動的解像度 ViT 付き VLM の Feature Alignment における画像構成要素の学習

壹岐太一　西田京介　齋藤邦子
NTT 株式会社 人間情報研究所
{taichi.iki,kyosuke.nishida}@ntt.com

概要

VLM 学習では文書画像や風景写真の全体を対象に学習させることが多い．一方，人には漢字の書き取りや図鑑の閲覧など視覚情報の構成要素を学習する機会がある．本研究では VLM 学習における Feature Alignment を画像の構成要素の学習と全体的な学習に分け，分離の効果を検証する．実例として日本語 OCR を含めた学習レシピで実験を行い，文書画像全体の OCR における学習停滞の解消や性能改善を示唆する結果を報告する．

1　はじめに

言語モデルと視覚エンコーダを接続し視覚言語モデル（VLM）を作る手法 [1, 2, 3] は，テキスト化されていない文書の理解 [4]，GUI の汎用的操作 [5]，ロボットの行動生成 [6] といった応用を可能にする基盤技術である．VLM の作成に利用可能な学習データセットの整備が進んでいる [1, 7, 8, 9, 10, 11, 12]．それらの既存研究は多様で高品質なデータを多量に収集することを主な焦点とする．データの種類や投入順序といった学習レシピの精緻化が，VLM 学習の効率化に向けた次の課題であると考える．

本研究は動的解像度 ViT 付き VLM の Feature Alignment の分解を提案する．画像の構成要素を学習する Atomistic Alignment (AA) と全体的な特徴を学習する Holistic Alignment (HA) に分け，順番に学習する（図 1）．要素を認識可能にしてから全体を学習することで学習の円滑化が期待される．また，合成データを用いる AA によって，単語など多様性のある画像要素を低コストで網羅的に学習できる．

応用の観点から日本語 OCR を含むデータセットを構築し検証した．画像から単語を読み取る単語 OCR と短文による物体説明で AA を学習後，文書全体の OCR と写真全体の詳細説明で HA を学習した結果，少ない計算コストで文書全体 OCR の学習停滞の抑制や性能改善に役立つことが示唆された．

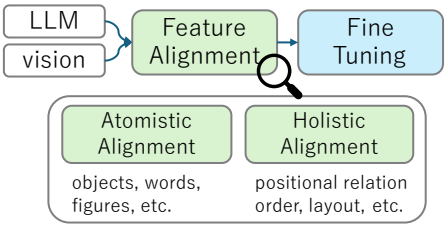

図 1　典型的な VLM 学習レシピの分解．本研究では VLM 学習における Feature Alignmnent を精緻化する．

2　関連研究

VLM 学習　プロジェクタ（画像特徴量の LLM 埋め込みへの変換）学習，プロジェクタと LLM の指示遂行への微調整を 2 段階で学習する LLaVA [1] や後者を画像ペア学習と指示遂行に分けて 3 段階で学習する VILA [3] が普及している．本研究では LLaVA における初期段階を Feature Alignment と定め，役割に応じた分解によって最適化を目指す．

局在パターンの積極的な学習　画像の構成要素など局在パターンの学習は重要である．[13] は VLM は小さな物体の認識に弱いことを実験的に示した．[14] は ViT を Local Alignment で訓練し，VLM の性能を改善した．本研究は局在パターンの学習を VLM 学習の中に陽に組み込む．

VLM の高解像度画像への対応方式　画像を視覚エンコーダの既定解像度 (300px 前後) で分解する画像タイル方式 [15, 16, 17]，画像タイルの特徴量を全体画像の特徴量に集約する画像タイル集約方式 [18, 19]，動的解像度に対応した ViT を用いる動的解像度 ViT 方式 [20, 21, 22] など多様な方式がある．本研究は動的解像度 ViT 方式の学習を対象にする．動的解像度 ViT 方式では既定解像度より小さな画像を使った入力長の削減や実際の画像内で出現する大きさに近い画像の学習が可能である．

表 1 データセットの概要．$\overline{\text{tok}}$ は学習設定の条件で前処理したときのトークン数の中央値を示す．

	データセット	タスク	事例数	$\overline{\text{tok}}$	評価指標	概要
AA	WS	単語 OCR	ja:652k	22	EM	NotoSansJP; 12, 15, 18em; 白地に黒; 余白最小
	WL	単語 OCR	ja:652k	25	EM	NotoSansJP; 21, 24, 27em; 白地に黒; 余白最小
	WR	単語 OCR	ja:652k	28	EM	ランダム（グリフ，サイズ，色，余白）
	WH	単語 OCR	ja:645k	29	EM	ランダム（手書き文字パターン，サイズ，色，余白）
	OBJ	物体説明	ja:30k, en:30k	93	Rouge1 F	CalTech256 [23] の画像，短文説明（合成）
HA	OCR-PDF	全体 OCR	ja: 50k	2210	ANLS	JDocQA [24] の PDF 画像 (ページ単位)
	OCR-SD	全体 OCR	ja:25k, en:25k	1479	ANLS	SynthDoG [25] のサブセット
	CAP-OI	写真説明	ja:50k, en:50k	1107	Rouge1 F	Open Images Datasets [26] の画像，長文説明（合成）

3 Atomistic Alignment

単語や物体など画像の構成要素を学習するステージである．動的解像度 ViT に対して小さな画像のまま構成要素を入力し，多量の事例を学習する．

3.1 学習データ

表 1 にデータセット概要，図 2 にデータ例を示す．AA では単語 OCR と物体説明を用いる．システムプロンプトは用いず，ユーザプロンプトには画像だけを含め，同一テキストの入力を最小限にとどめる．

3.1.1 単語 OCR

画像の単語を答えるタスクである．レンダリング条件によって 4 種類に分け，汎化性能を見る．使用する単語はデータセットで共通して MeCab IPA 辞書[1]の見出し語から重複を取り除いた 217,453 語である．各単語は各データセットで 3 回使用する．

WS (Word Small) 小さな文字サイズ（12，15，18em）と固定グリフ（NotoSansJP）を用いる．

WL (Word Large) 大きな文字サイズ（21，24，27em）と固定グリフ（NotoSansJP）を用いる．

WR (Word Random) Web で収集したグリフ（太さなどの詳細な違いを同一視すると 73 グループ），色，文字サイズ，余白をランダムにサンプルする．

WH (Word Hand Writing) 日本語の手書き文字パターンコレクション nakayoshi [27] から 100 人分をフォント化したグリフ，色，文字サイズ，余白をランダムにサンプルする．文字パターンに含まれない文字を含む単語は除外する．

3.1.2 物体説明

AA ステージで異なるタスクを同時に学習する影響を調べるために物体説明を学習する．

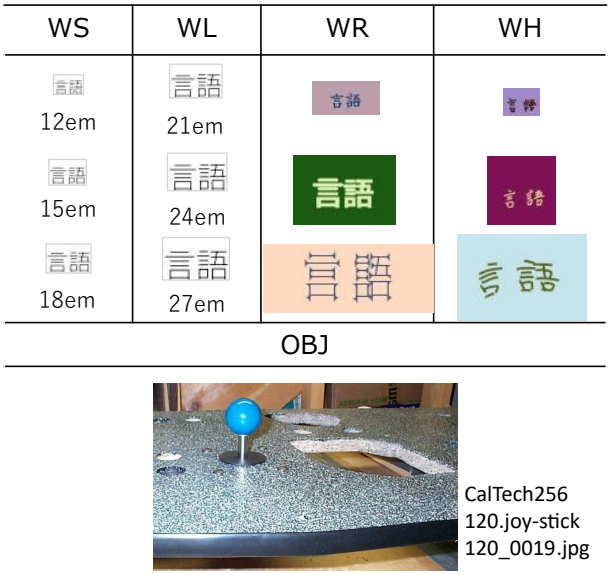

図 2 AA ステージにおける学習データ例．

OBJ (Object) 256 物体クラスからなる画像認識データセット CalTech256 [23] の画像を短文で説明する．教師データは Phi-3.5-vision[2]で生成した画像の英語による説明と，Mistral-Nemo-Japanese[3]で日本語訳したものを用いる．短文という条件のみ指示し，位置関係等の混入を厳格に制限してはいない．プロンプトは付録 A に示す．説明文が英語では 320 字，日本語では 160 字を超えた事例は棄却する．Phi-3.5-vision に与える画像の寸法は 320px 四方に収まるよう縮小し，実際のモデル入力に近付ける．

3.2 学習設定

モデル 動的解像度エンコーダ Pixtral-ViT，空間マージ付きプロジェクタ，LLM からなる

1) https://taku910.github.io/mecab/

2) https://huggingface.co/microsoft/Phi-3.5-vision-instruct

3) https://huggingface.co/cyberagent/Mistral-Nemo-Japanese-Instruct-2408

表 2　AA ステージの学習結果（テストデータ）

model	datasets	#training	FLOS	WS	WL	WR	WH	OBJ-ja	OBJ-en
AA-WS	WS	0.64M	0.81E	0.80	0.66	0.31	0.10	0.01	0.00
AA-..WL	WS, WL	1.28M	2.03E	0.91	0.97	0.61	0.25	**0.03**	0.00
AA-..WR	WS, WL, WR	1.92M	3.70E	0.92	**0.98**	0.83	0.59	0.01	0.00
AA-..WH	WS, WL, WR, WH	2.55M	5.12E	**0.93**	**0.98**	0.85	**0.76**	0.01	0.00
AA-..OBJ	WS, WL, WR, WH, OBJ	2.61M	9.99E	**0.93**	**0.98**	**0.86**	**0.76**	**0.03**	**0.54**

Mistral3 構造[4]を改変する．Pixtral-ViT の重み初期値は `ministral-3-8B-Instruct-2512`[5]の学習済みエンコーダから抽出する．プロジェクタの重みはランダムに初期化する．LLM の重み初期値は `llm-jp-3-7.2b-instruct3`[6]から抽出し，画像トークン，画像トークン列改行，画像トークン列末尾を示すトークンを追加する．ViT のパッチサイズは 14px，空間マージは 2 である．また，視覚トークンを扱うために chat template の拡張を行う（付録 B）．

主なハイパーパラメータ　画像入力は長辺が 224px 以下になるよう縮小する．224px 四方の画像は 64 トークンに相当する．入出力長を 128 トークンに制限する．学習対象は ViT（畳み込み層，正規化層を含む），プロジェクタ，新しく LLM に追加したトークン埋め込みとし，ViT の線形層には LoRA [28]（r=64）を適用する．目的関数はモデル出力部分に関する Causal Language Modeling（CLM）ロスを用いる．最適化器には重み減衰 0.1 の AdamW を用い，学習率は 3%の warm up 後，プロジェクタと LLM の埋め込みは 1e-3，ViT は 1e-5 で一定とする．グローバルバッチサイズを 2048 とする．データの 98%を学習，1%を検証，1%をテストに分け，1 エポック学習する．

評価指標　単語 OCR では事例ごとに出力と正解の完全一致を 1，不一致を 0 とし平均をとる（Exact Match; EM）．物体説明では事例ごとに出力と参照文の間で `rouge_score` ライブラリ[7]の rouge1 の F 値を算出し平均をとる（単語分割は付録 C 参照）．学習に要した計算量の見積りとして transformers Trainer の total_flos（Floating Operations）を報告する．

3.3　学習結果

表 2 に AA ステージの学習結果を示す．各モデルは学習データセットにおいて高いスコアを達成した．AA-..OBJ モデルの OBJ-ja スコアがほぼ 0 なのは言語指示を省いたことで英語で説明するモデルになったからである．OBJ-en のスコアは 0.54 であり，説明文の生成自体は学習できていると分かる．

要素の見た目に関して汎化が小さい　未学習のデータセットではスコアが低かった．例えば，WS（小さな文字）で学習した AA-WS は WS で 0.80 だが，WL（大きな文字），WR（ランダム），WH（手書きパターン）と見た目が離れるほど 0.66，0.31，0.10 と低下した．これは VLM の要素の見た目に汎化しにくい傾向を示唆する．多様な要素を低コストで学習する AA はこの傾向に対して効果的である．

データセット間の干渉は見られない　OBJ-ja を除いて，データセットを増やしても既存スコアは低下しなかった．物体説明と単語 OCR 間（AA-..OBJ とその他モデル）でもこの傾向は成り立っていた．少なくとも調査範囲でデータセット間の干渉は見られず，AA ステージでは可能な限り多様な画像要素を多量に学習する方針が望ましいと示唆している．

4　Holistic Alignment

文書のレイアウトや物体の位置関係など全体的な特徴を学習するステージである．本セクションでは AA の事前学習が HA に及ぼす効果を検証する．

4.1　学習データ

表 1 の HA に示した全体 OCR と写真説明を学習する．システムプロンプトで出力言語を指定し，ユーザプロンプトには画像のみを含める（付録 D）．

OCR-PDF　JDocQA [24] で収集された PDF 文書のページを OCR する．画像は pdfium2[8]でレンダリングする．文字が読めるように長辺を 2048px にスケールする．正解テキストは pdfium2 で抽出したテキストから作成する．文字に付与された座標に基づいて，ルールベースで行，行の空間的なまとまり，読み順を再構成して単線テキスト化する．

OCR-SD　風景などの写真に重畳したテキストを OCR する．合成 OCR データセット SynthDoG [25]

4）https://huggingface.co/docs/transformers/en/model_doc/mistral3
5）https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512
6）https://huggingface.co/llm-jp/llm-jp-3-7.2b-instruct3
7）https://pypi.org/project/rouge-score/
8）https://github.com/pypdfium2-team/pypdfium2

表 3　HA ステージの学習結果（テストデータ）．3 つのデータセットを同時に学習．

model	total FLOS	OCR-PDF	OCR-SD-ja	OCR-SD-en	CAP-OI-ja	CAP-OI-en
HA-direct	**13.7E**	0.002	0	0	0.501	0.476
HA-from-HA-direct	27.4E	0.006	0	0.002	0.551	0.528
HA-from-AA-WS	14.5E	0.589	0.937	0.970	0.596	0.590
HA-from-AA-..WL	15.7E	0.603	0.933	0.981	0.600	0.588
HA-from-AA-..WR	17.4E	0.618	0.949	**0.985**	0.590	0.589
HA-from-AA-..WH	18.8E	0.620	0.952	0.980	0.594	0.588
HA-from-AA-..OBJ	23.7E	**0.623**	**0.955**	0.976	**0.605**	**0.594**

から日本語と英語のデータをサンプリングする．

CAP-OI　Open Images Datasets [26] の画像を詳細説明する．教師データは OBJ と同様，Phi-3.5-vision で生成した英語説明文と Mistral-Nemo-Japanese で日本語訳した文を使う．プロンプトは付録 E に示す．

4.2　学習設定

モデル　AA を学習せずに HA を学習する HA-direct をベースラインとする．AA 学習後のモデルに HA を追加学習する HA-from-<AA ステージモデル>，HA 学習後に HA を追加で 1 エポック学習する HA-from-HA-direct を比較する．AA ステージモデル，HA-direct は LoRA をマージしてから学習する．

主なハイパーパラメータ　入力画像長辺を 1344px に，入出力長を 4096 トークンに制限する．学習対象は ViT，プロジェクタ，LLM の線形層とする．ViT と LLM の線形層には LoRA （r=192）を適用する．Warm up 後の学習率はプロジェクタと LLM は 1e-4，ViT は 1e-5 とする．グローバルバッチサイズを 512 とする．その他は AA ステージと同様とする．

評価指標　全体 OCR では正規化 Levenshtein 類似度の事例平均（ANLS）を使う．[29] に従って切り捨て閾値 τ を 0.5 とする．写真説明では物体説明と同様に rouge1 F 値の事例平均を使う．AA ステージと合算した計算量見積もりを合計 FLOS に示す．

4.3　学習結果

AA は効率的に HA の性能を改善する　表 3 は HA ステージの学習結果を示す．最も軽量な AA を実施した HA-from-AA-WS（HA-direct の合計 FLOS に対して約 6%増）を含め，AA を学習したモデルは一貫して OCR，説明文生成で HA-direct を上回った．いずれのモデルも 2 エポック分学習した HA-from-HA-direct より小さなコストで性能向上を実現しており，AA は効率的である．また，物体説明も学習した HA-from-AA-..OBJ の方が単語 OCR だけの HA-from-AA-..WH に比べて，僅かだが全体的にスコアが改善した．AA ステージは OCR 以外のタスクにも役立つ可能性が高い．

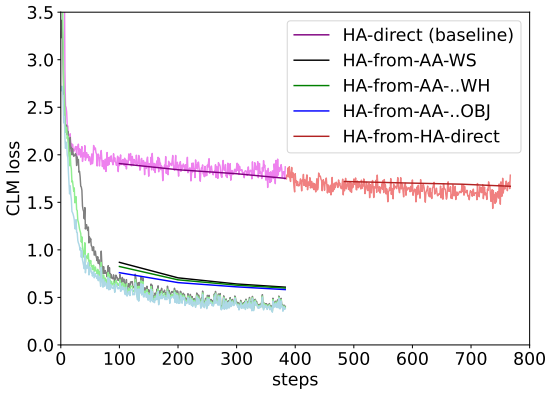

図 3　学習曲線の比較．各モデルにおいて色が薄い線は学習ロスを，色が濃い線は検証ロスを示す．

AA は HA の学習停滞を解消する　図 3 に学習曲線を示す．HA-direct は約 20 ステップから学習が停滞し，HA-from-HA-direct でも解消しなかった．AA を学習したモデルはその後もロスが低下し続けた．OCR のスコア差と合わせると，AA の単語 OCR 学習が全体 OCR の学習停滞を抑制したと推測される．

5　おわりに

動的解像度 ViT 付き VLM の Feature alignment を画像の構成要素を学習する Atomistic Alignment (AA) と全体的な特徴を学習する Holistic Alignment (HA) に分け，順番に学習することを提案した．単語 OCR を中心とする AA が HA における全体 OCR の学習停滞の解消と性能改善に効果的であると実証した．Feature alignment に焦点を当てた本研究の知見は効率的な VLM 学習の可能性を拓くものである．

本研究の限界　検証したモデル数が限られ，Feature Aligment 分解の効果がどの程度一般性を持つか明らかではない．特に，本研究の実験は VLM の一部として一度学習された ViT を用いた点に注意を要する．また，指示遂行の学習までを含めた学習レシピにおける効果の検証は残された課題である．

参考文献

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **NeurIPS**, Vol. 36, pp. 34892–34916, 2023. https://arxiv.org/abs/2304.08485.

[2] 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**, pp. 19730–19742, 2023. https://proceedings.mlr.press/v202/li23q.

[3] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In **CVPR**, pp. 26689–26699, 2024. https://openaccess.thecvf.com/content/CVPR2024/html/Lin_VILA_On_Pre-training_for_Visual_Language_Models_CVPR_2024_paper.html.

[4] Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In **AAAI**, Vol. 38, pp. 19071–19079, 2024. https://ojs.aaai.org/index.php/AAAI/article/view/29874.

[5] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In **ACL**, pp. 9313–9332, 2024. https://aclanthology.org/2024.acl-long.505/.

[6] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. **arXiv:2506.01844**, 2025. https://arxiv.org/abs/2506.01844.

[7] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. **TMLR**, 2025. https://openreview.net/forum?id=zKv8qULV6n.

[8] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. **NeurIPS**, Vol. 37, pp. 87310–87356, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

[9] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. **arXiv:2501.14818**, 2025. https://arxiv.org/abs/2501.14818.

[10] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. **arXiv:2510.17269**, 2025. https://arxiv.org/abs/2510.17269.

[11] Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. Constructing multimodal datasets from scratch for rapid development of a japanese visual language model. In **NAACL**, pp. 470–484, 2025. https://aclanthology.org/2025.naacl-demo.38/.

[12] Issa Sugiura, Shuhei Kurita, Yusuke Oda, Daisuke Kawahara, Yasuo Okabe, and Naoaki Okazaki. Waon: Large-scale and high-quality japanese image-text pair dataset for vision-language models. **arXiv:2510.22276**, 2025. https://arxiv.org/abs/2510.22276.

[13] Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. Exploring perceptual limitation of multimodal large language models. **arXiv:2402.07384**, 2024. https://arxiv.org/abs/2402.07384.

[14] Ian Connick Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In **ICLR**, 2025. https://openreview.net/forum?id=qssVptHTPN.

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **CVPR**, pp. 26296–26306, 2024. https://arxiv.org/abs/2310.03744.

[16] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. **arXiv:2503.19786**, 2025. https://arxiv.org/abs/2503.19786.

[17] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. **arXiv:2508.18265**, 2025. https://arxiv.org/abs/2508.18265.

[18] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In **ACL**, pp. 5817–5834, 2025. https://aclanthology.org/2025.acl-long.291/.

[19] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In **CVPR**, pp. 4122–4134, 2025. https://arxiv.org/abs/2412.04468.

[20] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. **arXiv:2410.07073**, 2024. https://arxiv.org/abs/2410.07073.

[21] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. **arXiv:2511.21631**, 2025. https://arxiv.org/abs/2511.21631.

[22] Junbo Niu, Yuanhong Zheng, Ziyang Miao, Hejun Dong, Chunjiang Ge, Hao Liang, Ma Lu, Bohan Zeng, Qiahao Zheng, Conghui He, et al. Native visual understanding: Resolving resolution dilemmas in vision-language models. **arXiv:2506.12776**, 2025. https://arxiv.org/abs/2506.12776.

[23] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007. https://authors.library.caltech.edu/records/5sv1j-ytw97.

[24] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. Jdocqa: Japanese document question answering dataset for generative language models. In **LREC-COLING**, pp. 9503–9514, 2024. https://aclanthology.org/2024.lrec-main.830/.

[25] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In **ECCV**, pp. 498–517. Springer, 2022. https://link.springer.com/chapter/10.1007/978-3-031-19815-1_29.

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. **IJCV**, Vol. 128, No. 7, pp. 1956–1981, 2020. https://link.springer.com/article/10.1007/s11263-020-01316-z.

[27] Kaoru Matsumoto, Takahiro Fukushima, and Masaki Nakagawa. Collection and analysis of on-line handwritten japanese character patterns. In **ICDAR**, pp. 496–500. IEEE, 2001. https://web.tuat.ac.jp/~nakagawa/pub/2001/pdf/matsu0109a-e.pdf.

[28] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **ICLR**, 2022. https://openreview.net/forum?id=nZeVKeeFYf9.

[29] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In **ICCV**, pp. 4291–4301, 2019. https://openaccess.thecvf.com/content_ICCV_2019/html/Biten_Scene_Text_Visual_Question_Answering_ICCV_2019_paper.html.

A　OBJ データセットの文生成

英語説明文　Phi-3.5-vision に temperature=0 の下，以下のプロンプトを与えた．ただし、<|image_1|>は画像トークンを表す．

```
<|user|>\n<|image_1|>\nPlease describe the image in a short sentence.\n<|end|>\n<|assistant|>\n
```

日本語訳　Mistral-Nemo-Japanese に temperature=0.3 の下，以下のプロンプトを与えた．ただし、{caption}は英語の説明文を表す．また，システムプロンプトは「あなたは親切な AI アシスタントです。」とした．

```
次の英文を日本語に訳してください。訳したテキストだけ答えて下さい。
\n===== 英文=====\n{caption}
```

B　作成モデルの Chat Template

llm-jp-3-7.2b-instruct3 の chat template を mistral3 の画像プロセッサが視覚トークンを扱う形式に拡張した．具体的にはメッセージの content を出力する際，その種類で分岐し，文字列の場合はそのまま出力し，リストの場合にはそのリストの要素でループするように変更した．ループ内では要素の type が text の場合はそのまま出力し，image の場合は画像トークンをひとつ出力する．また，system プロンプトを使用するために，メッセージの role が system のとき固定文字列ではなく content の内容を出力するように変更した．

C　Rouge1 の単語分割

参照文が英語の場合は語幹変形有りのデフォルトトークナイザを，日本語の場合は文字単位分割を用いた．

D　HA ステージのプロンプト

HA ステージのシステムプロンプトには「Please make a detailed {lang} caption for the images provided by users.」，「ユーザーから提供された画像に対して、{lang}で詳細なキャプションを作成してください。」をランダムに用いた．

{lang}に入れる言語はタスクによって異なる．CAP-OI では正解データの言語を用いた．OCR-SD では，正解データを「Text is placed on the front of the photograph. The content is as follows:\n"""{ocr_text}"""\n」または「写真の前面に文字が配置されています。内容は以下の通りです:\n"""{ocr_text}"""\n」のテンプレートで作成し，このテンプレートの言語を用いた．OCR される言語とテンプレートの言語は一致するとは限らない．OCR-PDF でも同様に，正解データを「The image shows a Japanese document. It reads as follow:\n"""{ocr_text}"""\n」または「画像は日本語の文書です。次のように書かれています:\n"""{ocr_text}"""\n」のテンプレートで作成し，このテンプレートの言語を用いた．テンプレート処理は OCR タスクをキャプション生成の一種として扱うために適用した．なお，OCR タスクの評価時には"""で囲われた部分を正規表現で抽出し，その部分に対して ANLS を算出した（抽出できなかった場合のスコアは 0 とした）．

E　CAP-OI データセットの文生成

英語説明文　Phi-3.5-vision に temperature=0 の下，以下のプロンプトを与えた．ただし、<|image_1|>は画像トークンを表す．

```
<|user|>\n<|image_1|>\nDescribe the image in detail.<|end|>\n<|assistant|>\n
```

日本語訳　Mistral-Nemo-Japanese に temperature=0.3 の下，以下のプロンプトを与えた．ただし、{caption}は英語の説明文を表す．また，システムプロンプトは「あなたは親切な AI アシスタントです。」とした．

```
次の英文を日本語に訳してください。カタカナ語の使用は避けてください。訳したテキストだけ答えて下さい。
\n===== 英文=====\n{caption}
```