Large-scale Vision Language Model による芸術作品に対する説明の生成

林和樹 [1] 坂井優介 [1] 上垣外英剛 [1] 林克彦 [2] 渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 北海道大学
{hayashi.kazuki.hl4, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

Large-scale vision-language models（LVLMs）は，ユーザーから入力された画像と指示に基づき文章を生成する巨大言語モデルである．これらのモデルを画像の創作支援に利用する際には，画像の構図や工夫，他作品との比較，歴史的背景，深い芸術的な知識に基づく説明の生成が要求される．しかし LVLM が芸術作品の説明に必要な知識，及び複雑な知識間の関連をどの程度理解し，それらを統合して説明に応用できるかは，明らかにされていない．本研究では，芸術作品に関する深い知識の理解と利用を定量的に評価するための新たなタスク，データセット及び評価尺度を提案し，さらに LVLM が芸術に関する知識を伴う説明を学習するための訓練データセットも公開する．このタスクは，芸術作品の画像とタイトルからの説明生成，及び画像のみを使用した説明生成の二つで構成され，LVLM の言語に基づく知識と視覚に基づく知識の両方を評価する．検証の結果，LVLM は元の LLM と同等の言語に基づく知識を保持しているものの，視覚に基づく知識の獲得は限定的であることが判明した．

1 はじめに

Vision & Language (V&L) では，大規模言語モデル（LLMs；Large Language models) [1, 2, 3] に，視覚エンコーダを組み合わせ学習し，視覚と言語を統合した Large scale vision language models (LVLMs) [4, 5, 6, 7] が V & L のベンチマーク [8, 9, 10, 11] において成功を収めている．これらのモデルは，図 1 にあるように絵画や写真などの創作支援に応用された際、芸術作品の主題，その歴史的な背景，関連する他の作品や芸術の流れなどの知識を体系的に活用して説明を生成する必要があり，知識を単に個別に認識するだけでは不十分である．このように創作支援において

図 1 LVLM を用いた創作支援の具体例

は芸術に関する知識の相互関係を深く理解し，それらを総合的に説明に結びつける能力が重要となる．

既存研究として VQA[12] の枠組みで芸術作品を対象とした知識を問うタスク [13, 14, ?] が存在するが，これらはいずれも独立した知識を扱うものに限られ，知識を組み合わせて説明を生成する能力を評価するための新たなタスクと評価指標が求められる．

上記の問題を解決するために本研究では，LVLM の芸術作品に関する説明生成能力を測定する新しいタスクと評価尺度を提案する．提案タスクにおいて LVLM は入力された芸術作品の画像とそのタイトルに対し，与えられた指示に従う説明を生成しなければならない．我々は提案タスクを遂行するために英語版 Wikipedia の artwork に該当する infobox を持つ記事 1 万件からデータセットを構築し，LVLM が芸術的な知識を伴う説明を学習するための訓練データセットも公開する．さらに，現在様々な V & L のベンチマークにおいて最高精度を達成している 3 つの LVLM を用いて検証を行った．検証の結果，LVLM は元となった LLM が保持する芸術に関する知識を保持していることが判明した．その一方で，芸術に

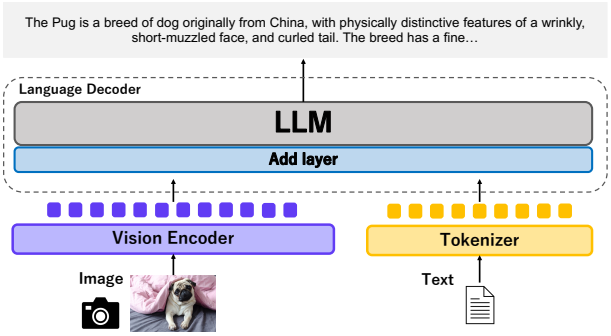

図 2 LVLM の構成．Vision Encoder と LLM を追加学習により統合する．

関する知識が視覚から与えられる情報と適切に対応づいていないということも明らかとなった．

2 LVLMs

本研究の対象となる LVLMs [4, 5, 6, 7] は図 2 にあるように対照学習によって学習された視覚に関する情報を司る Vision Encoder [4] と既存の LLM [1, 2, 3] を結合して視覚と言語の統合を目的とした追加学習を行うことで実現されている．これにより、パラメータが十倍以上の事前学習モデル [15, 16] と比較しても優れた性能が得られ，大規模なデータセットによる事前学習の必要性が再考されることとなった．一方でこの結合は部分的なネットワークの追加によって行われるため，LLM で獲得された知識と Vision Encoder が獲得された知識とが適切に対応づけられているかは定かではない．特に本研究が対象とする芸術に関する知識を伴う説明の生成には Vision Encoder 中の知識と LLM 中の知識を体系的に対応づけて運用する必要があり，既存の LVLMs では対処が難しいことが予想される．

3 提案タスクと評価指標

本節では芸術作品に対する LVLM の説明文生成能力を検証するタスクの設定 (§3.1) とそれに付随して提案する評価尺度 (§3.2) について説明する．

3.1 提案タスク

提案タスクにおいて LVLM は画像やタイトルを参照しながら与えられた指示に従う説明を生成しなければならない．表 1 に使用する指示文の例を示す．この例のように各指示文は 3 種類の詳細度 (Section, Subsection, Sub subsection) に分けられ，これらは指示文と対応する説明が Wikipedia 記事中のどの階層から抽出されたかにより決定される．提案タスクは

表 1 提案タスクにおける指示文の例．青は指示文に対応する説明が抽出された記事のタイトルを，赤はその説明に対応する記事中の節名を表す．

Type	Prompt
Section	Explain the History of this artwork, **Mona Lisa**.
Subsection	Explain the Creation and date about the History of this artwork, **Mona Lisa**.
Sub subsection	Explain the Creation about the Creation and date of the History in this artwork, **Mona Lisa**.

このようなプロンプトを用いてタイトルの有無により異なる次の設定を扱う:

タイトルあり LVLM が創作支援をする際に，タイトルにはその作品に対する筆者の意図が含まれており，その意図も考慮しながら説明を生成することが望ましい．この設定では画像と共にそのタイトルを入力とすることで，LVLM が言語に基づく情報と視覚に基づく情報を踏まえて適切な説明の生成が可能であるかを検証する．

タイトルなし 図 1 にあるように，制作過程などではタイトルが存在しないケースが存在する．そのような場合でも，画像からの視覚に基づく情報のみを用いて適切な生成が可能であるか検証する設定である．またタイトルの有無に基づく精度の差分を通じて、LVLM の純粋な視覚に基づく知識を分析することが可能である．

上記に加え，LVLM の汎化性能をより詳細に検証するため: (1) 訓練に含まれた画像を対象にした評価 (**seen**); (2) 訓練に含まれない画像を対象とした評価 (**unseen**); の両設定についての比較も行う．

3.2 評価尺度

評価にはモデルが生成した文章が参照とする説明にどれほど近いかを測るため，自然言語生成タスクの評価に広く用いられている BLUE，ROUGE，BERTScore [17] を利用する．さらに芸術作品の説明生成能力に焦点を絞り評価を行うために次の三つの評価尺度を提案する[1]:

Entity Coverage 生成された文章が参照説明中の芸術作品に関するエンティティをどれだけ正確に含んでいるかを、完全一致と部分一致 [18] の二つの設定で評価する．さらに、BLEU スコアで使用されるブレビティペナルティ [19] を適用し、適切な長さでの知識の正確性を検証する．

1) 各評価尺度の数式については付録 A に記載．

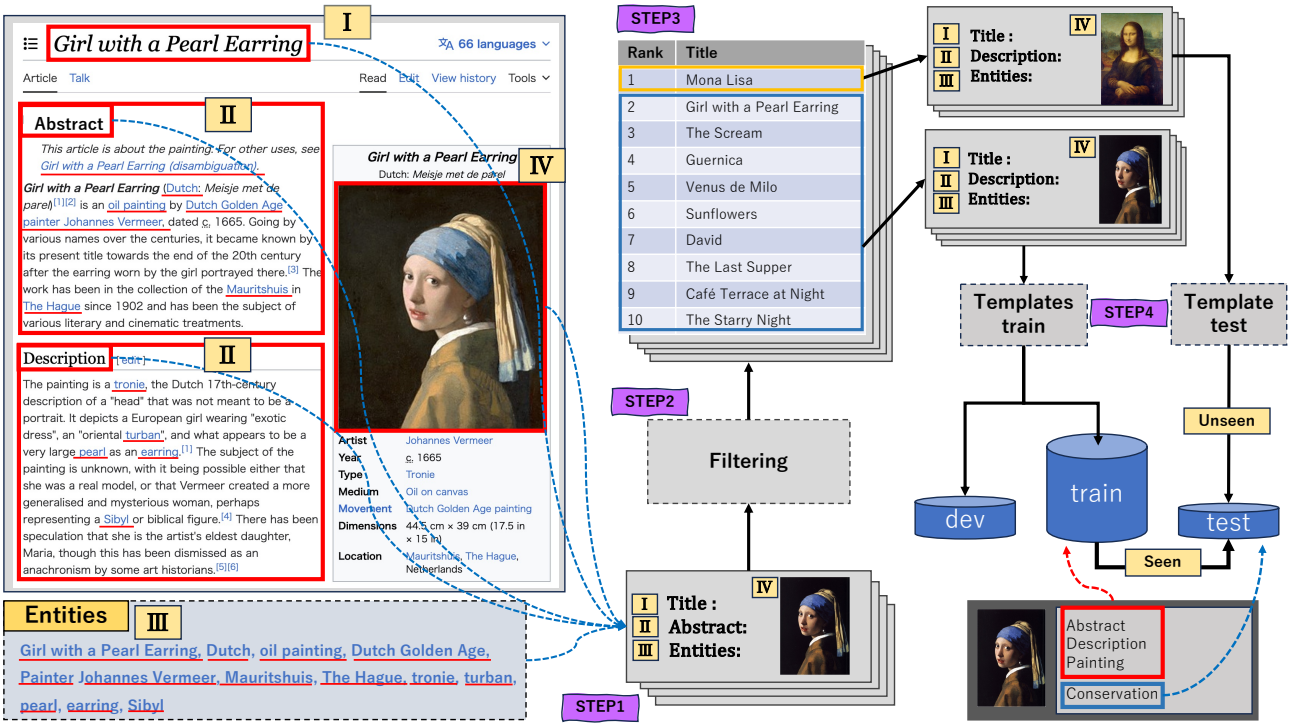

図 3　データ作成の概要図.

Entity F1　生成された説明と参照とする説明中に含まれる芸術作品に関するエンティティの出現頻度を評価する．ROUGE を参考にして生成された説明または参照とする説明中に含まれるエンティティのいずれかの出現頻度のうち最大のものを出現頻度の上限とし，再現率と適合率を計算することで、エンティティの適切な使用頻度を評価する．

Entity Cooccurrence　この尺度は，独立したエンティティのカバー率だけでなく，それらの間の関連性を文脈的にどのように組み合わせて全体の説明を形成しているかを評価する．具体的には、ある文のエンティティと前後 n 文中で出現するエンティティとを共起したとみなし，その共起のカバー率を評価することで，モデルが知識の関連性をどの程度理解し統合しているかを明らかにする．n の値を生成された説明中の文の数以上に設定することで，文章全体にわたるエンティティのペアの共起を考慮することが可能になる．なお，説明が長くなることによるカバー率の上昇を防ぐために，Entity Coverage と同様にブレビティペナルティを導入した．

4　データセットの作成

図 3 にデータ作成の概要を示す．データ作成の各ステップは次のように行われる:

STEP1　英語版 Wikipedia の artwork に該当する infobox を持つ記事全て（約 10,000）を収集し，節ごとに分割して説明文を作成した．さらに記事中でハイパーリンクになっている文字列を芸術に関するエンティティとして抽出した．各説明文は対応するタイトル，含まれていた節の階層情報 (Section, Subsection, Sub subsection)，画像，上記のエンティティの四つの情報を伴う．

STEP2　芸術に関する説明とは無関係な節，画像のない記事，固有の芸術作品でない記事に含まれる説明を除外した．

STEP3　芸術作品が LVLM の学習データに含まれるなどの知名度によって生じる偏りを防ぐため，データの並び替えを実施した．まず，ページ閲覧数，リンク数，編集回数，参照数，言語版数，記事の長さの六つの指標でデータをランク付けし，これらの平均ランキングが維持されるよう，テスト，開発，トレーニングデータを 1:1:8 の比率で均等に分割した．

STEP4　分割されたデータをさらに第 3.1 節で紹介したような指示文に変換した．変換に際して訓練データには多様性を確保するため，7 つの異なるテンプレートを用意した．

表 2 LVLM におけるタイトルあり設定とタイトルなし設定の検証結果

	評価設定	BLUE	ROUGE			BertScore	Entity Cov.		Entity F1	Entity Cooccurrence			
			1	2	L		exact	partial		n=0	n=1	n=2	n=∞
***With Title* (言語情報＋視覚情報)**													
mPLUG-Owl2	Unseen	0.97	26.2	5.60	16.5	83.2	11.6	19.0	12.6	1.46	1.21	1.12	1.08
LLaVA-1.5	Unseen	1.52	20.6	5.13	12.8	81.5	13.2	18.9	14.3	1.81	1.74	1.63	1.55
Qwen-VL	Unseen	1.53	27.2	6.38	16.7	83.2	15.3	23.2	16.9	1.71	1.46	1.41	1.35
Qwen-VL (FT)	Unseen	3.46	25.5	10.2	20.1	84.4	17.5	24.4	19.2	4.16	3.85	3.79	3.77
mPLUG-Owl2	Seen	1.10	26.3	5.78	16.9	83.2	11.5	19.0	12.9	1.33	1.23	1.23	1.21
LLaVA-1.5	Seen	1.62	20.1	5.12	12.9	81.4	11.6	17.4	13.2	1.50	1.23	1.21	1.19
Qwen-VL	Seen	1.55	27.8	6.64	17.2	83.4	15.2	23.4	17.1	1.40	1.26	1.29	1.24
Qwen-VL (FT)	Seen	3.95	27.3	11.3	21.6	84.9	19.2	26.2	21.6	5.27	4.83	4.63	4.50
***Without Title* (視覚情報)**													
mPLUG-Owl	Unseen	0.20	22.8	3.55	14.6	82.3	2.88	8.9	2.42	0.29	0.28	0.27	0.25
LLaVA-1.5	Unseen	0.18	18.0	2.69	11.8	81.2	2.12	7.04	1.49	0.10	0.12	0.13	0.13
Qwen-VL	Unseen	0.44	24.0	4.36	14.9	82.4	6.15	12.6	5.91	0.57	0.64	0.60	0.57
Qwen-VL (FT)	Unseen	1.87	23.6	7.56	18.0	83.9	12.6	17.7	12.7	2.09	1.93	1.82	1.83
mPLUG-Owl	Seen	0.14	22.6	3.39	14.7	82.2	1.68	7.78	1.46	0.09	0.11	0.11	0.11
LLaVA-1.5	Seen	0.14	17.2	2.48	11.4	81.2	0.96	5.90	0.74	0.09	0.07	0.07	0.05
Qwen-VL	Seen	0.38	24.4	4.29	15.2	82.4	4.57	11.2	4.45	0.22	0.25	0.25	0.25
Qwen-VL (FT)	Seen	1.89	24.3	7.81	18.5	84.1	12.7	18.3	13.4	2.05	1.96	1.87	1.82

表 3 LLM における検証結果 (unseen) なお，LLM は視覚情報を扱わないためタイトルあり設定にて実施.

	Entity Cov.		Entity F1	Entity Cooccurrence			
	exact	partial		n=0	n=1	n=2	n=∞
Llama2 [1]	16.7	24.6	18.9	2.83	2.39	2.21	2.13
Vicuna [2]	17.4	25.3	19.3	2.77	27.7	26.2	25.2
Qwen [3]	6.14	20.3	18.9	2.83	2.39	2.21	2.13

5 実験

設定 検証には mPLUG-Owl2 [7], LLaVA-1.5 [5], Qwen-VL [6] の三つのモデル，および Qwen-VL を本研究で作成したデータで追加学習したモデル Qwen-VL (FT) を使用した．実験設定の詳細については付録 D に記載した．

結果 表 2 に結果を示す．BLUE, ROUGE, BERTScore, Entity Coverage においては，Qwen-VL が一貫して最も高い性能を達成している．モデル間やタスク設定による大きな差が見られない．このことより本タスクにおける提案尺度の有効性を読み取ることができる．また Qwen-VL (FT) が全ての尺度で最高精度を達成していることから我々が作成したデータの有効性が確認された．さらに、タイトルの有無を比較した場合、全てのモデルでタイトルが無い場合の方が精度が低く，LVLMs は画像情報のみから知識を活用することが難しく，テキストから獲得される知識に頼っていることが示唆される．

この点について我々はさらなる検証を実施した．表 3 は各 LVLM の元となる LLM においてタイトル有り設定で説明を生成した際の結果を示す。これにより、視覚と言語の統合学習によってどの程度 LLM の知識が継承されているかを分析することができる。表 3 からは、芸術作品に関する知識において、Llama2 と Vicuna の性能が良く、Qwen の性能が低いことが分かる．本研究の提案する尺度に注目して LLM と LVLM を比較すると，mPLUG-Owl2 と LLaVA-1.5 では視覚と言語の統合学習によって芸術に関する知識が損なわれていることが分かる．一方で，Qwen-VL では統合によりタイトル有り設定で 10 % も性能が向上しており，視覚と言語の間の知識の統合がうまく行われていることが分かる．これは，Qwen-VL がタイトル無し設定で他モデルより優れていることとも符合する．

6 まとめ

本研究では，芸術作品に関する深い知識の理解と利用を定量的に評価するためのタスクとデータセット及び評価尺度を提案した．芸術作品の画像とタイトルからの説明生成，及び画像のみからの説明生成の二つで構成される提案タスクにより LVLM を検証した結果，LVLM は元の LLM が持つ芸術に関する知識を保持しているが，視覚に基づく芸術に関する知識の獲得は限定的であることが判明した．

謝辞

本研究は JSPS 科研費 JP23H03458 の助成を受けたものです.

参考文献

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023. arXiv:2309.16609.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. 2023. arXiv:2309.16609.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. **arXiv preprint arXiv:2301.12597**, 2023.

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. **arXiv preprint arXiv:2310.03744**, 2023.

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. **arXiv preprint arXiv:2308.12966**, 2023.

[7] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. **arXiv preprint arXiv:2311.04257**, 2023.

[8] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models, 2023.

[9] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. **arXiv preprint arXiv:2307.16125**, 2023.

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. **arXiv preprint arXiv:2306.13394**, 2023.

[11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? **arXiv preprint arXiv:2307.06281**, 2023.

[12] Ahmed Frikha Yezi Yang Denis Krompass Gengyuan Zhang Jindong Gu Volker Tresp Yao Zhang, Haokun Chen. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering. Vol. 2211.10567, , 2022. Computer Vision and Pattern Recognition (cs.CV).

[13] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In **Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II**, Vol. 16, pp. 92–108. Springer, 2020.

[14] Eva Cetinic. Towards generating and evaluating iconographic image captions of artworks. **J. Imaging**, 2021. To be published in Proceedings of the European Conference in Computer Vision Workshops 2020.

[15] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 23716–23736, 2022.

[16] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. **arXiv preprint arXiv:2303.03378**, 2023.

[17] 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.

[18] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. MultiSpanQA: A Dataset for Multi-Span Question Answering. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1250–1260, 2022.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

A　評価尺度の数式

下記に §3.2 で提案した評価尺度の詳細について数式を用いて説明する．なお，下記説明では生成された n 文からなる説明を $G = \{g_1, \cdots, g_n\}$，m 文からなる参照説明を $R = \{r_1, \cdots, r_n\}$ とする．また入力されたテキストに含まれるエンティティを抽出して返す関数を $Entity(\cdot)$ として定義する．なお，$|G|$ は生成された説明に含まれる全トークン数を，$|R|$ は参照説明に含まれる全トークン数をそれぞれ表す．

Entity Coverage　(EC) は次にように計算される:

$$EC(G,R) = BP(G,R) \times Cov(G,R) \tag{1}$$

$$BP(G,R) = \exp(\max(0.0, |G|/|R| - 1)) \tag{2}$$

ここで $Cov(G,R)$ は R 中のエンティティが G により カバーされた割合を返す関数である．なお部分一致を行う際には Lowest Common Subsequence (LCS) を用いて参照エンティティの長さに対して生成された説明中で一致する最長の長さの割合を一致率とした．

Entity F1　(EF_1) は次のように計算される:

$$EF_1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

$$P = \frac{\sum_{e_i \in G} \mathrm{Count}_{\mathrm{clip}}(e_i, G, R)}{\sum_{e_j \in G} \#(e_j, G)} \tag{4}$$

$$R = \frac{\sum_{e_i \in R} \mathrm{Count}_{\mathrm{clip}}(e_i, G, R)}{\sum_{e_j \in R} \#(e_j, R)}, \tag{5}$$

ここで $\#(e_j, G), \#(e_j, R)$ はそれぞれエンティティ e_j が G 及び R において何回出現したかを表す関数であり，$\mathrm{Count}_{\mathrm{clip}}(e_i, G, R)$ は e_i の G または R における少ない方の出現頻度を返す関数である．

Entity Cooccurrence　(ECooc) は式 (2) の BP を用いて次のように計算される:

$$ECooc(G,R) = BP(G,R) \times Cov(Co(G), Co(R)), \tag{6}$$

ここで関数 $Co(\cdot)$ はある文とその前後 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みを返却する．文への分割には nltk の文分割器を用いた．

B　タイトル生成の精度

	mPlug_Owl2	LLaVa-1.5	Qwen-VL
タイトル数	15/962	0/962	9/962

表 4　image 情報からの title 生成の精度

本研究で提案するタスクでは，タイトルあり設定とタイトルなし設定の差を分析の基盤としている．そのため，どちらの設定においても，エンティティに芸術作品のタイトルが含まれないようにしている．しかし、タイトルは芸術作品において最も直感的な知識の一つである．したがって，芸術作品の画像とプロンプトに”Please answer the title of this artwork” と与えた際に，モデルがどの程度純粋な画像情報のみからその芸術作品のタイトルを生成できるかの追加検証を行った．

表 4 のように、画像情報からタイトルを生成することは可能であるものの、テストセットにおける生成精度は約 1 %前後にとどまっている．これは、LVLMs（Language-Vision Latent Models）において画像と知識の結びつきが弱いことを示唆している．この結果から，改めて LVLMs が画像から知識生成する過程において、まだ改善の余地があると考えられる．

C　作成したデータセットの詳細

	訓練	開発	テスト (seen)	テスト (unseen)
画像数	7696	962	2167	962
データ数	134890	2906	2577	2806

表 5　作成データセットの画像とデータの数

D　実験設定の詳細

mPlug-Owl2，Llava-v-1.5，Qwen-VL は性能比較が実装の違いにより不公平にならないように，実験は全て RTX 6000 Ada の 1GPU で生成させた．生成 token を統一させるために max_token_length を 1024 で統一した．Fine tunning は RTX 6000 Ada を用いて，ハイパラメータは公開されている

表 6　実験に使用したハイパーパラメータ．記載していないパラメータについては初期設定を用いている．[4, 5, 6, 7]

Hyper Parameter	値
lora_alpha	256
lora_dropout	0.1
bottleneck_r	64
batch size	4
epoch	1
torch_dtype	bf16
lr_scheduler_type	cosine
learning_rate	1e-5
seed	42
model_max_length	2048