})	4.49 (7.43×10^{-6})	5.81 (8.57×10^{-9})	2.53 (1.28×10^{-2})	3.20 (1.38×10^{-3})

Table 11: Results of two-sided paired t-tests using InternVL2-2b, reported as t-statistics with p-values in parentheses.

Prompt	You are tasked with determining whether the provided question-answer pairs are examples of extractive question answering (Extractive QA). **You have been provided with the following:** 1. The document image. 2. A list of question-answer pairs. **Here are the questions and answers:** {Question_Answering} **Definition of Extractive QA** In the domain of document understanding, Extractive Question Answering (Extractive QA) refers to systems that analyze and comprehend both the visual and textual information within a document to directly extract answers to user queries from the document's existing content. The answers are typically located in specific sections of the document, eliminating the need for complex reasoning or the generation of new content. Extractive QA emphasizes precise localization and extraction of information to ensure the accuracy and verifiability of the answers. **Non-Extractive QA Question Types:** 1. **Counting Questions:** These require the system to count specific elements or occurrences within the document, such as "How many times is the term 'machine learning' mentioned in the report?" 2. **Comparing Questions:** These involve evaluating and contrasting different pieces of information within the document, such as "Which department had a higher budget allocation in Q2, Marketing or Sales?" 3. **Causal Reasoning:** These questions require understanding cause-effect relationships within the document, such as "What caused the increase in operational costs?" 4. **Synthesis Questions:** These require summarizing or aggregating information from the document, such as "Summarize the key findings of the annual report." 5. **Inference Questions:** These ask for conclusions based on implicit information within the document, such as "What can be inferred about the company's market strategy from the sales data?" **Your Task** For each question in the list, determine whether it is an example of extractive QA based on the definition provided. **Important:** - **Do not include any explanatory content in your response.** - **Respond in the following format for each question:** - If the question is extractive QA, respond with: "Yes". - If the question is not extractive QA, respond with: "No". **Example Response:** Q1: Yes Q2: No Q3: Yes
Slots	Question_Answering List of question-answering annotations for the given images.

Table 12: Prompt for using GPT-4o to filter the questions that do not directly reference the text.

Prompt	You are tasked with identifying the locations of answers to multiple questions about a document image. **You have been provided with the following:** 1. The document image. 2. A list of questions along with their corresponding answers. 3. Text extracted from the document image using an Optical Character Recognition (OCR) engine by a third party. **Here are the questions and answers:** {Question_Answering} **Here is the text extracted by the OCR engine:** {OCR_Text} **Your task:** For each question in the list, first determine whether the answer text can be found within the document image based on the OCR-extracted text. If the answer is present, identify the box ID(s) that contain the correct answer. Each answer appears **only once** in the document image and may be entirely within a single box or span multiple adjacent boxes, either horizontally or vertically. Include all relevant box IDs that collectively constitute the answer. If the answer text cannot be found in any box, indicate this as well. **It is important to emphasize that you should identify only the boxes that contain the correct answer text, not the boxes that are relevant to answering the question.** In other words, even if a question explicitly mentions a specific box, if the answer text does not appear in that box, it should not be considered. Keep in mind that you need to find the box that semantically matches the answer, not just the box with the answer text. This means you should fully consider all the information from the document image, including images, text, layout, and style. **Important:** - **Do not include any explanatory content in your response.** - **Respond in the following format for each question:** - If you find the box(es) containing the true answer, respond with: "Found [Box IDs]" - If you cannot find any boxes containing the true answer, respond with: "Not Found" **Example Response:** Q1: Found [9, 12] Q2: Not Found Q3: Found [15]	
Slots	Question_Answering	List of question-answering annotations for the given images.
	OCR_Text	JSON-formatted OCR text for the given images.

Table 13: Prompt for using GPT-4o to locate Box based on the OCR annotations of given image x_I.

Prompt	You are given several images with the page number indicated in the top left corner. You will also receive a number of independent question-answer pairs. For each question, your task is to identify which numbered page provide the information needed to arrive at the given answer. Note: - Please identify which page these key-value pairs are most likely to appear on. - Output only question-answer pair id and its corresponding number. Format: Q1:number {Question_Answering}
Slots	Question_Answering List of question-answering annotations for the given images.

Table 14: Prompt for using GPT-4o to identify the correct page for information extraction on DeepForm.

Prompt	You are asked to answer questions asked on a document image. The answers to questions are short text spans taken verbatim from the document. This means that the answers comprise a set of contiguous text tokens present in the document. Question: {Question} Directly extract the answer of the question from the document with as few words as possible. Answer:
Slots	Question The question about the given image.

Table 15: Prompt for evaluating close-source MLLMs on cognitive task in DocVQA and DUDE.

Prompt	You are now working on DeepForm, a dataset for extracting text from visually structured political ad receipts. This dataset focuses on five key fields: 1. **contract_num**: Contract number (multiple documents can share the same number if a contract is revised) 2. **advertiser**: Advertiser name (often a political committee, but not always) 3. **flight_from / flight_to**: Start and end air dates for the ad (also known as "flight dates") 4. **gross_amount**: Total amount paid for the ads The answer always appears in the document, but it may not match the exact words of the question or field name. Provide a contiguous text span from the form, and include no additional explanation besides the answer. Question: {Question} Answer:
Slots	Question The question about the given image.

Table 16: Prompt for evaluating close-source MLLMs on cognitive task in DeepForm.

Prompt	You are now working on FUNSD, a dataset for form understanding in scanned documents. These documents often contain text arranged in various sections, tables, or multi-line blocks, and your goal is to extract the text that directly answers each question. Your task is to return the contiguous text snippet from the document that fully answers each question. The answer is guaranteed to be present in the form image, so do not refuse. If the relevant text spans multiple lines or rows in a table, ensure you include all of them exactly as they appear. Avoid adding explanations or summarizing the text; simply return a contiguous text snippet from the form that best addresses the question. Question: {Question} Answer:
Slots	Question The question about the given image.

Table 17: Prompt for evaluating close-source MLLMs on cognitive task in FUNSD.

Prompt	You are analyzing a chart that may include numeric data, textual labels, and visual features (e.g., bars, lines, colors). Below are some example questions and answers from other charts—these examples are not from this chart. When answering the current question, rely solely on the information in the chart you are analyzing, and provide a concise answer based strictly on the chart's data. Avoid outside knowledge or extra explanations. Additionally, the question is guaranteed to have an answer found in the chart. For numeric answers, remove any commas or symbols (e.g., "%") unless specifically asked for. For instance, "37,133" should be written as "37133" and "32.4%" should be written as "32.4." Question: {Question} Answer:
Slots	Question The question about the given image.

Table 18: Prompt for evaluating close-source MLLMs on cognitive task in ChartQA.

Prompt	Analyze the provided image, which has a **single red box** containing text. **Extract only** the text inside this box, preserving the **original line order** from **top** to **bottom**. If there are multiple lines, output them **separately**; if there's just one line, output it **as is**. **Do not** include any text or descriptions from outside the red box, and **do not** add any extra punctuation, commentary, or code block markers. Return **only** the exact text inside the red box.

Table 19: Prompt for evaluating close-source MLLMs on perceptual task.

Prompt	**Task Description** You are tasked with generating potential OCR (Optical Character Recognition) error results based on the provided list of question-answer (QA) pairs. **Provided Content:** **List of QA Pairs:** {Question_Answering} **Your Task** For each QA pair, provide **3 possible OCR error results for the answer (A)**. **Each error result must maintain a similar format, contain different content, must not be identical to the original answer (A), and must be distinct from the other error results.** **Output Format** Please respond in **JSON** format according to the structure provided below. Note that "error1," "error2," and "error3" are merely placeholders.	
Slots	Question_Answering	List of question-answering annotations for the given images.

Table 20: Prompt for using GPT-4o to generate y_C^- (Section 4).