

GDI-Bench: A Benchmark for General Document Intelligence with Vision and Reasoning Decoupling

Siqi Li^{1,2} Yufan Shen¹ Xiangnan Chen^{1,2} Jiayi Chen³ Hengwei Ju^{1,4}
 Haodong Duan¹ Song Mao¹ Hongbin Zhou¹ Bo Zhang¹ Pinlong Cai¹
 Licheng Wen¹ Botian Shi¹ Yong Liu^{2,†} Xinyu Cai^{1,*,†} Yu Qiao¹
¹ Shanghai Artificial Intelligence Laboratory ² Zhejiang University
³ School of Science and Engineering, The Chinese University of Hong Kong
⁴ Fudan University

Abstract

The rapid advancement of multimodal large language models (MLLMs) has profoundly impacted the document domain, creating a wide array of application scenarios. This progress highlights the need for a comprehensive benchmark to evaluate these models’ capabilities across various document-specific tasks. However, existing benchmarks often fail to locate specific model weaknesses or guide systematic improvements. To bridge this gap, we introduce a General Document Intelligence Benchmark (GDI-Bench), featuring 1.9k images across 9 key scenarios and 19 document-specific tasks. By decoupling visual complexity and reasoning complexity, the GDI-Bench structures graded tasks that allow performance assessment by difficulty, aiding in model weakness identification and optimization guidance. We evaluate the GDI-Bench on various open-source and closed-source models, conducting decoupled analyses in the visual and reasoning domains. For instance, the GPT-4o model excels in reasoning tasks but exhibits limitations in visual capabilities. To address the diverse tasks and domains in the GDI-Bench, we propose a GDI Model that mitigates the issue of catastrophic forgetting during the supervised fine-tuning (SFT) process through an intelligence-preserving training strategy. Our model achieves state-of-the-art performance on previous benchmarks and the GDI-Bench. Both our benchmark and model will be open source.

1 Introduction

Rapid progress of large language models (LLMs) [1, 2] has placed multimodal large language models (MLLMs) [3, 4, 5] as a foundation of artificial intelligence, advancing document intelligence to a general stage. Cross-domain and multi-scale document understanding and extraction challenges are increasingly critical in real-world applications. With the rise of MLLMs, a series of more complex benchmarks have emerged to provide comprehensive frameworks for document understanding tasks [6, 7, 8, 9, 10, 11, 12]. However, given that document understanding involves multiple modalities, errors in MLLM outputs may arise from inaccurate visual recognition, limited language organization, or both. Consequently, a decoupled evaluation of MLLMs’ document processing abilities is essential.

To address these challenges, we propose the General Document Intelligence Benchmark (GDI-Bench), which aims to perform a decoupled evaluation of model performance in document tasks, thereby contributing to the identification of the model’s weaknesses. The GDI-Bench introduces three key improvements over existing benchmarks: (1) developing a cross-domain, multi-task benchmarks to ensure task diversity and fine-grained difficulty levels; (2) introducing complexity decoupling,

* project leader, † corresponding author

dividing multimodal document understanding into visual complexity and reasoning complexity, and for the first time, establishing a difficulty grading mechanism; (3) supporting the evaluation of MLLMs, OCR+LLM-level systems, and document parsing tools, offering comprehensive guidance for practical application solutions.

We evaluate 2 open-source and 4 closed-source large models using the GDI-Bench and find their performance to be suboptimal. To enhance the model’s performance on GDI-Bench, we further conduct supervised fine-tuning (SFT) on the InternVL3-8B model to explore whether a data-driven approach could help the model acquire extensive domain-specific document knowledge. However, we observe that supervised fine-tuning tends to cause the issue of catastrophic forgetting. To address this problem, we propose the Layer-wise Adaptive Freezing Tuning (LW-AFT) method, which alleviates the impact of catastrophic forgetting and enhances the model’s cross-domain and cross-task capabilities. Specifically, during the SFT process, LW-AFT freezes most of the parameters, with only a small subset of domain-sensitive parameters participating in gradient updates.

Our contributions are summarized as follows.

- We propose the GDI-Bench, which encompasses a broad range of tasks in the document domain. By decoupling complexity and grading difficulty, it helps the model identify its weaknesses and guides subsequent optimization.
- We propose Layer-wise Adaptive Freeze-Tuning, a training method that effectively alleviates catastrophic forgetting in document task SFT by parameter-freezing, improving performance on specific tasks while maintaining generalization capabilities.
- We propose a GDI model that achieves state-of-the-art (SOTA) performance on multiple document domain benchmarks as well as the GDI-Bench, demonstrating high generalization capabilities suitable for real-world applications.

2 Related Works

2.1 Document Benchmark

Early benchmarks in document understanding predominantly targeted single-domain data and specific tasks. For instance, DocVQA[13] focused on industrial document QA, VisualMRC[14] addressed web-based document comprehension, and ChartQA[15] specialized in chart-based question answering. Despite their contributions, these benchmarks lacked difficulty stratification and cross-domain generalization, largely due to the limited generalization capabilities of models available at the time. With the rise of multimodal large models, a series of more complex benchmarks emerged [6, 7, 8, 9, 10, 11, 12], aiming to provide a more comprehensive framework for document understanding tasks. OCRBenchV2[16] provides multiple OCR-related subtasks, expanding the scope of OCR evaluation. Fox[17] introduced region-of-interest (ROI)-based document parsing methods, enhancing the detailed analysis of document structures, while OmniDocBench[18] extended the evaluation to cross-modal tasks, covering a broader range of document parsing challenges. While these benchmarks have broadened the scope of evaluation, they still lack explicit difficulty grading, which limits their ability to assess model performance across tasks of varying complexity.

2.2 Document Understanding Model

Optical Character Recognition (OCR) has long been a fundamental task in computer vision. Existing OCR models can be broadly categorized into component-based and end-to-end approaches. Component-based methods [19, 20, 21] adopt a modular pipeline that assembles multiple expert-designed components such as layout analysis [22], text detection [23, 24, 25, 26], region extraction, and content recognition [27, 28, 29]. In contrast, end-to-end OCR models [30], especially those driven by Large Vision-Language Models (LVLMs)[31, 3, 32, 33, 34, 35, 36], aim to unify perception and reasoning within a single architecture. Most LVLM-based OCR systems utilize CLIP [37] as the vision backbone, while coupling it with a language model to jointly process visual and textual information in a unified framework. Recent works [34, 35, 38] adopt sliding-window strategies that partition the image into patches to cope with long and high-resolution inputs like PDFs.

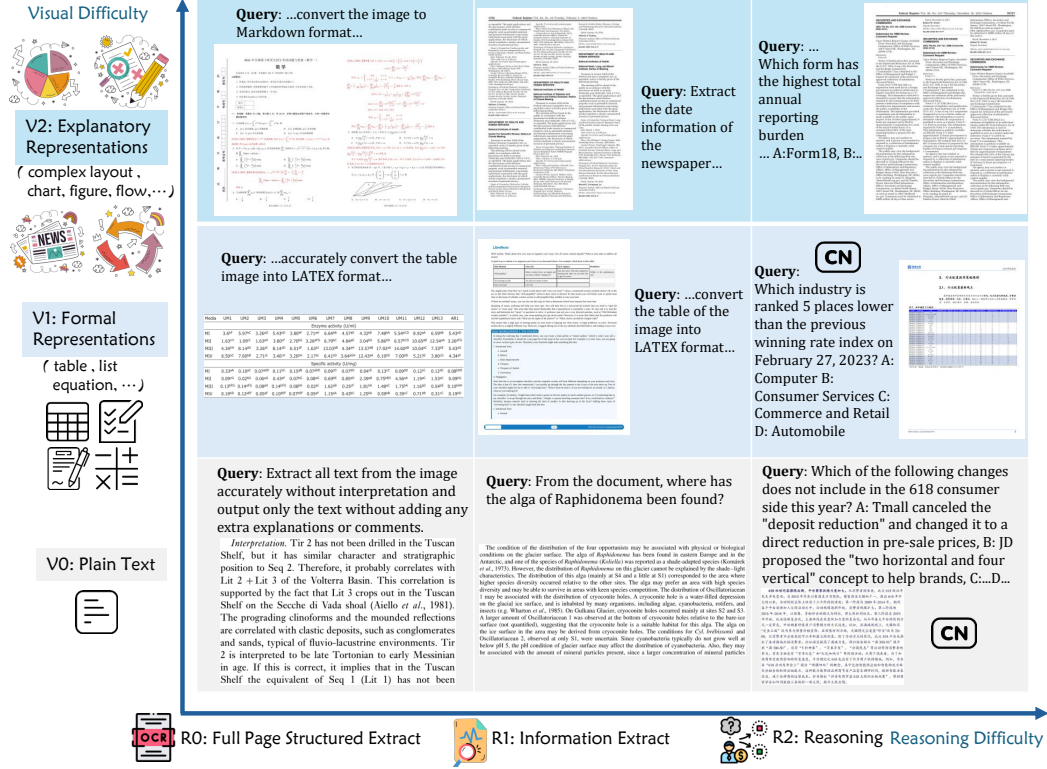


Figure 1: The overview of GDI-Bench.

2.3 Continual Learning of LLMs

Continual learning in large language models (LLMs) faces the core challenge of catastrophic forgetting. Existing research mainly adopts two strategies: data replay [39, 40] and parameter freezing [41, 42, 43]. Data replay strategies mitigate forgetting by revisiting samples from previous tasks during the training of new tasks. A typical approach is LAMOL[39], which generates pseudo-samples using the language models to avoid storage-based replay. Other approaches, such as experience replay and interleaved task training, have also been used to incorporate old data into the training loop. The parameter freezing method, on the other hand, protects existing knowledge by restricting parameter updates, such as through fine-tuning lightweight modules like LoRA[44], adapters [45], or prompt tokens [46]. Classical regularization-based methods like Elastic Weight Consolidation (EWC) [47] protect crucial weight parameters through importance constraints. More recent methods, such as Task Vector [48] and Gradient Projection [49], operate on the gradient or representation level to minimize interference between tasks and preserve generalization. In contrast to these methods, which generally rely on a limited number of pre-training datasets, leading to incomplete domain coverage and significant computational costs, our method presents a more comprehensive and efficient alternative.

3 Benchmark

3.1 Complexity Decoupling

As illustrated in Figure 1, a difficulty-decoupled evaluation protocol named GDI-Benchmark is introduced to comprehensively assess MLLMs' capabilities in visual information comprehension and reasoning. Distinct from VisualSimpleQA[50]'s structural decomposition of fact-seeking question answering tasks into visual recognition and knowledge dimensions through VLLM model architecture analysis, the proposed framework characterizes task complexity from fundamental difficulty perspectives by decoupling it into visual complexity and reasoning complexity components. Notably,

knowledge popularity is explicitly subsumed within the reasoning complexity dimension in this formulation.

3.1.1 Visual Complexity

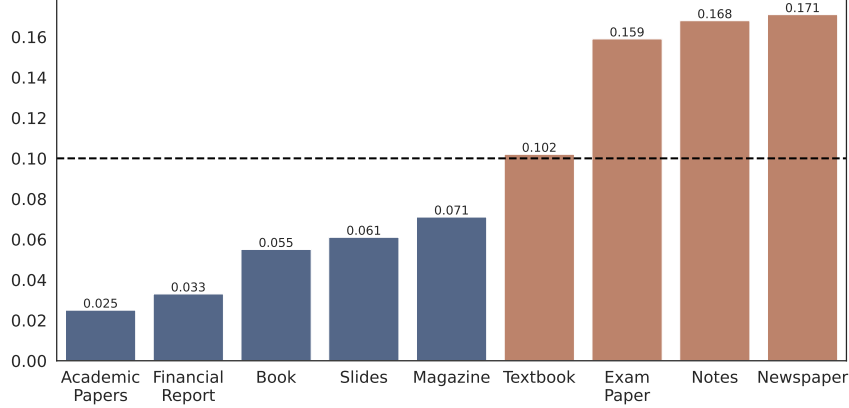


Figure 2: Distribution of visual complexity scores.

The visual complexity dimension is operationalized through a hierarchical categorization of document images into three levels: V0 (plain text), V1 (formal representations), and V2 (explanatory representations). V0 exclusively contains unstructured textual elements such as headings and paragraphs. Multimodal tasks on V0 documents typically achieve satisfactory performance via OCR-LLM pipeline architectures. As demonstrated in Figure 2, systematic analysis of the Omnidocbench benchmark reveals statistically significant performance gaps. This benchmark covers nine document categories. Current pipeline tools and VLLMs show notably worse performance on textbook, exam paper, notes, and newspaper types compared to other forms. These performance drops directly correlate with complex features like multi-column layouts and graphical elements. Based on these observations, we designate these four challenging categories as V2 and classify the remaining five as V1. This creates a data-driven taxonomy for visual complexity characterization.

3.1.2 Reasoning Complexity

The reasoning complexity characterization is formulated through a behavior-driven taxonomy. Three distinct levels are defined to progressively evaluate document understanding capabilities. It is specifically categorized into R0: Full Page Structured Extract, R1: Information Extract, and R2: Reasoning. Among these, the R0 task in V0 is more similar to OCR, and requires the ability to understand layout information such as tables, formulas, and complex page structures in V1 and V2 type images. For R1 tasks, the model needs not only to recognize visual content but also to understand the task itself in order to accurately extract relevant information, this often involves interpreting lines, tables, labels, and other visual elements present in V1 and V2 images. R2 tasks go a step further by requiring deeper reasoning capabilities, including the comprehension of logical inference and the ability to synthesize information across different modalities or layouts.

3.2 Annotation Process

3.2.1 Data Source

For the data acquisition phase of the General Document Intelligence (GDI) Benchmark, documents were primarily sourced from Omnidocbench[18], a comprehensive dataset containing document images from 9 domains. The dataset was further supplemented with an in-house collection of various document types, including exam papers, reports, newspapers and so on. This multi-source integration resulted in a well-rounded and representative dataset that captures the multifaceted nature of real-world document understanding tasks.

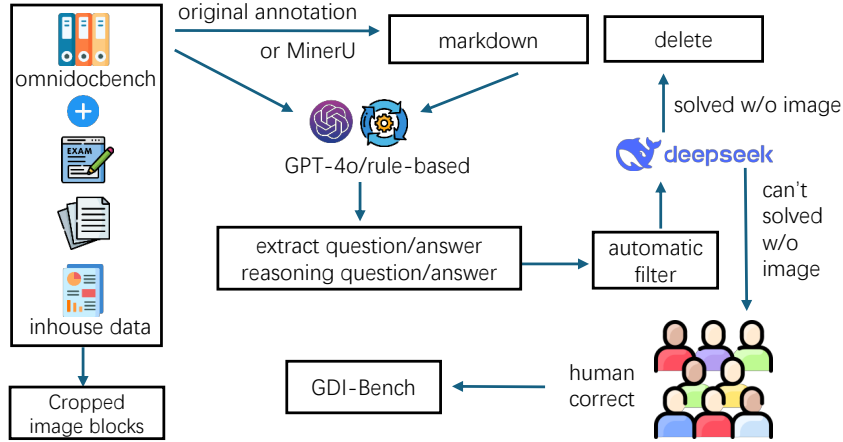


Figure 3: Annotation Process of GDI-Bench.

3.2.2 Data Construction

As shown in Figure 3, the data construction process begins with cropping single-layout sub-images from Omnidocbench[18] and inhouse documents to form the V0 raw image set. Based on end-to-end edit distance scores from state-of-the-art models and pipeline tools on Omnidocbench, domains with scores above 0.142 are categorized as V2, indicating high visual complexity. The remaining samples are assigned to V1.

For task construction, R0 tasks are generated using the original annotations or by synthesizing Markdown representations through MinerU[19]. To create R1 and R2 tasks, both the Markdown and corresponding images are input into GPT-4o to generate extractive and reasoning-based question-answer pairs. In addition, rule-based tasks are designed manually. These include extracting questions with fixed numbering, such as identifying problem numbers in exam papers.

To ensure high data quality, all synthetic cases are first evaluated by models. Cases that can be answered correctly by DeepSeek without requiring document input or that are flagged as low quality are filtered out. Finally, a team of PhD-level annotators reviews and verifies the remaining instances to ensure accuracy and correctness in the final benchmark dataset.

The final GDI-Benchmark contains a total of 2,989 test cases, distributed across different visual complexity levels (V0, V1, V2) and task types (R0, R1, R2), as summarized in Table 1.

Table 1: Complexity Distribution in the GDI Benchmark

| | v0 | v1 | v2 | Total |
|--------------|-----------|-----------|-----------|--------------|
| r0 | 379 | 528 | 453 | 1 360 |
| r1 | 223 | 407 | 276 | 906 |
| r2 | 133 | 311 | 279 | 723 |
| Total | 735 | 1 246 | 1 008 | 2 989 |

3.3 Evaluation Metrics

Metrics are defined per task type. R0 uses Average Edit Distance (AED) for character-level recognition accuracy. R1 adopts Average Normalized Levenshtein Similarity (ANLS) for entity and field extraction evaluation. For R2, Accuracy is the primary metric, as most tasks require single-choice answers. A small number of structured reasoning tasks in R2 require models to generate structured outputs, for which AED is used to assess similarity to reference answers.

Table 2: Comparison of GDI-Bench with Other Document Understanding Benchmarks

| Benchmark | Scenario | Task | Image | Difficulty Grading |
|-------------------|----------|------|-------|--------------------|
| Seed-bench-2-plus | 8 | 1 | 0.6k | ✗ |
| MMTab-eval | 1 | 9 | 23k | ✗ |
| MMC | 1 | 9 | 1.7k | ✗ |
| OCRBenchV2 | 31 | 23 | 9.5k | ✗ |
| Fox | 2 | 9 | 0.7k | ✗ |
| GDI-Bench | 9 | 19 | 1.9k | ✓ |

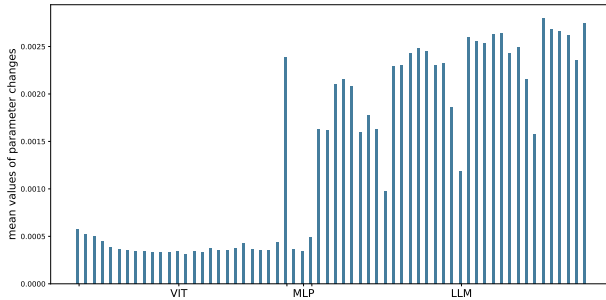
Table 3: Composition of the training set.

| Domain | Task | Training data |
|---------------------|---|---------------|
| Newspaper | Full text extraction of sliced newspaper pages. Extract header information of newspaper pages..... | 29k |
| Scientific Paper | Extract the author information of the paper. Extract full-text page content and organize it in markdown format. Answer reasoning questions for a paragraph on a page scientific paper page (/ multiple choice)..... | 108k |
| Infograph | Extract the maintitle. Answer reasoning questions for a infographic page (/ multiple choice)..... | 32k |
| Exam paper | Full text extraction of exampaper pages. Extract the content of the exam paper based on the type of question and the question number..... | 12k |
| Financial report | Extract the table on a page and convert to LaTeX format. Extract the content of the exam paper based on the type of question and the question number. | 3k |
| Handwritten Content | Recognize handwritten content in English and Chinese. | 12k |

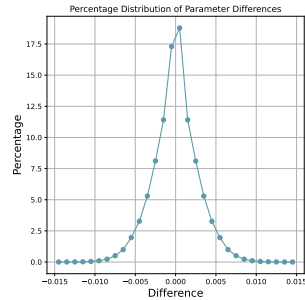
3.4 Comparison with Existing Document Intelligence Benchmarks

As shown in Table 2, most existing document intelligence benchmarks are limited in either scenario coverage, task diversity, or data scale. In contrast, the GDI-Bench offers a balanced combination of 9 diverse document scenarios and 19 representative tasks, based on a curated set of 1.9k images. Notably, it introduces difficulty grading , a feature absent in all other compared benchmarks.

4 Methodology



(a) The mean values of parameter changes per layer before and after full-parameter SFT.



(b) Distribution of parameter differences before and after full-parameter SFT.

GDI-Bench presents unique challenges for model designs, motivating the application of supervised fine-tuning to improve task-specific performance. To tackle this, we develop a multi-source training set that includes tasks structurally aligned with the benchmark, but drawn from diverse data domains, as shown in Table 3. We perform SFT on the InternVL3-8B model using this training set. However, subsequent evaluation of the model reveals significant catastrophic forgetting, wherein the model loses essential knowledge acquired during pre-training.

To investigate the phenomenon, we conduct a parameter update analysis of models undergoing full-parameter supervised fine-tuning. Our experimental results reveal significant sparsity in learned

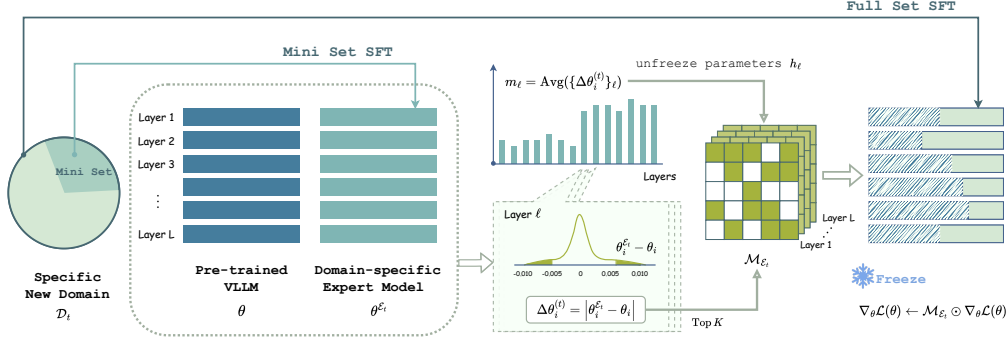


Figure 5: Overview of the Layer-wise Adaptive Freeze-Tuning method.

parameter adaptations: over 95% of parameters exhibit minimal changes, while a critical sparse subset (5%) undergoes substantial modification (>0.005), as depicted in Fig. 4b. This finding aligns with and extends the Lottery Ticket Hypothesis [51], leading to our formal hypothesis: For any pre-trained MLLM with parameters $\Theta = \{\theta_i\}_{i=1}^n$, there exists a sparse task-salient sub-network $\hat{\Theta} = \{\hat{\theta}_j\}_{j=1}^m$ ($m < n$) and a domain-specific transformation ψ , such that $\psi(f_{\hat{\Theta}}(x))$ achieves performance comparable to the full model $f_{\Theta}(x)$ on the target domain \mathcal{D} .

$$\forall x \in \mathcal{D}, \mathcal{P}(f_{\Theta}(x)) \approx \mathcal{P}(\psi(f_{\hat{\Theta}}(x))) \quad (1)$$

where $\mathcal{P}(\cdot)$ is the metric for evaluating the performance of the model, and the subnetwork $\hat{\Theta}$ consists of parameters exhibiting high update sensitivity during SFT. Since only a small subset of parameters is relevant to task-specific adaptation, we propose performing gradient updates solely on these parameters during SFT, while freezing the majority of the remaining parameters to preserve the model’s generalization and essential knowledge as much as possible.

We formalize this approach as Layer-wise Adaptive Freeze-Tuning (LW-AFT), which identifies and updates only critical parameters through layer-wise sensitivity analysis. By maintaining stable base network dynamics and avoiding large-scale parameter shifts, LW-AFT mitigates catastrophic forgetting while enabling efficient adaptation to downstream tasks. The overview of our method is shown in Fig. 5. For a specific new domain \mathcal{D}_t , the domain-specific expert model (denoted as \mathcal{E}_t) is fine-tuned based on the pre-trained parameters Θ , using only a fraction $\frac{1}{\alpha}$ of the full dataset. For each domain-specific expert model, the element-wise absolute change in parameters is computed as:

$$\Delta\theta_i^{(t)} = \left| \theta_i^{\mathcal{E}_t} - \theta_i \right|. \quad (2)$$

We perform a comprehensive layer-wise analysis of parameter updates $\Delta\theta_i^{(t)}$ before and after SFT. Fig. 4a illustrates the average magnitude of these updates per layer, revealing distinct trends between the vision and language components of the model. Specifically, vision layers exhibit more parameter modifications compared to their language layers. Intuitively, we assume that the layers undergoing substantial parameter updates are critical for domain adaptation, as their dynamics closely align with task-specific knowledge transfer. Conversely, layers with minimal updates appear less specialized to domain-specific features, and preserving their stability supports cross-domain generalization.

Therefore, we implement a layer-adaptive unfrozen parameter allocation strategy. Partitioning the model into L architectural layers with parameter counts $W = \{w_1, \dots, w_L\} \subseteq \mathbb{R}$, the distribution of trainable parameters across layers follows a proportional allocation mechanism: Given a global budget of H unfrozen parameters, each layer ℓ receives a number of unfrozen parameters h_ℓ determined by

$$h_\ell = \frac{m_\ell \cdot w_\ell}{\sum_{i=1}^L m_i \cdot w_i} \cdot H \quad (3)$$

where $m_\ell = \text{Avg}(\{\Delta\theta_j^{(t)}\}_\ell)$ is the average absolute change of the ℓ -th layer.

Table 4: The performance of different training methods on cross-domain and cross-task scenarios.

| Model | T1 | T2 | T3 | T4 | | | |
|----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | | | date | editor | email | phone |
| InternVL 3-8B | 0.216 | 0.383 | 0.840 | 0.101 | 0.884 | 0.398 | 0.398 |
| Full-Parameter Fine-Tuning | 0.040 | 0.913 | 0.600 | 0.864 | 0.992 | 0.783 | 0.989 |
| LoRA Fine-Tuning | 0.102 | 0.473 | 0.300 | 0.093 | 0.606 | 0.171 | 0.172 |
| LW-AFT (Ours) | 0.096 | 0.365 | 0.200 | 0.010 | 0.316 | 0.058 | 0.153 |

After determining the number of unfreezing parameters for each layer ℓ , we first sort the absolute changes $\{\Delta\theta_j^{(t)}\}_\ell$ in descending order, and then select only the top h_ℓ parameters for further updates. This selection is performed via the function:

$$\phi_{h_\ell} : \mathbb{R}^{w_\ell} \rightarrow \{0, 1\}^{w_\ell}, \quad \phi_{h_\ell}(\{\Delta\theta_j^{(t)}\}_\ell) = \begin{cases} 1, & \text{if } j \in \arg \text{TopK}_k(\{\Delta\theta_k^{(t)}\}_\ell, h_\ell) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The binary masks obtained for the expert \mathcal{E}_t is denoted as $\mathcal{M}_{\mathcal{E}_t} = \left\{ \phi_{h_\ell}(\{\Delta\theta_j^{(t)}\}_\ell) \right\}_{\ell=1}^L$. In the subsequent fine-tuning on a target domain, the gradient update is modified so that only the parameters with $\mathcal{M}_{\mathcal{E}_t}$ are updated:

$$\nabla_\theta \mathcal{L}(\theta) \leftarrow \mathcal{M}_{\mathcal{E}_t} \odot \nabla_\theta \mathcal{L}(\theta), \quad (5)$$

where \odot denotes the Hadamard product. This approach not only reduces the number of parameters to be trained but also preserves full-domain generalization without resorting to experience replay or other sophisticated fine-tuning techniques.

5 Experiments

The experiments in this section consist of two parts. In Section 5.1, we discuss the impact of our proposed Layer-wise Adaptive Freeze-Tuning (LW-AFT) method on cross-domain and cross-task generalization. In Section 5.2, we perform a comprehensive benchmark evaluation on the GDI-Bench.

5.1 Layer-wise Adaptive Freeze-Tuning.

5.1.1 Cross-domain and cross-task evaluation.

To validate the generalization capability of the LW-AFT method in cross-domain and cross-task scenarios, we designed the following experiments.

Cross-domain experiments with the same task.

- T1: Train the paragraph starting position OCR task in the academic paper domain and test the same task in the Infograph domain.
- T2: Train the information formatting and organization (JSON) task in the exam and academic paper domains, and test it in the Infograph domain.
- T3: Train the reasoning and question-answering task in the academic paper domains, and test it in the financial report domain.

Cross-task experiments with the same domain.

- T4: In the newspaper domain, only train the header information extraction and its formatting organization (JSON) task, and test the extraction task for header detail information (including extraction of newspaper date, editor, email, and phone number).

We compare the performance of full-parameter fine-tuning, LoRA fine-tuning, and our method under both settings, as shown in Table 4, which displays the edit distance for each task. As can be seen, our

method demonstrates strong cross-domain and cross-task capabilities, significantly outperforming the LoRA fine-tuning.

5.1.2 Performance on general datasets.

We evaluate the performance of full-parameter fine-tuning, LoRA fine-tuning, and our method on a range of general datasets, as shown in Table 5. It can be seen that our method preserved the model’s general capabilities and avoided catastrophic forgetting after SFT.

Table 5: The performance of different training methods on general datasets.

| Model | DocVQA↑ | ChartQA↑ | AI2D↑ | | TextVQA↑ | Ocrbench↑ | InfoVQA↑ | Seed2plus↑ | MMBench↑ | |
|----------------------------|-------------|-------------|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|
| | val | test | test | no mask | test | - | val | - | en | cn |
| InternVL 3 8B | 92.0 | 86.6 | 85.2 | 92.6 | 80.2 | 880 | 75.6 | 69.7 | 85.5 | 85.6 |
| Full-Parameter Fine-Tuning | 39.4 | 33.3 | 43.1 | 49.4 | 22.7 | 228 | 25.9 | 44.8 | 15.6 | 14.6 |
| LoRA Fine-Tuning | 91.3 | 85.8 | 85.4 | 93.1 | 81.6 | 862 | 74.4 | 69.7 | 85.2 | 84.8 |
| LW-AFT (Ours) | 91.6 | 85.8 | 85.6 | 93.2 | 82.1 | 871 | 75.8 | 70.6 | 85.2 | 84.8 |

Table 6: The performance of different training methods on general datasets.

| Model | R0V0 | R0V1 | R0V2 | R1V0 | R1V1 | R1V2 | R2V0 | R2V1 | R2V2 | Overall |
|-------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Claude3-5V-Sonnet | 0.10 | 0.74 | 0.78 | 0.60 | 0.52 | 0.91 | 0.18 | 0.52 | 0.62 | 0.58 |
| Gemini-2.0-Flash | 0.05 | 0.42 | 0.52 | 0.62 | 0.25 | 0.55 | 0.05 | 0.31 | 0.69 | 0.39 |
| GPT4o-2024-11-20 | 0.04 | 0.40 | 0.57 | 0.59 | 0.40 | 0.58 | 0.12 | 0.48 | 0.68 | 0.43 |
| Grok-2-Vision | 0.33 | 0.80 | 0.84 | 0.80 | 0.71 | 0.70 | 0.23 | 0.54 | 0.62 | 0.66 |
| Qwen2.5-VL-72B-Instruct | 0.03 | 0.35 | 0.44 | 0.41 | 0.24 | 0.33 | 0.11 | 0.36 | 0.50 | 0.31 |
| GDI-Model | 0.03 | 0.35 | 0.44 | 0.14 | 0.21 | 0.24 | 0.08 | 0.23 | 0.34 | 0.25 |

5.2 Benchmark Evaluation Results.

To assess the effectiveness of our GDI model on generalized OCR tasks, we compare it against several state-of-the-art vision-language models—namely Qwen2.5-VL-72B, Gemini-2.0-Flash, GPT-4o-2024-11-20, Claude-3-5-Sonnet, and Grok-2-Vision—on GDI-Bench’s multi-level challenge suite (v0-r0 through v2-r2). All evaluations were conducted with the same preprocessing pipeline.

On **v0-r0**, our 8B model matches or slightly exceeds performance of Qwen2.5-VL-72B. On the held-out generalization set, GDI-Model retains a slight advantage over the much larger models (70 B+), demonstrating superior resource-efficiency and robustness. These results confirm that, despite its smaller parameter count, our model achieves performance on par with or exceeding that of much larger open- and closed-source competitors.

6 Conclusion

We introduce the GDI-Bench, a document-domain benchmark with extensive domain coverage that pioneers a systematic difficulty grading system. Utilizing a complexity decoupling mechanism, the GDI-Bench breaks down multimodal document tasks into two orthogonal dimensions: visual complexity and reasoning complexity. It establishes a quantifiable, hierarchical standard that aids in identifying model issues and guiding optimization efforts. In addition, we propose the Layer-wise Adaptive Freeze-Tuning method and the corresponding GDI model. LW-AFT mitigates the problem of catastrophic forgetting during SFT, retains 99% of the model’s general capabilities during the SFT process, and enhances its performance on cross-domain and cross-task scenarios. The GDI model has demonstrated outstanding performance across multiple document-domain benchmarks, including the GDI-Bench, as well as general datasets. We hope that the GDI-Bench can help the advancement of future models and perhaps inspire new theoretical insights.

References

- [1] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- [4] OpenAI. Gpt-4 technical report, 2023.
- [5] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [6] 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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.
- [7] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models, 2024.
- [8] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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 (CVPR)*, pages 14375–14385, June 2024.
- [9] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.
- [10] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.
- [11] Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37:7185–7212, 2024.
- [12] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, pages 1–14, 2024.
- [13] Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. DuReader_{vis}: A Chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, 2022.
- [14] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021.
- [15] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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, 2022.
- [16] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024.
- [17] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.
- [18] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations, 2024.

- [19] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.
- [20] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.
- [21] Yuning Du, Chenxia Li, Ruoyu Guo, Cheng Cui, Weiwei Liu, Jun Zhou, Bin Lu, Yehua Yang, Qiwen Liu, Xiaoguang Hu, et al. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. *arXiv preprint arXiv:2109.03144*, 2021.
- [22] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.
- [23] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, pages 56–72. Springer, 2016.
- [24] Minghui Liao, Baoguang Shi, Xiang Bai, Cong Wang, Tong Lu, and Tao Mei. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [25] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [26] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019.
- [27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [28] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2006.
- [29] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023.
- [30] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model, 2024. URL <https://arxiv.org/abs/2409.1704>.
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [32] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.
- [33] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.
- [34] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [35] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
- [36] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [38] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.
- [39] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. Lamol: Language modeling for lifelong language learning. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2020.

- [40] Kasidis Kanwatchara, Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Boonserm Kijsirikul, and Peerapon Vateekul. Rational LAMOL: A rationale-based lifelong learning framework. 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 2942–2953, 2021.
- [41] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics ACL*, pages 1932–1945, 2024.
- [42] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*, 2023.
- [43] Minqian Liu and Lifu Huang. Teamwork is not always good: An empirical study of classifier drift in class-incremental information extraction. *arXiv preprint arXiv:2305.16559*, 2023.
- [44] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [45] Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. Continual sequence generation with adaptive compositional modules. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3653–3667, 2022.
- [46] Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Preserving in-context learning ability in large language model fine-tuning. 2022.
- [47] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [48] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proceedings of the Eleventh International Conference on Learning Representations, ICLR*, 2023.
- [49] Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2021.
- [50] Yanling Wang, Yihan Zhao, Xiaodong Chen, Shasha Guo, Lixin Liu, Haoyang Li, Yong Xiao, Jing Zhang, Qi Li, and Ke Xu. Visualsimpleqa: A benchmark for decoupled evaluation of large vision-language models in fact-seeking question answering, 2025.
- [51] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.