大規模視覚言語モデルにおける視覚情報の伝播経路と Registering の考察

坂上 温紀 [1,2] Zhi Qu[1] 上垣外 英剛 [1] 高村 大也 [2] 谷中 瞳 [3] 渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 産業技術総合研究所 [3] 東京大学
sakajo.haruki.sd9@naist.ac.jp taro@is.naist.jp

概要

大規模視覚言語モデルは大規模言語モデルに視覚エンコーダを統合することで，視覚情報とテキスト情報を要する様々なタスクにおいて高性能を達成してきた．視覚情報とテキスト情報の統合方法はモデルによって様々であるが，どの手法を採用する場合も，視覚情報は数百の画像トークンとしてモデルに入力されており，推論効率の低下の一因となっている．一方で，大規模視覚言語モデルは入力画像を無視してテキストを生成することもあり，視覚情報をモデルがどのように扱っているのかは自明でない．そこで本研究では，モデルがどのように視覚情報を扱っているのかを分析した上で，大規模言語モデル側での画像トークンの削減と視覚情報の伝播経路の明確化を実現する手法を提案する．

1 序論

大規模視覚言語モデル（Large-scale Vision Language Models; LVLMs）[1, 2, 3] は大規模言語モデル（Large Language Models; LLMs）に視覚エンコーダーを付加することにより，LLM の知識や推論能力を活用して視覚情報を要するタスクにおいて高い性能を示してきており [3, 4]，その能力は自動運転 [5] や画像診断 [6] といった領域での活用も模索されている．LVLM では，通常，画像や動画は数百のトークン列として LLM に入力されることから，視覚情報を扱う際の推論効率はテキストのみを扱う場合と比べると低くなる．

これまでの研究で，LVLM を含むマルチモーダル言語モデルにおいて，視覚情報がどのようにテキスト側に伝播するのかが明らかにされてきた [7, 8, 9]．例えば画像とマルチモーダルモデルには視覚情報を単一トークンを介してテキスト側に伝播させるものと，個々の画像トークンから直接テキスト側にその情報を伝えるものがあることが明らかになっている．このような分析はモデルの解釈性の観点でも意義を持つものであるが，その分析結果がモデルの性能向上や推論効率の向上に活用されることは多くはない．そこで，本研究では，モデルの画像情報の伝播経路を分析した上で，個々のモデルの振る舞いから LVLM の性能改善・推論効率の向上を図る．

実験の結果，LVLM は画像トークン・画像終端の特殊トークン両方を通じてテキスト側に視覚情報を渡している可能性が高いことが示された．そこで，LVLM の LLM 部分での視覚情報の扱いをより効率良くするために，特殊トークンに視覚情報を書き込むことができるような構成を提案する．提案手法では，テキストトークンが個々の画像トークンの代わりに，この特殊トークンを参照するような Attention 構造を持つ．これによって，この特殊トークンの数を減らすことを目標とすることで，推論効率の向上が期待される．

2 背景

大規模視覚言語モデル 大規模視覚言語モデル（LVLM）は通常，大規模言語モデル（LLM）に視覚エンコーダーを統合することによって実現される．視覚エンコーダーによって画像や動画といった視覚情報をベクトル化し，視覚エンコーダと LLM を繋ぐプロジェクターを介して，入力テキストとともに LLM に入力される．このとき，多くのモデルでは <画像トークン><画像トークン><画像終端の特殊トークン><テキストトークン><テキストトークン> というように，数百からなる画像トークンに続けて画像の終端を表す特殊トークン，さらに入力テキストという順の入力列を受け取る．自己回帰型 LLM を基にして構築される LVLM の Attention mask に関しては，画像トークンが入力に含まれる場合でも基本的には因果マスクが用いられるが，Gemma3 [10]

のように，画像トークン同士では双方向 Attention を用いるものもある．

LVLM の分析　LVLM は視覚情報とテキスト情報を組み合わせて推論しており，モデルの解釈性の観点では，視覚情報がどのようにテキスト側に伝わり，利用されるのかが重要となる．このような観点での研究は主にモデルのアーキテクチャごとに行われてきた．例えば，画像・テキスト両方を生成可能なモデルでは，視覚情報は画像終端の特殊トークン（End of Image; EOI トークン）を介してテキスト側に伝播し，画像とテキストを受け取りテキストのみを生成するようなモデルでは，EOI トークンではなく，個々の画像トークンから直接テキスト側に視覚情報が伝わることが明らかになっている [9]．また，視覚情報とテキスト情報の統合は中盤から後半の層にかけて行われ [8]，視覚情報の要約段階の有無もアーキテクチャによって異なる [7]．このように，LVLM はアーキテクチャ，特に視覚情報と画像情報の統合方法の違いにより，振る舞いが異なる．本研究では，自己回帰型 LLM を基とし，かつ EOI トークンが存在するモデルを分析の対象とする．

3　LVLM における画像情報の伝播

LVLM において，視覚情報がテキスト側に伝播し得るのは各 Transformer ブロックにおける Attention モジュールである．つまり，視覚情報がどのようにテキスト側に伝わるかは，Attention モジュール内でテキストトークンがどれほど画像トークンに注意を向けているのかによって評価される [9]．さらに，テキストトークンから画像トークンへの注意を無効化（Attention Knockout）することによる推論結果の変化を観察することによって，テキストから画像への注意がどれほど重要であるのかを確認する．

3.1　実験設定

長さ n の入力トークン系列において，画像トークンのインデックス列を $\mathcal{I}$，画像トークン直後の特殊トークン（EOI トークン）のインデックス（列）を $\mathcal{S}$，画像のあとに現れるテキストトークンのインデックス列を $\mathcal{T}$ とする．また，層 ℓ の Attention Head h における Attention weights を $\boldsymbol{A}^{\ell,h} \in \mathbb{R}^{n\times n}$，$\boldsymbol{A}^{\ell,h}_{i,j} \in [0,1]\ (\forall i,j)$ とする．

画像に対するテキストの注意割合　テキストトークンがどれほど画像トークン・EOI トークンに注意を向けているかの指標を f とし，テキストトークンから画像トークンへの注意の場合，以下のように定義する．

$$f_{\mathcal{I}} = \sum_{i\in\mathcal{T}}\sum_{j\in\mathcal{I}} \boldsymbol{A}_{i,j} \tag{1}$$

テキストトークンから EOI トークンへの注意の場合も同様に定義する．$f_{\mathcal{I}}$，$f_{\mathcal{S}}$ を用いて，画像トークンに対するテキストの注意割合を

$$\frac{f_{\mathcal{I}}}{f_{\mathcal{I}} + f_{\mathcal{S}}} \tag{2}$$

と定義する．EOS トークンに対するテキストの注意割合も同様に定義する．

テキストの画像トークンへの参照割合　テキストトークンが特定の画像トークンを強く参照するのか，多くの画像を等しく参照するのかを明らかにするために，各画像トークンがテキストから受ける注意の大きさの順に累積和を計算する．

Attention Knockout　Attention Knockout 後の Attention weights を $\boldsymbol{A}'$ とすると，テキストから画像への Attention Knockout は

$$\boldsymbol{A}'^{\ell,h}_{i,j} = \begin{cases} 0 & (i \in \mathcal{T},\ j \in \mathcal{I}\ \text{の場合}) \\ \boldsymbol{A}^{\ell,h}_{i,j} & (\text{それ以外}) \end{cases} \tag{3}$$

によって実現し，テキストから EOI トークンへの Attention Knockout は

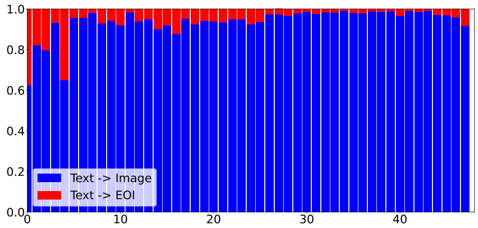
$$\boldsymbol{A}'^{\ell,h}_{i,j} = \begin{cases} 0 & (i \in \mathcal{T},\ j \in \mathcal{S}\ \text{の場合}) \\ \boldsymbol{A}^{\ell,h}_{i,j} & (\text{それ以外}) \end{cases} \tag{4}$$

によって実現する．本研究では全層に対して Attention Knockout を適用する．

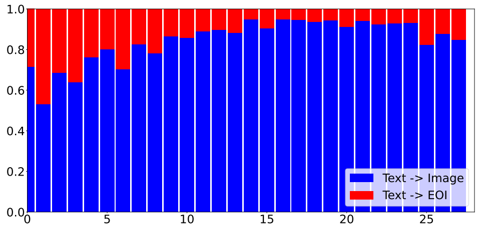
モデル・データセット　Qwen2.5-VL 3B/7B [4]，Gemma3 4B/12B [10] を対象とする．いずれも指示チューニング済みモデルを用いる．MMBench [11] を用いて，画像に対するテキストの注意割合と Attention Knockout による正答率，出力の変化を分析する．MMBench のプロンプトは付録 A.1 を参照．

3.2　結果

画像に対するテキストの注意割合　画像に関するトークン（画像トークン，画像終端の特殊トークン）に対するテキストトークンの注意の割合を図 1 に示す．いずれのモデルも，テキストトークンは単独の画像終端の特殊トークンに注意を向けた上で，主には画像トークンに強く注目していることがわかる．Gemma3 4B と Qwen2.5-VL 3B の結果は付録 B に示す．



(a) Gemma3 12B



(b) Qwen2.5-VL 7B

図 1: 画像関連トークンに対するテキストトークンの注意の割合

テキストの画像トークンへの参照割合　図 2，3 に各画像トークンがテキストから受ける注意の累積和を示す．この図から，Gemma3 12B ではテキストは全体の 20%以下の画像トークンを強く参照するのに対し，Qwen2.5-VL 7B では約 50%の画像トークンを参照しており，いずれのモデルでも層に依らないことがわかる．Gemma3 4B と Qwen2.5-VL 3B の結果は付録 B に示す．

Attention Knockout　Attention Knockout の結果を表 1 に示す．テキストから画像・EOI いずれかへの注意を無効にすると，いずれの場合も性能が著しく低下することがわかる．そこで実際にどのように出力が変化したのかを観察すると，Attention Knockout の方向に依らず，モデルは指示に従わない出力をしたり，出力の終端を示すトークン（EOS トークン）のみを出力するようになった．具体的は，Gemma3 4B では text → img，text → EOI どちらの方向の Attention Knockout でもほとんどの入力に対して EOS トークンを出力し，Gemma3 12B では指示に従った生成が行われなかった．Qwen2.5-VL 3B では text → img の Attention Knockout では EOS トークンのみを出力し，Qwen2.5-VL 7B では text → EOI では指示追従しない生成が大多数であった．このことから，これらのモデルでは，テキストから画像あるいは EOI への注意が指示追従性能に因果的に影響することが明らかになった．これは LLM 部分の Attention 重みに指示追従に関するニューロン [12, 13] が存在することを示唆する．

表 1: Attention Knockout の結果．EOI は画像終端の特殊トークンを示す．また，() 内の数字は標準偏差を表す．

モデル	Knockout の方向	正答率
Gemma3 4B	–	.356 (.479)
	text → img	.169 (.375)
	text → EOI	.224 (.418)
Gemma3 12B	–	.388 (.487)
	text → img	.277 (.447)
	text → EOI	.270 (.444)
Qwen2.5-VL 3B	–	.392 (.488)
	text → img	.176 (.381)
	text → EOI	.207 (.405)
Qwen2.5-VL 7B	–	.410 (.492)
	text → img	.086 (.281)
	text → EOI	.001 (.026)

3.3　現状の LVLM の課題

現状の LVLM には (1) 画像トークンにも注意が向いているため簡単には画像トークン単位での枝刈りができない，(2) 画像トークンは数百トークンとなるため LVLM の推論効率はメモリ・速度両方の観点で非効率である，という課題がある．

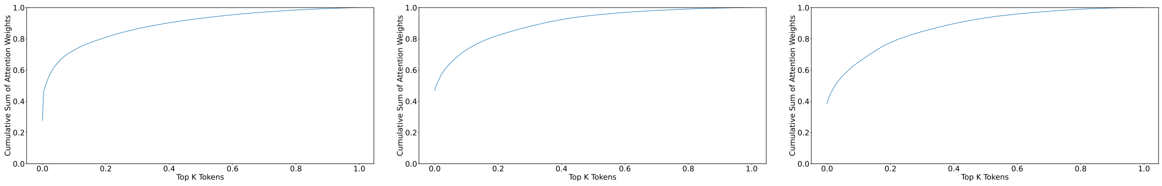
4　視覚情報のための Registering

視覚情報を LVLM の LLM 部分で効率的に扱うために，Gisting [14] や Registering [15] といった手法で採られるような，一部の情報をいくつかの特殊トークンに書き込ませるような構成を考える．第 3 節で観察したように，EOI トークンは既に視覚情報の一部をテキスト側に伝える役割を担っていることから，EOI トークンを Register トークンとし，通常の Attention mask に加え，テキストトークンから画像トークンへの注意を無効にするような Attention mask によって学習・推論を行う（図 4）．これによって，LLM 内部での視覚情報の明確にしつつ，Register トークンの分析や削減によって解釈性や推論効率の向上が期待される．

4.1　実験設定

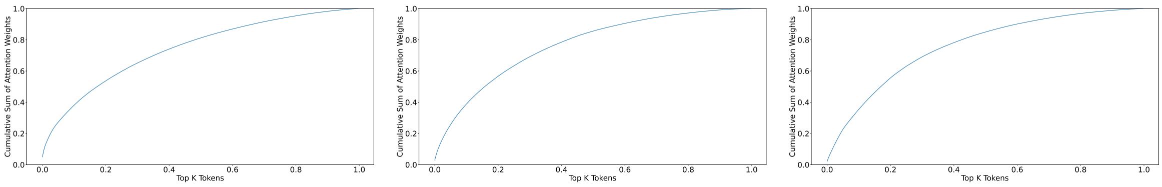
実験対象のモデルは Qwen2.5-VL 3B とし，MiniGPT-4 [16] の指示チューニングで用いられたデータセット[1]によってモデルをファインチュー

1） https://hf.co/datasets/Vision-CAIR/cc_sbu_align



(a) Layer 12 (b) Layer 24 (c) Layer 35

図 2: Gemma3 12B の画像トークンがテキストから受ける注意の累積.



(a) Layer 7 (b) Layer 14 (c) Layer 20

図 3: Qwen2.5-VL 7B の画像トークンがテキストから受ける注意の累積.

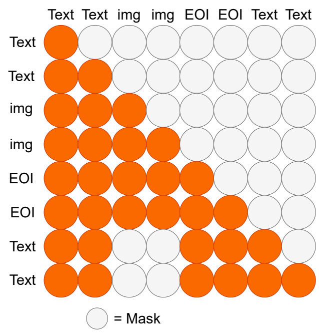


図 4: Register トークンと Attention Mask．ここでは EOI トークンを Register トークンとして扱う．

ニングする．データセットを訓練セットと検証セットに分割し，訓練セットは 16,880 件，検証セットは 1,876 件からなる．

ファインチューニングは LVLM の視覚エンコーダーとプロジェクターを凍結し，LLM 部分のみをフルパラメーターファインチューニングする．ベースラインとして，Registering なしで，このデータセットを用いて訓練したモデルを用意し，Registering を用いて訓練したモデルを，MMBench での正答率の観点から比較する．Registering を用いる場合は，Register トークンを 16 個使う場合と，256 個使う場合の 2 通りを実験する．ハイパーパラメーターは付録 A.2 を参照のこと．

4.2 結果

表 2 に，Register トークンを用いてファインチューニングしたモデルの MMBench における正答率を示す．この結果から，16 トークンの Register トークンを用いた場合には，表 1 の text → img 方向での Attention Knockout よりも性能は向上している一方で，ベースラインと比較すると，依然として正答率は低い．実際の出力を分析すると，Register トークンを用いたモデルは入力中の指示に従わない出力が大部分を占めており，これによって低い正答率となったと予想される．第 3 節が示唆するように，テキストトークンから画像トークンへの注意機構には指示追従性能に関するニューロンが含まれていることからも，今回の実験設定ではこの問題を解消するようにはモデルを訓練できなかったことを示唆する．より多様なデータセットと多くのデータを用いることでこの問題は解決される可能性がある．

表 2: Qwen2.5-VL 3B のファインチューニングの結果．() 内は標準偏差を表す．

	正答率
Fine-tuned	.392 (.488)
w/ registering (16 トークン)	.246 (.431)
w/ registering (256 トークン)	.253 (.435)

5 結論

本研究では LVLM における視覚情報の伝播経路を分析し，その結果，モデルは非効率な方法で視覚情報をテキストに伝えていることが明らかになった．そこで，視覚情報を LLM 部分で少数の特殊トークンに書き込むようなモデルの設計を考案した．実験の結果，提案手法は性能に限界があることが判明したが，訓練データの追加やデータ分布の調整によって性能が向上する可能性がある．

謝辞

国立研究開発法人産業技術総合研究所の令和 7 年度覚醒プロジェクトの支援を受けた．

参考文献

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In **Proceedings of the 36th International Conference on Neural Information Processing Systems**, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[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 **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[5] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. **IEEE Transactions on Intelligent Vehicles**, 2024.

[6] Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical VQA. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 19188–19205, Vienna, Austria, July 2025. Association for Computational Linguistics.

[7] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[8] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[9] Alessandro Pietro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in native multimodal models. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems**, 2025.

[10] Gemma Team. Gemma 3 technical report, 2025.

[11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model annbsp;all-around player? In **Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI**, p. 216–233, Berlin, Heidelberg, 2024. Springer-Verlag.

[12] Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. Language modeling with sparse product of sememe experts. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 4642–4651, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[13] Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley You Ren, Andrew Miller, Udhyakumar Nallasamy, and Jaya Narain. Do LLMs "know" internally when they follow instructions? In **The Thirteenth International Conference on Learning Representations**, 2025.

[14] Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[15] Zhi Qu, Yiran Wang, Jiannan Mao, Chenchen Ding, Hideki Tanaka, Masao Utiyama, and Taro Watanabe. Registering source tokens to target language spaces in multilingual neural machine translation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 21687–21706, Vienna, Austria, July 2025. Association for Computational Linguistics.

[16] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[17] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In **The Thirteenth International Conference on Learning Representations**, 2025.

表 3: 訓練時のハイパーパラメーター

Parameter	Value
Batch size	32
Epochs	10
Sequence length	8,192
Learning rate	2×10^{-7}
Learning rate scheduler	cosine
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1×10^{-8}
Precision	bf16
Max pixel	50,176
Min pixel	784
Seed	42

A　実験設定（詳細）

A.1　プロンプト

MMBench のプロンプトは [17] と同様に以下のものを用いる．

MMBench

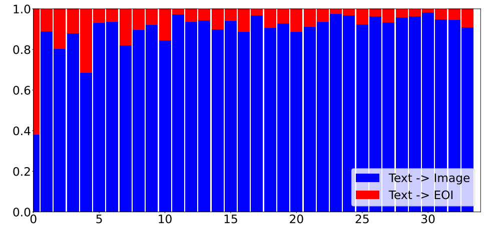
{Hint}
{Question}
A. {Choice A}
B. {Choice B}
C. {Choice C}
D. {Choice D}
Answer with the option's letter from the given choices directly.

A.2　ハイパーパラメーター

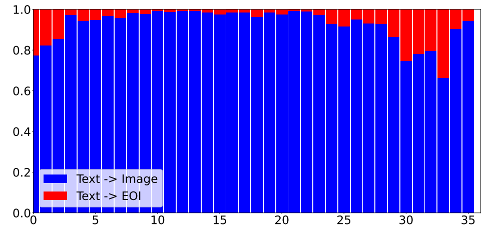
表 3 にファインチューニング時のハイパーパラメーターを示す．

B　視覚情報の伝播経路

図 5 に Gemma3 4B と Qwen2.5-VL 3B における画像関連トークンに対するテキストトークンの注意の割合を示す．図 6，7 にはそれぞれのモデルにおいて，画像トークンがテキストから受ける注意の割合の累積を示す．

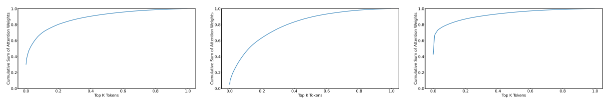


(a) Gemma3 4B



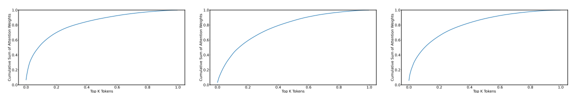
(b) Qwen2.5-VL 3B

図 5: 画像関連トークンに対するテキストトークンの注意の割合



(a) Layer 8　(b) Layer 17　(c) Layer 25

図 6: Gemma3 4B の画像トークンがテキストから受ける注意の累積．



(a) Layer 9　(b) Layer 16　(c) Layer 25

図 7: Qwen2.5-VL 3B の画像トークンがテキストから受ける注意の累積．