複数タスク・複数項目に跨ったマルチモーダル自動評価手法

大井 聖也 [1] 金子 正弘 [2,1] 岡崎 直観 [1,3,4] 井上 中順 [1]
[1] 東京科学大学 [2] MBZUAI [3] 産業技術総合研究所 [4] NII LLMC
{masanari.ohi@nlp., okazaki@, inoue@}comp.isct.ac.jp
masahiro.kaneko@mbzuai.ac.ae

概要

視覚言語モデル（Vision-Language Model; VLM）は与えられた画像と指示文に基づいて文を生成できる能力を持つ。しかし、VLM の出力文を評価する既存手法は、文の総合的な品質を測定することのみに注力しているため、結果の解釈性が乏しいことに加え、必要な評価項目を網羅できていない可能性がある。本研究では、文の評価項目ごとの質を網羅的にスコア付けし、それらのスコアを元に総合スコアを決定する自動評価手法 HarmonicEval を提案する。構築した人手評価データセット MMHE における実験により、HarmonicEval の人手評価との相関は既存手法を上回ることを示す[1)]。

1 はじめに

画像キャプション生成や画像質問応答などの視覚言語タスクにおける VLM の性能を測定する際、VLM が生成した文の自動評価が必要である [1, 2]。これまでに、BLEU [3] や CIDEr [4] などの *n*-gram ベースの手法に加え、CLIPScore [1] や BERTScore [5] などの深層ニューラルモデルに基づく手法などが利用されてきた。しかし、これらの既存手法は文の品質を表す総合的なスコアのみを出力するように設計されており、スコアの解釈性に欠ける。例えば、図 1 (a) に示す通り、総合スコアが 5 点満点中 4 点の結果からでは、文が曖昧・不自然であるという問題点が特定できない。近年ではスコアの理由を追加で生成する手法が提案されているが [2, 6]、テキストによる説明では一貫した分析や定量的な比較が困難であり、評価項目毎に定量化することが望ましい。

さらに、先行研究 [7, 8] で指摘されているように、既存手法は総合的な品質を測定する際に必要な評価項目を網羅できていない可能性がある。例えば、

図 1 既存手法の問題点と複数項目における評価の利点を示す図。各スコアは 1-5 点の 5 段階で評価される。(a) 参照表現生成タスクにおける評価で、既存手法の評価結果から文の問題点の特定が困難な例。(b) 画像質問応答タスクにおける評価で、既存手法が簡潔性を軽視している例。

図 1 (b) に示す通り、質問に対する回答生成では、正確性や完全性の評価項目が重視される一方で、簡潔性が軽視される可能性がある。

これらの問題に対処するため、本研究では事前に定義された評価項目ごとに生成文の質を測定し、それらの点数に基づいて総合的に生成文の品質を測定する自動評価手法 HarmonicEval を提案する。HarmonicEval では、VLM を評価器として文の評価を行う（これを VLM 評価器と書く）。まず、VLM 評価器に文の評価を指示するプロンプトを与え、評価項目ごとに評価スコアの出力を促す（項目別評価）。次に、スコアの分布を平滑化するために、VLM 評価器がスコアを出力する確率を基にスコアの期待値を計算し、項目ごとの評価スコアとして採用する（スコア平滑化）。最後に、スコアの分散が小さい評価項目により大きな重みを与え、評価項目ごとのスコアの重み付き平均を計算し、総合的な文の質を表す評価スコアとして用いる（スコア集計）。以上の

1) 我々のコードとデータセットは https://github.com/stjohn2007/HarmonicEval で公開されている。

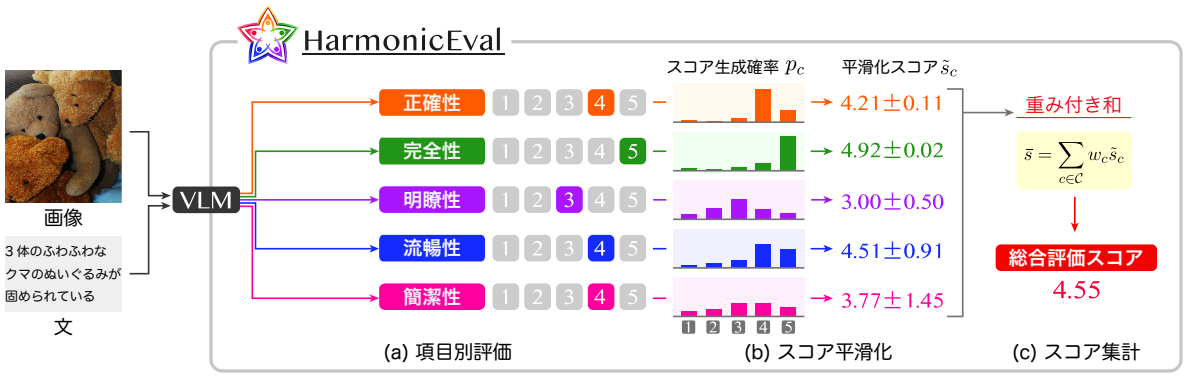

図 2　HarmonicEval の概要

手順により、HarmonicEval は解釈性に富み、評価項目を網羅した評価を実現する。

さらに、複数の視覚言語タスク・複数の評価項目における HarmonicEval と既存手法の性能を評価するために、我々は複数タスク・複数項目における人手評価データセット Multi-task, Multi-criteria, Human Evaluation (MMHE) を構築する。MMHE は、参照表現生成 (Referring Expression Generation; REG)、画像質問応答 (Visual Question Answering; VQA)、画像文書理解 (Visual Document Understanding; VDU)、画像キャプション生成 (Image Captioning; IC) の 4 タスクにおける VLM の出力文を、5 つの評価項目に関して人手で評価したデータ約 18,000 件で構成される。

MMHE を用いた実験の結果、HarmonicEval の人手評価との相関は、殆どのタスク・評価項目において既存手法を上回った。さらに、評価の際に既存手法が特定の評価項目を考慮できていない可能性を実験的に示した。

2　HarmonicEval

HarmonicEval の概要を図 2 に示す。まず、VLM 評価器に文の評価を指示するプロンプトを与え、各評価項目のスコアを独立に出力させることで項目別の評価を行う（2.1 節）。次に、評価スコアの出力確率を用い、項目ごとのスコアを平滑化する（2.2 節）。最後に、スコアの分散から算出した重みを用いて項目ごとのスコアの重み付き平均を計算し、総合的な評価スコアとする（2.3 節）。

2.1　項目別評価

本研究では、VLM 評価器 f にプロンプト $\boldsymbol{I_c}$ を与えることで、評価項目 $c \in \mathcal{C}$ における文 $\boldsymbol{t}$ [2] と画像 $\boldsymbol{x}$ の評価スコア $s_c = f([\boldsymbol{I_c}, \boldsymbol{t}], \boldsymbol{x})$ を生成させる。ここで、評価スコアは 5 段階 $s_c \in \{1, 2, 3, 4, 5\}$ であり、$\mathcal{C}$ は評価項目の集合を表す。自然言語生成タスクと視覚言語タスクにおける先行研究 [8, 9, 10, 11, 12, 13] に基づき、我々は $\mathcal{C}$ の要素を 5 つに決定した。

- 正確性: 文に含まれる情報が入力画像・入力文の内容に対して正確か
- 完全性: 文が入力画像・入力文の内容を十分に考慮できているか
- 明瞭性: 文の記述が曖昧でないか
- 流暢性: 文が文法的に正確で流暢か
- 簡潔性: 文が冗長でなく簡潔か

2.2　スコア平滑化

より正確な評価を実現するために、本研究ではスコア s_c の平滑化を行う。具体的には、先行研究 [2, 14] に倣い、VLM 評価器がスコアを出力する確率に基づいてスコアの期待値 $\tilde{s}_c$ を計算する。

$$\tilde{s}_c = \sum_{s=1}^{5} s p_c(s) \tag{1}$$

ここで、$p_c(s) = P(f([\boldsymbol{I}_c, \boldsymbol{t}], \boldsymbol{x}) = s)$ は VLM 評価器が評価項目 c においてスコア s を出力する確率である。$\tilde{s}_c$ が評価項目 c における最終的なスコアとして採用される。

2.3　スコア集計

本研究では、評価項目ごとのスコア $\{\tilde{s}_c \mid c \in \mathcal{C}\}$ の重み付き平均を計算することで、文の総合的な質を表す総合評価スコア $\bar{s}$ を得る。

$$\bar{s} = \sum_{c \in \mathcal{C}} w_c \tilde{s}_c \tag{2}$$

各項目の重み w_c は、スコアの標準偏差 σ_c に基づき、次式で計算する。

$$w_c = \frac{1}{H} \sigma_c^{-2(1-\gamma)/\gamma} \tag{3}$$

2)　例えば、$\boldsymbol{t}$ は画像キャプション生成においてはキャプション、質問応答生成においては質問と回答のペアとなる。

σ_c はスコアの出力確率を用いて以下のように計算される。

$$\sigma_c = \sqrt{\sum_{r=1}^{5} (r - \tilde{s}_c)^2 \, p_c(r)} \tag{4}$$

式 3 は σ_c の大きな項目に小さな重みを与え、σ_c の小さな項目に大きな重みを与えるため、計算される総合評価スコアに対する統計的変動の影響が抑えられ、より正確な評価が行えると期待している。また、$H = \sum_c \sigma_c^{-2(1-\gamma)/\gamma}$ は重みの範囲を $0 \le w_c \le 1$ にするための定数である。$0 < \gamma \le 1$ は σ_c の影響を調節するためのハイパーパラメータであり、実験では $\gamma = 0.75$ を用いる。γ の値を決定するための議論を付録 A に記す。

3 MMHE データセット

MMHE データセットは、複数の視覚言語タスクにおける VLM の出力文を、複数の評価項目及び総合的な質において人手で評価したデータ 18,000 件で構成される。データセットの事例を付録図 3 に示す。

3.1 視覚言語タスク

MMHE では、以下の 4 つの視覚言語タスクを採用する。

- REG（参照表現生成）：画像のうち長方形で囲まれた部分を記述する表現を生成するタスク
- VQA（画像質問応答）：画像に関する質問文に対して回答を生成するタスク
- VDU（画像文章理解）：文書を含む画像に関する質問文に対して回答を生成するタスク
- IC（画像キャプション生成）：画像のキャプションを生成するタスク

3.2 データセット構築

MMHE は (1) 入力の選定、(2) 出力の生成、(3) 人手評価 の手順で構築した。

入力の選定 まず、タスクの入力を既存のデータセットから収集した。入力は、REG・IC においては画像、VQA・VDU においては画像と質問文である。我々は REG では RefCOCO [15]、VQA では OK-VQA [16]、VDU では VisualMRC [17]、IC では MSCOCO [18] の検証もしくは評価サブセットから 100 事例ずつをランダムに取得した。

出力の生成 次に、複数の VLM を用いて、各タスクの入力に対する出力文を生成した。

表 1 MMHE における総合評価の一致率 (%)。太字は最も高い一致率を、「平均」は 4 つのタスクの平均値を表す。

手法	REG	VQA	VDU	IC	平均
BLEU	45.3	29.4	57.3	46.8	44.7
ROUGE	49.0	30.8	56.0	47.9	45.9
CIDEr	42.5	25.0	62.1	42.7	43.1
METEOR	44.4	29.4	59.7	53.6	46.8
BERT-S	46.2	33.8	62.1	53.1	48.8
BART-S	56.4	20.5	60.9	57.8	48.9
CLIP-S	60.1	39.7	60.9	52.0	53.2
FLEUR	62.9	**76.4**	60.9	73.9	68.5
HarmonicEval	**66.6**	**76.4**	**73.4**	**77.0**	**73.4**

具体的には、LLaVA-1.5-7/13B [19]、InstructBLIP-Vicuna-7/13B [20]、Qwen-VL [21]、Qwen2-VL-Instruct-7/72B [22]、CogVLM-Chat [23]、GPT-4o-mini、GPT-4o [24] の 10 個の VLM を用いた。

人手評価 最後に、5 人のアノテーターが出力文に対して項目毎と総合的な質の評価を行った。項目毎の評価では、各評価項目の定義に基づいて、出力文の質を 1 から 5 点の 5 段階で評価した。総合評価では、同じ入力に対して得られた 3 つの異なる生成文のうち、最も良い文を選択する形式を採用した。人手評価時の偏り [25] を防ぐため、総合評価においては各事例ごとの 5 段階評価は実施しなかった。

4 実験

4.1 設定

比較手法 本研究では 8 つの既存手法を HarmonicEval と比較する。n-gram ベースの手法として BLEU [3]、ROUGE [26]、METEOR [27]、CIDEr [4] を、深層ニューラルネットワークの手法として BERTScore [5]、BARTScore [28]、CLIPScore [1] を、そして VLM ベースの手法として FLEUR [2] を用いる。

評価指標 総合評価の性能を測定するための指標には一致率 (%) を、項目別評価の性能を測定するための指標にはケンドールの順位相関係数を用いた。

実装の詳細 HarmonicEval で使用する VLM 評価器には GPT-4o を用いた。GPT-4o の出力をできるだけ固定するため、推論時の温度パラメータは `temperature` = 0 を指定した。

4.2 実験結果

総合評価 表 1 に、HarmonicEval と既存手法の総合評価性能を比較した結果を示す。HarmonicEval は REG (66.6)、VQA (76.4)、VDU (73.4)、IC (77.0) の全

表 2　MMHE における項目別評価の相関係数。太字は最も高い相関係数を、黒の下線は手法ごとにタスクの中で最も高い相関係数を、灰色の下線は手法ごとにタスクの中で最も高い相関係数を表す。

手法	REG					VQA					VDU					IC				
	正確	完全	明瞭	流暢	簡潔	正確	完全	明瞭	流暢	簡潔	正確	完全	明瞭	流暢	簡潔	正確	完全	明瞭	流暢	簡潔
BLEU	6.0	6.9	3.9	1.2	6.1	-1.3	-10.4	-11.0	-19.3	4.1	19.8	12.9	14.9	14.3	21.2	4.4	4.5	5.9	0.3	11.3
ROUGE	2.3	5.7	4.4	-3.5	3.9	7.1	-2.8	-5.0	-8.1	10.2	20.0	14.7	16.2	17.9	22.7	5.2	6.5	9.0	4.4	9.7
CIDEr	6.4	3.4	2.4	-9.7	20.9	-27.8	-39.0	-19.5	-26.0	-3.8	23.7	15.8	19.3	18.0	23.8	0.7	-1.6	8.7	-3.8	14.5
METEOR	1.9	5.3	5.2	-5.1	-6.3	5.3	-3.9	-8.2	-8.5	2.7	17.8	18.0	16.9	20.5	14.9	6.8	12.1	7.3	-2.3	1.0
BERT-S	6.5	6.9	-6.5	-8.6	12.4	-2.8	-14.3	4.9	-10.0	6.1	21.0	17.4	20.4	21.6	23.9	12.3	11.1	6.4	4.7	10.5
BART-S	4.4	6.7	4.2	-7.8	3.1	-13.4	-20.2	-2.8	-16.6	1.6	22.4	21.3	21.6	17.9	14.7	4.8	4.3	4.3	2.2	3.2
CLIP-S	13.5	14.4	6.8	-0.9	-5.1	6.6	5.4	7.2	8.1	4.5	15.2	12.5	15.0	12.6	8.4	20.2	21.3	11.1	3.2	3.5
FLEUR	**29.3**	**30.8**	18.6	8.7	11.2	38.7	38.2	**39.9**	39.8	**44.7**	38.1	37.1	44.6	35.2	28.2	33.9	35.0	**25.9**	24.5	14.0
HarmonicEval	23.2	**30.8**	**24.0**	**20.7**	**23.8**	**53.5**	**50.6**	31.8	**51.9**	44.4	**60.0**	**48.8**	**47.9**	**51.2**	**45.8**	**44.7**	**50.3**	19.8	**36.4**	**22.8**

てのタスクにおいて、人手評価との相関が最も高かった。VLM ベースの既存手法である FLEUR は VQA で最も高い一致率を達成したものの、VDU で比較的低い一致率となった。これらの結果から、評価項目を網羅的に考慮して総合的な評価を行うことが複数のタスクにおいて有効である。

項目別評価　表 2 に、各手法による評価結果と人手評価の評価項目ごとの相関係数を示す。既存手法は項目ごとに評価するように設計されていないため、全ての項目で同じスコアを用い、相関を計算している。一方で、HarmonicEval は項目ごとに評価スコアを予測し、個別のスコアを用いて相関を計算している。表の太字で示されているように、HarmonicEval は殆どの評価項目において最も高い相関を達成した。この結果は、HarmonicEval が項目別評価を適切に実施できていることを示唆している。

既存手法のタスクごとの特徴　タスクごとに既存手法が重視している / していない評価項目に着目し、既存手法の評価結果を分析する。表 2 において、手法・タスクごとに最も高い相関を黒の下線で、最も低い相関を灰色の下線で表した。REG においては、ほとんどの手法において完全性が最も高い相関を示した。この結果は、長方形で囲まれた部分を特定するために十分な情報が必要であるからと解釈できる。VQA と VDU においては、ほとんどの手法において完全性が最も低い相関を示しており、情報が不十分な出力文に対しても既存手法が高いスコアをつけてしまうことが分かる。IC においては、殆どの手法において流暢性が最も低い相関を示しており、既存手法が文法的に誤りのある出力文に対しても高いスコアをつける可能性を示唆している。

アブレーション実験　表 3 に、HarmonicEval における項目別評価・スコア平滑化・スコア集計それぞれのアブレーション実験の結果を示す。項目別評価を除いた実験では、VLM 評価器に総合評価スコアを直接予測するようにプロンプトを与え、平滑化したスコアを総合評価スコアとして採用した。また、スコア平滑化を除いた実験では、VLM 評価器が生成したスコア s_c を用いて総合評価スコアを計算した。さらに、スコア集計を除いた実験では、各項目におけるスコアの平均値を総合評価スコアとして用いた。項目別評価を除いたことにより、REG、VQA、IC の評価性能が低下しており、評価項目を網羅的に考慮することが総合評価に良い影響を与えていることが明らかとなった。同様に、スコア平滑化・スコア集計を除くと、ほとんどのタスクにおける評価性能が低下しており、各機構が評価性能の向上に寄与していることが分かる。

表 3　アブレーション実験の結果。

手法	REG	VQA	VDU	IC	平均
HarmonicEval	66.6	**76.4**	73.4	**77.0**	**73.4**
- 項目別評価	62.0	73.5	**75.9**	76.5	72.0
- スコア平滑化	**67.5**	70.5	70.4	72.4	70.2
- スコア集計	65.7	75.0	73.4	76.5	72.6

5　おわりに

本研究では、複数の視覚言語タスクにおける生成文を複数の評価項目、および総合的な質に関して評価可能な自動評価手法 HarmonicEval を提案した。また、我々は 4 つの視覚言語タスクと 5 つの評価項目における人手評価データセット MMHE を構築した。MMHE を用いた実験の結果、HarmonicEval は既存手法を超える人手評価との相関を示した。今後は、few-shot 推論や思考の連鎖 [29] などを用い、評価性能の向上や、HarmonicEval に内在する評価バイアス [14, 30, 31] の検証に取り組みたい。

謝辞

本研究は JSPS 科研費 22K12089 の助成を受けたものです。また、本研究成果は、国立研究開発法人情報通信研究機構（NICT）の委託研究（22501）により得られたものです。本研究は、東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施しました。

参考文献

[1] Jack Hessel, Ari Holtzman, Maxwell Forbes, and et al. CLIPScore: A reference-free evaluation metric for image captioning. In **Proc. of EMNLP**, pp. 7514–7528, 2021.

[2] Yebin Lee, Imseong Park, and Myungjoo Kang. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In **Proc. of ACL**, pp. 3732–3746, 2024.

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

[4] Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In **Proc. of CVPR**, pp. 4566–4575, 2015.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, and et al. BERTScore: Evaluating text generation with BERT. In **Proc. of ICLR**, 2020.

[6] David Chan, Suzanne Petryk, Joseph Gonzalez, and et al. CLAIR: Evaluating image captions with large language models. In **Proc. of EMNLP**, pp. 13638–13646, 2023.

[7] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, and et al. SummEval: Re-evaluating summarization evaluation. **TACL**, Vol. 9, pp. 391–409, 2021.

[8] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, and et al. Transparent human evaluation for image captioning. In **Proc. of NAACL**, pp. 3464–3478, 2022.

[9] Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In **Proc. of IJCNLP**, pp. 343–348, 2017.

[10] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, and et al. Neural text summarization: A critical evaluation. In **Proc. of EMNLP-IJCNLP**, pp. 540–551, 2019.

[11] Markus Freitag, George Foster, David Grangier, and et al. Experts, errors, and context: A large-scale study of human evaluation for machine translation. **TACL**, Vol. 9, pp. 1460–1474, 2021.

[12] Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. FineSurE: Fine-grained summarization evaluation using LLMs. In **Proc. of ACL**, pp. 906–922, 2024.

[13] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. Image understanding using vision and reasoning through scene description graph. **CVIU**, Vol. 173, pp. 33–45, 2018.

[14] Yang Liu, Dan Iter, Yichong Xu, and et al. G-eval: NLG evaluation using gpt-4 with better human alignment. In **Proc. of EMNLP**, pp. 2511–2522, 2023.

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In **Proc. of EMNLP**, pp. 787–798, 2014.

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In **Proc. of CVPR**, pp. 3190–3199, 2019.

[17] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **Proc. of AAAI**, 2021.

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, and et al. Microsoft coco: Common objects in context. In **Proc. of ECCV**, 2014.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **Proc. of NeurIPS**, 2023.

[20] Wenliang Dai, Junnan Li, DONGXU LI, and et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In **Proc. of NeurIPS**, pp. 49250–49267, 2023.

[21] Jinze Bai, Shuai Bai, Shusheng Yang, and et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. **arXiv preprint arXiv:2308.12966**, 2023.

[22] Peng Wang, Shuai Bai, Sinan Tan, and et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. **arXiv preprint arXiv:2409.12191**, 2024.

[23] Weihan Wang, Qingsong Lv, Wenmeng Yu, and et al. Cogvlm: Visual expert for pretrained language models. In **Proc. of NeurIPS**, 2024.

[24] OpenAI. Gpt-4o system card, 2024.

[25] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, and et al. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024.

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

[27] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, 2005.

[28] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In **Proc. of NeurIPS**, Vol. 34, pp. 27263–27277, 2021.

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, and et al. Chain-of-thought prompting elicits reasoning in large language models. In **Proc. of NeurIPS**, Vol. 35, pp. 24824–24837, 2022.

[30] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and et al. Judging LLM-as-a-judge with mt-bench and chatbot arena. In **Proc. of NeurIPS**, Vol. 36, pp. 46595–46623, 2023.

[31] Masanari Ohi, Masahiro Kaneko, Ryuto Koike, and et al. Likelihood-based mitigation of evaluation bias in large language models. In **Proc of. ACL**, pp. 3237–3245, 2024.

[32] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. **Journal of Artificial Intelligence Research**, Vol. 47, pp. 853–899, 2013.

[33] Peter Anderson, Basura Fernando, Mark Johnson, et al. Spice: Semantic propositional image caption evaluation. In **Proc. of ECCV**, pp. 382–398, 2016.

[34] Ming Jiang, Qiuyuan Huang, Lei Zhang, and et al. TIGEr: Text-to-image grounding for image caption evaluation. In **Proc. of EMNLP-IJCNLP**, pp. 2141–2152, 2019.

[35] Sijin Wang, Ziwei Yao, Ruiping Wang, and et al. FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation. In **Proc. of CVPR**, pp. 14050–14059, 2021.

[36] Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In **Proc. of ACL**, pp. 13559–13568, 2024.

[37] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, and et al. UMIC: An unreferenced metric for image captioning via contrastive learning. In **Proc. of ACL and IJCNLP**, pp. 220–226, 2021.

[38] Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. InfoMetIC: An informative metric for reference-free image caption evaluation. In **Proc. of ACL**, pp. 3171–3185, 2023.

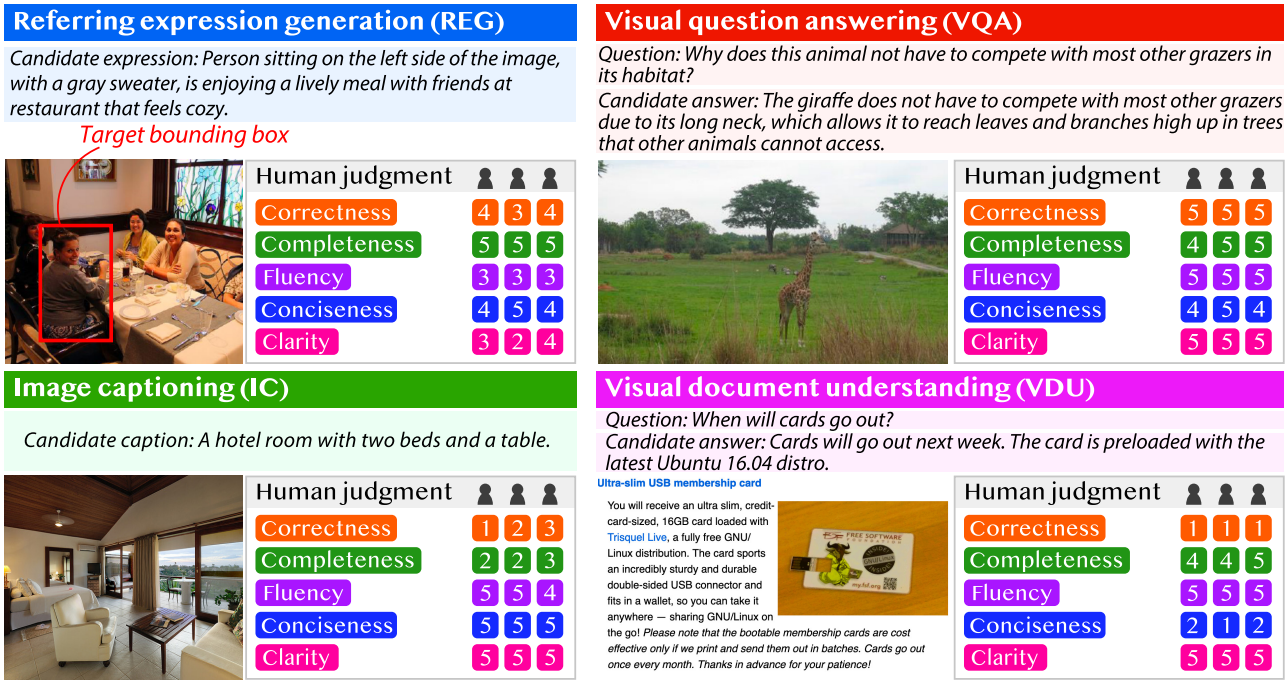

図 3　MMHE データセットの事例。

A　HarmonicEval における γ の選択

式 3 における γ の適切な値について議論する。

$\gamma = 1$ の場合　$w_c = 1/|\mathbb{C}|$ となり、各スコアの平均値が総合評価スコアとして採用される。スコアの統計的変動を考慮していないため、適切な選択ではないと考えられる。

$\gamma = 0.5$ の場合　重みはスコアの分散の逆数に比例するように計算され（$w_c \propto \sigma_c^{-2}$）、$\bar{s}$ の分散は最小となる。σ_c が統計的変動にのみ影響されると仮定した場合、$\bar{s}$ に対する統計的変動の影響は最小となり、理想的な状態になる。しかし、σ_c は統計的変動以外に各項目に内在する分散の影響を受けると考えられるため、$\bar{s}$ に対する統計的変動の影響は最小とならず、最良の選択でない可能性がある。

$\gamma \to 0$ の場合　最も小さい標準偏差を持つ項目 $c = \mathrm{argmax}_{c \in \mathbb{C}} \sigma_c$ のスコアが総合評価スコア $\bar{s}$ として採用される。一つの項目だけを考慮するため、適切な選択ではないと考えられる。

以上の議論より、$\gamma = 0.5$ は統計的変動のみを考慮しており、$\gamma = 1$ は統計的変動を全く考慮していないため、$0.5 \leq \gamma \leq 1$ の範囲の γ が妥当な値であると推測できる。本研究では、0.5 と 1 の中間の値が最適な値だと仮定し、$\gamma = 0.75$ を実験で用いた。

B　追加の実験

γ の探索実験　表 4 に、γ の値を 0.01, 0.50, 0.75, 1.00 の 4 通りに変化させた際の総合評価の精度を示す。付録 A で推測した通り、$\gamma = 0.75$ が最も高い精度を達成した。

既存データセットにおける評価性能　IC タスクの既存の人手評価データである Flickr8k-EX と Flickr8k-CF [32] を用いて、MMHE 以外のデータセットにおける MMHE の評価性能を検証した。先行研究 [1, 2] に倣い、ケンドールの順位相関係数 τ_b を評価指標に用いた。表 5 に示す通り、HarmonicEval は既存手法と比較可能、もしくは上回る性能を示した。

表 4　γ の探索実験の結果。

γ	REG	VQA	VDU	IC	Avg.
0.01	47.2	69.1	61.4	53.1	57.7
0.50	**66.6**	**76.4**	**73.4**	76.5	73.2
0.75	**66.6**	**76.4**	**73.4**	**77.0**	**73.4**
1.00	65.7	75.0	**73.4**	76.5	72.6

表 5　IC タスクの既存データセットにおける評価結果。

	手法	Flickr8k-EX	Flickr8k-CF
Reference-based	BLEU	30.8	16.9
	ROUGE	32.3	19.9
	METEOR	41.8	22.2
	CIDEr	43.9	24.6
	SPICE [33]	44.9	24.4
	BERT-S	39.2	22.8
	BERT-S++ [5]	46.7	–
	TIGEr [34]	49.3	–
	ViLBERTS-F [12]	50.1	–
	FAIEr-4 [35]	52.6	35.4
	RefCLIP-Score [1]	53.0	36.4
	Polos [36]	**56.4**	37.8
	RefFLEUR [2]	51.9	38.8
Reference-free	UMIC [37]	46.8	–
	FAIEr-r [35]	50.1	32.4
	CLIP-S	51.5	34.4
	InfoCLIP [38]	32.6	23.5
	InfoMetIC [38]	54.2	36.3
	InfoMetIC$^+$ [38]	55.5	36.6
	FLEUR	53.0	38.6
	HarmonicEval	53.1	**39.2**