

# 半自己回帰的に拡散モデルを活用するトレースベースの意図反映キャプション生成

平野理子<sup>1</sup> 小林一郎<sup>1</sup>

<sup>1</sup> お茶の水女子大学

{hirano.satoko, koba}@is.ocha.ac.jp

## 概要

本研究ではユーザが画像をなぞった軌跡（トレース）が示すユーザ固有の説明意図を生成文に反映する画像キャプション生成手法を提案する。トレースの密集度から関心領域を特定し、座標変化に基づいて説明順序を決定、滞在時間を用いて各領域への関心度を評価する。これらの情報を基に拡散言語モデルを半自己回帰的に用いることで、文長や説明順序を制御しつつ、流暢性の向上を図った。実験の結果、提案手法は自己回帰モデルを含む既存手法より10%以上高い精度を達成し、内容・順序・詳細さの3観点でユーザの意図を忠実に反映できることを確認した。

## 1 はじめに

画像キャプション生成は、画像の内容を自然言語で説明するマルチモーダル理解において重要なタスクである。近年、画像認識と自然言語処理の発展により、高品質なキャプション生成が可能となった。しかし生成されるキャプションは多くの場合、画像全体の概要説明にとどまり、ユーザーが注目する特定の領域や詳細について、より精緻な説明を得ることは難しい。この課題に対し、生成キャプションを制御可能にする Controllable Image Captioning の研究が進められているが、既存手法の多くはバウンディングボックスや文長など限定的な制御にとどまる [1, 2, 3, 4]。本研究では、ユーザが画像上をなぞったトレース情報を制御信号として活用し、座標や滞在時間などの多様な情報を基に、個人の嗜好や視点を反映した柔軟なキャプション生成を提案する。近年、拡散過程を用いた生成モデルが従来の敵対的生成ネットワークを超える性能で高品質な画像生成を実現している [5, 6]。Liら [7] は、連続的な情報を扱う拡散モデルを離散的な自然言語生成に適

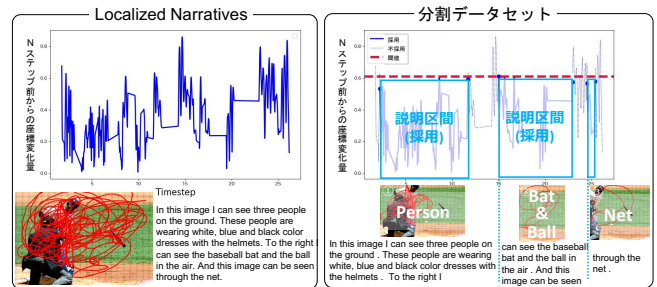


図1 Localized Narratives と構築した分割データセットの例

用し、条件付きテキスト生成において優れた性能を示した。この成果を契機に拡散言語モデルを自然言語処理タスクに応用する研究が急速に進展している [8, 9, 10, 11]。また非自己回帰型生成を基本とする拡散言語モデルの拡張により、半自己回帰および自己回帰型生成が可能となり [12, 13]、トークン間の依存関係の考慮による生成文の流暢性向上が期待されている。さらに、拡散モデルが状態遷移を伴う潜在変数モデルである特性を活かし、分類器などの補助モデルから得られる勾配を利用することで、言語モデルの追加学習を行わずに多様な属性を持つテキスト生成が可能である [14, 12, 10]。本研究では、拡散言語モデルを用いた半自己回帰型キャプション生成手法を提案する。この手法では、トレース情報に基づく領域指定と、文脈を考慮した段階的な生成により、ユーザーが注目する特定領域に関する詳細かつ自然なキャプション生成を実現する。また、自己回帰型と非自己回帰型を組み合わせたハイブリッド設計により、流暢性と制御性の両立を目指す。

## 2 提案手法

非自己回帰型手法 [15] における流暢性不足を改善するため、トレースが示すユーザの関心領域ごとに順次生成を行う半自己回帰型手法を提案する。まずは提案モデルの学習および推論に使用する独自の

データセットを構築する。

## 2.1 分割データセット構築

マウスで画像をなぞりながら口頭で説明を加える注釈が行われた画像アノテーションデータセットである Localized Narratives(LN) [16] を基に「分割データセット」を構築する (図 1)。構築の流れは以下の通りである:

1. 各タイムステップにおいて  $N$  ステップ前からのペン先の座標変化量を計算。  $N$  は全座標データ数の 10 分の 1 に設定。
2. 座標変化量が全体の 90 パーセント値を下回る区間は特定領域の説明区間とみなし採用。一方、上回る区間は領域間の移動と判断し不採用とする。
3. 極端に短い採用区間は情報が乏しいため、不採用とする。
4. 最終的な採用区間について四隅の座標を算出し、長方形の BBox(Bounding Box) を切り出す。
5. 各 BBox に、区間冒頭までに発語された単語群を対応するキャプションセグメントとして割り当てる。

表 1 各データセットの精度

	CLIPScore [17]
Localized Narratives	22.09
分割データセット	<b>22.40</b>

**分割データセットの評価** 構築した分割データセットを定量的に評価するため CLIPScore [17] を導入する。CLIPScore は CLIP(言語と画像のマルチモーダルモデル) [18] を用いて画像とテキストの一致度を評価する指標である。元のデータセット LN については、画像全体と発話キャプション全体間の CLIPScore を全画像について計算し、その平均値を用いて評価した。一方、分割データセットについては、抽出した各 BBox と対応するキャプションセグメント間の CLIPScore を計算し、その平均値を各画像のスコアとした上で全データの平均値を求めて評価する。表 1 より、分割データセットにおける CLIPScore は LN を上回っており、トレース情報に基づくデータの分割によって画像とキャプションの関連性が強化されたことが確認できる。

## 2.2 提案モデル

提案モデルは拡散過程に基づく言語モデルと分類器の二つの要素から構成される。二つの要素は同時

に学習され、生成時は拡散言語モデルによる自然言語文生成過程を分類器で制御することで、画像の内容に応じた自然言語文、つまりキャプションの生成を実現する。

### 2.2.1 拡散言語モデル

DLM(Diffusion Language Model) は拡散過程を用いた非自己回帰型の言語モデルであり、標準的な連続状態を扱う拡散モデルに対して、埋め込みと丸め込みの過程を導入することで構築される。DLM の生成過程は以下の式 (1) で表され、完全なノイズから徐々にノイズを取り除くことでデータを生成する潜在変数モデルである。

$$p(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^T p(\mathbf{x}_{t-1} | \mathbf{x}_t) \quad (1)$$

**半自己回帰型拡散言語モデル SSD-LM** (Semi-autoregressive Simplex-based Diffusion Language Model) [12] は半自己回帰的にテキストを生成する拡散言語モデルで、トークンのまとまりを左から右に順番に生成する。まとまり内で前後のトークンを考慮した状態遷移を繰り返すため、自己回帰モデルと拡散モデルの両方の利点を兼ね備えている。概念的には、SSD-LM は拡散モデルを使用して完全なノイズと前文脈  $w^{context}$  が与えられたら、その続きに相当する長さ  $B$  のトークンのまとまりを生成し、新しい文脈  $w^{context+B}$  として次の生成ステップに渡す。

### 2.2.2 分類器

分類器の役割は、拡散言語モデルが推論過程で生成する潜在変数を勾配更新し、最終的に生成される自然言語文が条件を満たすように制御することである。ノイズを徐々に除去しながらサンプリングする過程を条件  $c$  に基づいて制御する分布  $p(\mathbf{x}_{t-1} | \mathbf{x}_t, c)$  は以下のように分解できる [14]。

$$p(\mathbf{x}_{t-1} | \mathbf{x}_t, c) \propto p(\mathbf{x}_{t-1} | \mathbf{x}_t) \cdot p(c | \mathbf{x}_{t-1}) \quad (2)$$

右辺第 1 項は各タイムステップでのノイズ除去を表し DLM によってパラメータ化される。一方、第 2 項はノイズの乗った状態  $\mathbf{x}_{t-1}$  から制御条件  $c$  への変換を表し、分類器の学習対象である。

### 2.2.3 学習過程

分類器は拡散言語モデル内で得られる値を用いるため、両者は構築した分割データセットを使って同

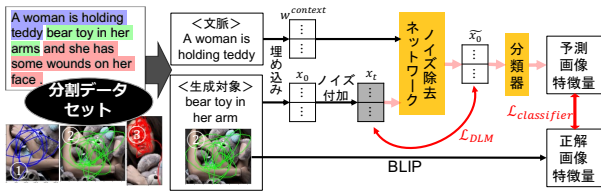


図2 拡散言語モデルと分類器の同時学習の流れ

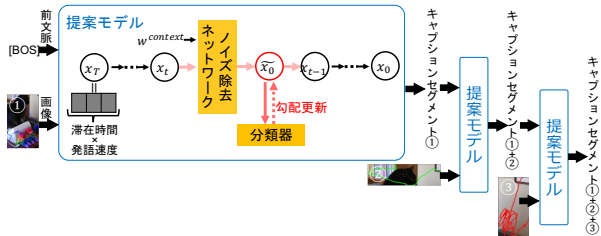


図3 半自己回帰型意図反映キャプション生成の流れ

時に学習される (図 2). 分割データセット内の任意の BBox と対応するキャプションセグメントを生成対象のテキスト, それまでに発語されたキャプションを前文脈  $w^{context}$  と設定する. まず, この前文脈と生成対象をトークナイザーを通して埋め込み表現を生成し, 生成対象埋め込みのみにサンプリングされたタイムステップ  $t$  によって決められる量のノイズを乗せ, ノイズの乗った状態  $x_t$  を得る. これらをノイズ除去ネットワーク  $f(x_t, t, w^{context})$  に入力し元の状態を復元  $\hat{x}_0$  し, 以下の損失関数 (3) で拡散言語モデルと分類器を同時に最適化する:

$$L = \sum \|x_0 - f(x_t, t, w^{context})\|^2 + \sigma \|c - \hat{c}\|^2 \quad (3)$$

ここで,  $\hat{c}$  は  $f$  によって予測された言語特徴量  $\hat{x}_0$  から分類器が線形回帰した対応する BBox の画像特徴量,  $c$  は正解画像特徴量,  $\sigma$  は言語モデルと分類器の損失を調整する重みである.

## 2.2.4 生成過程

本手法では, 各 BBox を説明するキャプションを半自己回帰的に生成する (図 3). まず生成キャプションの長さを個人の特性に応じて決定する. 発語キャプションの総単語数を採用トレース区間の滞在時間合計で割った値をアノテーターごとに算出し, トレース 1 秒あたりの発語単語数を解析する. この値と各 BBox での滞在時間の積を基に生成単語数を決定し, その長さに対応するガウシアンノイズを初期状態  $x_T$  として用意する. 半自己回帰生成の初期前文脈は開始を示すスペシャルトークンとし, トークナイザーに入力して埋め込み表現に変換する. 初期状態  $x_T$  を基点とし, タイムステップ  $t = T$  から 1

まで順次ノイズ除去を行う. 具体的には  $x_t$  と前文脈  $w^{context}$  をノイズ除去ネットワークに入力し, ノイズが除去された状態  $\hat{x}_0$  を推定する. この  $\hat{x}_0$  を分類器に入力し, 制御条件である BBox の画像特徴量  $\hat{c}$  を予測し, 正解画像特徴量  $c$  との損失から勾配を計算して  $\hat{x}_0$  を更新する. この更新状態を基に次の状態  $x_{t-1}$  をサンプリングし, これを繰り返すことで条件画像領域に沿ったキャプションを生成する. 各 BBox に対して生成されたキャプションは, 次の BBox のキャプション生成時の前文脈として逐次的に蓄積される. 最終的に, ユーザの言及した物体を指定した順序通りに, 重きを置いた物体に関してはより詳細に説明する形で, ユーザの説明意図を反映したキャプション生成を実現する.

## 3 実験

### 3.1 実験設定

**データセット** まず Microsoft COCO<sup>1)</sup> の画像キャプションデータセットを事前学習用に用いる. 続いて, Microsoft COCO の画像と, それをなぞりながら口頭で説明したキャプションで構成される Localized Narratives [16] を基に構築した分割データセットを使用する (節 2.1).

**実装詳細** RoBERTa トークナイザーを使ってテキストを埋め込み, DLM 内のノイズ除去ニューラルネットワークには RoBERTa<sub>BASE</sub> モデルを使用する [19]. 各 BBox からの特徴量抽出には BLIP [20] を採用する.

**評価指標** 本手法は生成されるキャプションが 3 つの観点 (内容, 順序, 詳細さ) でユーザの説明意図を反映することを目的とする. まず, 内容の評価では, 各説明対象物体について適切に記述されているかを確認するため, 以下の 4 つの指標を用いる. 文字列一致度を測定するために BLEU-1 [21] と ROUGE-L [22] を導入し, 意味的類似度を評価するために BERTScore [23] を用いる. また, トレースとの一致, すなわち画像領域との対応を確認するために CLIPScore [17] を算出する. 次に, 順序の評価では, トレースの順番通りに各物体が説明されているかを 0 と 1 の配列で表し F 値を計算する. 最後に, 詳細さの評価では, 各 BBox への説明の長さがトレースの滞在時間に比例しているかを確認する. これはキャプション全体ではなく, 発語キャプショ

1) <https://cocodataset.org/#home>



ンと生成キャプションの各セグメント同士を上記の指標を用いて比較することで評価する。

表 2 実験結果

	事前学習	生成長設定	BLEU-1		ROUGE-L		BERT Score		CLIP F 値		len
AR	√	-	0.0950	0.1135	0.4991	23.64	0.0141	30.63			
NAR	√	特化	0.1866	0.2109	0.5578	23.41	<b>0.2568</b>	38.10			
SAR		特化	0.1885	0.2099	0.5675	<b>24.05</b>	0.1266	34.86			
	√	共通	0.2061	0.1987	0.5579	23.78	0.1328	41.23			
	√	特化	<b>0.2122</b>	<b>0.2267</b>	<b>0.5733</b>	23.70	0.1546	39.01			

### 3.2 実験結果

実験の結果を表 2 に示す。自己回帰型 (AR) モデルである GIT [24] と、拡散言語モデルを非自己回帰的 (NAR) に用いて生成する手法 [15] を比較対象ベースラインとして使用する。結果より、構築した分割データセットを用いて半自己回帰的に生成を行う提案手法が正解キャプションとの一致を測る 3 つの指標において最も高い精度を達成した。拡散言語モデルを非自己回帰から半自己回帰的型生成に改良することで、流暢性の向上や前文脈を考慮したテキスト生成が可能になることが示唆される。F 値に関しては非自己回帰手法には及ばない結果となったが、流暢性と制御性のトレードオフや評価方法に起因する問題が考えられる。評価対象となる単語が限定されるため、提案手法が画像内容を的確に説明していても、特定の単語の出現がないために低い F 値が算出される場合があった。また生成を個人の発語速度から決定する場合としない場合で精度に優位な差があり、パーソナライズ化が精度向上に寄与することが示されている。学習データの増加は流暢性の向上に大きく貢献しており、拡散言語モデルにおいて重要な要素であることが確認できた。

### 3.3 分析

生成キャプションからも (図 4)、提案手法による生成例で内容・順序・詳細さの観点からユーザ固有の説明意図が流暢性を持って表現されていることがわかる。例 1 の正解発語キャプションを見ると、ユーザはまず画像中央の馬に着目し、徐々に背景や水辺など上の方に視点を移動させたことがわかる。提案手法も同様に各領域をトレースの順序通りに説明しているが、「horse」という単語でなく「animal」と表現したため、F 値の評価に繋がらなかった点を確認できる。特定の単語だけでなく類似性の高い単語の出現も評価に含めるよう指標の改善が必要であ

		発語キャプション	非自己回帰	半自己回帰
例 1	BBox1	In this image, group of <b>horses</b> are walking on the <b>sand</b> . The bottom, we can see few plants. In the middle, we	In this image we can see group of <b>horses</b> on the ground. We can see some stones on the water. In the background there is	In the center of the image, we can see water, trees and there are some animals on the <b>sand</b> . In the background
	BBox2	can see a <b>sea</b> . There is a <b>sky</b> on the top of the image.	<b>sky</b> in the water	, there is a <b>sky</b> . On top there are
例 2	BBox1	in this image there is a <b>person</b> is sitting on the <b>chair</b> and wearing a maroon	In this image, there is an inside view of a house. In	In this image there is a <b>person</b> sitting on a <b>chair</b> and holding a mobile
	BBox2	t shirt and blue jeans	the foreground, we can	phone and wearing a
	BBox3	and a house has <b>table</b> , chairs, plants and many	also see a person and a sitting on the chair.	t shirt and beside to the <b>table</b> there is a chair
	BBox4	thing are there.	In the background, we can see a house, a chair and a floor.	and on the table there is a table, on the table there is a bottle

図 4 生成キャプション例

る。また、BBox1 での滞在時間の方が BBox2 より長くなっているが、その点を生成単語数に反映できていることが非自己回帰、半自己回帰両者の生成キャプションから確認された。さらに提案手法は、半自己回帰型生成による文法的整合性と自然な文章構造の維持が見受けられる。これに対し、非自己回帰手法では特に例 2 において文法的誤りや不自然な構文が観察された。一方で、提案手法が例 2 の BBox4 に対して作成したキャプションでは同一単語「table」の複数回の出現が目立ち、一度に生成するブロック内のトークンの依存関係の考慮に関しては拡散言語モデル特有の問題点を確認された。

## 4 おわりに

本研究は、各ユーザ特有の説明意図を含むトレースを使った特化型キャプション生成手法を開発した。トレースを解析することで、ユーザが画像中のどの領域にどの程度関心を持ち、どの順序で説明したかを明らかにし、その情報に基づいて拡散言語モデルに半自己回帰的にキャプションを生成させる。実験の結果、提案手法は他手法よりも発語キャプションとの高い一致度を示す、多様性に富む結果を生成できることを確認した。今後の課題として、より適切にトレース情報を解析した分割データセットの構築に取り組む必要がある。現在の分割データセットには、一つの画像から類似した BBox が複数切り出されたり、物体を含まない BBox が存在したりしており、これが全ての手法の精度に影響を及ぼしている可能性がある。座標変化量だけでなく曲率や速度の変化を考慮した構築手法を検討したい。

## 謝辞

本研究は科研費 23K28143 の助成を受けたものです。

## 参考文献

- [1] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. **Pattern Recognition**, Vol. 138, p. 109420, 2023.
- [2] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning, 2020.
- [3] Shunqi Mao, Chaoyi Zhang, Hang Su, Hwanjun Song, Igor Shalyminov, and Weidong Cai. Controllable contextualized image captioning: Directing the visual narrative through user-defined highlights. In **Proceedings of the 18th European Conference on Computer Vision (ECCV)**, 2024.
- [4] Chen Cai, Kim-Hui Yap, and Suchen Wang. Towards attribute-controlled fashion image captioning. **ACM Trans. Multimedia Comput. Commun. Appl.**, 2024.
- [5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.
- [6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [7] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-lm improves controllable text generation. **ArXiv**, 2022.
- [8] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to sequence text generation with diffusion models. In **International Conference on Learning Representations, ICLR**, 2023.
- [9] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Seqdiffuseq: Text diffusion with encoder-decoder transformers. **ArXiv**, 2022.
- [10] Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer, 2024.
- [11] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. DiffuSum: Generation enhanced extractive summarization with diffusion. Toronto, Canada, 2023. Association for Computational Linguistics.
- [12] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. Toronto, Canada, 2023. Association for Computational Linguistics.
- [13] Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. Ar-diffusion: Auto-regressive diffusion model for text generation, 2023.
- [14] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.
- [15] 平野理子, 小林一郎. 拡散過程に基づくモデルによるトレースからユーザの意図を反映したキャプション生成への取り組み. 人工知能学会全国大会論文集, Vol. JSAI2024, , 2024.
- [16] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In **ECCV**, 2020.
- [17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. Association for Computational Linguistics, 2021.
- [18] 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, 2021.
- [19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- [20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of Machine Learning Research. PMLR, 2022.
- [21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Association for Computational Linguistics, 2002.
- [22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [23] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.
- [24] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. **arXiv preprint arXiv:2205.14100**, 2022.
- [25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2015.
- [26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017.
- [28] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.
- [29] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.

## A 参考情報

### A.1 実験詳細

**データセット** 実験に使用した2つのデータセットは Karpathy 分割手法 [25] に従い、画像を 113,287 枚を学習用, 5,000 枚を評価用, 残り 5,000 枚をテスト用データに分割して使用した. それぞれの画像とキャプションのペアデータ数を表 3 に示す.

表 3 各データセットの規模

ペアデータ数	学習用	評価用	テスト用
COCO	566,747	25,010	25,010
分割データセット (LN)	130,831	5,903	6,090

**モデル設定** 学習時のハイパーパラメータの設定は表 4 に示す通りで, AdamW[26] を最適化に使用した. Tesla V100 の GPU3 台を使って, 学習と生成を行った.

表 4 学習時のハイパーパラメータ

データセット	batch size	learning rate	# epochs
COCO	24	2e-4	5
分割データセット	24	2e-4	6

**比較手法** GIT [24] は, CLIP のビジョンエンコーダー [18] を活用し, 画像に基づく文生成を行うデコーダーのみの自己回帰型 Transformer モデル [27] である. 本研究では, GIT モデルに BBox を順次入力し, 生成されるキャプションをプロンプトとし次の BBox を処理する. この手順を繰り返し, 最終的に得られるキャプションを基に損失や精度を算出した.

**生成決定手法** 各 BBox を記述する単語数を, 個人の発語速度とトレース滞在時間の積から決定するプロセスの精度を表 5 に示す. 各 BBox に対応する発語キャプションセグメントの単語数と予測単語数の一致度を二つの指標で評価した. 結果から, 個人の発語速度に基づく生成決定プロセスが高い精度を示すことがわかる. また, 全員の発語速度の平均値を共通で使用する場合と比較しても, 個人特化型の方が精度がわずかに上回る.

表 5 生成決定手法の精度

	コサイン類似度	ピアソン相関係数
特化	<b>0.9704</b>	<b>0.7826</b>
共通	0.9703	0.7837

**F 値** F 値算出のためのトレースの順番通りに各物体が説明されているかを 01 で表す説明有無配列の作成方法について記述する. まず物体検出モデ








	発語キャプション	提案手法(SAR)	
ユーザ A	BBox1 	Here in this picture we can see three people standing on the ground	In this image we can see two people in front of
	BBox2 	and in the middle of them we can see a person lying on the surfboard present on the ground and in the far we can see huts present and we can also see plants and	the water and they are surfing and at the right side of the image we can see a water. Left side of the image we can see a person
	BBox3 	trees covered	standing and surfing and we
	BBox4 	all over there .	can see two persons standing on the
ユーザ B	BBox1 	In this image we can see a person on a skateboard which is placed on the ground . We can also see shoes ,	In this image I can see a person on surfing, in the middle few other people are standing and holding sticks.
	BBox2 	some people standing , some houses with roof , poles , ropes ,	
	BBox3 	plants and a group of trees .	

図 5 同一画像に対する異なるトレースからの生成例

ル (DETR) [28] と意味的セグメンテーションモデル (SegFormer) [29] に一連の BBox を入力し, 検出される物体のラベル名を取得する. 次に発語キャプションの各セグメントが各ラベル名を含むか判定する. 含む場合は 1, 含まない場合は 0 として, 検出ラベルの数に対応する長さの 01 配列を各キャプションセグメントごとに作成し, それらを連結して正解説明有無配列を構成する. 生成キャプションについても同様に, 各セグメントが各ラベル名を含むか調べ説明有無配列を作成する. 最終的に, この2つの説明有無配列間で F 値を計算する.

### A.2 生成結果

同じ画像への異なるトレースが生成キャプションに与える影響を調査した結果を図 5 に示す. トレースと発語キャプションから, ユーザ A はまず周囲の立っている人に注目した後, 地面のサーフボードに寝転んでいる人に関心を移している. 一方, ユーザ B は最初にサーフボードに寝転んでいる人について記述した後, 続いて周囲の状況を説明している. 提案手法による生成キャプションは完全に正確ではないものの, ユーザ A のトレースに従い立っている人を先に言及し, その後サーフボードについて説明する構成となっている. また, ユーザ B のトレースを基に構築された分割データセット内の BBox1 では, 中央のサーフボードに乗った人の姿がより中心的に描画されている. このようにトレースは各ユーザの関心や興味を反映しており, それが発語キャプションにも対応していることが示された. トレースを活用することで提案手法は同じ画像から各ユーザ固有の意図を反映した多様なキャプションを生成できる可能性が示唆された.