知識グラフの反復的な探索による画像の詳細な説明文の生成

加藤 優汰 [1]　尾崎 慎太郎 [2]　林 和樹 [2]
坂井 優介 [2]　上垣外 英剛 [2]　林 克彦 [1,2]　渡辺 太郎 [2]
[1] 東京大学　[2] 奈良先端科学技術大学院大学 (NAIST)
{ukato6209, katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp
ozaki.shintaro.ou6@naist.ac.jp
{hayashi.kazuki.hl4, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

大規模視覚言語モデル（LVLM）は，画像に基づく文章生成や視覚質問応答において高い性能を示している．しかし，画像に含まれる対象に関連する実体や概念（エンティティ）間の事実関係を網羅的かつ正確に記述することは依然として困難である．本研究では，LVLM による詳細かつ正確な画像説明生成を目的として，知識グラフから検索拡張生成（RAG）を用いて事実情報を効率的に利用する枠組みを提案する．具体的には，回答生成と知識グラフ検索を反復的に行い，正誤判定に基づいて探索を制御することで，必要十分な事実情報を効率よく取得する．また，画像とエンティティの対応関係が明確な芸術作品を対象とした知識グラフ（ExpArt-KG）を構築した．提案手法を構築した知識グラフに適用して実験を行った結果，提案手法により芸術作品における説明文の詳細度が向上し，固定回数の反復と同水準の生成品質を維持しつつ，外部知識の検索コストを削減できることが示された．

1　はじめに

大規模視覚言語モデル (LVLM) [1, 2, 3] は画像に基づく文章生成や視覚質問応答において高い性能を示しており [4, 5, 6]，その一つに，画像を詳細に説明する説明生成能力がある． このような場面において，LVLM は単に画像に含まれる人物や物体を認識するだけではなく，それを認識した対象の周辺知識と結びつけることが求められる．しかし，LVLM は認識した対象に関連するエンティティ間の事実関係を網羅的かつ正確に説明することが困難であることが指摘されている [7].

このような LVLM の限界を補う実用的な手法として，外部知識を利用する検索拡張生成 (RAG) [8] が挙げられる．この枠組みは，検索された信頼性の高い外部情報を手掛かりに回答を生成することで，幻覚を抑制し，モデルに詳細な知識を補完することを可能にする．特に，エンティティ間の複雑な事実関係を網羅的かつ正確に扱う上では，体系的な構造を持つ知識グラフを活用し，エンティティ間の事実関係を明示的に与える手法が有効である．追加学習を行わずにモデルの知識を拡張できる一方で，知識グラフを利用した既存の手法は探索の深さや反復回数を固定的に設定しており，探索が浅ければ情報不足に，深ければ探索コストの増加につながるというトレードオフが存在する [9]． そのため，有用な事実関係を効率的に検索し，説明生成に適切に利用するための汎用的な枠組みは十分に整理されていない．

本研究では，LVLM による画像の詳細な説明生成を目的として，生成回答に対する大規模言語モデル (LLM) [10, 11, 12, 13] による正誤判定 [14] を導入し，その判定に基づいて知識グラフを動的に探索し回答を再生成する反復 RAG 枠組みを提案する．加えて，主要な画像とそのエンティティが一対一に定まる芸術作品ドメインを対象とした，LVLM の知識を拡張するための知識グラフ（ExpArt-KG) を構築する．

提案手法と ExpArt-KG を用いた際の LVLM の説明生成能力を複数観点の評価尺度により評価した結果，回答生成の反復により知識グラフが効率的に探索され，説明文の詳細度が向上することを確認した．また，正誤判定に基づく反復制御により，固定回数の反復と同水準の生成品質を維持しつつ，外部知識の検索コストを削減できることを示した．

2　提案手法

我々が提案する検索手法では，LVLM に与えるクエリに含まれるエンティティに対応するノードの局所的な探索と，取得した事実情報を用いた回答の

生成および検証を反復的に行うことで，外部知識の検索コストを抑えつつ有用な事実情報を取得する．LLM の自己検証と RAG を統合した反復回答により性能向上が確認された先行研究 [15] に倣い，LLM による回答の検証を導入することで，効率的な探索と高品質な説明生成の両立を図る．本手法で用いるプロンプトの詳細は付録 A に示す．

提案手法の具体的な手順は以下の通り:

(1) トリプルの取得　クエリに含まれるエンティティのうち，知識グラフのノードとして存在するものを抽出し，それらを含むトリプルを取得する．

(2) トリプルの選定　Step 1 で取得したトリプルに対して TF-IDF [16] に基づくランク付けを行い，各エンティティを含むトリプルから上位 10 件をそれぞれ選定する． TF-IDF によるランク付けでは，トリプルを構成する要素を語，特定のエンティティを含むトリプルの集合を文書とみなす． TF-IDF の詳細な算出方法は付録 A に示す．

(3) 回答生成　元のクエリに (2) で選定したトリプルを付加した拡張クエリを構成し，対象の作品（画像）とともに LVLM に与えて回答を生成する．

(4) 回答検証　元のクエリと Step 3 で生成された回答を LLM に与え，回答の正誤を検証する．LLM は Forced-decoding により「True」または「False」のいずれかを出力する． 「True」が得られた場合はその回答を LVLM の最終的な回答とし，「False」の場合はトリプルの取得および回答の生成を再度行う．

(5) トリプルの再取得　Step 4 で「False」が得られた場合，元のクエリまたは直前に生成された回答を用いて，Step 1 および Step 2 の手順でトリプルを再取得し，Step 3 および Step 4 により回答の再生成と検証を行う． この処理は，「True」が得られるか最大反復回数に達するまで繰り返す．

3　知識グラフの構築手法

画像を説明生成することに適した知識グラフを構築するための汎用的な手法を導入する．任意の質問応答データセットに含まれるテキストの情報に基づき，外部知識ベースを参照してエンティティとそれらの関係性を抽出することで知識グラフを構築する．具体的な構築の手順は下記の通りである．

(1) ノードの候補の抽出　質問応答データセットに含まれる質問文，参照説明文，および利用可能なメタデータ（画像のタイトルなど）に含まれる文字列から，知識グラフのノードの候補となる語句を抽出する．

(2) ノードの選定　(1) で抽出したノードの候補のうち，英語版 Wikipedia の記事のタイトルとして存在するものをエンティティとして採用する．さらに，対象ドメインに特化した密な知識グラフを構築するため，対応する Wikidata [17] の識別子が複数存在するような曖昧性の高い語や，一般概念を表す語句はノードから除外する．これにより，対象ドメインに固有かつ明確なエンティティのみを選定する．

(3) エッジの構築　(2) で選定された各ノードに対応する Wikidata のエンティティを参照し，それらの間に定義されている述語を取得する．取得した述語をノード間の意味的関係を表すエッジとして定義することで，エンティティ間の事実関係を構造化した知識グラフを構築する．

本研究では，この手法を ExpArt [7] に適用し，芸術作品に関する知識グラフである ExpArt-KG を構築した．

4　実験設定

モデル　画像の説明文生成を行う LVLM に Qwen 3-VL [18]，回答の検証を行う LLM（検証用 LLM と定義する）に Qwen 3 [10] を用いた．

データセット　ExpArt にあるテストデータ[1)]のうち，画像の URL[2)]が 2025 年 12 月時点で有効であり，かつ構築した ExpArt-KG のエンティティを少なくとも 1 つ以上回答文に含まれるデータを対象とし，その中から評価として用いることのできる約 25%を抽出して評価に用いた. 詳細は付録 A に記す.

入力設定　本実験では，各画像に対して以下の 2 種類の質問文の設定を用いた．

(1) **タイトルあり**: LVLM は画像とタイトルを含む質問文から説明文を生成する．検証用 LLM はタイトルを含む質問文と回答に基づいて正誤判定を行う．

(2) **タイトルなし**: LVLM は，画像の説明文の生成時に画像と，タイトルを含まない質問文が与えられる．検証用 LLM には，同様にタイトルを含まない質問文と回答のみが与えられるため，回答の論理的整合性のみを根拠に正誤判定が行われる．

れらの複数の設定を通じて，タイトルの有無による生成と検証の性能の差，および回答の正誤判定における画像のタイトルの効果を検証する．

1) https://huggingface.co/datasets/naist-nlp/ExpArt
2) https://commons.wikimedia.org/wiki/Main_Page

表 1　本実験における複数設定で得られたスコア．**太字**は評価指標における最高値を示す．着色された行のうち、青色は提案手法、灰色は Baseline を示す．

RAG の設定	TF-IDF	BLEU	ROUGE			BERT Score	Entity F1 (↑)	Entity Coverage (↑)		Entity Cooccurrence (↑)				系列長
			1	2	L			完全一致	部分一致	n=0	n=1	n=2	n=∞	
タイトルあり														
RAG-Loop5	PID	3.22	30.73	9.02	19.02	84.51	41.47	24.40	70.66	7.93	8.77	8.89	8.90	121.6
RAG-Loop5	PID-QID	3.12	30.89	9.01	18.97	84.53	43.24	25.33	70.84	8.16	9.11	9.53	9.45	121.2
RAG-Loop5	QID	3.22	**31.17**	**9.11**	**19.20**	84.54	**43.85**	25.77	70.50	**8.60**	**9.61**	**9.92**	**9.79**	123.8
RAG-Validate	PID	3.15	30.55	8.92	18.90	84.53	41.89	24.62	71.75	7.89	8.43	8.46	8.41	116.3
RAG-Validate	PID-QID	3.18	30.66	8.96	18.92	84.55	42.17	25.30	71.37	8.47	8.71	9.01	8.95	118.1
RAG-Validate	QID	**3.22**	30.85	9.07	18.96	**84.56**	42.96	**25.80**	71.33	8.53	9.19	9.52	9.45	120.1
Baseline	–	1.18	24.89	5.53	15.94	83.62	16.93	11.04	**74.09**	1.76	1.55	1.53	1.44	91.3
タイトルなし														
RAG-Loop5	PID	0.54	23.92	3.95	14.88	82.61	11.29	6.87	68.71	1.05	1.20	1.27	1.24	119.6
RAG-Loop5	PID-QID	0.60	24.02	**4.06**	14.91	82.64	**11.75**	**7.29**	68.60	**1.49**	**1.68**	**1.78**	**1.73**	120.7
RAG-Loop5	QID	**0.61**	**24.18**	3.98	**14.94**	82.60	11.12	7.08	68.23	1.41	1.60	1.73	1.66	123.8
RAG-Validate	PID	0.43	22.58	3.51	14.14	82.62	8.59	5.71	71.38	0.76	0.95	0.97	0.95	102.7
RAG-Validate	PID-QID	0.43	22.51	3.52	14.08	82.61	8.63	5.59	70.88	0.77	0.75	0.78	0.78	103.8
RAG-Validate	QID	0.44	22.52	3.52	14.10	82.60	8.65	5.71	70.84	0.69	0.86	0.90	0.89	104.9
Baseline	–	0.29	20.59	3.18	13.26	**82.65**	5.95	4.21	**72.72**	0.54	0.52	0.50	0.51	82.1

知識グラフの検索　トリプルの順位付けは TF-IDF を用いて 3 種類定義する．具体的には，(1) 述語を単語とし，そのスコアを用いる手法 (PID)，(2) 隣接するエンティティを単語とし，そのスコアを用いる法 (QID)，(3) その両方を単語とし，両方のスコアの和をスコアとする手法 (PID-QID) の設定を検証した．詳細は付録 A に示す．

RAG の設定　ExpArt-KG と提案手法 (RAG-Validate) の有効性を検証するために，RAG を使用せず回答を 1 度だけ生成する手法 (Baseline)，および RAG を使用して回答生成を 5 回繰り返す手法 (RAG-Loop5) との比較を行った．

評価指標　評価には，生成文と参照説明文の語彙的および意味的類似度を評価するため，自然言語生成タスクで標準的に用いられる BLEU [19]，ROUGE [20]，BERTScore [21] を使用した．加えて，芸術作品に関する説明の詳細度を測るため，Entity Coverage，Entity F1，Entity Cooccurrence [7] を用いた．これらの数式については付録 A に記載する．

(1) **Entity Coverage** 生成文が参照説明文中の芸術作品に関するエンティティをどの程度網羅しているかを，完全一致と部分一致の 2 つの設定で評価する．

(2) **Entity F1** 生成文と参照説明文が含む芸術作品に関するエンティティを比較し，エンティティの出現頻度および適切な使用頻度を評価する．

(3) **Entity Cooccurrence** 文全体を通じたエンティティ間の共起関係に着目し，共起のカバー率を評価する．簡潔ペナルティ [19] を採用し，冗長な説明文の過大評価を防止する．

5　実験結果および分析

各設定における評価指標のスコアを表 1 に示す．BERTScore と Entity Coverage の部分一致のスコアを除いたすべての評価指標で，RAG-Validate および RAG-Loop5 の Baseline に対するスコアの向上が見られた．BERTScore は RAG の有無に依らず高い水準である一方で，Entity F1，Entity Coverage，Entity Cooccurrence において顕著な改善が見られた．これらは，Baseline が既に文法的な流暢さを備えている一方で，ExpArt-KG および提案手法の利用によって説明文に含まれるエンティティの正確性と網羅性が大幅に改善され，画像の対象とその周辺知識に対する説明文の詳細度が向上しうることを示している．TF-IDF のスコアの算出方法に関しては，隣接するエンティティのみを語の構成要素として使用する算出方法において，最も大きなスコアの向上が見られた．トリプルの述語よりもエンティティを検索に用いる方が，より固有性の高いエンティティの識別能力を向上することが示唆される．また，作品のタイトルを含む質問文に対して，RAG-Validate は

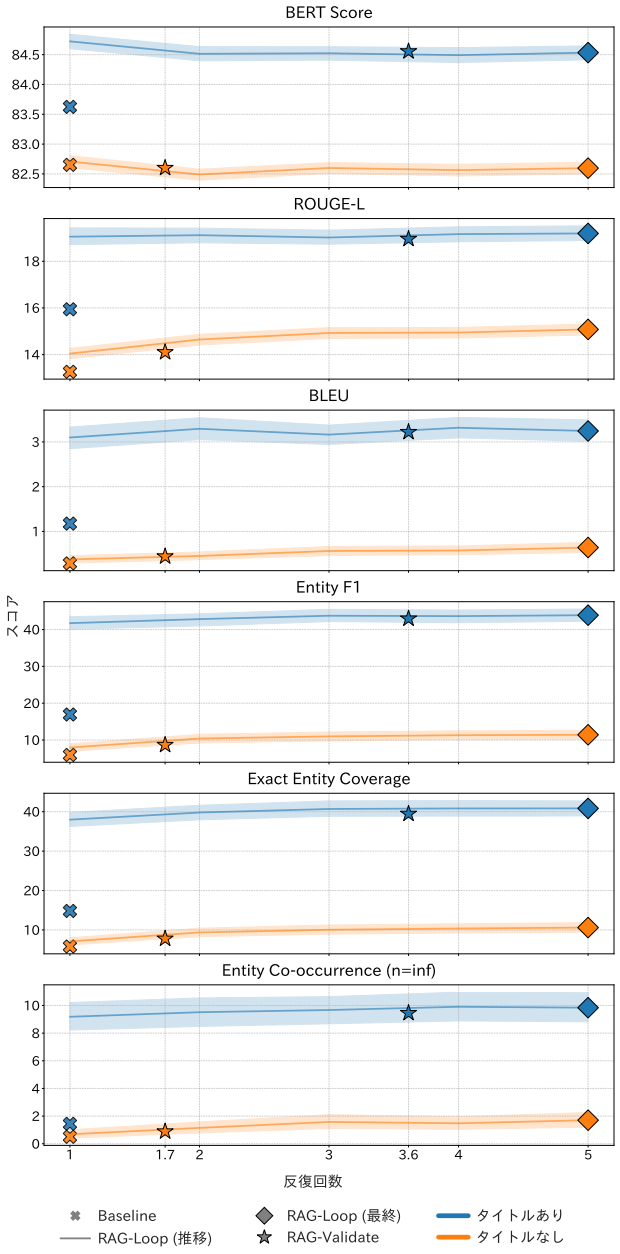

図 1　RAG-Loop5 の各試行と Baseline により得られる各スコアを示す．横軸は反復回数を表し，縦軸はスコアを表す． 実線は RAG-Loop5 のスコアの推移を表し，波線は Baseline のスコアを表す． 編みかけの領域は RAG-Loop5 のスコアの 95%信頼区間を示す．

RAG-Loop5 と同水準のスコアを記録している．

図 1 は Baseline，RAG-Loop5 および RAG-Validate における回答の反復生成回数とスコアの遷移を示す．タイトルなしの設定では，タイトル有りの設定の回答に比べて初回の回答のスコアが著しく低い値を示しているが，RAG-Loop5 による回答の反復生成によりスコアは継続的に上昇する傾向が確認された． 一方で，タイトルありの設定ではスコアは上昇するが，一定回数以降はスコアの上昇幅は減少する傾向が見られた．これは，知識グラフのノードを複数回辿ることにより，回答の生成に有用な事実情報が段階的に取得されることを示唆している．特に，初期回答が含む関連エンティティが少なく，探索の手がかりが少ない場合にもスコアの向上が確認された事実は，反復的な検索が知識グラフ上のマルチホップ推論として機能し，回答の生成に有用なエンティティを効果的に探索していることを裏付けている．また，RAG-Validate の回答の反復生成の平均回数はタイトルありの設定の場合 3.6 回であり，固定回数で反復を行う RAG-Loop5 に比べて少ない回数であるにも関わらず，回答のスコアは RAG-Loop5 と同程度の水準を維持している．この結果から，LVLM が生成した回答を検証用 LLM によって正誤判定を行い，妥当と判断された時点で反復を終了する提案手法は，生成品質を損なうことなく，一律に一定回数の反復を行う場合に生じる不要な探索を回避し，外部知識の検索コストを削減できることが示された．さらに，タイトルありの設定では RAG-Validate は RAG-Loop5 と同等のスコアを維持しているが，タイトルなしの設定では RAG-Loop5 よりも低いスコアとなった．タイトルなしの設定におけるスコアの低下は，検証器が正誤判定の基準となる知識を欠いたことで，誤った回答を適切に棄却できず，探索を早期に終了してしまったことに起因すると考えられる．これは，RAG-Validate における検証が，単なる生成文の流暢さではなく，入力されたタイトルと回答内容の事実関係の整合性に基づいた判断を行っていることを示している．

6　おわりに

画像の説明文の詳細度を向上させることを目的として，知識グラフを効果的に利用する検索拡張生成の手法を提案した．提案手法では，生成された回答の正誤判定と再生成を反復的に行い，知識グラフを動的に探索することで，説明生成に有用なエンティティを効率的に取得することを可能にした．加えて，芸術作品に関連するエンティティから構成される知識グラフである ExpArt-KG を構築した．実験結果から，提案手法は回答の妥当性に応じて反復回数を動的に制御することで，固定回数の反復生成と同程度の生成品質を維持しつつ総反復回数を削減できることが確認され，高品質な説明生成と外部知識の検索コストの削減を両立できることが示された．

謝辞

本研究は JSPS 科研費 JP23K28148, JP24K02993 の助成を受けたものです.

参考文献

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **Advances in neural information processing systems**, Vol. 36, pp. 34892–34916, 2023.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. **arXiv preprint arXiv:2502.13923**, 2025.

[3] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. **arXiv preprint arXiv:2412.08905**, 2024.

[4] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025.

[5] 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, 2023.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. **Advances in neural information processing systems**, Vol. 36, pp. 49250–49267, 2023.

[7] Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Towards artwork explanation in large-scale vision language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 705–729, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in neural information processing systems**, Vol. 33, pp. 9459–9474, 2020.

[9] Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In **ICLR 2025 Workshop on Foundation Models in the Wild**, 2025.

[10] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[11] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. **arXiv preprint arXiv:2507.06261**, 2025.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[15] Kazuki Hayashi, Hidetaka Kamigaito, Shinya Kouda, and Taro Watanabe. Iterkey: Iterative keyword generation with LLMs for enhanced retrieval augmented generation. In **Second Conference on Language Modeling**, 2025.

[16] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. **Information processing & management**, Vol. 24, No. 5, pp. 513–523, 1988.

[17] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledge base. **Communications of the ACM**, Vol. 57, pp. 78–85, 2014.

[18] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, and Chang Gao et al. Qwen3-vl technical report, 2025.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[21] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

A 付録

プロンプト 提案手法において使用するプロンプトを表 2 に示す．

表 2 提案手法で使用されるプロンプト．

用途	プロンプト
LVLM w/o KG	**System:** You are an assistant helping with visual question answering. **User:** Question: {question} Answer the question based on the provided image. Do not include any additional text other than the answer itself.
LVLM w/ KG	**System:** You are an assistant helping with visual question answering using external knowledge. **User:** Below are factual triples retrieved from a knowledge graph to help answer the question. Each triple is formatted as: [SUBJECT, PREDICATE, OBJECT] Retrieved Facts: {triplets} Question: {question} Answer the question by integrating the image, your internal knowledge, and any retrieved facts that are relevant to the question. Do not include any additional text other than the answer itself.
検証用 LLM	**System:** You are an assistant that validates answers. **User:** Question: {question} Answer: {answer} Judge whether the provided answer is correct and addresses the question. Respond with only "True" or "False". Do not provide any additional explanation or text.

使用モデルの詳細 本稿で言及したモデルの略称と，Hugging Face のモデル ID の対応を表 3 に示す．

表 3 使用モデルと Hugging Face ID の対応．

本稿での略称	Hugging Face Model ID
Qwen-VL	Qwen/Qwen3-VL-8B-Instruct
Qwen	Qwen/Qwen3-4B-Instruct-2507

データセットの統計 データセットの各処理段階における質問数を表 4 に示す．なお，各質問にはデータの仕様として，タイトルの有無による 2 通りの設定が含まれている．

表 4 データセットの統計．

内訳	質問数
ExpArt のテスト用データ	5,227
フィルタリング後	4,823
サンプリング後（評価用データ）	1,199

TF-IDF に基づくトリプル選定の詳細 本手法では，エンティティ e に接続するトリプル集合を文書 D_e とし，トリプル $T \in D_e$ の構成要素（述語 p または隣接ノード n）を語 t とみなす．語 t の重要度を表す TF-IDF 値 $w(t, D_e)$ は，文書内の頻度 $f_{t,e}$, 全エンティティ数 $N = |\mathcal{E}|$, および語 t を含むエンティティ数 $n_t = |\{e' \in \mathcal{E} : t \in D_{e'}\}|$ を用いて次式で定義される．数値計算の安定性を考慮し，対数正規化および平滑化を行う．

$$w(t, D_e) = \log(1 + f_{t,e}) \cdot \log\left(\frac{N+1}{|\{e' \in \mathcal{E} : t \in D_{e'}\}| + 1}\right) \quad (1)$$

各設定におけるトリプル T のスコア $S(T)$ は以下のように算出される．(1) **PID**: $S(T) = w(p, D_e)$, (2) **QID**: $S(T) = w(n, D_e)$, (3) **PID-QID**: $S(T) = w(p, D_e) + w(n, D_e)$．このスコアに基づき，各エンティティにつき上位のトリプルを選定する．

エンティティ評価指標の定義 生成された n 文からなる説明を $G = \{g_1, \ldots, g_n\}$, 参照説明を $R = \{r_1, \ldots, r_m\}$ とする．テキストから抽出されたエンティティの集合を $E_G = Entity(G)$, $E_R = Entity(R)$ とおく．また，$|G|, |R|$ をそれぞれの全トークン数とする．冗長な生成文に対するペナルティ項 $BP(G, R)$ を次式で定義する．

$$BP(G, R) = \exp\left(-\max\left(0.0, \frac{|G|}{|R|} - 1\right)\right) \quad (2)$$

各評価指標は以下のように算出される．

(1) **Entity Coverage (EC)**: R に含まれるエンティティが G にカバーされた割合である．$Cov(G, R)$ をエンティティのカバー率を返す関数とし，次式で計算する．部分一致のカバー率の算出には最長共通部分列を用いる．

$$EC(G, R) = Cov(G, R) \quad (3)$$

(2) **Entity F1 (EF1)**: エンティティの出現頻度に基づく適合率 P と再現率 R_{ecall} の調和平均である．$\#(e, \cdot)$ を対象の文におけるエンティティ e の出現回数とし，$\mathrm{Count}_{\mathrm{clip}}(e, G, R)$ を G または R における少ない方の出現頻度とすると，次式で定義される．

$$EF_1 = \frac{2 \times P \times R_{ecall}}{P + R_{ecall}} \quad (4)$$

$$P = \frac{\sum_{e \in Entity(G)} \mathrm{Count}_{\mathrm{clip}}(e, G, R)}{\sum_{e \in Entity(G)} \#(e, G)} \quad (5)$$

$$R_{ecall} = \frac{\sum_{e \in Entity(R)} \mathrm{Count}_{\mathrm{clip}}(e, G, R)}{\sum_{e \in Entity(R)} \#(e, R)} \quad (6)$$

(3) **Entity Cooccurrence (ECooc)**: 文とその前後 n 文のコンテキスト窓において共起したエンティティの組を返す関数を $Co(\cdot)$ とし，次式で計算する．文への分割には nltk の文分割器を用いた．

$$ECooc(G, R) = BP(G, R) \times Cov(Co(G), Co(R)) \quad (7)$$