Visual Question Answering における 視線情報を用いた質問の曖昧性解消

稲積 駿 [1,2]　河野 誠也 [2,1]　湯口 彰重 [3,2]　川西 康友 [2,1]　吉野 幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学
[2] 理化学研究所ガーディアンロボットプロジェクト [3] 東京理科大学
inazumi.shun.in6@naist.ac.jp akishige.yuguchi@rs.tus.ac.jp
{seiya.kawano,yasutomo.kawanishi,koichiro.yoshino}@riken.jp

概要

ユーザ (話者) と対話システムの会話には，指示語や省略に起因した曖昧さが含まれる．こうした曖昧さは，話者の視線や指差しといった情報で補完することが可能である．本研究では視線情報を用いた話者の質問の曖昧性解消を，Visual Question Answering (VQA) のタスクにおいて実現する．このため，視線情報と曖昧な質問が紐づいた視線情報付き VQA データセット (GazeVQA)，および注視対象の情報を活用した質問応答モデルを提案する．GazeVQA による実験では，提案モデルが注視対象の推定を用いない既存モデルと比較して優れた性能を達成した．

1　はじめに

実世界の事物を考慮してユーザ (話者) と共同作業が可能な対話システムの実現に向けて，画像を交えた質問応答タスク（VQA:Visual Question Answering) [1, 2, 3] がこれまで提案されてきた．既存の VQA における質問はその意図が明確であり，その回答は所与の画像と組み合わせることで一意に決定できる．しかし，実際の話者とシステムの会話には，指示語・省略に起因した曖昧性が含まれており [4, 5]，これらの曖昧性を VQA では陽に扱っていない．こうした問題は，とくに日本語では顕著である [6, 7]．例えば，「それ取ってきてくれる?」という質問は「それ」という指示語が原因で複数の解釈を持つ可能性がある．こうした指示語や省略の曖昧性は実世界の情報を参照することで解消可能な場合が多い．例えば，話者の視線情報 [8] や指差し [9]，あるいは共同注視 [10] は，指示語・省略の参照先を明らかにする重要な手がかりである．

本研究では，視線情報を用いた質問の曖昧性解消を目的として，視線情報付き VQA データセット (GazeVQA) を提案する．図 1 に示すように，GazeVQA は画像内の人物を話者とみなし，話者が発する指示語・省略が含まれる曖昧な質問に対し，話者の視線情報を考慮してシステムが回答する状況を想定する．Gazefollow [11] に含まれる一般物体認識用の画像データセット (COCO) [12] を対象に，注視対象に関する質問と回答をクラウドソーシングにて収集した．

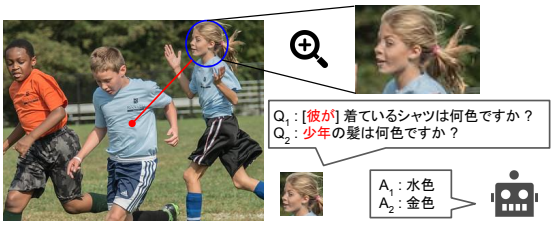

図 1　視線情報付き VQA データセットの質問と回答の例．角括弧は省略された物体名．視線元に対応する視線先の点が複数個付与されている．

本研究ではさらに，画像・質問に加え視線情報を利用するモデルを提案する．テキスト・画像をクエリとした物体セグメンテーションの研究 [13] に着想を得て，既存モデル [14] に線形層のアダプタ [15] を追加する．アダプタは注視対象を表す関心領域と全体画像を統合する役割を持ち，モデルは注視対象部分に焦点を置くことが可能となる．実験の結果，提案モデルは GazeVQA における特定の質問タイプに精度良く回答ができ，ベースラインと比較して優れた性能を達成した．

2　関連研究

2.1　文脈情報付き VQA データセット

VQA は画像に関する質問に対してシステムが回答を与えるタスクである [1, 2, 3]．本研究は視線情

報を持たない場合に曖昧になるような質問を研究対象としており，注視対象の物体名が明記されていない質問を意図的に収集した．

画像と質問に加え多様な文脈情報を収録したVQA データセットが提案されてきた [16, 17, 18, 19, 20]．VQA-HAT [16] や VQA-MHUG [17] は，質問を解く際に生じる人間の主観的な注視情報を用いることで，VQA タスクの精度改善を図った．Point and Ask [18] は，画像に与えられた点情報を用いて，代名詞を含む質問の曖昧性解消に焦点を当てている．GazeVQA は，追加の文脈情報として画像情報から得られる話者の注視情報を利用する点と日本語特有の主語や目的語の省略を含む質問を収録した点において，これらの先行研究とは異なる．

2.2 注視対象推定

注視対象推定は画像に映る人物の頭部画像から，その人物の注視先を推定するタスクである [11, 21, 22]. Gazefollow [11] は，人物の視線元と視線先のアノテーションが付与された注視対象推定のデータセットである．COCO [12] を含む様々な画像データセットから収集した人物を対象にしている．Gazefollow に含まれる視線情報は，視線先が常に物体あるいは注視対象の具体的な名称と紐付いていない．そこで本研究では，Gazefollow の COCO サブセットの物体アノテーションを利用して，視線先の物体に関する質問と回答を収集した．また，実際に注視対象推定を行う際は，Gazefollow の視線元に対応付けられた人物の頭部画像を利用した [21].

3 視線情報付き VQA データセット

3.1 タスク設定

GazeVQA では質問をする話者がシステムの一人称視点画像に映っていることを想定している．GazeVQA のタスクは画像，話者の質問，および話者の視線情報を考慮して回答を生成することを目標とする．本タスクを以下のように定義する．

視線情報付き VQA タスク　全体画像 I，質問 $\mathbf{q}$，および注視領域 I_s から，回答 $\mathbf{y}$ を出力する．以後，このタスクを GazeVQA タスクと呼ぶ．

I_s を取得するための注視対象推定タスクを次のように定義する．

注視対象推定タスク　全体画像 I と発話者の頭部画像 I_h から I_s を取得する．

3.2 データ収集

Gazefollow の COCO サブセットに含まれる画像 I と物体情報から質問 $\mathbf{q}$ と回答 $\mathbf{y}$ をクラウドソーシングにより収集した．収集手順を以下に示す．

Step1: 画像と視線情報の選定　質問と回答を収集する前に，前処理として視線情報の選別を行い，機械的に 14,000 件の画像・視線情報ペアを選定した．COCO の物体セグメンテーションを利用し，視線先が物体を差していないケースを除いた．

Step2: 質問と回答の収集　クラウドソーシングにより，26,296 件の質問・回答を収集した．視線情報が付与された画像と COCO の物体ラベルをもとに，ワーカは視線先の物体に関する質問とその回答を作成する．視線先の物体名は質問文中に含めないようワーカに依頼することで，回答に視線情報を要する質問を収集した．

Step3: 質問と回答の選別　画像と対応していない質問を除くべく，17,276 件の質問・回答を選定した．Step2 の作業中に，設問 (「本項目には何も入力しないでください」) を配置し，この設問に対して回答を与えたワーカの質問を手作業で確認した．この結果，作業を依頼した 246 名のワーカから，27 名のワーカのアノテーションを全て除去した．

Step4: テストセットの整備　GazeVQA の訓練・開発・テストセットを，13,785・1,811・1,860 (0.8 : 0.1 : 0.1) とした．先行研究 [1] にならって，回答の多義性・多様性を GazeVQA の評価で考慮するためテストセットに回答を 10 件割り付けた．具体的には，視線元と曖昧な質問 (AQ) のみをワーカに与え，テストセットの質問 1 件につき，9 人のワーカが回答を追加で作成した．さらに，アノテータが，質問・回答・視線情報を参考に，注視対象の名称や特徴が補完された明確な質問 (CQ) を，テストセットの質問全てに割り付けている．

3.3 従来データセットとの比較

表 1 より，GazeVQA[1)] と日本語 VQA データセット (VQA-ja) [3] の統計量を比較する．GazeVQA のユニークな質問の割合 (46.46%) は VQA-ja の割合 (45.21%) よりも多く，質問の平均文字長は GazeVQA の方が長い．また，GazeVQA のユニークな回答の割合 (33.87%) は日本語 VQA (17.10%) より多く，回答の平均文字長は GazeVQA の方が長い．

1) 3.2 節の Step3 で得た質問と回答を対象とする．

表 1　GazeVQA と日本語 VQA [3] (VQA-ja) の統計情報.

	GazeVQA	VQA-ja
画像数	10,760	99,208
質問・回答ペア	17,276	793,664
ユニークな質問数	8,628	358,844
ユニークな回答数	5,853	135,743
質問の平均文字長	15.37	14.82
回答の平均文字長	4.92	4.56

4　ベースラインと提案モデル

GazeVQA タスクのベースラインとなる ClipCap [14]，ベースラインにアダプタを追加した提案モデル (ClipCap + Adapter)，および注視領域を得るための注視対象推定プロセスを解説する.

4.1　ベースライン： ClipCap

ClipCap は事前学習済みの画像エンコーダ [23]・言語デコーダ [24] で構成された画像キャプションモデルであり [14]，VQA タスクに対しても適用事例がある [25].

画像エンコーダ　ClipCap の画像エンコーダは，RGB の全体画像 $I \in \mathbb{R}^{W\times H\times 3}$ を入力として，画像系列 $\mathbf{r} = \{r_1, \ldots, r_n\}$ を出力する．ただし，n は $\mathbf{r}$ の系列長であり，$\mathbf{r}$ の要素 r_i は質問 $\mathbf{q}$ のトークン埋め込みと同次元である．CLIP [23] の画像エンコーダと単一の線形層 f により，I を画像系列 $\mathbf{p} = \{p_1, \ldots, p_n\}$ に変換する (式 1).

$$\{p_1, \ldots, p_n\} = f(\mathrm{CLIP}(I)) \tag{1}$$

多層の Transformer ブロック [26](F) により $\mathbf{p}$ を同次元の画像系列 $\mathbf{r}$ に変換する (式 2).

$$\{r_1, \ldots, r_n\} = F(\{p_1, \ldots, p_n\}) \tag{2}$$

CLIP の画像エンコーダとデコーダを結ぶ，これらの Transformer ブロックを Mapping Network と呼ぶ.

言語デコーダ　言語デコーダは，質問 $\mathbf{q} = q_1, \ldots, q_m$ と画像系列 $\mathbf{r}$ から構成した入力系列 (式 3) から，回答系列 $\mathbf{y}$ を自己回帰的に出力する.

$$\{r_1, \ldots, r_n, [\mathrm{SEP1}], q_1, \ldots, q_m, [\mathrm{SEP2}]\} \tag{3}$$

ここで，[SEP1] と [SEP2] は，それぞれ “質問:”と “回答:”であり，デコーダへのプロンプトを表す.

4.2　提案モデル： ClipCap + Adapter

提案モデルは，Mapping Network の各ブロックの先頭に挿入されたアダプタ [15] により，全体画像

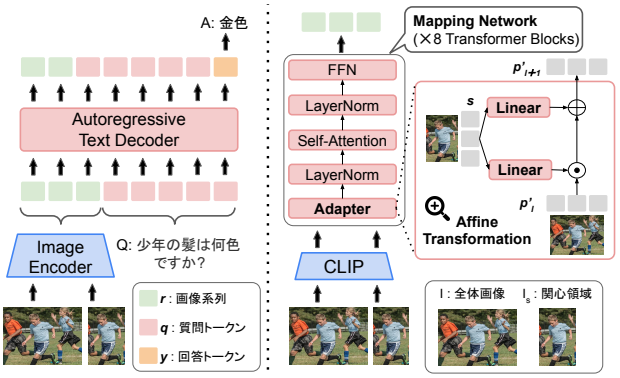

図 2　左: 提案モデルの概要図．右: 提案手法の詳細.

I と注視領域 I_s を統合する．I_s を考慮した画像系列を言語デコーダの入力として，視線情報を要する曖昧な質問に対し，適切な回答を得ることを期待する．具体的には，ClipCap の画像エンコーダの処理 (式 1) と同様に，I と I_s から，画像系列 $\mathbf{p}$ と注視領域の画像系列 $\mathbf{s} = \{s_1, \ldots, s_n\}$ を構成する．$\mathbf{p}$ と $\mathbf{s}$ の要素ごとのアフィン変換[2] を計算し，それぞれの系列を Mapping Network に入力する (式 4).

$$\mathbf{p}'_{l+1} = g(\mathbf{s}) \odot \mathbf{p}'_l \oplus h(\mathbf{s}), \tag{4}$$

ただし，g と h は 1 層の線形層，$\mathbf{p}'_l$ は l 層目の Transformer ブロックの入力[3] である.

4.3　注視対象推定

アダプタへの入力となる注視領域 I_s を視線元の頭部画像 I_h から取得する．CNN ベースのモデル [22] により，I と I_h から，顕著性マップ S を出力する．閾値 0 として S を二値化し，注視対象のバウンディングボックスを求めて I_s を取得する[4] [27].

5　実験

5.1　実験設定

データセット　日本語 VQA データセット (VQA-ja) [3]，日本語画像キャプションデータセット (STAIR) [28] で VQA モデルを事前学習し，GazeVQA でファインチューニングした.

評価指標　回答の多義性・多様性を考慮した VQA スコア (Acc) [1] と回答フレーズの類似を考慮した BERT スコア (Bs) [29] を評価に用いた.

2)　要素ごとの積 ($\odot$) と和 ($\oplus$) を意味する.
3)　1 層目のブロックの入力は，$\mathbf{p}'_l = \mathbf{p}$ である.
4)　S の各要素が全て 0 で I_s の取得が困難である場合は擬似的に I を I_s とみなす.

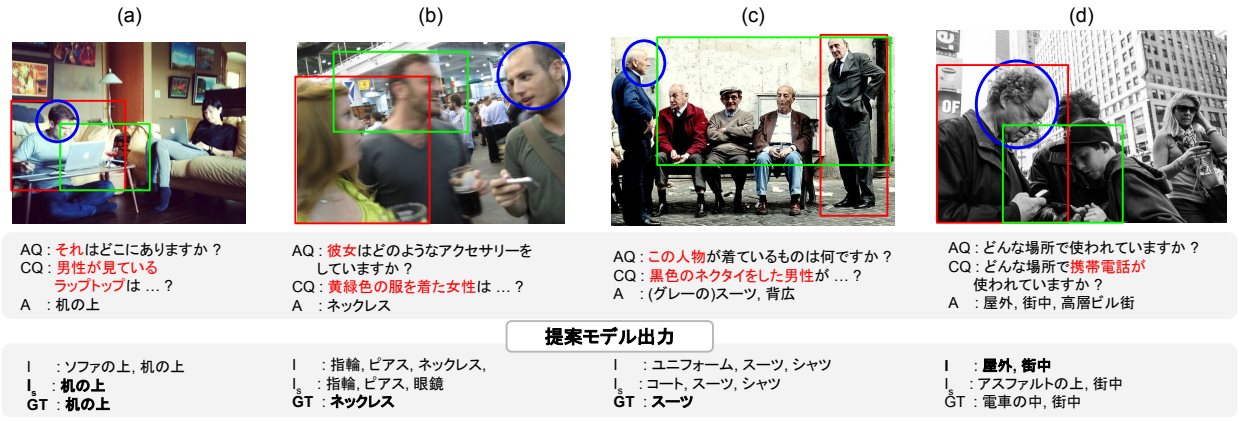

図 3 提案モデルとベースラインの出力例．AQ，CQ，A はそれぞれ曖昧な質問，明確な質問，人手回答．VQA スコアが最も高いモデルの出力を太字で，*GT* と I_s の範囲をそれぞれ赤枠，緑枠で示す．

5.2 定量評価

提案モデル 対 ベースライン 表 2 より，アダプタに関心領域 (I_s) を入力した提案モデル (ClipCap + Adapter (I_s)) とベースライン (ClipCap) を比較する．Mapping Network と言語デコーダを訓練した場合，提案モデルはベースラインの性能を下回った．Mapping Network のみ，Adapter のみを訓練した場合，提案モデルの性能がベースラインの性能を上回った．ベースラインと比較して，提案モデルは 16M 程度のパラメータ更新で，曖昧な質問に対して精度良く回答を生成できることが判明した．

GazeVQA の精度向上に寄与する要素 表 2 より，Mapping Network のみを訓練した提案モデルとベースラインを比較する．アダプタに画像 (*I*) を入力した提案モデル (ClipCap+ Adapter(*I*)) は，ベースラインと比較して性能が良く，関心領域やその正例[5] (*GT*) を入力とした提案モデル (ClipCap+ Adapter(I_s, *GT*)) と比較して性能に差が無い．アダプタを追加することによる訓練パラメータの増加が，GazeVQA タスクの精度向上の一因であると言える．

5.3 定性評価

アダプタへの入力の差異が提案モデルの結果に与える影響を確認する．アダプタに *GT* を入力した場合，物体の形状や名称など，注視対象の属性を問う曖昧な質問に対して，その回答が一意に定まる傾向にあった．正確な回答を与えるに資する視覚情報が *GT* に存在する場合，その傾向が顕著に現れる (図 3(a,b))．注視対象を絞り込めていない場合，モデルは一貫性の無い回答を出力する (図 3(c))．アダプタに *I* を入力した場合，画像全体の理解を要する質問に対し正確な回答を与える傾向にあった (図 3(d))．

5) COCO [12] のバウンディングボックスを意味する．

表 2 提案モデルとベースラインの評価結果．GazeVQA での学習・評価を 5 回繰り返した際の平均値を報告する．$|\theta|$ はモデルの訓練可能なパラメータ数 [M].

Model	$\|\theta\|$	Acc	Bs
Fine-tuned Text Decoder & Mapping Network			
ClipCap	410	**36.80**	**81.75**
ClipCap + Adapter (*I*)	426	34.78	81.39
ClipCap + Adapter (I_s)	426	34.15	81.28
ClipCap + Adapter (*GT*)	426	34.72	81.33
Fine-tuned Mapping Network			
ClipCap	74	35.83	81.21
ClipCap + Adapter (*I*)	90	**38.45**	**81.74**
ClipCap + Adapter (I_s)	90	38.11	81.71
ClipCap + Adapter (*GT*)	90	38.01	81.70
Fine-tuned Adapter Only			
ClipCap + Adapter (*I*)	16	40.06	81.91
ClipCap + Adapter (I_s)	16	39.03	81.92
ClipCap + Adapter (*GT*)	16	**40.09**	**82.01**

6 おわりに

本研究では，人間の会話に生じる曖昧さの問題に視線情報を用いて対処することを目的として，視線情報付き VQA データセット，および視線情報を活用するモデルを提案した．GazeVQA の質問は，指示語・省略に起因する曖昧さを含み，回答に話者の視線情報を要する．GazeVQA による実験の結果，提案モデルがベースラインよりも GazeVQA タスクの性能が良いことが示された．とりわけ，視線情報を用いた提案モデルは，注視対象の属性に関する曖昧な質問に対して正確な回答を与えることができた．

謝辞

本研究はJSPS科研費JP22H04873の助成を受けた.

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In **ICCV**, pp. 2425–2433, 2015.

[2] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In **CVPR**, pp. 6904–6913, 2017.

[3] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In **COLING**, pp. 1918–1928, 2018.

[4] Tadahiro Taniguchi, Daichi Mochihashi, Takayuki Nagai, Satoru Uchida, Naoya Inoue, Ichiro Kobayashi, Tomoaki Nakamura, Yoshinobu Hagiwara, Naoto Iwahashi, and Tetsunari Inamura. Survey on frontiers of language and robotics. **Advanced Robotics**, Vol. 33, No. 15–16, pp. 700–730, 2019.

[5] Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Natural deictic communication with humanoid robots. In **IROS**, pp. 1441–1448, 2007.

[6] Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In **COLING**, 2002.

[7] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In **COLING**, pp. 769–776, 2008.

[8] Nathan J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. **Neuroscience & Biobehavioral Reviews**, Vol. 24, No. 6, pp. 581–604, 2000.

[9] Shu Nakamura, Yasutomo Kawanishi, Shohei Nobuhara, and Ko Nishino. DeePoint: Visual pointing recognition and direction estimation. In **ICCV**, 2023.

[10] Roberta Rocca, Mikkel Wallentin, Cordula Vesper, and Kristian Tylén. This and that back in context: Grounding demonstrative reference in manual and social affordances. In **CogSci**, 2018.

[11] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In **NIPS**, Vol. 1, pp. 199–207, 2015.

[12] Tsung Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In **ECCV**, pp. 740–755, 2014.

[13] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In **CVPR**, pp. 7086–7096, 2022.

[14] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: CLIP prefix for image captioning. arXiv:2111.09734, 2021.

[15] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. **Distill**, Vol. 3, No. 7, p. e11, 2018.

[16] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In **EMNLP**, pp. 932–937, 2016.

[17] Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. In **CoNLL**, pp. 27–43, 2021.

[18] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. arXiv:2011.13681, 2020.

[19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In **CVPR**, pp. 326–335, 2017.

[20] Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In **ACL**, pp. 8182–8197, 2020.

[21] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In **ECCV**, pp. 397–412, 2018.

[22] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In **CVPR**, pp. 5395–5405, 2020.

[23] 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. In **ICML**, Vol. 139, pp. 8748–8763, 2021.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[25] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In **ECCV**, pp. 146–162, 2022.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, pp. 6000–6010, 2017.

[27] Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola. Saliency based image cropping. In **ICIAP**, Vol. 8156, pp. 773–782, 2013.

[28] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In **ACL**, Vol. 2, pp. 417–421, 2017.

[29] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **ICLR**, 2020.

[30] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In **ICML**, Vol. 97, pp. 6105–6114, 2019.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **ICLR**, 2019.

表 3 GazeVQA の質問タイプの類型. 各質問は日本語表現に含まれる単語にもとづいて分類した.

質問タイプ	日本語表現	質問数
What	-	14,141
is/are/do/does	何, どんな, どの	7,215
color	+ 色	3,626
condition	+ 状態	1,240
kind	+ 種類	903
shape	+ 形	703
others	-	454
Where	どこ	1,085
How	どれ, いくつ	996
Which	どちら, どっち	295
Others	いつ, 誰, なぜ	875

A GazeVQA の分析

A.1 統計的分析

表 3 に GazeVQA に含まれる質問タイプの類型を示す. 表 3 より，GazeVQA に含まれる “what”形式の質問の割合は 81.85%である. Shimizu らの文献 [3] の Table 1 より，VQA-ja に含まれる “what”形式の質問の割合は 67.10%であることから，GazeVQA は VQA-ja よりも “what”形式の質問の割合が 1 割ほど多い. 物体の位置関係を問う “where”形式の質問や物体の個数を問う “how”形式の質問の割合は，あわせて 12.04%である. GazeVQA には，その他に，現在時刻を問う “when”タイプの質問や人物に関する “who”タイプの質問も収録されている.

A.2 定性的分析

GazeVQA における曖昧な質問とその回答の特徴を明確にするため，GazeVQA のテストセットに対する事例分析を実施する. 図 4 に GazeVQA のテストセットに含まれる質問と回答の実例を示す. 指示語による曖昧さを持つ質問 (図 4 (a))，日本語特有の省略による曖昧さを持つ質問 (図 4 (b)) であってもワーカ間の回答は一致する. さらに，注視対象の候補が 2 つ以上ある場合でも，質問の内容を考慮することで，その回答は一つに決定される (図 4 (c)). しかし, 視線情報や質問文の内容を考慮しても, 複数ワーカ間の回答が一致しない質問は一部存在する (図 4(d)).

B 実験設定の詳細

B.1 訓練対象

表 2 では，Mapping Network と言語デコーダのパラメータ[6] を訓練の対象とした結果を報告した. ベースラインは 410M，アダプタは 16M，提案モデルは 426M の訓練可能なパラメータを持つ. また，明示的にアダプタの重みを更新するべく，Mapping Network のみを訓練する場合およびアダプタのみを訓練する場合の結果も報告した. Mapping Network のみを訓練した場合，ベースラインは 74M，提案モデルは 90M の訓練可能なパラメータを持つ.

6) 機械学習モデルにおけるテンソルの要素数の総和を指す.

図 4 GazeVQA のテストセットの実例.AQ・太字の回答は 3.2 節の Step3 で得た質問・回答である.CQ はアノテータの作業により明確化した質問である.

B.2 実装詳細

CLIP の画像エンコーダは，ResNet ベースの $RN \times 4$ [30] を利用した[7]. 画像 I と関心領域 I_s は，CLIP の正規化と同様の処理を行う. このため，CLIP の入力は，縦横 224 画素にリサイズされた画像であり，出力は 640 次元のベクトルである. Mapping Network を 8 層の Transformer ブロックで構成し，画像系列 $\boldsymbol{p}$，$\boldsymbol{s}$，$\boldsymbol{r}$ の系列長 $n = 10$ とする. 言語デコーダは事前学習済み GPT-2 [24] を利用した[8].

実験では，STAIR [28] に収録された 123,287 枚の画像と 616,435 件のキャプション,VQA-ja [3] に収録された 99,208 枚の画像と 793,664 件の質問・回答ペア，GazeVQA によりモデルを学習した. 具体的には，バッチサイズを 32，学習率を 2e-5, オプティマイザを AdamW [31] として，各データセットで 10epochs ずつ学習した. GazeVQA の回答を推論する際は，デコーディングにサイズ 10 の Beam Search を用いた.

B.3 評価指標

式 5 に VQA スコア [1] の定義を示す.

$$Acc = \min(\frac{cnt}{3}, 1) \tag{5}$$

ここで,cnt は回答セットに含まれる 10 件の人手回答とモデルの予測が完全一致した回数を意味する. BERT スコアに関して，本研究では多言語 BERT[9] の分散表現を評価に利用した.

7) https://github.com/openai/CLIP
8) https://huggingface.co/rinna/japanese-gpt2-medium
9) https://huggingface.co/bert-base-multilingual-cased