InstructDoc: 自然言語指示に基づく視覚的文書理解

田中涼太 [1,2]　壱岐太一 [1]　西田京介 [1]　齋藤邦子 [1]　鈴木潤 [2]
[1] 日本電信電話株式会社　NTT 人間情報研究所
[2] 東北大学
{ryota.tanaka, taichi.iki, kyosuke.nishida, kuniko.saito}@ntt.com,
jun.suzuki@tohoku.ac.jp

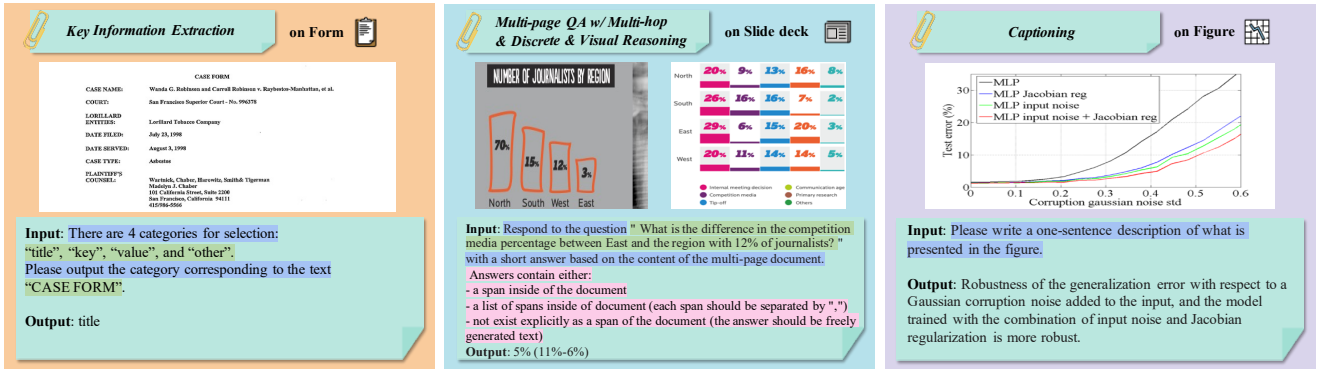


図 1　InstructDoc の例．文書画像を知識源とし，指示（*intent*，*query and options*，*answer style*）を基に回答を出力．

概要

自然言語指示に基づいて，文書を視覚的に理解するための基盤データセットである InstructDoc を提案する．InstructDoc は 12 種類の視覚的文書理解（VDU）タスクから構成されており，多様な自然言語指示を提供した最大規模のデータセットである．さらに，大規模言語モデル (LLM) の推論能力を活用し，視覚的文書理解を行う新たなモデルを提案する．実験により，我々のモデルは自然言語指示を基に未知の VDU タスクに適応できることを示し，従来のマルチモーダル LLM の性能を凌駕することを確認した．

1　はじめに

我々が扱う文書はテキストや視覚要素（図表など）を含み，多様な種類・形式が存在する．こうした実世界の文書を読解し理解する技術の実現は，AI 分野における重要課題の一つである．この実現に向けて，視覚的文書理解（**VDU**: Visual Document Understanding）に関する数多くの研究が，文書画像質問応答 [1, 2, 3] や情報抽出 [4, 5] など幅広いタスクに取り組んできた．また，モデルが画像を理解し指示に従うことで汎化能力を高める Visual Instruction Tuning [6, 7] が提案されているが，主にシーン画像内の視覚（非テキスト）オブジェクトの理解に焦点を当てており，任意の文書及び VDU タスクを統一的に理解・遂行可能なモデルは未だ実現していない．

本研究では，様々な VDU タスク・データセット（30 種類のデータセット・12 種類のタスク）をカバーする最大規模の指示チューニングデータセット **InstructDoc**[1] を提案する．図 1 に示す様に，InstructDoc を構成する各データセットは，人手でアノーションされたユーザの意図や回答のスタイルなどを含む多様な指示が提供されている．また，自然言語指示を基に文書レイアウトの理解，視覚要素理解，算術演算など，様々な推論能力を必要とする．

さらに，大規模言語モデル（LLM）の推論能力を活用し，テキスト・レイアウト・視覚要素を同時に理解可能な **Instruct**ion-based **D**ocument **r**eading and understanding model (**InstructDr**) を提案する．実験により，我々のモデルは自然言語指示を基に未知の VDU タスクに適応できることを示し，従来のマルチモーダル LLM（mLLM）の性能を凌駕することを確認した．また，指示チューニング済みの本モデルの重みを初期値として Fine-Tuning することで，複数の VDU タスクで世界最高性能を達成した．

1）我々のデータセットとコードは https://github.com/nttmdlab-nlp/InstructDoc で公開されている

2 InstructDoc

2.1 問題定義

InstrunctDoc を構成する全てのタスクは，「自然言語指示 T と文書画像 I が与えられ，回答 A を出力するタスク」と定義できる．各タスクは，1 件以上のデータセットから構成されている．各データセット $\mathcal{D}$ は K 件の指示 $\mathcal{T}^{\mathcal{D}} = \{T_1^{\mathcal{D}}, ..., T_K^{\mathcal{D}}\}$ が付与されており，N 件のインスタンス $\{(\mathcal{T}^{\mathcal{D}}, I_j, A_j)\}_{j=1}^{N}$ が含まれる．また，インスタンスごとに指示文は $\mathcal{T}^{\mathcal{D}}$ から無作為に選択される．

主に zero-shot でモデルを評価する．具体的には，特定のタスク集合で学習したモデルを，以下で定義する未知のデータセットで評価する:

- **Test**$_{\text{Cross-Dataset}}$: データセットは未知だが，タスクと文書の種類は学習データに含まれる
- **Test**$_{\text{Cross-Task}}$: データセットとタスクは未知だが，文書の種類は学習データに含まれる
- **Test**$_{\text{Cross-Domain}}$: データセット，タスク，文書の種類が未知

2.2 データ収集

ソースデータ収集 Web 上で利用可能な 30 件の VDU データセットを収集した．従来研究 [8, 9] で定義されるタスク分類方法に倣って，収集したデータセットを 12 件のタスクに分類した．詳細は，付録に掲載する．

Query 修正 情報抽出タスク (FUNSD [4], CORD [5]) において，曖昧性を含む Query に対して，Query の修正を行った（例: `menu.vatyn` → `menu_whether_price_tax_included`）．

指示アノテーション 各データセットに対して，人手で 5-10 件の異なる指示テンプレートを作成した．収集された QA タスクは，回答の形式が複数存在する．例えば，DocVQA [1] の回答は文書内のテキストから “スパン抽出”する形式である一方，VisualMRC [2] の回答は “要約”形式である．そのため，図 1 で示す様に，指示文は，*intent*（タスクの解き方）や *answer style*（モデルに要求する回答形式）が含まれる様にアノテーションを行なった．また，データセットが *query and options*（クエリと選択肢）を提供する場合，指示テンプレートに代入する．

データ分割 InstructDoc を 23 件/7 件の held-in/-out データセットに分割した．以下の held-out データセットで zero-shot 性能を測る．(i) **Test**$_{\text{Cross-Dataset}}$: FUNSD [4], CORD [5], (ii) **Test**$_{\text{Cross-Task}}$: ChartQA [10], InfoVQA [3], TabFact [11], (iii) **Test**$_{\text{Cross-Domain}}$: DUDE [12], SlideVQA [13]. 他データセットは held-in（学習用）として使用する．

表 1 データセットの比較. IT は指示テンプレート.

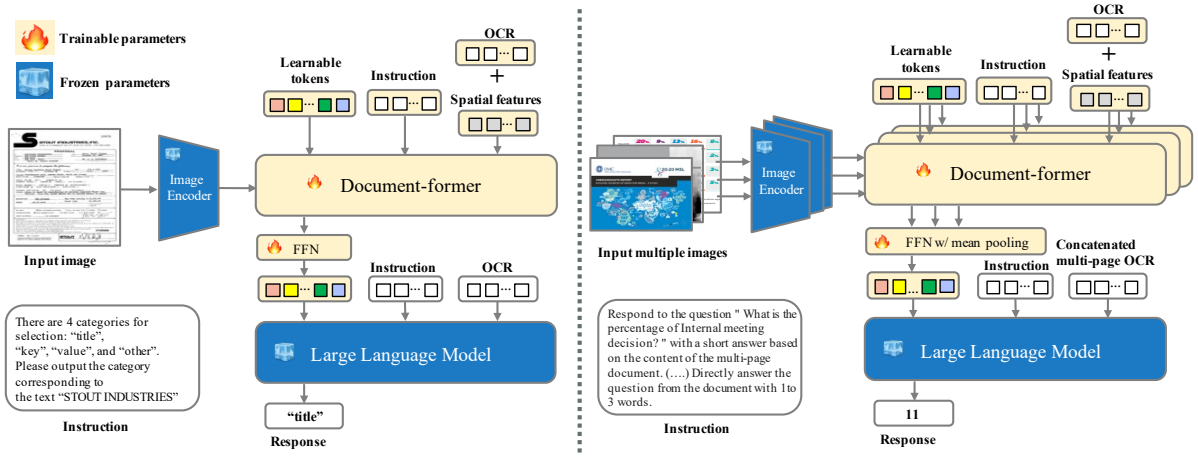
	LLaVAR	DocOwl	InstructDoc
Both Single/Multi-page			✓
Instruction annotation		✓	✓
Answer style annotation			✓
# Document types	8	7	Open
#Seed datasets	1	8	30
#Task clusters	1	3	12
#Avg.$_{\pm\text{Std.}}$ IT words	-	$5_{\pm 0}$	$20.3_{\pm 11.2}$
#Avg.$_{\pm\text{Std.}}$ IT	-	$1_{\pm 0}$	$7.4_{\pm 2.4}$
#Avg.$_{\pm\text{Std.}}$ OCR words	$52.5_{\pm 93.1}$	$270.1_{\pm 807.2}$	$443.2_{\pm 1442.8}$
#Avg.$_{\pm\text{Std.}}$ Answer words	$34.5_{\pm 27.5}$	$1.9_{\pm 2.7}$	$5.88_{\pm 17.7}$

2.3 統計情報および従来研究との比較

表 1 に統計情報と従来の文書画像を対象とした指示チューニングデータセット [14, 15] との比較を示す．InstructDoc は主に 3 つの特長を持つ．1) InstructDoc は複数ページの文書を含むオープンな文書種類・形式を扱った最初のデータセットである．また，OCR トークン数の標準偏差が LLaVAR（93.1 語）や DocOwl（807.2 語）と比べて大きい（1442.8 語）ことから，InstructDoc がより困難な設定であることを示している．2) InstructDoc は DocOwl と比較して 4 倍のタスク数を提供し，最も幅広いタスクをカバーしている．3) InstructDoc はより広範な指示（20.3 語，7.4 件のテンプレート）及び回答スタイルに関するアノテーションを提供する．一方，DocOwl の指示は限定的であり（5 語，1 件のテンプレート），LLaVAR は自動生成された指示のみのため，汎化性能には限界があることが報告されている．

3 提案モデル

提案モデル InstructDr を図 2 に示す．IstructDr は，指示チューニングされた FlanT5 [16] を LLM にもつ最先端の mLLM である BLIP-2 [17] をベースとする．InstructDr の貢献は，1) マルチモーダル情報を考慮して文書画像を LLM の表現に変換可能な Document-former を持つ点，2) 指示に基づく統一的なフォーマットによる同時学習を行う点，3) 複数ページで構成される文書を理解可能な点である．



(a) Our InstructDr for a single-page document　(b) Our InstructDr for a multi-page document

図 2　提案モデル．Document-former によって文書をエンコードし，その結果と指示文に基づいて LLM は回答を生成．

3.1　文書画像埋め込み

エンコーディング　CLIP 画像エンコーダ [18] を利用し，文書画像を視覚特徴 $\mathbf{z}^{\text{vis}}$ にエンコードする．更に，文書画像に対して OCR 及びトークナイズを適用することで，OCR 系列 $\{s_i\}_{i=1}^{M}$ と矩形領域 $\{(x_i^1, y_i^1, x_i^2, y_i^2)\}_{i=1}^{M}$ を獲得する．ここで，(x^1, y^1)，(x^2, y^2) は矩形領域の左上および右下の座標を表す．学習パラメータ $\mathbf{W}^{\{s,x,y,h,w\}}$ を用いて，OCR 埋め込み $\mathbf{z}_i^{\text{ocr}} = \mathbf{z}_i^{\text{word}} + \mathbf{z}_i^{\text{bbox}}$ を獲得する．ここで，$\mathbf{z}_i^{\text{word}} = \mathbf{W}^s(s_i)$，$\mathbf{z}_i^{\text{bbox}} = \mathbf{W}^x(x_i^1, x_i^2) + \mathbf{W}^y(y_i^1, y_i^2) + \mathbf{W}^h(y_i^2 - y_i^1) + \mathbf{W}^w(x_i^2 - x_i^1)$ である．同様に，$\mathbf{W}^s$ を用いて，指示文を $\mathbf{z}^{\text{ins}}$ に変換する．

Document-former　文書のマルチモーダル情報を考慮し，画像エンコーダと LLM を接続する学習可能なモジュールである Document-former を導入する．Document-former の構造は，複数層の cross-attention 層付き Transformer である．文書画像の表現を LLM の空間に写像するために，学習可能な m 個の d 次元トークン $\mathbf{z}^{\text{token}} \in \mathbb{R}^{m \times d}$ を用いる．これらのトークンは，cross-attention 層を通じて $\mathbf{z}^{\text{vis}}$ と相互作用し，self-attention 層を通じて $\mathbf{z}^{\text{ins}}$ と $\mathbf{z}^{\text{ocr}}$ と相互作用することで，$\mathbf{z}^{\text{doc}}$ を獲得する．そして，Feed-Forward Network（FFN）層を通じて，LLM の入力埋め込み次元 d^{LLM} を持つ $\mathbf{h}^{\text{doc}} \in \mathbb{R}^{m \times d^{\text{LLM}}}$ に変換される．

3.2　マルチモーダル大規模言語モデル

文書画像埋め込みと LLM の接続　LLM は文書画像埋め込み $\mathbf{h}^{\text{doc}}$，自然言語指示，OCR 系列を入力とし，回答 $\mathbf{A}$ を出力する．LLM のパラメータは，指示チューニング済み FlanT5 を初期値として使用する．

マルチタスク学習　全てのタスクを指示に基づく系列変換タスクとして解く．モデルを効率的に学習するために，Document-former（$\mathbf{W}^{\{s,x,y,h,w\}}$ を含む）とその後段の FFN 層のパラメータを更新し，他のパラメータは凍結する．学習は，予測系列の負の対数尤度を最小化することで，モデルを最適化する．

複数ページ文書理解　図 2b に示す様に，各ページは画像エンコーダと Document-former によって個別に処理され，LLM に入力される前に mean-pooling を実施する．LLM への OCR 入力は，各ページの OCR 系列をページ順に連結した．

4　評価実験

主に，$\text{Test}_{\text{Cross-Dataset}}$，$\text{Test}_{\text{Cross-Task}}$，$\text{Test}_{\text{Cross-Domain}}$ の 3 つの zero-shot 条件で評価する．更に，タスクに特化した Fine-Tuning の設定でも評価する．

ベースライン　Zero-shot 設定では，**LLaVA** [7] を含む 5 つの最先端の mLLM [14, 19, 17, 9] を比較対象とする．提案モデルのベースとなる BLIP-2 に対して，InstructDoc で学習したモデルとも比較する．回答長を制御するために，長さを制御するフレーズ（例: *use 1 to 3 words to answer*）を指示文に追加した．Fine-Tuning 設定では，LayoutT5 [2] を含む教師あり学習を行なった最先端の VDU モデル（**Supervised SOTA models**）[13, 12] を比較対象とした．

評価指標　各データセットの評価プロトコルに倣って，InfoVQA と DUDE では **ANLS** [20]，SlideVQA では **EM**，ChartQA では Relaxed Accuracy (**RAcc.**)，FUNSD と CORD では entity F1 (**eF1**)，TabFact では Accuracy (**Acc.**)，VisualMRC では **ROUGE-L** を用いた．更に，**F1** を zero-shot の設定で用いた．

表 2 Zero-shot 性能. T/L/V はテキスト/レイアウト/画像特徴を表す． #TuP は学習パラメータ数. I.Doc は InstructDoc.

Model	Modal	#TuP	$\text{Test}_{\text{Cross-Dataset}}$		$\text{Test}_{\text{Cross-Task}}$			$\text{Test}_{\text{Cross-Domain}}$		
			FUNSD eF1/F1	CORD eF1/F1	ChartQA RAcc./F1	InfoVQA ANLS/F1	TabFact Acc./F1	DUDE ANLS/F1	SlideVQA EM/F1	Held-out Avg.
LLaVA [7]	TV	13B	12.0/1.3	0.2/ 5.1	0.0/1.7	3.4/3.5	0.0/0.0	6.5/5.9	0.0/2.3	3.1/2.8
LLaVAR [14]	TV	13B	12.0/2.0	0.1/10.8	0.0/3.0	6.2/4.6	0.0/2.1	8.1/5.1	0.0/6.2	3.8/4.8
MiniGPT-4 [19]	TV	3.1M	12.0/2.2	0.2/ 2.1	0.0/0.4	4.3/0.5	0.3/0.2	5.9/1.1	0.0/0.4	3.2/1.0
InstructBLIP [9]	TV	103M	16.8/15.0	4.9/9.5	3.3/7.2	8.7/7.3	33.6/33.7	11.0/8.8	5.2/9.0	11.9/12.9
BLIP-2 [17]	TV	103M	19.6/19.6	32.0/51.9	23.6/21.5	48.2/36.7	58.6/58.6	39.8/35.4	28.3/38.8	35.7/37.5
BLIP-2 trained on I.Doc	TV	103M	26.0/26.1	33.8/54.7	24.7/21.2	47.8/35.4	58.8/58.8	43.9/40.4	30.1/38.8	37.9/39.3
InstructDr (Ours)	TLV	103.1M	**38.2/38.1**	**46.0/62.7**	**29.4/22.3**	**50.9/37.6**	**59.4/59.4**	**45.2/41.6**	**31.9/40.2**	**43.0/43.1**

表 3 モデル構造と指示文に関するアブレーション評価. Mean pooling をベクトル結合（concatenate）に変更した.

Model	CORD eF1	TabFact Acc.	DUDE ANLS	Held-out Avg.
InstructDr	**46.0**	**59.4**	**45.2**	**43.0**
w/o Document-former	38.5	58.8	44.6	40.2
w/o Spatially OCR features	33.8	58.8	43.9	37.9
w/o Mean pooling (concat.)	-	-	43.8	-
w/o Instructions	0.4	3.7	24.4	21.3
w/o Query rephrasing	30.9	-	-	-
w/o Answer style annotation	-	-	44.2	-

表 4 Held-in/out における Fine-tuning の性能.

Model	VisualMRC ROUGE-L	DUDE ANLS	SlideVQA EM	SlideVQA F1
Supervised SOTA models	52.2	46.1	33.5	41.7
BLIP-2	60.5	45.6	36.9	46.5
InstructDr	**61.1**	**46.8**	**37.7**	**47.3**

4.1 評価結果と分析

提案モデルは従来の mLLM の性能を上回るか？ 表 2 に示す様に，我々のモデルは全データセットで最も高い性能であった．これは，InstructDoc の指示チューニングが未知の VDU データセット・タスク・ドメインでの性能を向上させることを示している．一方，BLIP-2 に指示チューニングを行なった InstructBLIP は，BLIP-2 よりも性能が低かった．これは，学習時に InstructBLIP が画像内にテキストが含まれることを想定していないことが原因だと考えられる．InstructDoc で学習された BLIP-2 は，InstructDr と比較して性能が劣っており，InstructDr の方が多様な実世界文書の理解に適していることを示している．これは，表 3 に示した結果からも言える．Document-former やレイアウト情報，複数ページの情報集約方法が有効であることを確認した．

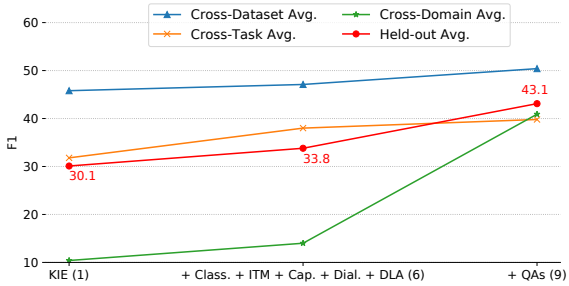


図 3 学習で用いたタスク数と性能．(·) はタスク数.

指示文は性能向上に寄与するか？ 表 3 に示す様に，指示文を除去した場合（入力は文書画像と *query and options* のみ），zero-shot 性能が低下した．さらに，我々がアノテーションを行なったクエリの修正や回答スタイルは，性能の向上に寄与した．

タスクの多様性は性能向上に寄与するか？ 図 3 に示す様に，タスク数の増加に伴い，zero-shot 性能が向上することが確認できる．

提案モデルの重みはタスク特化学習を行う初期値として有効か？ 表 4 に，held-in（VisualMRC）と held-out（DUDE，SlideVQA）タスクにおける指示チューニング済み InstructDr の Fine-Tuning 性能を示す．InstructDr は 3 つのデータセットにおいて世界最高性能を達成し，タスク特化学習のための初期値として優れていることが検証された．

5 おわりに

自然言語指示に基づく VDU を実現するための大規模指示チューニングデータセット InstructDoc を提案した．また，LLM の推論能力を活用し，視覚的文書理解を行う InstructDr を提案した．提案モデルは様々な VDU データセット・タスク・ドメインに対して，指示に基づいて汎化することを確認した．視覚表現された文書を基に QA を行う技術や Web 検索など産業上重要なサービスの発展に貢献できる．

参考文献

[1] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In **WACV**, pp. 2200–2209, 2021.

[2] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **AAAI**, pp. 13878–13888, 2021.

[3] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In **WACV**, pp. 1697–1706, 2022.

[4] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In **ICDARW**, 2019.

[5] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. In **Workshop on Document Intelligence at NeurIPS**, 2019.

[6] Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In **ACL**, pp. 11445–11465, 2023.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **arXiv:2304.08485**, 2023.

[8] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **ICLR**, 2021.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. **arXiv:2305.06500**, 2023.

[10] 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 **ACL Findings**, pp. 2263–2279, 2022.

[11] Lukasz Borchmann, Micha l Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Micha l Turski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In **NeurIPS**, 2021.

[12] Jordy Landeghem, Rubén Tito, Lukasz Borchmann, Micha l Pietruszka, Pawe l Józiak, Rafa l Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). **arXiv:2305.08455**, 2023.

[13] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In **AAAI**, pp. 13636–13645, 2023.

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

[15] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding. **arXiv:2307.02499**, 2023.

[16] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. **arXiv:2301.13688**, 2023.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In **ICML**, 2023.

[18] 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 **ICML**, pp. 8748–8763, 2021.

[19] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. **arXiv:2304.10592**, 2023.

[20] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In **ICCV**, pp. 4290–4300, 2019.

[21] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. DocILE benchmark for document information localization and extraction. **arXiv:2302.05658**, 2023.

[22] Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. Spatial dual-modality graph reasoning for key information extraction. **arXiv:2103.14470**, 2021.

[23] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In **ICDAR**, pp. 1516–1520, 2019.

[24] Xingyu Chen, Zihan Zhao, Lu Chen, Jiabao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. In **EMNLP**, pp. 4173–4185, 2021.

[25] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In **ICDAR**, pp. 947–952, 2019.

[26] Oliver Tüselmann, Friedrich Müller, Fabian Wolf, and Gernot A Fink. Recognition-free question answering on handwritten document collections. In **ICFHR**, pp. 259–273, 2022.

[27] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In **ACMM**, pp. 4857–4866, 2022.

[28] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In **NeurIPS**, 2021.

[29] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In **ECCV**, pp. 235–251, 2016.

[30] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In **NeurIPS**, 2022.

[31] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In **EMNLP Findings**, pp. 3258–3264, 2021.

[32] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In **UIST**, pp. 498–510, 2021.

[33] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In **ICDAR**, pp. 991–995, 2015.

[34] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In **COLING**, pp. 949–960, 2020.

[35] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In **KDD**, p. 3743–3751, 2022.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **arXiv:1711.05101**, 2017.

[37] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. **arXiv:2306.01733**, 2023.

[38] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. **arXiv:2305.18565**, 2023.

[39] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In **ACMM**, pp. 4083–4091, 2022.

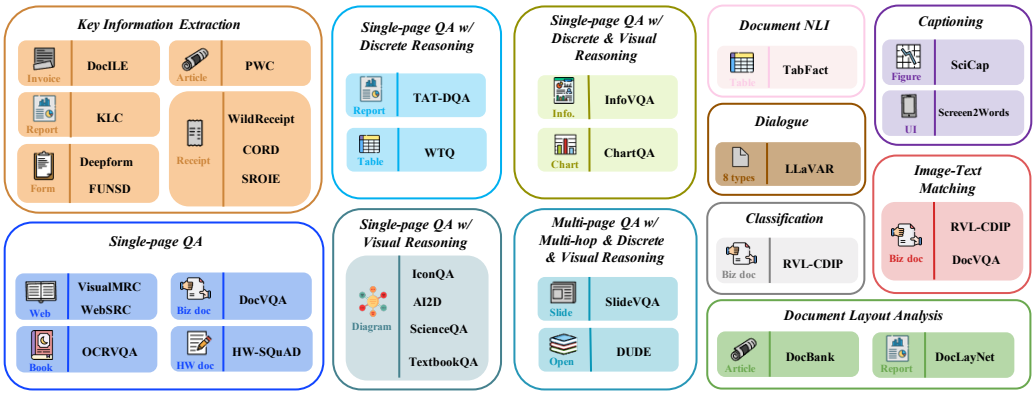
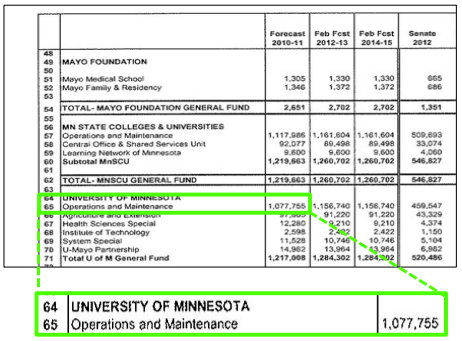


図 4　InstructDoc を構成するタスクとデータセット．広範な視覚的文書理解タスクや文書の種類・形式を扱う．

表 5　Zero-shot 性能．*は異なる split での評価結果を示す．

Model	Modal	$\text{Test}_{\text{Cross-Dataset}}$		$\text{Test}_{\text{Cross-Task}}$			$\text{Test}_{\text{Cross-Domain}}$		
		FUNSD eF1/F1	CORD eF1/F1	ChartQA RAcc./F1	InfoVQA ANLS/F1	TabFact Acc./F1	DUDE ANLS/F1	SlideVQA EM/F1	Held-out Avg.
Supervised SOTA models	TLV	92.1/-	97.7/-	72.3/-	54.8*/-	83.2*/-	46.1*/-	33.5/41.7	-/-
ChatGPT (`gpt-3.5-turbo-0613`)	T	21.8/21.2	30.4/49.3	16.0/16.8	37.8/29.5	52.5/52.4	34.5/32.3	11.7/23.8	29.2/32.2
GPT-4	T	47.5/47.5	69.4/81.7	20.9/27.6	49.9/46.5	68.8/68.8	46.3/45.1	21.0/36.4	46.3/50.5
InstructDr	TLV	38.2/38.1	46.0/62.7	29.4/22.3	50.9/37.6	59.4/59.4	45.2/ 41.6	31.9/40.2	43.0/43.1



You must extract the answer to the question "**What was the total forecast for University of Minnesota operations and maintenance for 2010-2011?**" after (…) or summing values. If you could not answer the question, the answer is 'none'.

ChatGPT: none
BLIP-2: -$220,590
InstructDr: 1,077,755

図 5　生成例．出力は correct/sufficient と incorrect/insufficient な回答に大別できる．(...) は省略．

A　付録

ソースデータ収集　図 4 で示す様に，VDU に関する 12 種類のタスク・30 件のデータセットを収集した．

- **Key Information Extraction (KIE)**: 文書内の各単語を意味ラベルにアサインするタスク [21, 4, 22, 5, 23].
- **Single-page QA**: 一ページ文書のテキスト/レイアウトに関する QA タスク [2, 24, 25, 1, 26].
- **Single-page QA w/ Discrete Reasoning**: 算術推論（四則演算，ソート，カウント）が必要な QA タスク [27].
- **Single-page QA w/ Visual Reasoning**: 物体検出や常識理解など視覚推論が必要な QA タスク [28, 29, 30, 29].
- **Single-page QA w/ Discrete & Visual Reasoning**: 算術推論と視覚推論が必要な QA タスク [3, 10]
- **Multi-page QA w/ Multi-hop & Discrete & Visual Reasoning**: 複数ページの文書に対して，マルチホップ推論，算術推論，視覚推論が必要な QA タスク [13, 12].
- **Document NLI**: 文書の含意関係認識タスク [11]
- **Dialogue**: 文書を基に対話を行うタスク [14].
- **Captioning**: 文書の内容を説明するタスク [31, 32].
- **Classification**: 文書分類タスク [33].
- **Document Layout Analysis (DLA)**: 文書の要素を特定するタスク [34, 35]
- **Image-Text Matching (ITM)**: OCR と画像が対応している否かを予測するタスク [33, 1].

実装　従来研究 [8] に従い，各データセットについて最大 5k インスタンスのサンプリングを行なった．AdamW [36] を重み減衰 0.05 で使用した．学習可能なトークンの数 m を 32 に設定した．モデル入力の画像は全て 224×224 にリサイズした．8 台の A100（40G）で 3 エポックの学習を行なった．各データセットが OCR を提供していない場合は，Google Vision API を用いて抽出した．

API ベースの LLM や教師あり学習モデルと比べて性能はどうか？　表 5 に示す様に，提案モデルは全データセットで ChatGPT の性能を上回った．さらに，InstructDr は，複数の推論能力（算術推論，視覚推論，マルチホップ推論など）を必要とする DUDE や SlideVQA データセットにおいて，教師あり SOTA モデル [37, 38, 39, 12] や GPT-4 と競合する性能を達成した．これは，我々のモデルが InstructDoc を用いた指示チューニングにより，多様な能力を効果的に学習できることを示している．

生成例　生成例を図 5 に示す．ChatGPT はテキスト情報しか考慮できないため不正解となった．BLIP-2 は構造化されたテキストを理解できないため項目の抽出に失敗しているのに対し，InstructDr は指示を理解し文書のマルチモーダル情報を考慮して項目の抽出に成功している．