CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

画像キャプション生成タスクにおけるマルチモーダル大規模言語モデルのための Preference Optimization

西村 喬行[1,a)]　小槻 誠太郎[1]　松田 一起[1]　飯岡 雄偉[1]　杉浦 孔明[1,b)]

概要

本研究では，画像キャプション生成タスクにおける，既存の Multimodal Large Language Model（MLLM）の性能改善を目的とする．既存の MLLM は，しばしば人間が重要視する情報を欠落させることで画像との整合性がとれていないキャプションを生成することがある．そこで，本研究では，MLLM に対する画像キャプション生成性能を改善するためのフレームワークを提案する．本フレームワークは既存の MLLM に対して，大規模かつ多様な preference data を作成し，画像キャプション生成に対する自動評価尺度を用いて preference optimization を行う．標準的なベンチマークで評価を行った結果，提案フレームワークによる学習を行った MLLM は既存の fully open MLLM と比較して多くの尺度で上回る結果を得た．

1. はじめに

Multimodal Large Language Model（MLLM）は，ロボティクス分野や医療画像理解など幅広い応用先において著しい発展を遂げている [24]．これらの各分野への応用拡大に伴い，MLLM のマルチモーダル言語理解能力および生成能力には更なる向上が求められている．特に，入力された視覚情報を自然言語で記述する画像キャプション生成について性能が高い MLLM は有用性が高い．しかし，それらの MLLM は，しばしば画像との整合性が十分に保たれないキャプションを生成するという課題が存在する．

既存の MLLM に対する画像キャプション生成性能向上を目的とした preference optimization は，大規模で多様なキャプションを含む preference data が必要であるため難しい [7]．Preference optimization には数十万サンプル規模のデータセットが必要であるが，4.1 節で詳述するように，そのようなデータセットは存在せず既存のデータセット [23, 19] から生成すると多様性や規模の観点で不十分である．そのため，多くの MLLM はそのような preference optimization を行っておらず，しばしば人間が重要視する情報を欠落させることがある．

図 1　入力画像の代表例

そこで，本研究では大規模かつ多様な preference data を作成し，画像キャプション生成に対する自動評価尺度 Polos [23] を用いて preference optimization を行うフレームワークを提案し，MLLM のキャプション生成性能を改善する．これにより，MLLM が人間の注目する箇所について過不足ないキャプションを生成できるようになると考えられる．

提案手法の新規性は以下の通りである．

- 大規模な画像群に対して生成した多様なキャプションから，MLLM および Polos に基づき，preferece data を作成する Automatic Preference Data Creation (APDC) モジュールを導入する．
- 人間の選好を反映したキャプションを生成するために，Polos スコアを用いて preference optimization を行う Preference Optimization for Image captioning (POIC) モジュールを導入する．

2. 問題設定

本研究は，既存の fully open MLLM における画像キャプション生成タスクの性能改善を目的とする．本タスクでは，与えられた画像に対して短文キャプションまたは長文キャプションを生成することが求められる．短文キャプションは，シーンの重要な部分を簡潔に説明することに重点を置き，長文キャプションは，詳細で文脈に富んだ説明を提供することを目的としている．

図 1 の画像が与えられた際，短文キャプションとしては “A dog chasing a red frisbee in an open field.” と，長

[1] 慶應義塾大学
[a)] t-nishimura@keio.jp
[b)] komei.sugiura@keio.jp

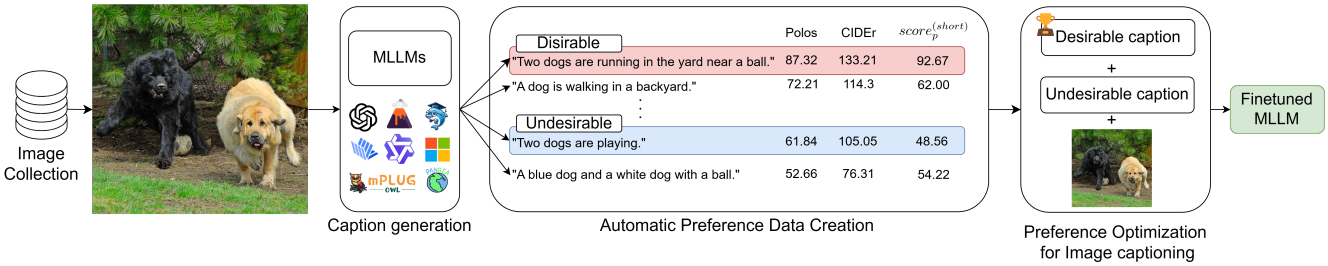

図 2 提案フレームワークの構成図

文キャプションとしては “This image is taken outdoors. At the bottom of the image we can see the grass on the ground. In the middle of the image a dog is running on the ground to catch a frisbee. In the background we can see the plants.” と生成することが期待される．本タスクの入力は画像であり，出力は画像キャプション y_{cap} である．また，本研究は，fully open MLLM を扱うことを前提とする．

3. 提案手法

本研究では，fully open MLLM の画像キャプション生成能力向上を行うためのフレームワークを提案する．本フレームワークは，画像キャプションに対する自動評価尺度を活用して，KTO [7] を拡張したものであり，追加の fine-tuning が可能な MLLM に対して広く適用できる．

図 2 に提案手法の構成図を示す．図に示す通り，提案フレームワークは主に Automatic Preference Data Creation(APDC) および Preference Optimization for Image captioning(POIC) の主に 2 つから構成される．提案フレームワークの入力は画像群 $\boldsymbol{x}_{\text{img}}$ である．

3.1 Automatic Preference Data Creation

Automatic Preference Data Creation モジュール (APDC) は，大規模な画像群に対して多様なキャプションを生成し，自動評価尺度を用いてスコアを付与することで，preference optimization に用いる preference data を作成する．ここで preference data は，desirable キャプションと undesirable キャプションのペアのことを指す．多くの preference optimization では，数十万サンプル規模の preference data が必要である [7]．一方，我々の知る限り，多様な画像およびキャプションを含むそのようなデータセットは存在しない．そこで，APDC では，MLLM を用いてキャプションを生成し，自動評価尺度を用いたスコアに基づいて，大規模な画像群に対する preference data を作成する．

APDC では，$\boldsymbol{x}_{\text{img}}$ を入力し，各画像に対する preference data を出力する．まず，MLLM を用いてキャプション群を生成する．この際に，9 つの MLLM [6, 1, 26, 3, 10, 25, 27, 5, 9] を使用する．また，事前実験の結果に基づき，MLLM に入力するプロンプトとして，短文キャプションには “Provide a one-sentence caption for the provided image.”，長文キャプションには “Describe the image in detail.” を使用する．また，正規化のためにそれらのキャプションに含まれる英数字を英語表記に変換した．次に，生成された各キャプションに対して自動評価尺度を用いてスコアを付与する．短文キャプションにおいては，人間の評価との相関が高く推論時間が短い Polos [23] および評価尺度を報酬として学習を行う既存手法 [17] で良好な結果が得られている CIDEr [21] を併用して評価スコアを付与する．長文キャプションにおいては，Polos が CLIP [15] 由来の特徴量を用いており，77 トークン以上の入力を想定していないため，Polos の代わりに VELA [28] を用いる．VELA は特に長文キャプション における人間の評価との相関が高い．

最後に，POIC において求められる形式で計算される評価スコアに基づき，各画像に対して preference data を作成する．まず，各画像に対して評価スコアが最も高いキャプションを desirable キャプション，最も低いキャプションを undesirable キャプションとする preference data を生成する．次に，N 番目に高いキャプションを desirable キャプション，N 番目に低いキャプションを undesirable キャプションとする preference data を生成する操作を n_{dup} 回繰り返す．これらの処理を画像群全体に適用し，合計 n_{pd} 個の preference data を生成する．

3.2 Preference Optimization for Image captioing

Preference Optimization for Image captioing モジュール (POIC) は，人間の選好を反映したキャプションを生成するために，前述の方法で得た preference data を用いた preference optimization を行う．人間の選好を反映させる学習手法に RLHF [12]，DPO [16]，KTO [7] などがある．しかし，これらは指示に従う能力の向上に重点を置いているため，これらを直接適用するだけでは人間の選好を反映したキャプションを生成するような学習としては不十分である．そのため，POIC では KTO を拡張した KTO-IC を導入する．KTO-IC は，キャプション自動評価尺度を用いたスコアに基づく損失関数により対象モデルが人間の選好を反映したキャプションを生成するような学習手法である．

本モジュールの入力は 2 つのキャプション y_1 および y_2 からなる preference data (y_1, y_2) であり出力は損失である．まず，あるキャプション y に対する評価スコア S_p を以下のように定義する．

表 1　テスト集合における提案手法とベースライン手法の定量的比較結果

	nocaps					Flickr30k					LN-Flickr30k				
手法	P	P-S	B@4	C	M	P	P-S	B@4	C	M	P	P-S	B@4	C	M
LLaVA-1.5-7B -pretrained [11]	66.87	85.36	23.94	57.91	26.81	69.50	84.11	5.98	23.49	17.60	58.34	86.08	3.82	8.72	12.51
LLaVA-1.5-7B [11]	70.94	85.30	31.74	105.61	31.34	67.59	84.35	19.28	80.34	27.35	59.10	84.79	1.72	4.58	15.26
Pangea-7B [27]	72.47	84.88	**46.48**	117.16	30.41	**75.20**	83.62	29.94	82.36	24.52	59.67	85.59	2.22	0.30	**16.72**
提案手法	**74.36**	**86.10**	42.99	**121.10**	**32.84**	64.63	**85.55**	**30.24**	**90.97**	**28.43**	**64.50**	**87.94**	**5.23**	**11.15**	14.94

$$\mathrm{S}_p(y)=\begin{cases}\mathrm{Polos}(y)+\lambda_{\mathrm{c2p}}\mathrm{CIDEr}(y), & \text{if } y \text{ is a short caption}\\ \mathrm{VELA}(y)+\lambda_{\mathrm{c2m}}\mathrm{CIDEr}(y), & \text{otherwise}\end{cases}$$

このとき Polos(・), CIDEr(・), VELA(・) は, それぞれ Polos スコア, CIDEr スコア, VELA スコアを計算する関数である．また, λ_{c2p}, λ_{c2m} は, それぞれ Polos スコアおよび VELA スコアに対する CIDEr スコアの重みづけを行うハイパーパラメータである．次に, y に対する preference optimization 適用前のモデルと最適化対象のモデルの予測の違いを評価する報酬関数 $r_\theta(x,y)=\log(\boldsymbol{\pi}_\theta(y\mid x)/\boldsymbol{\pi}_{\mathrm{ref}}(y\mid x))$ を定義する．ここで，$\boldsymbol{\pi}_{\mathrm{ref}}$ および $\boldsymbol{\pi}_\theta$ は，preference optimization 適用前のモデルと最適化対象のモデルの出力分布である．また, プロスペクト理論 [20] における人間の損失回避性を反映させるために，参照点 $z_0=\mathrm{KL}\big(\boldsymbol{\pi}_\theta(y'\mid x)\,||\,\boldsymbol{\pi}_{\mathrm{ref}}(y'\mid x)\big)$ からの逸脱を損失と見なす．このとき，$\boldsymbol{\pi}_{\mathrm{ref}}(y'|\boldsymbol{x})$ および $\boldsymbol{\pi}_\theta(y'|\boldsymbol{x})$ は，出力空間全体 $\mathcal{Y}$ で定義された条件付き確率分布でありそれぞれ preference optimization 適用前のモデルおよび最適化対象モデルの出力分布である．なお $y'\in\boldsymbol{\mathcal{Y}}$ であり，$\boldsymbol{\mathcal{Y}}$ は全ての可能な出力の集合を表す．しかし，定義通りに $\boldsymbol{\pi}_\theta$ から y' を全てサンプリングすることは現実的でないため，KTO [7] と同様に z_0 を近似する．次に，S_p に基づき価値関数を以下のように定義する．

$$v(\boldsymbol{x},y_1,y_2)=\begin{cases}\lambda_d\sigma\Big(\beta\cdot\big(r_\theta(\boldsymbol{x},y_1)-z_0\big)\Big), & \text{if } \mathrm{S}_p(y_1)>\mathrm{S}_p(y_2)\\ \lambda_u\sigma\Big(\beta\cdot\big(z_0-r_\theta(\boldsymbol{x},y_1)\big)\Big), & \text{otherwise}\end{cases}$$

ここで β，λ_d，λ_u は，それぞれ損失回避性に基づく重み付けのハイパーパラメータ，desirable キャプションおよび undesirable キャプションに対する重みづけのハイパーパラメータである．また，σ はそれぞれロジスティックシグモイド関数である．これらを用いて，$L_{\text{KTO-IC}}$ は以下のように定義される．

$$L_{\text{KTO-IC}(\boldsymbol{x},y_1,y_2)}=\mathbb{E}_{(\boldsymbol{x},y_1,y_2)\sim D}\Big[\lambda_y-v(\boldsymbol{x},y_1,y_2)-v(\boldsymbol{x},y_2,y_1)\Big]$$

ここで，λ_y は $\lambda_d+\lambda_u$ である．

4. 実験

4.1 実験設定

本研究では，KTO-IC による preference optimization を行うために preference dataset を構築する．そのためには，多様な画像群およびキャプションが含まれる数十万サンプル規模の preference data が必要である [7]．画像キャプションに対して評価が付与されたデータセットとして FOIL [19]，Polaris [23] があるが，これらは上記のような preference optimization を行うには不十分である．しかし，FOIL データセットはキャプションとハルシネーションを含む FOIL キャプションのペアからなる．しかし，FOIL キャプションは元のキャプションに対してルールベースにより単語の置換を行うことで生成されており文の多様性に欠ける．また，Polaris データセットは，1 万枚の画像しか含まれていないため，画像の多様性に欠ける．そのため，16 万枚の画像およびそれらに付与した 170 万個のキャプションを用い，159,000 サンプルの preference data からなる Polos FB データセットを作成した．Polos FB データセットの作成方法については 3.1 節に詳述する．

このデータセットは, 119,000 サンプルの英語キャプションを含む．そのうち，79,000 サンプルは短文キャプションであり，語彙サイズは 15,441 語，総単語数は 2,057,457 語である．desirable キャプションと undesirable キャプションの平均文長は，それぞれ 11.50 語と 14.45 語である．残りの 40,000 サンプルは長文キャプションであり，語彙サイズは 25,913 語，総単語数は 6,151,963 語である．desirable キャプションと undesirable キャプションの平均文長は，それぞれ 58.65 語と 94.10 語である．

実験において，訓練集合は 113,050 サンプル，検証集合は 5,950 サンプルであった．テスト集合には，このタスクで標準的に用いられるデータセットである，nocaps val split [2](nocaps), Flickr30k test split [14](Flickr30k), Localized Narratives Flickr30k subset [22](LN-Flickr30k) を使用した．各データセットのサンプル数は，それぞれ 5,000，4,500，1,000，1,000，125,436 である．パラメータ更新には訓練集合をハイパーパラメータの選択には検証集合を使用し，モデルの性能評価にはテスト集合を使用した．

本研究では，Pangea-7B モデル [27] に対して提案手法による訓練を行った．モデルは AdamW を用いて訓練され，学習率，β_1，β_2，ϵ はそれぞれ 1e-6，0.9，0.98，1e-6 に設定した．学習は，5 エポック行った．また，ランク，アルファスケーリング，ドロップアウト率をそれぞれ 128，256，

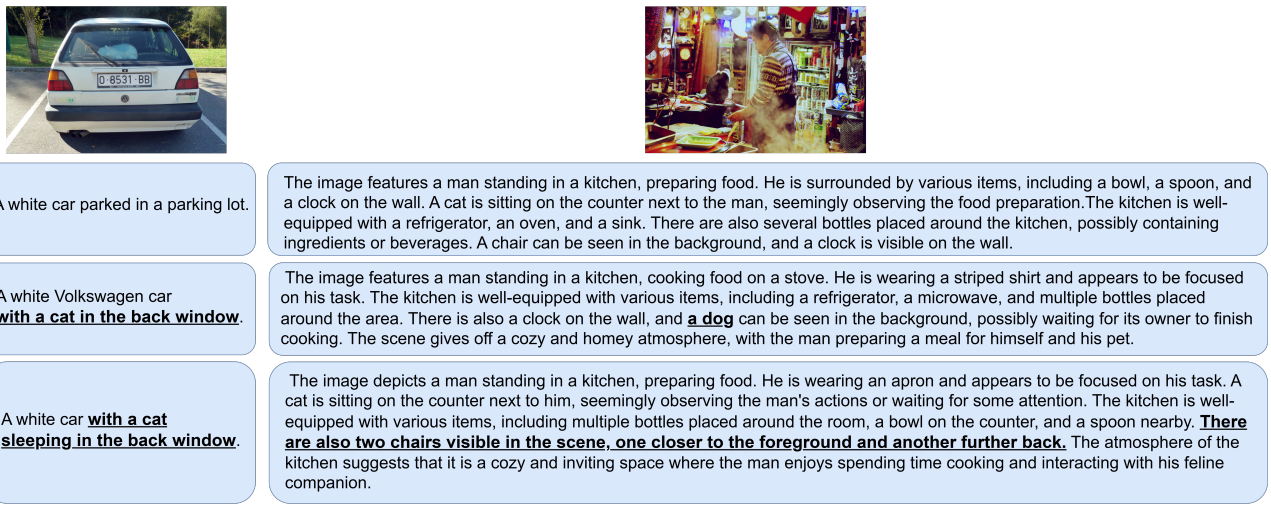

図 3 テスト集合における提案手法の成功例

0.1 と設定した LoRA [8] を適用した．さらに，画像エンコーダは，フリーズして訓練を行い，モデルの学習可能パラメータ数と積和演算数はそれぞれ約 76 億，約 96 億 2 千万であった．8 基の A100(80GB) を用いて学習を行った．

4.2 実験結果

表 1 に，各データセットにおけるベースライン手法との定量的比較の結果を示す．また，各評価において最も高い数値を太字で示す．表中の P, P-S, B@4, C, M はそれぞれ Polos [23], PAC-S [18], BLUE-4 [13], CIDEr [21], METEOR [4] を指す．ベースライン手法として，LLaVA-1.5-pretrained [11], LLaVA-1.5 [11] および Pangea-7B [27] を選んだ．ここで LLaVA-1.5 pretrained は，LLaVA-1.5 の visual instruction tuning を行う前のモデルを指す．LLaVA-1.5 は標準的な fully open MLLM であり，Pangea は fully open MLLM として良好な結果が得られているため選んだ．評価尺度として，本タスクにおいて標準的な BLEU，CIDEr，METEOR を用いた．さらに人間の評価との相関が高く，推論時間が短い Polos，PAC-S も用いた．なお，本研究では評価尺度の内 Polos を主要尺度とした．

表 1 に示すように，nocaps において，LLaVA-1.5-pretrained，LLaVA-1.5，Pangea，提案手法の Polos スコアは，それぞれ，66.87，70.94，72.47，74.36 であり，提案手法はベースラインのうち Polos および CIDEr のスコアが最も高い Pangea と比較してそれぞれ 1.89，3.94 ポイント上回った．Flickr30k において，LLaVA-1.5-pretrained, LLaVA-1.5, Pangea, 提案手法の順に Polos スコアは 69.50, 67.59，75.20，64.63 であり CIDEr スコアは 23.49，80.34，82.36，90.97 であった．Flickr30k において，提案手法の CIDEr スコアは，最も良い結果が得られた Pangea と比較して 8.61 ポイント向上した．一方，Polos スコアは 10.57 ポイントの減少であった．これについて定量的にエラー分析を行った結果，生成文において記述すべき物体に関する説明や参照表現が不足しているエラーが最も多くみられた．LN-Flickr30k では，LLaVA-1.5-pretrained，LLaVA-1.5, Pangea, 提案手法の順に Polos スコアは，58.34, 59.10, 59.67，64.50 であり，CIDEr は 8.72，4.58，0.30，11.15 であった．提案手法は，最も良いベースラインと比較して Polos，CIDEr において 4.83，2.43 ポイント向上した．

図 3 に提案手法の成功例を示す．図 3 において，(a) は入力画像を示す．また同様に図 3 において，(b), (c) および (d) はそれぞれ，提案手法，LLaVA-1.5, Pangea による生成文を示す．左列の (c), (d) において，全てのベースラインは車中にある動物のぬいぐるみを “cat” と誤ってキャプションを生成しているのに対し，提案手法は適切に “A white car parked in a parking lot.” と生成した．右列 (c) の LLaVA-1.5 においては，“There is also ⋯ a dog” とキャプションを生成しており，物体名称が誤っている．右列 (d) の Pangea は，“There are also two chairs visible in the scene, one closer to the foreground and another further back.” と生成し，存在しない物体について言及している．一方，提案手法は，上記のようなハルシネーションを含まない詳細なキャプションを生成した．

5. おわりに

本研究では，既存の fully open MLLM における画像キャプション生成タスクの性能改善を扱った．

将来研究として，情報の欠落を低減するために，画像エンコーダーを複数利用し，空間関係理解タスクを含むデータセットによる追加学習を行い，画像から多様な情報を画像特徴量として抽出し統合する能力および空間理解能力を高めることが期待される．また，人間の preference は多様性を有することから，Polos に対して個々人ごとに preference optimization を実施し得られた複数の Polos を利用することで，多様な preference の表現が可能になると考えられる．さらに，ノイズ制御パラメータを導入することで，その多様性を一層高めることが期待される．

謝辞

本研究の一部は，JSPS 科研費 23K28168，JST ムーンショットの助成を受けて実施されたものである．

参考文献

[1] Abdin, M., Aneja, J. et al.: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, *arXiv preprint arXiv:2404.14219* (2024).

[2] Agrawal, H., Desai, K., Wang, Y. et al.: nocaps: novel object captioning at scale, *ICCV*, pp. 8947–8956 (2019).

[3] Bai, J. et al.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, *arXiv preprint arXiv:2308.12966* (2023).

[4] Banerjee, S. et al.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *ACL*, pp. 65–72 (2005).

[5] Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J. et al.: Sharegpt4v: Improving large multi-modal models with better captions, *ECCV*, pp. 370–387 (2024).

[6] Chen, Z., Wu, J., Wang, W. et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks, *CVPR*, pp. 24185–24198 (2024).

[7] Ethayarajh, Kawin and others: Model alignment as prospect theoretic optimization, *ICML* (2024).

[8] Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models, *ICLR* (2022).

[9] Hurst, A., Lerer, A. et al.: GPT-4o System Card, *arXiv preprint arXiv:2410.21276* (2024).

[10] Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B. et al.: LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models, *ICLR* (2025).

[11] Liu, H., Li, C., Wu, Q. and Lee, Y. J.: Visual Instruction Tuning, *NeurIPS*, Vol. 36, pp. 34892–34916 (2023).

[12] Ouyang, L., Wu, J., Jiang, X. et al.: Training Language Models to Follow Instructions with Human Feedback, *NeurIPS*, Vol. 35, pp. 27730–27744 (2022).

[13] Papineni, K., Roukos, S., Ward, T. and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, *ACL*, pp. 311–318 (2002).

[14] Plummer, B. A., Wang, L. et al.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, *ICCV*, pp. 2641–2649 (2015).

[15] Radford, A., Kim, J. W., Hallacy, C. et al.: Learning Transferable Visual Models from Natural Language Supervision, *ICML*, pp. 8748–8763 (2021).

[16] Rafailov, R., Sharma, A. et al.: Direct Preference Optimization: Your Language Model is Secretly a Reward Model, *NeurIPS*, Vol. 36, pp. 53728–53741 (2023).

[17] Rennie, S. et al.: Self-critical Sequence Training for Image Captioning, *CVPR*, pp. 7008–7024 (2017).

[18] Sarto, S. et al.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, *CVPR*, pp. 6914–6924 (2023).

[19] Shekhar, R., Pezzelle, S., Klimovich, Y. et al.: FOIL it! Find One Mismatch Between Image and Language caption, *ACL*, pp. 255–265 (2017).

[20] Tversky, A. and Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and uncertainty*, Vol. 5, pp. 297–323 (1992).

[21] Vedantam, R., Zitnick, L. and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, *CVPR*, pp. 4566–4575 (2015).

[22] Voigtlaender, P., Changpinyo, S., Pont, J. et al.: Connecting Vision and Language with Video Localized Narratives, *CVPR*, pp. 2461–2471 (2023).

[23] Wada, Y., Kaneda, K., Saito, D. and Sugiura, K.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning, pp. 13559–13568 (2024).

[24] Wu, J., Gan, W., Chen, Z. et al.: Multimodal large language models: A survey, *BigData*, pp. 2247–2256 (2023).

[25] Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H. et al.: Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks, *CVPR*, pp. 4818–4829 (2024).

[26] Ye, Q., Xu, H., Ye, J. et al.: mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration, *CVPR*, pp. 13040–13051 (2024).

[27] Yue, X., Song, Y., Asai, A., Kim, S., Nyandwi, J., Khanuja, S. et al.: Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages, *ICLR* (2025).

[28] 松田一起, 和田唯我, 小槻誠太郎, 杉浦孔明: LLM-Hybrid-as-a-Judge に基づく長文画像キャプション向け自動評価尺度, *MIRU* (2025).