大規模視覚言語モデル内部における ダイアグラムの表現形成過程

吉田遥音 [1]　工藤慧音 [1]　青木洋一 [1]　田中涼太 [2]　斉藤いつみ [1]　坂口慶祐 [1]　乾健太郎 [1]
[1] 東北大学　[2]NTT 株式会社　人間情報研究所
yoshida.haruto.p1@dc.tohoku.ac.jp

概要

大規模視覚言語モデル（LVLM）は高いダイアグラム理解性能を示すが，ノードと有向エッジで表される要素間の関係理解には課題がある．この根本的な原因を探るため，本研究では有向グラフに基づく合成ダイアグラムデータセットを構築し，モデルの内部表現をプロービングした．その結果，ノードの情報と大域的な構造的特徴は，視覚エンコーダの単一の隠れ状態にすでに線形に表現される一方で，エッジの情報は言語モデルのテキスト部分で初めて線形に表現されることを明らかにした．この知見は，線形分離可能な表現の形成段階が視覚情報の種類ごとに異なることを示唆し，LVLM が要素間の関係理解に苦戦する理由を説明する一助になりうる．

1　はじめに

ダイアグラムは視覚的な表現を通じて複雑な情報を伝達し，人々の理解を促すことから，学術研究 [1, 2] やビジネス [3, 4] などの領域において重要なコミュニケーション媒体となっている．大規模視覚言語モデル（LVLM） [5, 6, 7] は高いダイアグラム理解性能を示している [8, 9] が，先行研究では，LVLM が関係情報，特に矢印や線で示される要素間の接続の理解に依然として苦戦することが報告されている [10, 11]．これらの関係情報の理解はダイアグラムの構造と意味を捉える上で重要であり，その認識の失敗がボトルネックになっている．

この問題をより深く理解するため，本研究ではダイアグラム特有の視覚要素が LVLM 内部でどのように認識されているかを調査する．具体的には，ノード，エッジ，および大域的構造に関する情報が，視覚エンコーダと言語モデルのどこで，いつアクセス可能になるかに焦点を当てる．自然画像を対象とした分析 [12, 13] とは対照的に，本研究では構造化された関係を表現するために記号的表現に依存するダイアグラムを対象とする．また制御された分析のために，合成ダイアグラムデータセットを構築し，プロービングによって，視覚エンコーダと言語モデル内部におけるノード，エッジ，および大域的な構造情報の線形分離可能性を検証する．

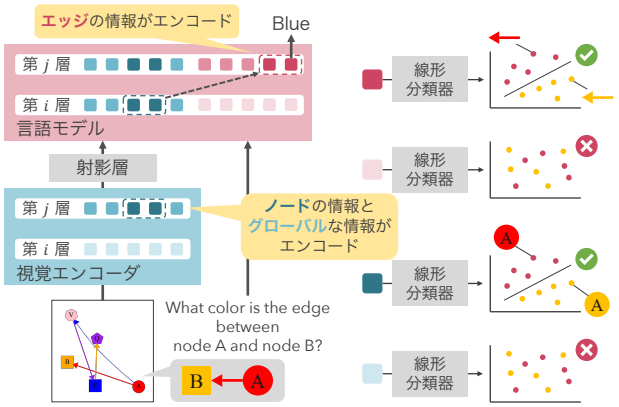

図 1　本研究の概要．合成ダイアグラムを用いたプロービングにより，LVLM の内部表現を分析する．ノード情報（例：ノードの色）と大域的情報（例：ノード数）は視覚エンコーダ内の単一の画像パッチに線形に表現されるが，エッジ情報（例：エッジの色）は言語モデル内の単一のテキストトークンに線形に表現されることを明らかにした．

プロービングの結果，ノードおよび大域的な構造情報は視覚エンコーダの単一の画像パッチにすでに線形に表現されているのに対し，エッジの情報は言語モデルの単一のテキストトークンで初めて線形に表現されることを示した．この知見は，LVLM 内部で視覚情報がどのように処理，表現されているかを明らかにし，ダイアグラム理解システムの設計と分析に対する指針を提供する．

2　データセット構築

ノードの色やエッジの向きなどの分析対象の要素を正確に制御するため，ノードとエッジを持つ有向

表 1　分析対象の観点とラベルの対応例.

観点 a	ラベル y
ノードの色	{ 赤, 緑, 黄, 青, 茶, 橙, 桃, 紫 }
エッジの向き	{ A→B, B→A }
ノードの数	{ 1, 2, 3, 4, 5 }

グラフの合成データセットを構築する.

2.1　ダイアグラムの仕様

各ダイアグラムは 5 つのノードからなり，各ノードには識別子として単一のアルファベット文字が記載されている．また，各ノードは色と形，各エッジは色と種類の属性を持つ.

2.2　評価観点

先行研究 [14, 15, 16, 17, 10, 11] に基づき，ダイアグラムにおける基本的な視覚情報として 11 の観点を定義する．また本稿では，代表的な観点として表 1 に示すノードの色，エッジの向き，ノードの数について議論する．各観点の分析にあたり，識別子 A を持つノード（ノード A）と，ノード A と B を繋ぐエッジ（エッジ AB）に焦点を当てる．例えばノードの色の観点では，ノード A の色に注目する．詳細は付録 A に記載した.

2.3　データセットの定義

観点 a が認識可能かどうかを検証するためのデータセット $\mathcal{D}^{(a)}$ を以下のように定義する：

$$\mathcal{D}^{(a)} = \{(x_i, y_i)\}_{i=1}^{N}. \tag{1}$$

x_i は i 番目のダイアグラム画像，y_i はそのラベルである．具体的には，表 1 に示すような，各観点についてのラベル y_i を含むデータセットである．さらに，各観点 a に対して，図 2 に示すような，ノード配置が異なる 2 種類のデータセットを構築する：

- $\mathcal{D}^{(a)}_{\text{rand}}$: 各ダイアグラムに対して，ノード配置が独立にランダムに決定される.
- $\mathcal{D}^{(a)}_{\text{fix}}$: すべてのダイアグラムに対して，各識別子を持つノードが同じ位置に配置される.

$\mathcal{D}^{(a)}_{\text{rand}}$, $\mathcal{D}^{(a)}_{\text{fix}}$ はともに，分類クラスあたり 100 件のデータからなる.

プロービング用訓練データセット　プローブの訓練データセットとして，ランダムなノード配置のダイアグラムデータセット $\mathcal{D}^{(a)}_{\text{train}}$ を構築する．プローブが訓練データ内の表層的な分布に基づいて学習することを防ぐため，対象ノードまたはエッジが存在しないダイアグラムからなるデータセット $\mathcal{D}^{(a)\perp}_{\text{rand}}$ を用意する．$\mathcal{D}^{(a)\perp}_{\text{rand}}$ は $\mathcal{D}^{(a)}_{\text{rand}}$ と同数のサンプルを含み，すべてのサンプルの正解ラベルは $y = \perp$（N/A の意）である．そして，プローブ訓練データセットを $\mathcal{D}^{(a)}_{\text{train}} = \mathcal{D}^{(a)}_{\text{rand}} \cup \mathcal{D}^{(a)\perp}_{\text{rand}}$ と定義する.

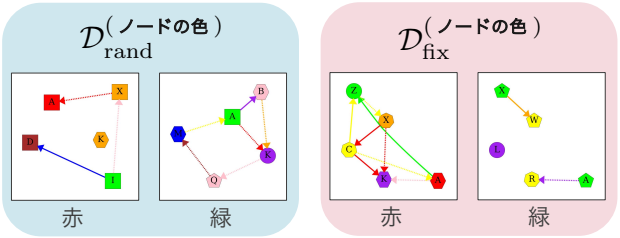

図 2　合成ダイアグラムの例.

表 2　VQA の正解率（%）.

	ノードの色	エッジの向き	ノードの数
Qwen3-VL 8B	91.4	49.3	40.3
チャンスレベル	12.5	50.0	20.0

評価データセット　プローブの評価データセットとして，固定されたノード配置のダイアグラムデータセット $\mathcal{D}^{(a)}_{\text{test}}$ を構築する．$\mathcal{D}^{(a)}_{\text{test}}$ は異なるノード配置に統一された 10 件の $\mathcal{D}^{(a)}_{\text{fix}}$ からなり，$\mathcal{D}^{(a)}_{\text{test}} = \{\mathcal{D}^{(a)j}_{\text{fix}}\}_{j=1}^{10}$ と表される．これを用いることで，得られた結果が特定のノード配置でのみ生じるものでないことを保証する.

3　予備実験：VQA

予備実験として，合成ダイアグラムに対する LVLM の VQA 性能を評価する．具体的には，モデルにダイアグラム x と質問 q を入力し，正解ラベル y を正しく予測できるかを評価する.

モデル　これ以降のすべての実験において分析対象のモデルは Qwen3-VL 8B [5] である.

評価指標　すべての評価データサブセット $d_{\text{fix}} \in \mathcal{D}^{(a)}_{\text{test}}$ にわたる平均正解率を用いる.

実験結果　表 2 に，観点ごとの Acc_{VQA} を示す．エッジの向き以外の正解率はチャンスレベルを上回り，エッジの向きの正解率はチャンスレベル程度であった．これは，モデルがエッジの向き以外を一定程度正しく認識しているが，エッジの向きの認識は不十分であることを示唆している.

4　プロービング

LVLM 内のどの層の，どの位置にダイアグラムの視覚情報が線形に表現されているかを特定する．そ

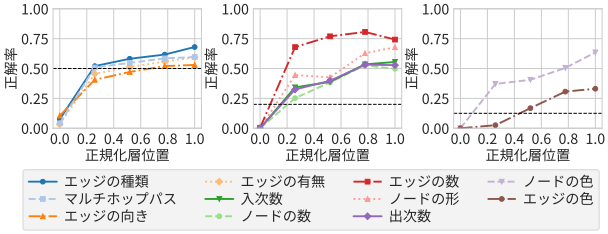

図 3　視覚エンコーダの各層の正解率．x 軸は層の相対位置，y 軸は正解率である．正解率は各層で最も正解率が高い画像パッチの正解率であり，黒の点線は閾値 τ を表す．

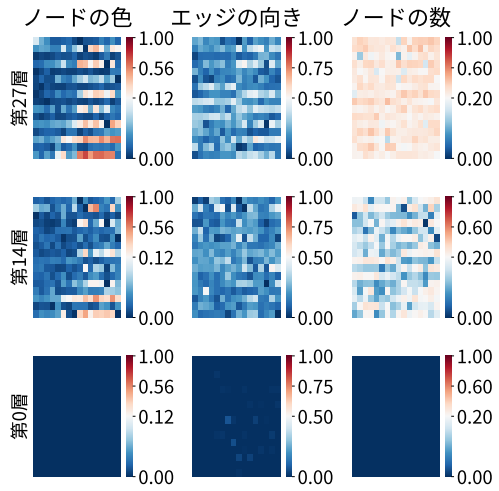

図 4　視覚エンコーダにおける位置ごとの正解率．評価ダイアグラムのノード配置は，図 2 の右側のダイアグラムと同じである．

のために，視覚エンコーダと言語モデルの各層に対してプローブを訓練し，隠れ状態からどの程度 VQA の回答を予測できるかを評価する．

4.1　プローブの定義

ダイアグラム x と質問テキスト q を入力したときの，視覚エンコーダまたは言語モデルの第 l 層，位置 t における隠れ状態を $h_{l,t}$ とする．隠れ状態から VQA タスクの回答または $\perp$ を予測するプローブ f を，線形分類器として以下のように定義する：

$$\hat{y} = f(h_{l,t}) = \arg\max_{\mathcal{Y}\cup\{\perp\}} \boldsymbol{W}\boldsymbol{h}_{l,t} + \boldsymbol{b}. \tag{2}$$

ここで，$W \in \mathbb{R}^{(|\mathcal{Y}|+1)\times|h_{l,t}|}$ は重み行列，$b \in \mathbb{R}^{(|\mathcal{Y}|+1)}$ はバイアス項である．

以下の 3 つの構成要素について，それぞれ独立にプロービングを行う：

視覚エンコーダ　位置 t の画像パッチに対応する隠れ状態 $h_{l,t}$ からラベル y を予測する．訓練には層 l におけるすべての位置 t の隠れ状態を使用し，各層 l に対してプローブ f_l を訓練する．

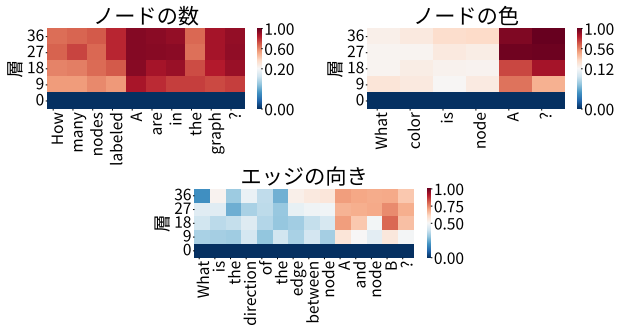

図 5　言語モデルのテキスト部分に対するプロービング結果．横軸は質問内のトークン位置，縦軸は層を示す．

言語モデル（画像部分）　位置 t の画像パッチに対応する隠れ状態 $h_{l,t}$ からラベル y を予測するプローブ f_l を訓練する．

言語モデル（テキスト部分）　位置 t のテキストトークンに対応する隠れ状態 $h_{l,t}$ からラベル y を予測する．層 l と位置 t における各隠れ状態に対してプローブ $f_{l,t}$ を独立に訓練する．プローブの対象位置 t は，質問 q 内のすべてのトークン位置である．

4.2　評価指標

評価サブセット $d_{\text{fix}} \in \mathcal{D}_{\text{test}}^{(a)}$ に対して，第 l 層の位置 t の隠れ状態を使用したときのプローブ f の正解率を $\text{acc}_{l,t}(d_{\text{fix}})$ とする．最終的な評価スコアは，すべてのサブセットにわたる平均正解率である：

$$\text{Acc}_{l,t}(\mathcal{D}_{\text{test}}^{(a)}) = \underset{d_{\text{fix}}\in\mathcal{D}_{\text{test}}^{(a)}}{\text{mean}} \text{acc}_{l,t}(d_{\text{fix}}). \tag{3}$$

ここで，$\text{mean}_{s\in\mathcal{S}}(\cdot)$ はすべての $s \in \mathcal{S}$ に対する値の算術平均である．

プロービングの成否は，正解率が閾値 $\tau = 1/|\mathcal{Y}|$ を超えたかで判定し，$\text{Acc}_{l,t} > \tau$ の場合，プロービングが成功したと結論づける．

4.3　実験結果

エッジの向きは視覚エンコーダで線形に表現されない．　図 3 に，視覚エンコーダの各層における正解率を示す．エッジの向き以外の正解率は，層の深さとともに徐々に増加し，深い層ほどこれらの観点に対して線形分離可能な表現を形成することを示している．一方で，エッジの向きの正解率は最終層においても低いままであり，単一の隠れ状態からは線形分離不可能であることを示唆している．

ノード情報と大域的な構造情報は視覚エンコーダで線形分離可能．　図 4 に，視覚エンコーダの各層

表 3　因果介入の結果．介入率は画像パッチの総数に対する介入対象の割合（%）．

	ノードの色	エッジの向き	ノードの数
介入なし	91.4	49.3	40.3
介入あり	11.7	48.8	25.7
コントロール	83.4	50.3	31.8
チャンスレベル	12.5	50.0	20.0
介入率（%）	19.1	0.625	89.6

の各位置における正解率を示す．ノードの色の観点は，層が進むにつれて右下の領域を中心に高い正解率を持つ領域が拡大した．対象ノード A は右下に配置されているため，ノード A に関する情報が主にその領域と近傍に表現されていることを示している．

視覚エンコーダは背景位置に対応する隠れ状態を効果的に活用する．　ノードの数の観点については，後半層においてほぼすべての位置で正解率が閾値を上回った．これは，ダイアグラム全体からの手がかりの統合を必要とする大域的情報が，背景領域を含む広範な隠れ状態に分散していることを示し，ViT [18] におけるレジスタトークン [19] や attention sink [20] との関連を示唆している．

テキストの条件付けが線形分離可能な表現を構築する．　図 5 に，言語モデルのテキスト部分に対する各層の各位置における正解率を示す．どの観点も，対象ノードまたはエッジを指定するトークンで正解率が増加した．これは，言語モデルが入力テキストを条件として，画像パッチからテキストの隠れ状態へ選択的に情報を集約していることを示唆している．また，ノードの数の観点は最初のトークンから比較的高い正解率を示し，大域的な情報は明示的な条件がなくとも集約されることを示唆している．

言語モデル（画像部分）での表現の変化は限定的であった．詳細は付録 B.2 に記載した．

5　因果介入

プロービングで検出した線形分離可能な表現を，モデルが実際に推論に使用しているかを検証する．そのために，視覚エンコーダの隠れ状態の一部を言語モデルに入力される前に破壊し，VQA 性能の変化として因果的寄与を測定する．

5.1　実験設定

対象層　言語モデルへ入力される視覚エンコーダのすべての層に介入を適用する．

対象位置　介入の対象位置は，プロービングの正解率が閾値 τ を超える位置すべてとする．つまり，データセット $d_{\text{fix}} \in \mathcal{D}_{\text{test}}$ と視覚エンコーダの第 l 層に対して，対象位置集合は以下で定義される：

$$\mathcal{S}^l_{d_{\text{fix}}} = \{t \in \{1, \ldots, T\} \mid \text{acc}_{l,t}(d_{\text{fix}}) > \tau\}. \tag{4}$$

ここで，T は位置の総数である．また，$\mathcal{S}^l_{d_{\text{fix}}}$ の補集合を $\bar{\mathcal{S}}^l_{d_{\text{fix}}}$ とし，これは正解率が τ 以下の位置の集合である．

介入操作　各入力 x_i に対して，介入対象の隠れ状態（位置 $t \in \mathcal{S}^l_{d_{\text{fix}}}$）を，それ以外のすべての隠れ状態（位置 $t \in \bar{\mathcal{S}}^l_{d_{\text{fix}}}$）の平均ベクトルで置き換える：

$$\tilde{h}_{i,l,t} = \begin{cases} \text{mean}_{t \in \bar{\mathcal{S}}^l_{d_{\text{fix}}}} \; h_{i,l,t} & (t \in \mathcal{S}^l_{d_{\text{fix}}}), \\ h_{i,l,t} & (t \notin \mathcal{S}^l_{d_{\text{fix}}}). \end{cases} \tag{5}$$

介入操作によって Acc_{VQA} が減少した場合，プロービングによって検出された情報が推論に因果的に寄与していることを示唆する．

コントロール実験　介入効果が真にプロービングの正解率が高い隠れ状態への介入に起因するかを明確にするため，コントロール実験として，プロービングの正解率が閾値 τ を超える位置と同数の位置をランダムに選択した場合の実験を行う．

5.2　実験結果

表 3 に因果介入の結果を示す．ノードの色，ノードの数の観点は，介入後に Acc_{VQA} が低下し，コントロールでの Acc_{VQA} の低下は限定的であった．すなわち，これらの観点はプロービングの正解率が高かった画像パッチの隠れ状態が推論に寄与していることを示している．一方で，エッジの向きの観点は，介入しない場合の Acc_{VQA} がすでにチャンスレベル程度であり，介入の影響は確認できなかった．

6　結論

本研究では，合成ダイアグラムを用いたプロービングにより，ダイアグラムの視覚情報が LVLM 内部のどこに表現されるかを分析した．その結果，ノードの情報と大域的な構造情報は視覚エンコーダの単一の画像パッチで線形に表現されるが，エッジの情報は言語モデルのテキスト部分で初めて線形に表現されることを明らかにした．この知見は，LVLM が異なる種類の視覚情報を異なる方法で処理することを示し，この違いがダイアグラムの要素間の関係認識における課題の一因である可能性を示唆する．

謝辞

本研究は JST CREST JPMJCR20D2，JSPS 科研費 JP25KJ0615，JSPS 科研費 JP25K03175，JST PREST JPMJFR242S，JSPS 科研費 JP24K20829，JST SPRING JPMJSP2114 の助成を受けたものです．本研究を進めるにあたり多大なご助言，ご協力を賜りました Tohoku NLP グループの皆様に感謝いたします．

参考文献

[1] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-Tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Kurt W Kohn and Mirit I Aladjem. Circuit diagrams for biological networks. **Mol. Syst. Biol.**, Vol. 2, No. 1, p. 2006.0002, January 2006.

[3] Gregor Jošt, Jernej Huber, Marjan Heričko, and Gregor Polančič. Improving cognitive effectiveness of business process diagrams with opacity-driven graphical highlights. **Decis. Support Syst.**, Vol. 103, pp. 58–69, November 2017.

[4] Sebastian Kernbach and Martin J Eppler. The use of visualization in the context of business strategies: An experimental evaluation. In **2010 14th International Conference Information Visualisation**, pp. 349–354. IEEE, July 2010.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. **arXiv [cs.CV]**, November 2025.

[6] 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. **arXiv [cs.CV]**, February 2025.

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 26296–26306, 2024.

[8] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATH-VERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In **Lecture Notes in Computer Science**, Lecture notes in computer science, pp. 169–186. Springer Nature Switzerland, Cham, 2025.

[9] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In **The Twelfth International Conference on Learning Representations**, October 2023.

[10] Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? In **Forty-second International Conference on Machine Learning**, June 2025.

[11] Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 30678–30701, Stroudsburg, PA, USA, 2025. Association for Computational Linguistics.

[12] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In **The Twelfth International Conference on Learning Representations**, 2024.

[13] Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. Probing multimodal large language models for global and local semantic representations. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 13050–13056, 2024.

[14] Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. Using color in visualization: A survey. **Comput. Graph.**, Vol. 35, No. 2, pp. 320–333, April 2011.

[15] Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin J Latecki. FlowLearn: Evaluating large vision-language models on flowchart understanding. https://par.nsf.gov/servlets/purl/10544557, August 2024. Accessed: 2025-8-15.

[16] Anna Sterzik, Nils Lichtenberg, Jana Wilms, Michael Krone, Douglas W Cunningham, and Kai Lawonn. Perception of line attributes for visualization. **IEEE Trans. Vis. Comput. Graph.**, Vol. 30, No. 1, pp. 1041–1051, January 2024.

[17] Chin Tseng, Arran Zeyu Wang, Ghulam Jilani Quadri, and Danielle Albers Szafir. Shape it up: An empirically grounded approach for designing shape palettes. **IEEE Trans. Vis. Comput. Graph.**, Vol. PP, No. 1, pp. 349–359, September 2024.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **International Conference on Learning Representations**, October 2020.

[19] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In **The Twelfth International Conference on Learning Representations**, October 2023.

[20] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In **The Twelfth International Conference on Learning Representations**, 2024.

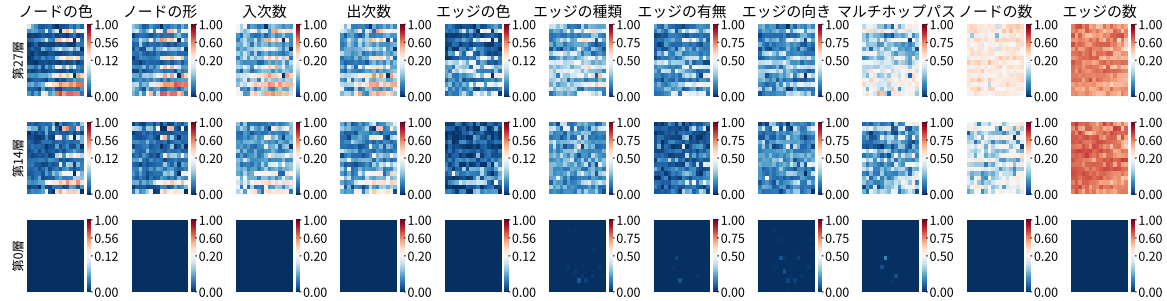

図 6　視覚エンコーダにおける位置ごとの正解率．評価ダイアグラムのノード配置は，図 2 の右側のダイアグラムと同じである．

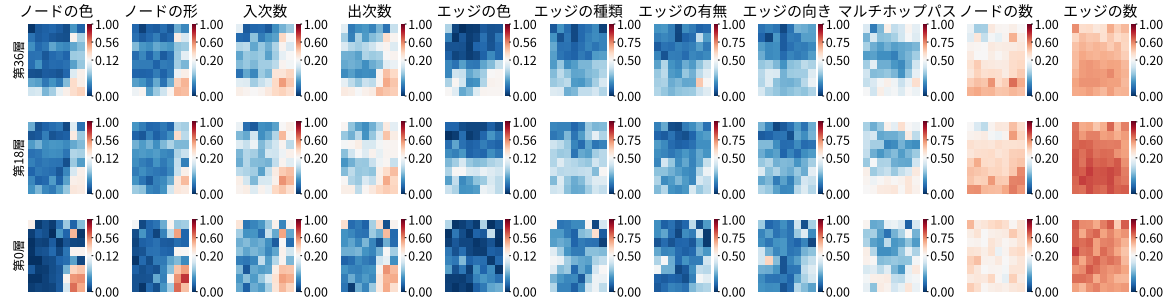

図 7　言語モデルにおける位置ごとの正解率．評価ダイアグラムのノード配置は，図 2 の右側のダイアグラムと同じである．

表 4　VQA の正解率（%）．エッジの向き以外のすべての観点で正解率がチャンスレベルを上回った．

	ノードの色	ノードの形	入次数	出次数	エッジの色	エッジの種類	エッジの有無	エッジの向き	パスの有無	ノードの数	エッジの数
Qwen3-VL 8B	91.4	76.6	40.3	34.7	57.3	73.5	69.6	49.3	58.3	40.3	21.6
チャンスレベル	12.5	20.0	20.0	20.0	12.5	50.0	50.0	50.0	50.0	20.0	20.0

A　データセットの詳細

すべての観点とそれに対応するラベルは以下の通りである．

- ノードの色: { 赤, 緑, 黄, 青, 茶, 橙, 桃, 紫 }
- ノードの形: { 円, 四角形, 五角形, 六角形, 七角形 }
- 入次数: { 0, 1, 2, 3, 4 }
- 出次数: { 0, 1, 2, 3, 4 }
- エッジの色: { 赤, 緑, 黄, 青, 茶, 橙, 桃, 紫 }
- エッジの種類: { 実線, 点線 }
- エッジの有無: { 有, 無 }
- エッジの向き: { A→B, B→A }
- パスの有無: { 有, 無 }
- ノードの数: { 1, 2, 3, 4, 5 }
- エッジの数: { 1, 2, 3, 4, 5 }

B　実験結果の詳細

予備実験として，合成ダイアグラムに対する LVLM の VQA 性能を評価する．具体的には，シンプルな VQA 設定において，ダイアグラム x と質問 q が与えられたときに，モデルが正解ラベル y を正しく予測できるかを評価する．

B.1　事前実験：VQA

すべての観点に対する VQA の正解率を表 4 に示す．

B.2　プロービング

すべての観点に対する視覚エンコーダの位置ごとの正解率を図 6 に，言語モデルの位置ごとの正解率を図 7 に示す．