視覚的文脈を利用した視覚言語モデルによる画像キャプション生成自動評価手法

前田 航希 ♣,♠ 栗田 修平 ♠ 宮西 大樹 ◇,♠ 岡崎 直観 ♣
♣ 東京工業大学, ♠ 理化学研究所, ◇ 国際電気通信基礎技術研究所
koki.maeda@nlp.c.titech.ac.jp, shuhei.kurita@riken.jp
miyanishi@atr.jp, okazaki@c.titech.ac.jp

概要

既存の画像キャプション生成の品質自動評価は，画像に映る物体への言及を過大評価するほか，物体の特性や関係といった細かい言及の間違いを見過ごすなど，人間の判断に近い評価をするには未だ不十分である．本稿では，視覚言語モデル（VLM）を用いた画像キャプション生成の自動評価手法として VisCE^2 を提案する．本手法は，VLM が持つ視覚理解能力によって，画像に写る物体の属性や関係を含めた視覚的文脈を明示的に言語化し，VLM の言語理解能力によって視覚的文脈を考慮した評価を可能にする．メタ評価実験を通じて，VisCE^2 は既存手法と同等以上の人手評価との相関を示し，視覚的文脈の追加が評価性能を向上させることを実証した．

1 はじめに

画像キャプション生成の自動評価は，視覚的な情報を言語化する過程での正確さや自然さを効率的に判断する上で重要な役割を果たす．これまでに，事前に用意した参照文と生成したキャプションの *n*-gram の重複を計測する手法 [1, 2]，視覚言語モデルから取得した画像とテキストの分散表現の類似度を計測する手法 [3] などが提案され，画像キャプション生成の自動評価の品質は大きく向上したものの，未だに人手評価の品質と比べて大きな差が存在している．また，CLIP-S [3] 視覚言語基盤モデルを用いた評価指標は，物体の属性や物体間の関係の誤りを見逃す傾向があると指摘されている [4]．

この問題を解決するために，我々は評価プロセスに**視覚的文脈**の抽出を導入する．視覚的文脈は，背景や細かな物体，および推測される事実を含む，画像を詳述する周辺情報である．画像から視覚的文脈を抽出して評価に用いることで，画像内の物体に加え，その状態や相互作用，意図を正確に記述したかを識別し，より包括的な評価が可能になる．

本研究では，視覚的文脈を活用した視覚言語モデルベースの画像キャプション生成の自動評価手法（**Vis**ion Language Model-based **C**aption **E**valuation Method leveraging **Vis**ual **C**ontext **E**xtraction; **VisCE**^{2}）を提案する．VisCE^2 は，VLM を用いて画像内の物体とその相互関係からなる視覚的文脈を明示的に言語化し，画像と言語化した視覚的文脈をもとに，生成したキャプション候補のスコアを計算することで，自動評価を行う．VLM の視覚理解能力と言語理解能力を用いて，視覚的文脈を考慮しながら言語および視覚の両観点から画像キャプション生成の評価を行うことで，評価品質の向上が期待できる．

我々は人手評価を用いたメタ評価実験を行い，**(i)VisCE^2 は人手評価と高い相関を示し，既存の手法と同等以上の評価性能を持つこと (ii) 評価時に画像の視覚的文脈を構成的な表現で与えることが評価性能の向上につながること**を実証した [1)]．

2 関連研究

2.1 画像キャプション生成の自動評価手法

画像キャプション生成の一般的な品質自動評価では，事前に用意した参照文と生成したキャプションの一致度を複数の指標 [1, 2, 5, 6, 7] で計測して比較を行う．しかし，参照文を用意するコストが大きく，人間の評価との相関が低いことが知られている．この問題を解決するために，参照文を用いずに自動評価する試みが為されている．CLIP-S [3] は，視覚と言語の基盤モデル CLIP [8] を用いて画像とテキストの埋め込み表現の cos 類似度を評価スコアとみなす．また，CLIP-S の訓練事例を拡張した

1) 本研究の実装は以下の URL で公開する予定である: https://github.com/nlp-titech/VLM_caption_eval

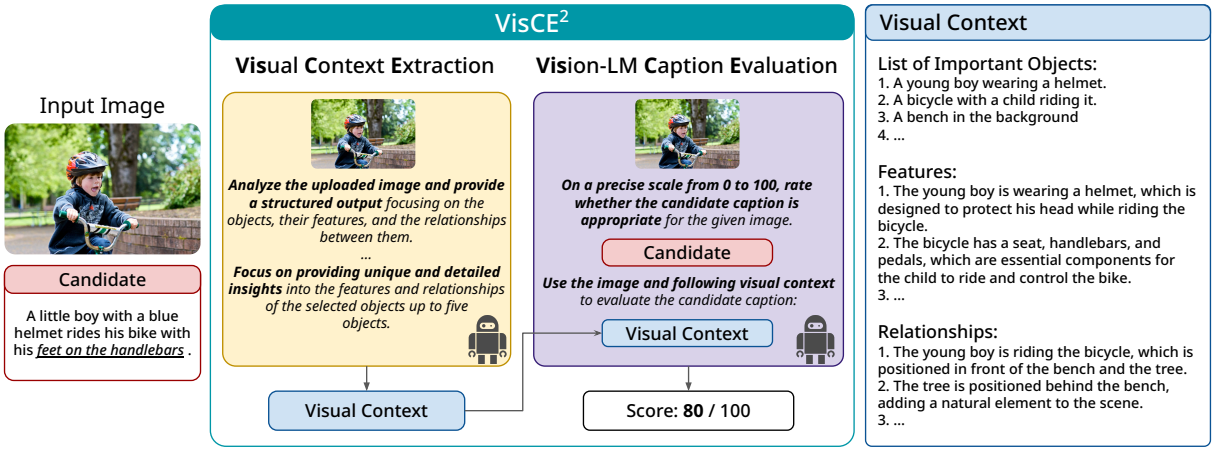

図 1　VisCE^2 によるキャプション品質自動評価の概要と入出力の例．まず VLM は画像から視覚的文脈を抽出し，テキストとして出力する．その後，画像および候補文に加えて得られた視覚的文脈を用いてキャプションを評価する．

PAC-S [9] や，生成文の正確性と画像に映る物体の網羅性を評価した InfoMetIC [10] などが提案された．CLAIR [11] は ChatGPT [12], Claude [13], PaLM [14] といった大規模言語モデルを画像キャプション生成の評価に活用し，人手評価との高い相関を示した．しかし，意味的な類似度を評価するための参照文作成のコストがかかるうえ，モデルが公開されておらず評価の透明性も低い．提案手法は，VLM が画像の視覚的文脈を明示的に出力して自身に与えることで，参照文付与のコストを低減し，周辺情報を考慮したきめ細かい評価ができる．

2.2　視覚言語モデル（VLM）

異なるモダリティの統合的な理解を目指して，視覚と言語の両潜在空間の距離学習を行う事前学習手法である CLIP [8] が提案された．その後，複数の視覚言語課題を同一のモデルで解くことを目的として，OFA [15], InstructBLIP [16] などが提案されてきた．言語モデルの発展と並行して提案された LLaVA [17] は，言語課題において高い性能を発揮した LLaMA [18] と Vicuna [19] を CLIP と統合し，微調整を行うことで多くのマルチモーダル課題での Zero-shot 性能を飛躍的に向上させた．また，モデルは非公開だが API を通じて利用可能な VLM として，GPT-4V [20] や Gemini [21] などがある．本研究では，結果の再現性の観点から，公開されたモデルの中で最も性能が高い LLaVA-1.5 [22] を採用した．

3　提案手法: VisCE^2

VisCE^2 は，画像キャプション生成の品質評価をマルチモーダルなテキスト補完課題とみなす．本手法では，画像，候補文，およびキャプション評価用に設計したプロンプトを入力として，VLM が続くテキストを補完し，キャプションの品質を示すスコアを含む文を生成する．

図 1 に VisCE^2 の概要を示す．VLM を用いた評価に際し，評価の基準となる参照情報を画像・テキストの両方で与えるために，VisCE^2 ではキャプションの品質を評価するプロセスを 2 段階に分ける．第一段階では，VLM は画像の視覚的文脈を抽出し，構造的に出力する．具体的には，VLM に画像を与え，画像に存在する物体，物体の特徴，および物体同士に成り立つ関係に分類し，それぞれ列挙する．これにより，物体に関するより詳細な情報が保存され，評価に活用することができる．また，画像だけでは第二段階では，第一段階で得られた視覚的文脈を用いてキャプションの評価を行う．画像，候補文と視覚的文脈をプロンプトに統合し，VLM に与える．キャプションの総合的な品質を 0 以上 100 以下の整数値で出力するよう VLM に指示し，キャプションの品質を反映したスコアを出力させる．

4　実験

本稿で提案した自動評価手法のメタ評価を行うため，VisCE^2 と画像キャプションへの人手評価がどの程度整合するかを検証した．評価に用いたデータセットの詳細は付録 A に詳述する．また，視覚的文脈の有効性を検証するために複数のプロンプトによる評価性能の比較を行った．VisCE^2 は利用する VLM に制限はないが，結果の再現性と透明性の観点から公開されている VLM の中で最も性能が高い LLaVA1.5-13B を採用した．

表 1 Flickr8k-Expert [23]，Composite [24] における人手評価との相関．**Ref.**: 参照文の利用の有無，**太字**: 最良スコア，†: 先行研究で報告されたスコアを示す．

Method	Ref.	Flickr8k-Expert	Composite
BLEU-4 [1]	✓	30.6	28.3
ROUGE [5]	✓	32.1	30.0
METEOR [6]	✓	41.5	36.0
CIDEr [2]	✓	43.6	34.9
SPICE [7]	✓	51.7	38.8
RefCLIP-S [3]	✓	52.6	51.2
†RefPAC-S [9]	✓	55.5	51.5
†CLAIR$_E$ [11]	✓	62.7	59.2
CLIP-S [3]	✗	51.1	49.8
†PAC-S [9]	✗	53.9	51.5
†InfoMetIC [10]	✗	54.2	**59.2**
VisCE^2	✗	**59.0**	55.0

4.1 人手評価との相関

我々は Flickr8k-Expert [23], Composite [24] を用いて，VisCE^2 のキャプション品質評価能力を検証した．自動評価と人手評価の相関を測るメタ評価指標には，Kendall の順位相関係数を用いた．表 1 は，VisCE^2 の人手評価との相関を既存の自動評価指標 [1, 2, 3, 5, 6, 7, 9, 10, 11] と比較したものである．

Flickr8k-Expert．VisCE^2 は参照文を用いない手法の中で最高性能を達成した．また，公開されたモデルを利用したものに限れば，参照文を用いた手法を含めた中で最も高い評価性能であった．これまでの最高性能であった InfoMetIC に大きく差をつけて上回り (+4.8pt)，従来の自動評価指標と比較して，提案手法は人間に沿ったキャプション品質評価ができることが示された．

Composite．VisCE^2 は高い水準の評価性能を示した．CLIP-S, PAC-S と比較して 3pt. 以上の性能向上が見られ，画像キャプション生成の自動評価に対する VLM の有効性が実証された．提案手法は InfoMetIC の性能を下回ったが，InfoMetIC は Composite を構成する 3 種類のデータセットで品質推定器を微調整しており，これが性能差の一因として考えられる．

一方で，VisCE^2 は大規模言語モデルと参照文を用いる CLAIR と比較して低い結果となった．利用する言語モデルの規模の差を鑑みると，性能の差に寄与する要素の特定にはさらなる検証が必要である．

4.2 キャプション対の比較

我々は，キャプション対の選好判断を収集した人手評価データセット Pascal-50S [2] を用いて評価指標の相対的な評価の性能を検証した．候補文の組は以下の 4 つのグループに分類されており，グループごとの人手評価との一致度を計測した．表中の記号はキャプションの組の性質を表し，それぞれ以下の内容を表す．

1. HC: 人間が書いたキャプションどうしの組
2. HI: 人間が書いた，一方は正しく，もう一方は誤った画像に付与されたキャプションの組
3. HM: 人間が書いたものと生成キャプションの組
4. MM: 自動生成されたキャプションどうしの組

表 2 の判定精度の結果から，VisCE^2 は CLIP-S と同等程度だが，これまでの参照文を用いない自動評価指標と比較して低い性能を呈した．CLAIR も同様の傾向を示しており，言語モデルに依存する手法に共通する欠点だと考えられる．

表 2 Pascal-50S [2] における判定精度．

Method	Ref.	PASCAL-50S				
		HC	HI	HM	MM	Mean
BLEU-4 [1]	✓	53.0	92.4	86.7	59.4	72.8
ROUGE [5]	✓	51.5	94.5	92.5	57.7	74.0
METEOR [6]	✓	56.7	97.6	94.2	63.4	77.9
CIDEr [2]	✓	53.0	98.0	91.5	64.5	76.7
SPICE [7]	✓	52.6	93.9	83.6	48.1	69.5
RefCLIP-S [3]	✓	64.9	99.5	95.5	73.3	83.3
†RefPAC-S [9]	✓	**67.7**	99.6	**96.0**	**75.6**	**84.7**
†CLAIR$_E$ [11]	✓	57.7	**99.8**	94.6	**75.6**	81.9
CLIP-S [3]	✗	55.9	99.3	96.5	72.0	80.9
†PAC-S [9]	✗	60.6	99.3	**96.9**	72.9	82.4
†InfoMetIC [10]	✗	**69.0**	**99.8**	94.0	**78.3**	**85.3**
VisCE^2	✗	60.7	99.6	93.6	69.3	80.8

表 3 THumB 1.0 における自動評価手法と人手評価との Pearson 相関係数．w/o, w/ Human はそれぞれ人間が書いたキャプションを候補文に含めるかを示している．

Method	Ref.	w/o human			w/ human		
		P	R	Total	P	R	Total
BLEU-4 [1]	✓	.21	.13	.25	.15	.04	.13
ROUGE [5]	✓	.26	.17	.31	.18	.07	.18
CIDEr [2]	✓	.27	.18	.33	.21	.11	.23
SPICE [7]	✓	.26	.15	.30	.20	.09	.21
RefCLIP-S [3]	✓	.34	.27	.44	.31	.26	**.41**
CLIP-S [3]	✗	.18	.27	.32	.17	.28	.32
†InfoMetIC [10]	✗	.22	**.30**	.37	.21	**.32**	.38
VisCE^2	✗	**.54**	.08	**.45**	**.49**	.06	**.39**

4.3 キャプション品質評価の特性

VisCE^2 のキャプション評価の特性を調べるために，キャプションに精度と想起スコアが付与された

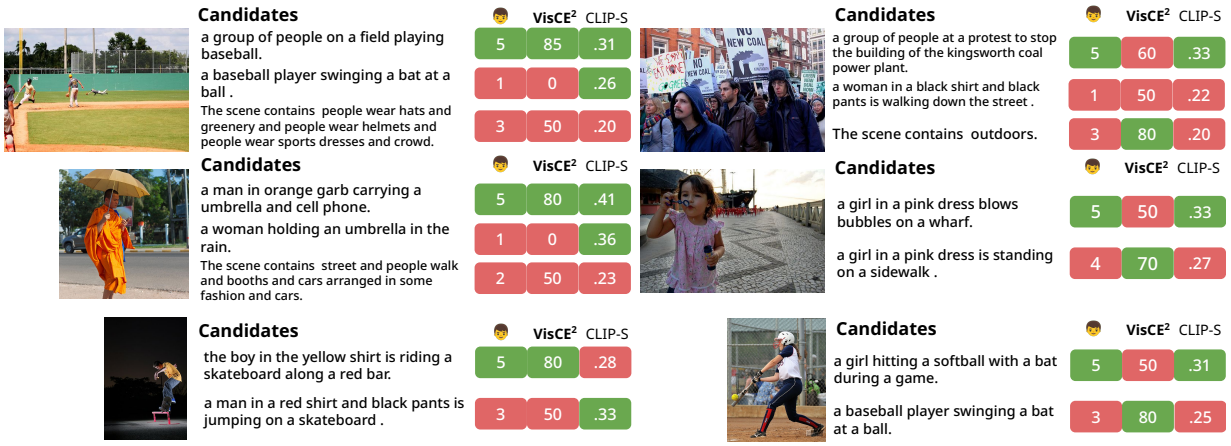

図 2　Composite データセットの画像，人手評価および候補文と，それに対する VisCE^2 と CLIP-S の評価スコアの比較．左側に VisCE^2 が正しく評価できた例，右側に誤って評価した例を示す．

表 4　異なるプロンプト戦略での人手評価との相関と精度スコア．THumB は w/ Human の設定での Precision との相関を示す．

Prompt	Ref	F8k-Exp.	Com.	P-50S	THumB
Vanilla	✗	55.9	52.4	80.5	.41
CoT	✗	54.6	54.5	79.2	.41
Description	✗	57.4	52.5	77.5	.47
VisCE^2	✗	**59.0**	**55.0**	**80.8**	**.49**

THumB1.0 [25] を用いてそれぞれのスコアとの相関を計測した．先行研究 [10, 25] に倣い，メタ評価指標には Pearson の相関係数を用いた．

表 3 に示すように，VisCE^2 は精度との相関で最高性能を達成した．RefCLIP-S と比較すると，総合スコアでは同程度の相関を示しているが，精度との相関は提案手法が大きく差をつけて上回り (+0.20/0.18pt.)，想起との相関は劣る (-0.19/-0.20pt.)．このことから，VisCE^2 は視覚的文脈の挿入による画像の周辺情報を基に，内容の網羅性よりも描写の正確性を重視することが明らかになった．

4.4　視覚的文脈の有無が及ぼす評価性能への影響

我々は，視覚的文脈の挿入が評価性能の向上に寄与するか検証するために，4 種類のプロンプトを用いて人手評価との相関を比較した（付録表 5）：

1. **Vanilla**: スコアのみを出力させる．
2. **Chain-of-Thought**: 理由とスコアを出力させる．
3. **Description**: 視覚的文脈として LLaVA-1.5 が生成したキャプションを利用する．
4. **VisCE^2（提案手法）**：視覚的文脈を構成的な形式で出力し評価に利用する．

表 4 に示すように，VisCE^2 は全ての項目で最も良い評価性能を示した．注目すべきことに，VisCE^2 は Description 設定と比較して性能が有意に向上した．この結果は，参照文よりも視覚的文脈を用いることが評価性能の向上に寄与することを示し，提案手法の有効性を実証するものである．

4.5　定性的評価

提案した評価尺度の品質を評価するために，VisCE^2 が出力したスコアを人手評価および CLIP-S と定性的な例を用いて比較した．図 2 の左列の例を見ると，CLIP-S は画像に含まれる物体を含んだキャプションを高く評価し，構成的な誤りを過小評価している．VisCE^2 はそのようなキャプションを正しく評価できることがわかる．しかし，VisCE^2 は画像の主題でない状況を正確に描写したキャプションに高得点をつける傾向にあり，右列のような誤りが見られる．この精度を重視する傾向は 4.3 節で見られた結果と整合する．

5　おわりに

我々は，視覚的文脈を抽出し評価に活用する VLM ベースの画像キャプション生成の自動評価手法である VisCE^2 を提案した．メタ評価実験では既存の指標と比較して人間の判断に沿った評価ができ，視覚的文脈の挿入が評価性能を向上させることを実証した．生成キャプションの画像の内容における網羅性評価と VisCE^2 と組合せることでさらなる評価性能の向上が期待される．また，VLM による評価を幻覚評価や他の視覚言語課題に応用する予定である．

謝辞. 本研究は JST さきがけ JPMJPR20C2，JSPS 科研費 22K17983，JSPS 科研費 JP20269633 の助成を受けたものです.

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.

[2] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **CVPR**, pp. 4566–4575, 2015.

[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In **EMNLP**, 2021.

[4] Saba Ahmadi and Aishwarya Agrawal. An examination of the robustness of reference-free image captioning evaluation metrics. **arXiv preprint arXiv:2305.14998**, 2023.

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

[6] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In **WMT**, pp. 376–380, 2014.

[7] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In **ECCV**, pp. 382–398, 2016.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. **arXiv preprint arXiv:2103.00020**, 2021.

[9] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In **CVPR**, 2023.

[10] Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. In **ACL**, pp. 3171–3185. Association for Computational Linguistics, 2023.

[11] David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating Image Captions with Large Language Models. In **EMNLP**, Singapore, Singapore, December 2023. Association for Computational Linguistics.

[12] OpenAI. Introducing ChatGPT, 2022.

[13] Yuntao Bai, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. **arXiv preprint arXiv:2204.05862**, 2022.

[14] Aakanksha Chowdhery, et al. PaLM: Scaling language modeling with pathways. **arXiv preprint arXiv:2204.02311**, 2022.

[15] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequencelearning framework. **arXiv preprint arXiv:2202.03052**, 2022.

[16] Wenliang Dai, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. **arXiv preprint arXiv:2305.06500**, 2023.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **arXiv preprint arXiv:2304.08485**, 2023.

[18] Hugo Touvron, et al. LLaMA: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[19] 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, March 2023.

[20] OpenAI, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[21] Gemini Team, et al. Gemini: A family of highly capable multimodal models. 2023.

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

[23] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. **J. Artif. Intell. Res.**, Vol. 47, pp. 853–899, 2013.

[24] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From images to sentences through scene description graphs using commonsense reasoning and knowledge. **arXiv preprint arXiv:1511.03292**, 2015.

[25] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. Transparent human evaluation for image captioning. In **NAACL**, pp. 3464–3478, Seattle, United States, July 2022. Association for Computational Linguistics.

[26] Xinlei Chen, et al. Microsoft COCO captions: Data collection and evaluation server. **arXiv preprint arXiv:1504.00325**, 2015.

[27] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. **TACL**, Vol. 2, pp. 67–78, 2014.

表 5　実験において利用したプロンプト．ただし，{caption} は候補文，{context} は第一段階で生成されたキャプションおよび視覚的文脈をそれぞれ挿入する．画像はいずれのプロンプトにおいても先頭に与えられる．

Method	Prompt
Vanilla	On a precise scale from 0 to 100, rate whether the candidate caption is appropriate for the given image. Candidate caption: {caption} Your rating must be a single digit between 0 and 100.
CoT	On a precise scale from 0 to 100, rate whether the candidate caption is appropriate for the given image. Candidate caption: {caption} Your rating must be a single digit between 0 and 100. Let's think step by step.
Description - Step 1	Generate a detailed description for the given image.
VisCE^2- Step 1	Analyze the uploaded image and provide a structured output focusing on the objects, their features, and the relationships between them. Select up to five of the most important elements. The output should be organized as follows: List of Important Objects (up to five): - Object 1: [Brief description] - Object 2: [Brief description] - (Continue as necessary, up to five objects) Features (Specific characteristics and attributes of each object, such as color, shape, size, and texture): - Features of Object 1: [Detailed description of features] - Features of Object 2: [Detailed description of features] - (Continue as necessary for each selected object) Relationships (The way objects interact or are positioned relative to each other, without using specific object names or symbols): - Description of a relationship: [General description] - Another relationship: [General description] - (Continue as necessary for each relevant relationship) Focus on providing unique and detailed insights into the features and relationships of the selected objects up to five objects.
Description - Step 2 VisCE^2- Step 2	On a precise scale from 0 to 100, rate whether the candidate caption is appropriate for the given image. Candidate caption: {caption} Use the image and following visual context to evaluate the candidate caption: Visual context: {context} Your final rating must be a single digit between 0 and 100.

A　評価データセットの詳細

Flickr 8k Expert [23] は 5,644 組の画像と自動生成キャプションが含まれている．各ペアは 3 人の専門家によって 1 から 4 でスコア付けされ，合計 17,000 の人手評価を含んでいる．ここで専門家とはイリノイ大学の学生であり，クラウドソーシングを用いて集めたデータセット (Flickr8k-CrowdFlower) との差異の明確化のための語用である．1 はキャプションが画像と全く関連しないことを示し，4 はキャプションが対応する画像を間違いなく説明していることを示す．

Composite [24] は，MSCOCO [26] (2,007 枚)，Flickr8k [23] (997 枚), Flickr30K [27] (991 枚) の 3 つのデータセットから合計 3,995 枚の画像を収集したデータセットである．各画像に 2 つの自動生成キャプションと 1 つの人間が書いたキャプションが付与され，1 から 5 の 5 段階の Likert 尺度でスコア付けされている．

PASCAL-50S [2] は，UIUC Pascal Dataset から抽出された 1000 の画像に対して，それぞれ少なくとも 50 個の参照キャプションが与えられる．比較対象となる候補文は人間が作成したキャプションと 5 つの自動生成手法を用いて生成したキャプションが対となり，4000 対集められている．人手評価は 3 名の評価の多数決で定められ，参照文とより類似していると判断されたキャプションが選好の人手評価として付与される．また，作業者は評価時に “類似” の概念は明確に指示されていなかった．

THumB 1.0 [25] は，MSCOCO [26] からランダムに抽出された画像 500 枚に，1 つの人間が書いたキャプションと 4 つの自動生成キャプションが付与したデータセットである．各組に，キャプションの正確さ (Precision)，顕著な情報の網羅性 (Recall)，総合的な品質 (Total) の 3 つのスコアが人手で付与されている．