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

松田 一起[1,a)]　和田 唯我[1]　小槻 誠太郎[1]　杉浦 孔明[1,b)]

概要

本研究では，マルチモーダル大規模言語モデル（MLLM）が生成する長文キャプションの自動評価に取り組む．既存の画像キャプション生成向け自動評価尺度の多くは，短文キャプションを対象に設計されており，長文キャプションの自動評価には適さない．一方，大規模言語モデル（LLM）や MLLM を用いた既存の LLM-as-a-Judge 手法は，自己回帰型の推論および画像の早期統合によって，評価が非常に低速である．そこで本研究では，高速かつ人間による評価に近い長文キャプション向け自動評価尺度 VELA を提案する．また，LLM を用いた画像に基づく高速な評価を可能とするため，新たなフレームワーク LLM-Hybrid-as-a-Judge を導入する．さらに，長文キャプション向け自動評価尺度のための，新たなデータセット LongCap-Arena を構築した．本データセットは，画像とそれに対応する長文の参照文および候補文に加え，3 つの観点（Descriptiveness, Relevance, Fluency）から候補文を評価した合計 32,246 個の人間による評価を含む．実験の結果，提案尺度は LongCap-Arena データセットにおいて，既存の自動評価尺度ならびに LLM-as-a-Judge 手法を上回る結果を得た．

1. はじめに

マルチモーダル大規模言語モデル（MLLM）は，高度な言語理解および生成能力により，ロボティクスや医療分野など幅広い領域で社会応用が進められている [1, 35, 23, 24, 21, 11, 5]．特に，画像キャプション生成において視覚情報に基づく詳細かつ正確な記述を生成可能な MLLM は有益である．長文の画像キャプションを生成可能な MLLM を効率的に開発するためには，人間による評価と高い相関を持つ自動評価尺度の構築が不可欠である．しかし，既存の自動評価尺度の多くは，短文キャプションの評価を前提に設計されており，MLLM が生成する平均 100 単語以上で構成される長文キャプションの評価には不適当である．また，自動評価において高い性能を達成している LLM-as-a-Judge [8, 18, 36, 42] は，長文キャプションの入力を扱うことができる一方，推論速度が非常に低速のため実用性に欠ける．

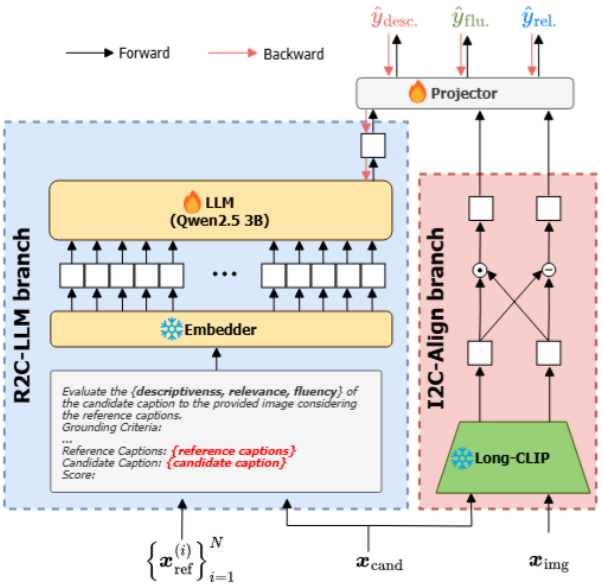

図 1　提案尺度 VELA の全体図

本研究では，MLLM が生成する長文キャプションの自動評価を扱う．本タスクでは，画像と，平均 100 語以上で構成される長文の候補文および人間が付与した参照文が与えられる．自動評価尺度は，画像および参照文をもとに，候補文の記述性（Descriptiveness），関連性（Relevance），流暢性（Fluency）の 3 つの観点から評価を行い，各観点ごとに評価値を出力する．

多くの自動評価尺度は短文キャプションを対象に設計されており，MLLM が生成する，平均 100 単語以上で構成された長文キャプションの評価には十分でない．古典的な評価尺度 [27, 38, 6, 4, 39, 20] は，局所的な表現一致によるスコアリングを行うため，長文キャプションにおいて重要である文の構成や一貫性など大域的な要素を適切に評価することが難しい．既存のデータ駆動型自動評価尺度 [14, 29, 31, 40, 26] は，主に短文キャプションで学習されているだけでなく，2 章でも後述する通り，入力系列長に制

[1] 慶應義塾大学
[a)] k2matsuda0@keio.jp
[b)] komei.sugiura@keio.jp

約があるアーキテクチャであることが多い．さらに，LLM や MLLM に基づく自動評価尺度（LLM-as-a-Judge）は，画像キャプション生成の評価において人間による評価との相関が高い一方で，推論速度が非常に低速で実用性に欠けるという問題がある [8, 18, 36, 42]．その原因の一つとして，LLM-as-a-Judge は LLM を自己回帰的に用いるため，計算コストが高いという点が挙げられる．さらに，MLLM を用いた評価手法 [18, 36, 42] では，画像情報を早期統合するため入力系列長が増大し，推論速度がさらに低下する．実際，これらの手法は標準的なベンチマーク（e.g., [22, 3]）の評価に 3 時間以上を要するため，実用性に欠ける．

そこで本研究では，高速かつ人間による評価に近い長文キャプション向け自動評価尺度 VELA を提案する．図 1 に VELA の全体図を示す．また，LLM を用いた画像に基づく高速な評価を可能とするため，新たなフレームワーク LLM-Hybrid-as-a-Judge を導入する．既存の MLLM [1, 35, 23, 24, 21, 11, 5] が画像情報を早期統合するのとは異なり，LLM-Hybrid-as-a-Judge では画像情報を後期統合することで入力系列長の増大を抑え，高速な評価を実現する．提案尺度は，本フレームワークに基づき，LLM を非自己回帰に用いるブランチと，画像の特徴抽出を行うブランチを後期統合することで，画像を用いた高速な自動評価を行う．さらに，長文キャプション向け自動評価尺度を評価および学習するため，新たなデータセット LongCap-Arena を構築した．LongCap-Arena は，画像とそれに対応する平均 131.4 単語の長文参照文，また平均 101.2 単語の長文候補文 に加え，Descriptiveness, Relevance, Fluency の 3 観点から候補文を評価した合計 32,246 個の人間による評価を含む．

提案手法における新規性は次の通りである．

- 長文キャプションに対して，3 つの観点から人間の評価と相関した評価値の出力を行う教師あり自動評価尺度 VELA を提案する．
- 画像に基づいた高速な LLM ベースの評価を実現するため，R2C-LLM（Reference-to-Candidate LLM）ブランチと，I2C-Align（Image-to-Candidate Alignment）ブランチを後期統合する LLM-Hybrid-as-a-Judge フレームワークを提案する．
- 長文キャプション向け自動評価尺度の学習および評価のためのデータセット LongCap-Arena を構築する．

2. 提案手法

本研究では，詳細な長文画像キャプションを 3 つの観点から評価する自動評価尺度 VELA を提案する．提案手法は，R2C-LLM ブランチ（Reference-to-Candidate）と I2C-Align ブランチ（Image-to-Candidate）の 2 つの主要なブランチで構成される．

2.1 入力

自動評価尺度への入力 $\boldsymbol{x}$ を以下のように定義する: $\boldsymbol{x} = \{\boldsymbol{x}_{\text{img}}, \{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}, \boldsymbol{x}_{\text{cand}}\}$．ここで，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times H\times W}$，$\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N} \in \{0,1\}^{N\times V\times L}$，$\boldsymbol{x}_{\text{cand}} \in \{0,1\}^{V\times L}$ は，それぞれ画像，参照文，候補文を表す．また，H，W，V，L はそれぞれ画像の高さ，幅，語彙数，入力トークン数を示す．自動評価尺度は $\boldsymbol{x}$ を入力として，Descriptiveness, Relevance, Fluency における評価値 $\hat{\boldsymbol{y}} = (\hat{y}_{\text{desc}}, \hat{y}_{\text{rel}}, \hat{y}_{\text{flu}}) \in \mathbb{R}^3$ を出力する．

2.2 R2C-LLM ブランチ

R2C-LLM ブランチでは，軽量な LLM を非自己回帰的に扱い，人間によって付与された $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ に対する $\boldsymbol{x}_{\text{cand}}$ の評価を高速に行う．本ブランチで LLM を用いる理由は，LLM は広範なドメインに基づくデータセットで事前学習されており，LLM-based の評価は LLM-free の評価よりも優れた性能が期待されるためである [8, 18, 36, 42]．ただし，LLM-as-a-Judge は自己回帰に推論し評価速度が遅いため，本ブランチでは非自己回帰型での評価をおこなう．さらに，MLLM を用いた評価手法は，画像情報を早期統合することにより，より推論速度が低下するため，本ブランチでは MLLM を用いない．

本ブランチではまず $\boldsymbol{x}_{\text{cand}}$ と $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ から，評価プロンプト $\boldsymbol{x}_{\text{prompt}}$ を作成する．当プロンプトは，既存研究 [15, 12, 36, 17] における指示文に基づき設計した．次に，$\boldsymbol{x}_{\text{prompt}}$ を LLM (Qwen2.5-3B) の入力として非自己回帰に推論を行い，$\boldsymbol{h}_{\text{lang}} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_M)$ を得る．ここで，$\boldsymbol{h}_{\text{lang}} \in \mathbb{R}^{M\times d}$ は最終層にの隠れ状態を示し，M は入力系列長，d は隠れ状態の次元を表す．先行研究 [7, 34, 33] では，LLM から埋め込みを得る方法として，最終層における隠れ状態の有効性が示されている．そのため，系列全体の平均プーリング特徴量と End-of-Sequence (EOS) トークンに対応する特徴量 $\boldsymbol{h}_M$ を結合した $\boldsymbol{g}_{\text{r2c}}$ を計算し，R2C-LLM ブランチの出力とする．

2.3 I2C-Align ブランチ

I2C-Align ブランチでは，長系列の入力が可能な Long-CLIP [43] を特徴量抽出器として用い，MLLM に依存せずに $\boldsymbol{x}_{\text{img}}$ に対する $\boldsymbol{x}_{\text{cand}}$ の評価を行う．本ブランチで $\boldsymbol{x}_{\text{img}}$ を入力とするのは，多くの既存研究において画像を用いない場合と比べて有効性が示されているためである [14, 29, 31, 30, 40, 26]．一方で，MLLM に依存しない設計としたのは，先述のとおり MLLM における画像情報の処理は計算コストが高いためである [8, 18, 36].

はじめに，Long-CLIP を用いて $\boldsymbol{x}_{\text{cand}}$ と $\boldsymbol{x}_{\text{img}}$ からそれぞれ $\boldsymbol{h}_{\text{cand}}$ と $\boldsymbol{h}_{\text{img}}$ を抽出する．既存の画像を用いる評価指標 (e.g. [14, 29, 31, 30, 40, 26]) で多く用いられている CLIP [28] は入力系列長に 77 トークンの上限があるため，長文キャプションの自動評価においてより有効な

表 1 ベースライン手法との定量的比較結果

	Metrics	TestA			TestB			Inference Time
		Desc.	Rel.	Flu.	Desc.	Rel.	Flu.	
		Kendall's τ_c						[ms]
Image captioning metrics	BLEU	28.6	2.4	25.5	32.0	-10.1	-3.5	0.46
	CIDEr	-7.0	6.7	4.4	4.0	-3.4	1.9	1.3
	CLIP-S	24.5	18.6	25.5	27.3	22.5	24.5	26
	CLIP-S$_{avg}$	-8.6	11.5	3.2	12.8	27.5	28.4	200
	RefCLIP-S	13.4	7.3	9.5	21.2	10.3	10.9	33
	PAC-S	24.8	14.7	23.6	27.6	25.7	23.0	48
	PAC-S$_{avg}$	-7.4	14.6	6.2	6.6	29.2	28.4	360
	RefPAC-S	22.6	19.1	24.9	40.7	29.2	27.9	52
	Polos	28.5	18.1	30.6	41.1	22.4	20.0	33
	DENEB	10.3	18.4	22.2	31.3	35.7	<u>32.6</u>	47
	PAC-S++	29.7	21.4	34.2	28.1	21.9	21.1	36
	PAC-S++$_{avg}$	-7.2	19.4	6.0	14.1	32.4	30.3	270
	RefPAC-S++	25.4	23.3	28.9	40.3	22.2	24.2	40
LLM-as-a-Judge	FLEUR	17.3	2.6	0.5	12.6	10.6	-3.1	1300
	RefFLEUR	21.3	10.3	7.2	28.1	12.3	17.5	1400
	G-VEval	28.3	22.5	18.2	38.1	22.2	19.2	1800
	GPT4o w/o references	<u>54.1±1.0</u>	<u>36.8±6.3</u>	20.9±1.0	43.6±2.0	<u>37.3±3.4</u>	25.2±1.0	1900
	GPT4o w/ references	47.0±1.1	26.2±2.2	<u>35.4±2.9</u>	<u>46.9±2.6</u>	30.4±2.3	25.1±4.3	2000
	VELA (Ours)	**56.4±1.3**	**40.0±1.1**	**57.4±1.3**	**54.0±0.4**	**52.3±1.1**	**39.0±2.3**	260
	Human performance	56.1	46.6	24.5	48.9	52.6	24.4	—

Long-CLIP を利用した．続いて，要素間の絶対差分とアダマール積を用いて，以下のように $\boldsymbol{h}_{\text{cand}}$ と $\boldsymbol{h}_{\text{img}}$ 間の類似度 $\boldsymbol{g}_{\text{i2c}}$ を算出する．

$$\boldsymbol{g}_{\text{i2c}} = [|\boldsymbol{h}_{\text{img}} - \boldsymbol{h}_{\text{cand}}|; \boldsymbol{h}_{\text{img}} \odot \boldsymbol{h}_{\text{cand}}]$$

最終的な評価値 $\hat{\boldsymbol{y}} \in \mathbb{R}^3$ は以下のように計算される．

$$\hat{\boldsymbol{y}} = (\hat{y}_{\text{desc}}, \hat{y}_{\text{rel}}, \hat{y}_{\text{flu}}) = \sigma(\boldsymbol{W}[\boldsymbol{g}_{\text{r2c}}; \boldsymbol{g}_{\text{i2c}}] + \boldsymbol{b})$$

ここで，σ はシグモイド関数を示し，$\boldsymbol{W}$ と $\boldsymbol{b}$ は学習可能なパラメータである．

3. LongCap-Arena データセット

長文キャプションを扱う本タスクでは，人間によって付与された高品質かつ詳細な参照文を評価の基準として用いることが重要である．また，自動評価尺度と人間による評価との相関も不可欠である．そのため，本タスクにおけるデータセットは人間によって付与された長文の参照文と人間による評価を有することが望ましい [2, 15, 17, 40, 26]．しかし，我々の知る限り，人間によって付与された長文の参照文を含み，長文キャプションに対する自動評価尺度の性能を測ることができるデータセットは存在しない [42]．また，人間による評価を含む既存のデータセットは主として短文キャプションに焦点を当てており，Composite [2]，Flickr8k-CF [15]，Polaris [40] における候補文の平均単語数はそれぞれ 12.6 語，11.4 語，9.4 語である．さらに，本分野における標準的なデータセット [2, 15, 32] では，単一の観点のみに基づいた人間による評価が付与されており，評価の解釈性が不十分である．

そこで，本研究では，3 つの観点に基づく人間による評価を含んだ，長文キャプション向け自動評価尺度のための LongCap-Arena データセットを構築した．本データセットは，画像，長文の候補文，人間によって付与された長文の参照文，および Descriptiveness，Relevance，Fluency の 3 観点それぞれにおける人間による評価から構成される．本データセットにおける参照文と候補文の平均単語数はそれぞれ 131.4 語，101.2 語であり，これは既存のデータセット [2, 15, 40, 26] と比べておよそ 10 倍から 15 倍の数である．

画像と参照文は，DCI データセット [37] に含まれるものを使用した．DCI データセットの参照文は SAM [16] が生成したサブマスクに基づく詳細な説明を付与されており，長文キャプションを扱う本データセットの構築に適切だと考えられる．また，候補文は DCI データセットの画像から，代表的な 10 個の MLLM を用いて生成した ([1, 11, 10, 25, 24, 13, 5, 9, 19, 41])．さらに，評価者は，前述の 3 観点において，各候補文を独立に評価した．

LongCap-Arena の訓練・検証集合には，DCI データセットの訓練集合を使用した．また，DCI データセットにおける検証・テスト集合の分割に合わせ，VELA のテスト集合を TestA および TestB セットの 2 つに分割した．TestA は DCI データセットの検証集合における全画像，TestB はテスト集合における全画像をそれぞれ含む．訓練集合，検証集合，TestA，TestB セットは，それぞれ 11,971，1,309，294，324 のサンプルで構成される．

4. 実験結果

4.1 定量的結果

表 1 に，TestA セットおよび TestB セットにおけるベー

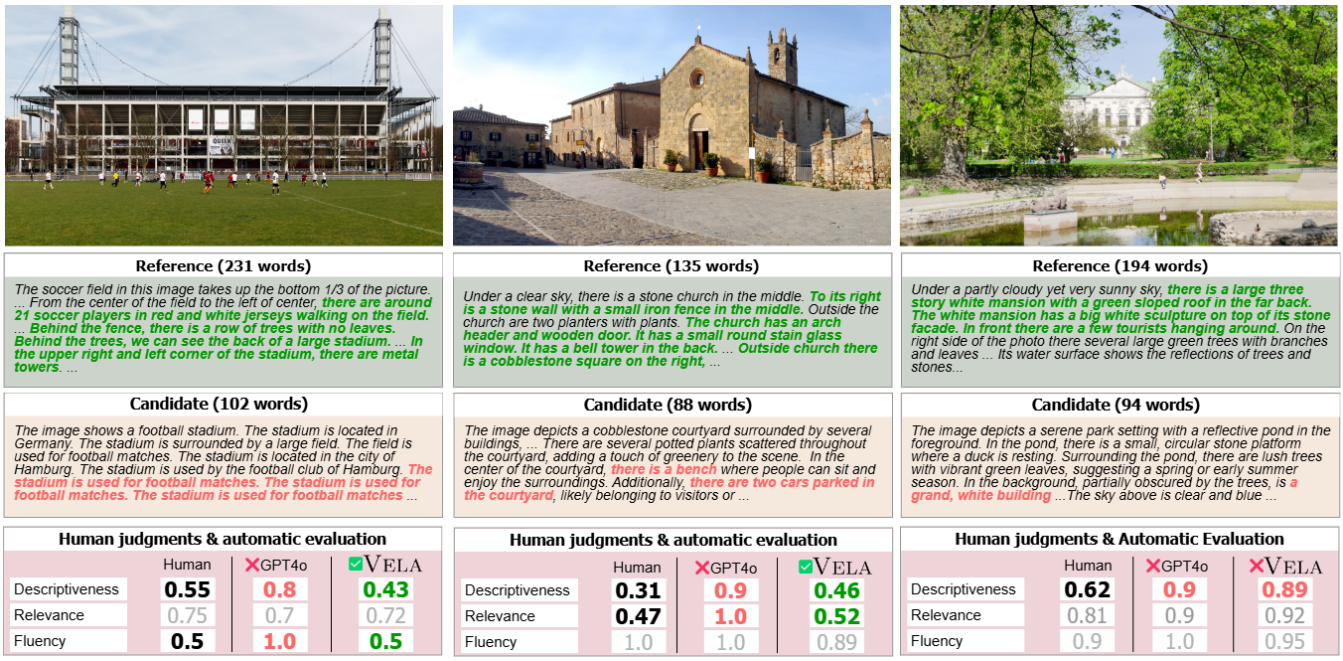

図 2　定性的結果

スライン尺度との定量的な比較結果を示す．本表における “Human performance” は，LongCap-Arena データセットにおける人間の評価性能を示す．画像キャプション生成の既存研究 [14, 29, 31, 30, 42] に従い，評価指標として Kendall’s τ_c を用いた．人手による性能の考察およびベースライン尺度の詳細については付録に記載している．

表 1 より，提案尺度は TestA セットにおいて，GPT-4o の参照文なし・あり設定と比較して，Descriptiveness で 2.3 ポイント，Relevance で 3.2 ポイント，Fluency で 22.0 ポイント上回った．TestB セットでは，すべてのベースライン尺度と比較して，Descriptiveness で 7.1 ポイント，Relevance で 15.0 ポイント，Fluency で 6.4 ポイント上回った．

また，表 1 に，GeForce RTX 3090 GPU および Intel Core i9-10900KF CPU を用いて LongCap-Arena データセット上で測定したサンプルあたりの推論時間を示す．FLEUR, RefFLEUR, G-VEval, GPT-4o といった既存の LLM ベースの尺度は，それぞれ 1280ms，1392ms，1812ms，1905ms と，いずれも 1000ms を超える長い推論時間を示した．これに対し，VELA は 258ms と，既存の LLM ベースの尺度より約 5 倍高速な評価速度であった．

4.2　定性的結果

図 2 に提案尺度の成功例および失敗例を示す．左図および中図は成功例を，右図は失敗例を示す．左図のサンプルにおいて，$\boldsymbol{x}_{\text{cand}}$ は画像における主要な要素を含む一方，$\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ と比較すると詳細な記述に欠けるため，Descriptiveness における人間による評価 y_{desc} は 0.55 であった．GPT-4o は当観点において 0.8 と誤った評価値を出力したのに対し，提案尺度は 0.43 とより適切な評価値を出力した．また，$\boldsymbol{x}_{\text{cand}}$ には冗長な表現 “The stadium is used for football matches” が含まれており，Fluency における人間による評価 y_{flu} は 0.50 であった．当観点において，GPT-4o は 1.0 と誤った評価値を出力したのに対し，提案尺度は 0.50 とより適切な評価値を出力した．

中図のサンプルでは，$\boldsymbol{x}_{\text{cand}}$ が画像の主要な要素を含むものの，$\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ と比較すると細部の記述が不足しているため，Descriptiveness における人間による評価 y_{desc} は 0.31 であった．GPT-4o は当観点において 0.9 と誤った評価値を出力したのに対し，提案尺度は 0.46 とより適切な評価値を出力した．これらの結果から，提案尺度は人間による評価により近い評価値を出力したといえる．右図の失敗例の詳細な分析は付録に記載した．

5.　おわりに

本研究では，マルチモーダル大規模言語モデル（MLLM）が生成する長文キャプションに対して，3 つの観点（Descriptiveness, Relevance, Fluency）から人間の評価と相関した評価値の出力を行う自動評価尺度 VELA を提案した．画像に基づいた高速な LLM ベースの評価を実現するため，R2C-LLM ブランチと I2C-Align ブランチを後期統合する LLM-Hybrid-as-a-Judge フレームワークを提案した．長文キャプションの自動評価尺度の学習および評価のためのデータセット LongCap-Arena を構築した．LongCap-Arena データセットにおいて，VELA が既存の画像キャプション自動評価尺度および LLM-as-a-Judge 手法を上回り，人間による評価との高い相関を示すことを確認した．

謝辞

本研究は，Apple 社の助成を受けて実施された．本研究で述べられた見解，意見，発見，結論および推奨は全て著者らのものであり，明示的または暗黙的を問わず，Apple 社の見解，方針または立場を反映するものではない．また，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] Achiam, J., Adler, S., Agarwal, S. et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2023).

[2] Aditya, S. et al.: From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge, *arXiv preprint arXiv:1511.03292* (2015).

[3] Agrawal, H., Desai, K. et al.: nocaps: Novel Object Captioning at Scale, *ICCV*, pp. 8948–8957 (2019).

[4] Anderson, P., Fernando, B., Johnson, M. et al.: SPICE: Semantic Propositional Image Caption Evaluation, *ECCV*, pp. 382–398 (2016).

[5] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C. and Zhou, J.: Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities, *ICLR* (2024).

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

[7] BehnamGhader, P., Adlakha, V., Mosbach, M. et al.: LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders, *COLM* (2024).

[8] Chan, D., Petryk, S. et al.: CLAIR: Evaluating Image Captions with Large Language Models, *EMNLP*, pp. 13638–13646 (2023).

[9] 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).

[10] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L. et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks, *CVPR*, pp. 24185–24198 (2024).

[11] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P. N. and Hoi, S.: InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning, *NeurIPS* (2023).

[12] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R. and Radev, D.: SummEval: Re-evaluating Summarization Evaluation, *arXiv preprint arXiv:2007.12626* (2020).

[13] Gong, T., Lyu, C., Zhang, S. et al.: MultiModal-GPT: A Vision and Language Model for Dialogue with Humans, *arXiv preprint arXiv:2305.04790* (2023).

[14] Hessel, J., Holtzman, A. et al.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning, *EMNLP*, pp. 7514–7528 (2021).

[15] Hodosh, M., Young, P. and Hockenmaier, J.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, *JAIR*, Vol. 47, pp. 853–899 (2013).

[16] Kirillov, A., Mintun, E., Ravi, N. et al.: Segment Anything, *ICCV*, pp. 3992–4003 (2023).

[17] Lee, H., Yoon, S. et al.: UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning, *ACL*, pp. 220–226 (2021).

[18] Lee, Y. et al.: FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model, *ACL*, pp. 3732–3746 (2024).

[19] Li, J. et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, *ICML*, pp. 19730–19742 (2023).

[20] Lin, C.: ROUGE: A Package For Automatic Evaluation Of Summaries, *ACL*, pp. 74–81 (2004).

[21] Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M. and Han, S.: Vila: On Pre-training for Visual Language Models, *CVPR* (2024).

[22] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R. et al.: Microsoft COCO: Common Objects in Context, *ECCV*, pp. 740–755 (2014).

[23] Liu, H., Li, C. et al.: Visual Instruction Tuning, *NeurIPS*, pp. 34892–34916 (2023).

[24] Liu, H., Li, C., Li, Y. and Lee, Y. J.: Improved Baselines with Visual Instruction Tuning, *CVPR*, pp. 26296–26306 (2024).

[25] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S. and Lee, J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (2024).

[26] Matsuda, K. et al.: DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning, *ACCV*, pp. 3570–3586 (2024).

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

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

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

[30] Sarto, S., Cornia, M. et al.: BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues, *ECCV*, p. 70–87 (2024).

[31] Sarto, S., Moratelli, N. et al.: Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training, *arXiv preprint arXiv:2410.07336* (2024).

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

[33] Springer, J. M., Kotha, S., Fried, D., Neubig, G. and Raghunathan, A.: Repetition Improves Language Model Embeddings, *ICLR* (2025).

[34] Su, H., Shi, W., Kasai, J. et al.: One Embedder, Any Task: Instruction-Finetuned Text Embeddings, *Findings of ACL*, pp. 1102–1121 (2023).

[35] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A. et al.: Gemini: A Family of Highly Capable Multimodal Models, *arXiv preprint arXiv:2312.11805* (2023).

[36] Tong, T. C., He, S. et al.: G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o, *arXiv preprint arXiv:2412.13647* (2024).

[37] Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V. and Romero-Soriano, A.: A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions, *CVPR*, pp. 26700–26709 (2024).

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

[39] Wada, Y. et al.: JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models, *CoNLL*, pp. 424–435 (2023).

[40] Wada, Y., Kanta, K. et al.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning, *CVPR*, pp. 13559–13568 (2024).

[41] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z. et al.: GIT: A Generative Image-to-text Transformer for Vision and Language, *TMLR* (2022).

[42] Yao, Z. et al.: HiFi-Score: Fine-Grained Image Description Evaluation with Hierarchical Parsing Graphs, *ECCV*, pp. 441–458 (2024).

[43] Zhang, B., Zhang, P., Dong, X., Zang, Y. and Wang, J.: Long-CLIP: Unlocking the Long-Text Capability of CLIP, *ECCV*, pp. 310–325 (2024).