複数領域に対するキャプション生成を用いた目の不自由なユーザ向けの画像理解支援

A System for Image Understanding Support for Visually Impaired Users Using Multi-Region Caption Generation

XU Yiling[1]　SHAN Junjie[1]　安尾 萌[2]　西原 陽子[1]
Yiling Xu[1], Junjie Shan[1], Megumi Yasuo[2], and Yoko Nishihara[1]

[1] 立命館大学　情報理工学部
[1] College of Information Science and Engineering, Ritsumeikan University
[2] 立命館グローバル・イノベーション研究機構
[2] Ritsumeikan Global Innovation Research Organization

Abstract: This study proposes an interactive support system designed to assist visually impaired users in achieving a deeper understanding of complex images. Conventional methods that generate a single holistic caption often fail to convey complex compositions. To address this issue, we propose a sub-image captioning approach. We implemented and evaluated two methods: the Simple Sub-image Captioning Method, and the Max Cover Dense Captioning Method. Experimental results demonstrated that the latter method achieved higher objective scores in eight out of 10 image categories, particularly for images with complex scenes and small but critical elements, and a similar trend was observed in the subjective evaluations. Conversely, for images where a single item occupies most of the frame, a holistic description sometimes proved more effective, indicating that the optimal strategy is content-dependent.

1　はじめに

視覚に不自由があることは単なる公衆衛生上の深刻な課題であるだけでなく，社会的な公正性と人類の福祉に関わる重要な議題でもある [1, 2]．そのため，目の不自由なユーザ向けユーザインタフェース（UI）の開発は，益々喫緊の課題となっている [3]．

生成 AI などの近年の技術革新に支えられ，これらのツールは彼らの情報化社会への参加に新たな機会を提供している．その中核となるアプローチは，画像内の視覚情報をテキスト記述に変換し，それを音声合成を介して読み上げることで，ユーザが周囲の世界や任意の画像を「聞く」ことを可能にするものである [4, 5]．

既存技術は単一文章の生成において大きな成功を収めているが [6, 7]，画像全体に対して包括的な説明文を生成するという主流のアプローチには，情報の伝達能力において深刻な限界が存在する．

実際のユーザを対象とした調査では，既存の AI ツールが生成する記述は「詳細さが不十分」であり [8]，物体の空間的な位置関係や色といった具体的な視覚情報，あるいは人物間の関係性のような文脈情報が欠落しているという不満が広く報告されている [9]．これは，従来の手法が複雑な画像情報を単一の文章に圧縮する過程で，ユーザが明確なメンタルイメージを構築するために不可欠な詳細情報を失ってしまうためである．さらに，情報は一方的に提示されるため，ユーザが個人的に興味のある領域を対話的に探索し，詳細を得ることもできない．したがって，多領域的かつ構造化された画像情報を対話的に提示する手法を開発することが，これらの問題を解決する鍵であり，本研究の主な焦点である．

本研究では，単一の説明文による限界を克服する画像理解支援システムを提案する．具体的には，画像を複数のサブ領域に分割し，各領域に対して独立した説明文を生成する，新たな「サブ画像記述」手法を導入する．この手法は，目の不自由なユーザがより豊かで詳細なメンタルイメージを構築し，脳内でより現実に近い視覚シーンを再構成できるよう支援することを目的としている．

2 関連研究

本章では，提案システムの研究に関連する既存研究を，(1) 目の不自由なユーザ向け支援システムの応用研究と，(2) 画像説明生成の基盤となるマルチモーダルモデルの技術研究，の 2 つの観点から概観する．それぞれの分野における既存研究の達成点と未解決の課題を明らかにすることで，本研究の位置付けを明確にする．

2.1 目の不自由なユーザ向け支援システムの既存研究

近年，目の不自由なユーザが画像を理解するために，どのような記述を提供すべきかについて活発な研究が行われている．Doore ら [10] は，特にアート作品の鑑賞という文脈において，ユーザの要求と AI モデルの性能を包括的に調査した．彼らはユーザ調査を通じて，短い一行の概要説明だけでは不十分であり，空間情報と主題情報の両方を含む詳細で多層的な記述が強く好まれることを明らかにした．

また，Fernando ら [8] は，日常的に利用される画像認識ツール（IRT）に関する広範なレビューとユーザ評価を行った．その調査結果によれば，ユーザが既存ツールに対して抱く主な不満点は「記述が不十分」であることであり，将来のツールに最も望む改善点として「より詳細な情報」の提供が挙げられている．

このように，芸術鑑賞という特殊な文脈と，日常利用という一般的な文脈の双方において，既存の単一文章による画像説明がユーザの「深い理解」へのニーズを満たせていないという共通の課題が確認された．この事実は，本研究で提案するサブ画像キャプション生成のような，より構造化され詳細な情報を提供するための新しいアプローチの必要性を強く示唆している．

2.2 画像説明生成マルチモーダルモデルに関する既存研究

2.1 節で述べた先進的な応用を実現するには，深層学習に基づく画像説明生成分野の飛躍的な進歩が不可欠である．現在，この分野の進化は，OpenAI によって開発された CLIP（Contrastive Language-Image Pre-training）[11] に代表される，大規模視覚言語モデルの登場により新たな段階に入った．

CLIP がもたらした新しいパラダイムを応用した代表的な研究として，Mokady らが提案した ClipCap[12] が挙げられる．本研究では，この ClipCap の高い効率性と汎用性に着目し，提案システムのアーキテクチャとして採用する．

また，画像の局所的な理解を深める研究として，Johnson らによる「Dense Captioning」および DenseCap モデル [13] が挙げられる．これは従来の「1 画像に 1 説明」という枠組みを超え，画像内の多数の重要領域を自動特定し，各領域に個別の説明文を生成するものである．さらに Delloul ら [14] は，DenseCap を応用し，RGB-D カメラの深度情報を活用して物体間の位置関係（左右・前方など）を明確に記述する手法を提案した．しかし，この手法は特殊なハードウェアに依存するため，一般的な RGB 画像には適用できないという制約がある．

これに対し，本研究で提案するシステムは，画像の部分領域から生成された多数のテキスト記述を最適化し，ユーザへ提示するものである．提案システムは，(1) 深度カメラのような特殊なハードウェアを必要とせず，あらゆる RGB 画像に適用可能な汎用性を持つ点，そして (2) 単なるアルゴリズムの提案に留まらず，実利用を想定したユーザインターフェースの実装と評価を含んでいる点において，より実用的な貢献を目指すものである．

3 提案システム

2 章で述べた課題に対処するため，本研究では図 1 に示すように，複数の記述を用いて画像の理解を支援するシステムを提案する．本章では，この共通基盤の上に構築された，アプローチの異なる 2 つの具体的な手法について詳述する．

3.1 提案手法 1：部分領域記述法

1 つ目の提案手法は,「部分領域記述法」である．図 2 にその処理例を示す．本手法では，入力画像を 3×3 の均等なグリッドで一律に分割し，9 つのサブイメージ（領域）を生成する．続いて，これら 9 つのサブイメージをそれぞれ独立した画像として扱い，ClipCap モデルに入力することで，各領域に対応する個別の説明文を生成する．本手法の利点は，画像の全領域を網羅的にカバーできる点，および「左上」「中央」「右下」といったユーザにとって直感的かつ予測可能な構造で情報を提供できる点にある．

3.2 提案手法 2：重ね領域最大法

2 つ目の提案手法は，意味的文脈を考慮した「重ね領域最大法」である．図 3 にその処理例を示す．本手法の目的は，単純な部分領域記述法が持つ構造的な網羅性と明快さを維持しつつ，より意味内容の豊かな領域に対して説明を割り当てることにある．単純なグリッ

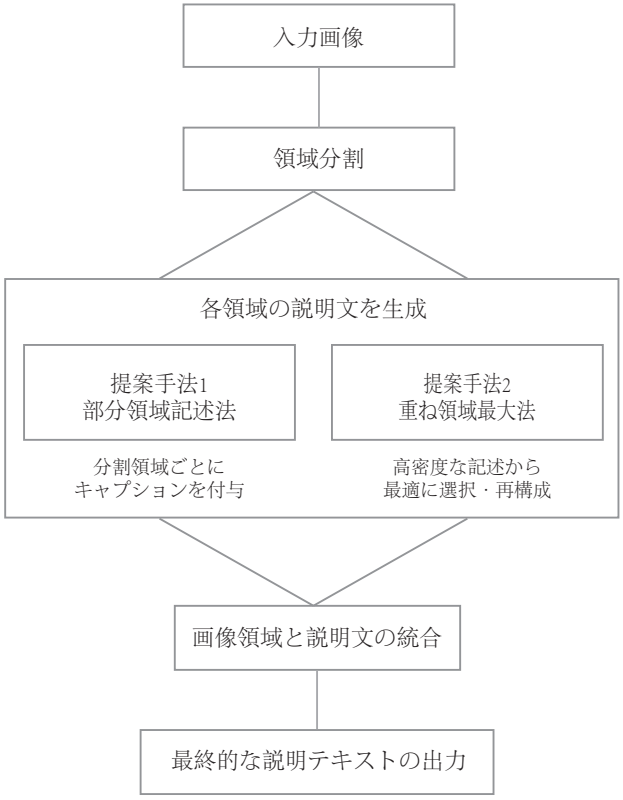

図 1: システムの概要

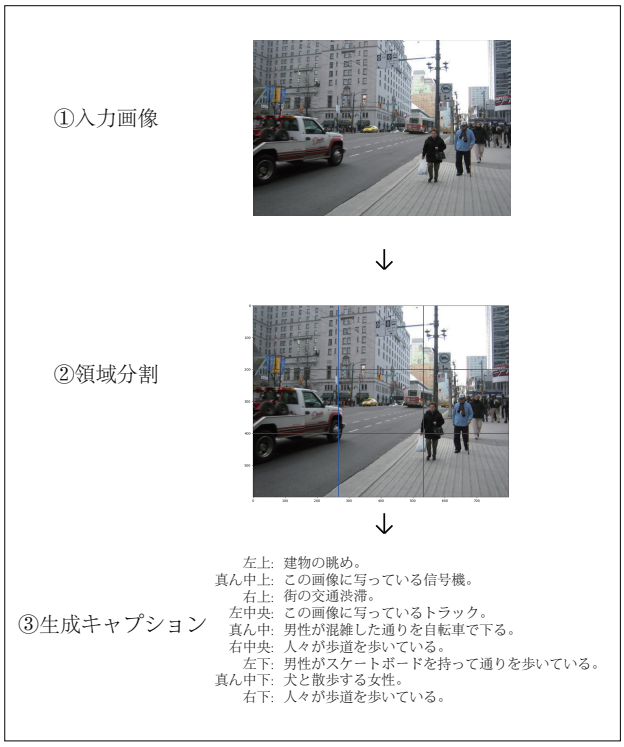

図 2: 提案手法 1 の実行例．画像は Visual Genome データセット [15] から引用．

ド分割では，物体が複数の領域に分断されたり，ある領域に意味のない背景のみが含まれたりするという課題がある．この問題を解決するため，本手法は以下の 2 段階のプロセスで構成される．すなわち，(1) 画像から詳細な領域記述候補を網羅的に生成する段階，および (2) 生成された候補の中からグリッド構造に合わせて最適な記述を選択する段階である．

第一段階では，Johnson らが提案した「Dense Captioning」のアプローチを適用する．まず，CNN を用いて画像から特徴マップを抽出する．次に，特徴マップ上の各位置を基準に様々なアスペクト比を持つバウンディングボックスを生成し，物体が存在し得る領域を提案する．モデル内の Localization Layer が各提案領域の信頼度を予測し，最終的に 1 枚の画像から数百個にも及ぶ「意味のある領域」と「説明文」のペアを候補として生成する．図 3 の②にその例を示す（可視化のため，スコア上位 50 件のみを表示している）．

第二段階では，第一段階で生成された無秩序な候補群の中から，提示に適した記述を抽出する．具体的には，画像を 3 × 3 のグリッドで覆い，9 つの各グリッドセルについて，第一段階で生成された全ての候補領域との空間的な重複度を IoU（Intersection over Union）を用いて算出する．IoU は，2 つのバウンディングボックスの重なり具合を 0 から 1 の値で評価する指標であり，2 つの領域の共通部分の面積を，和集合の面積で除算することで求められる．これを数式で表すと式 (1) の通りである．

$$IoU = \frac{Area(B_{grid} \cap B_{candidate})}{Area(B_{grid} \cup B_{candidate})} \tag{1}$$

式 (1) において，B_{grid} はグリッドセル，$B_{candidate}$ は候補領域のバウンディングボックスを示す．図 3 の③に計算の概念図を示す．各グリッドセル（図中の青枠）に対し，IoU が最大となる候補領域（図中の赤枠）に紐づく説明文を，そのセルを代表する最終的な記述として採用する．図 3 の④は，この選択プロセスを経た最終出力を示しており，各グリッドセルに対して選択された説明文と，その根拠となった最大 IoU 値が列挙されている．

この 2 段階のプロセスにより，本手法は最終的に 9 つの構造化された説明文を出力する．最終的な出力形式は部分領域記述法と同様であるが，各説明文が単純な矩形の切り抜きではなく，意味のある物体や部分を捉えた領域に基づいている点で質的に異なる．本手法は，Dense Captioning が持つ「意味的な詳細さ」と，グリッド構造による「提示の明確さ」の両立を実現するものである．

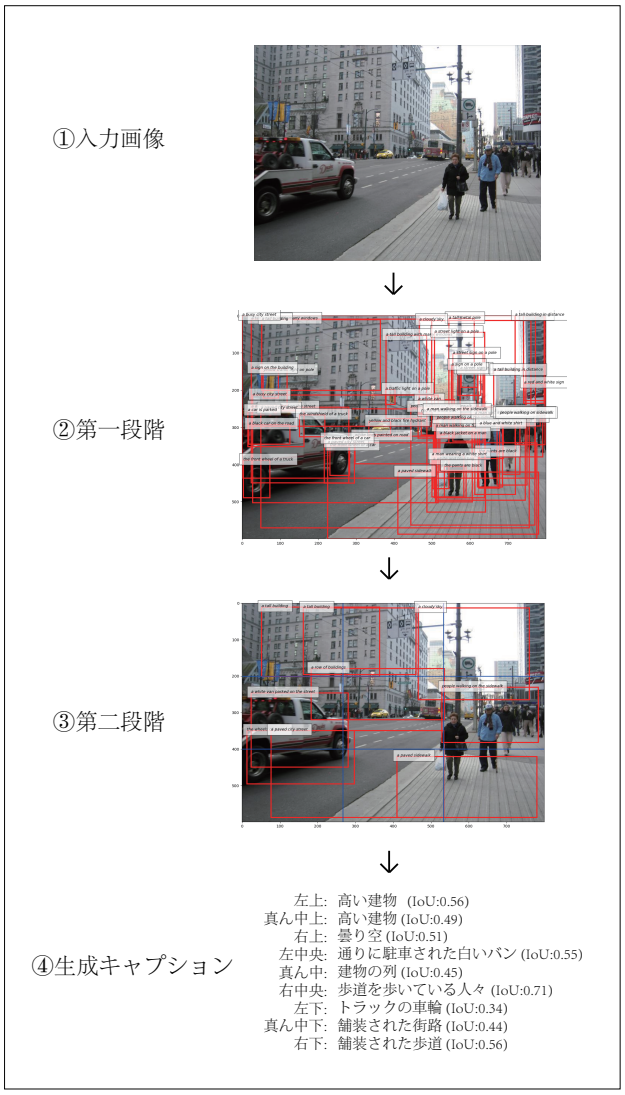

図 3: 提案手法 2 の実行例．画像は Visual Genome データセットから引用．

4 評価実験

4.1 実験の概要

提案した部分領域記述法と重ね領域最大法の有効性を評価するため，ユーザ参加型の比較実験を実施した．従来的なアプローチである単一の包括的な説明文を生成する手法をベースラインとして設定し，これら 3 つの手法が生成した説明文の理解しやすさを多角的に評価する．本実験では，20 名の参加者を募集し，提案手法により生成された画像説明文のみを頼りに，元の画像を想像してスケッチを描いてもらうタスクを課した．

実験評価として，客観的評価と主観的評価の 2 つの指標を記録した．客観的評価では，ユーザの「感知と理解」の度合いを測定するため，参加者が描いたスケッチと元の画像を大規模言語モデルに入力し，両者の類似度スコアを算出させた．これに加え，主観的評価として，参加者自身が各説明文からどれだけ鮮明に画像を想像できたかを「想像のしやすさ」として 5 段階で評価した．このように，本研究では客観的な大規模言語モデルによる類似度スコアと主観的なユーザ評価を組み合わせることで，各記述法がユーザの心的イメージ構築に与える影響を多角的に分析した．

4.2 実験環境・設定

4.2.1 実験用データ

実験には 10 カテゴリのデータを使用した．これらのカテゴリは，提案手法 2（重ね領域最大法）の基盤である Visual Genome データセット [13, 15] と，提案手法 1（部分領域記述法）の基盤である CLIP ベースのモデル [12, 16] が得意とするデータセットの 2 つのグループから抽出された．

具体的には，Visual Genome グループからは，「スキー」や「シマウマ」といった同データセットの主要カテゴリ（人物，スポーツ，動物）や，「信号機」のような一般的な物体認識能力を評価するカテゴリが選ばれた．CLIP ベースのグループからは，「車」や「料理」といった専門ドメインにおけるきめ細かい分類能力や，「風景」や「日常」のような広範なシーン認識・汎用性能を評価するカテゴリが選定された．

カテゴリは，単一の被写体を含む画像，複数のオブジェクトを含む複雑なシーン，および様々な専門ドメインを網羅し，内容の多様性を確保するように意図的に選定された．

4.2.2 実験用インタフェース

目の不自由なユーザが画像を直接見ることができないシナリオをシミュレートするため，我々は画像を表示せずに説明文を提供する UI を開発した（図 4 参照）．この UI は，画面上部の情報提示エリア，下部の描画エリア，そして右側の制御・評価エリアという 3 つの主要な領域で構成されている．参加者はこの UI を通じて各手法が生成した説明文を受け取り，タスクを遂行した．

情報提示の形式は，制御エリアの「方法」ボタンによって選択された手法に応じて変化する．「clip_1」（ベースライン）が選択された場合，画像全体を要約する単一の説明文が右側のテキストボックスに直接表示される．一方，「clip_2」（部分領域記述法）または「densecap」（重ね領域最大法）が選択された場合は，左上のエリアがインタラクティブな 3 × 3 グリッドとして機能する．参加者がこのグリッドのいずれかの区画をクリックす

図 4: 提案システム用の実験インタフェース

ると，その特定領域に対応する説明文が右側のテキストボックスに表示される．

参加者は提示された説明文に基づき，下部の「描画エリア」にスケッチを描く．描画されるスケッチの解像度は元の画像と完全に一致するよう制御され，後の客観評価（LLM による類似度スコア算出）の公平性を確保している．描画完了後，参加者は右下の「評価」ボタンを使い，説明文の「想像のしやすさ」を 5 段階で評価し，結果を送信する．

4.3 実験手順

本研究は，以下の手順で実験を実施した．

1. 合計 100 枚の画像に対し，ベースライン手法（単一記述法），提案手法 1（部分領域記述法），提案手法 2（重ね領域最大法）の 3 つの手法で説明文を生成し，合計 300 の説明文セットを作成した．
2. これら 300 個の説明文セットをランダムに 20 組へ均等に分割した．各組には 15 件の説明文（3 手法× 5 画像分）が含まれており，組番号は 1〜20 である．
3. 実験には 20 名の参加者を募集し，各参加者に 1 から 20 のいずれかの組番号を割り当てた．
4. 各参加者に対し，割り当てられた組に含まれる 15 件の説明文を 1 つずつ提示し，それぞれの説明文に基づいて簡易的なスケッチを描画してもらった．
5. 加えて，参加者に対し「説明文から想像しやすいかどうか（画面構成の理解容易性）」について 5 段階で主観的に評価してもらった．

4.4 評価方法

4.4.1 客観評価

我々は大規模言語モデル（LLM）を用いた独自の評価フレームワークを構築した．具体的には，まず参加者がテキスト記述のみに基づいて描いたスケッチを収集した．その後，LLM に対し，参加者のスケッチと元の参照画像との類似度スコアを算出させた．

この際，LLM にはプロンプトを与え，0 から 100 の間の類似度スコアを出力させた．このスコアを，本研究における「感知と理解」の客観的指標として採用した．

生成 AI の結果は，実行ごとに割り当てられるシード値によって僅かに変動する可能性があることを考慮し，本研究では詳細な採点基準を設計しただけでなく，各画像ペアに対する評価を 20 回繰り返した．そして，これらの評点の平均値を，最終的な「理解度」の客観スコアとして採用した．

4.4.2 主観評価

評価指標として「想像のしやすさ」を設定し，これは「説明文からどれだけ容易に画面全体の様子を想像できるか，また，構図を理解しやすいか」と定義される．参加者には，各説明文を評価した後，この「想像のしやすさ」について 1（非常に想像しにくい）から 5（非常に想像しやすい）までの 5 段階で評定してもらった．収集された評定値は，手法ごとに平均値や分散などの統計量を算出し，その分析を通じて各アプローチの優劣を比較した．

4.5 実験結果

4.5.1 描画結果の比較

各手法が生成した説明文から参加者が実際にどのような画像を思い描いたかを探るため，描画されたスケッチの典型的な例を提示する．図 5 は，同一の元画像に対し，3 つの異なる手法で生成された説明文に基づいて描かれたスケッチの比較例である．

4.5.2 客観評価（LLM による類似度スコア）の集計結果

全画像を通した手法ごとの全体的な性能と，画像カテゴリごとの性能を詳細に分析するため，それぞれ統計量を算出した．表 1 は，全カテゴリを対象とした各手法の平均スコアと分散を示す．重ね領域最大法が最も高い平均スコア（30.83）を示し，ベースラインは最も低い（24.48）．表 2 は，各カテゴリごとに手法別の

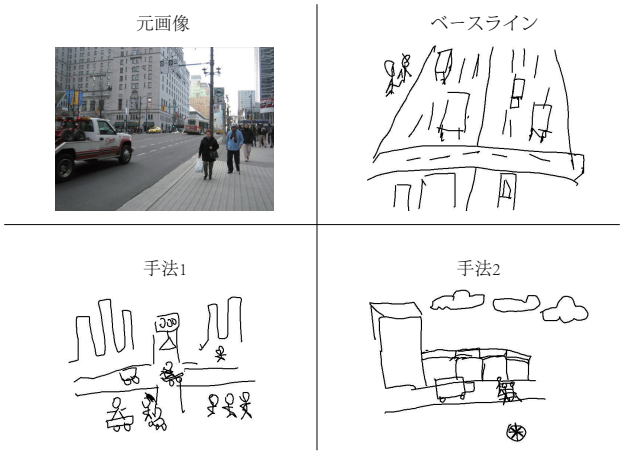

図 5: 描画されたスケッチの比較例．画像は Visual Genome データセットから引用．

表 1: 全カテゴリにおける各手法の平均スコアと分散

手法	平均スコア	分散
ベースライン	24.48	93.22
部分領域記述法	25.12	97.75
重ね領域最大法	**30.83**	117.81

平均スコアを示している.「06 車」カテゴリでは手法 2 が最も高いスコア（37.875）を記録し,「08 料理」ではベースラインが最も高い（21.390）など，カテゴリによって異なる傾向が見られる．

各手法間に統計的に有意なスコア差が存在するかを検証するため，全体およびカテゴリ別に対して一元配置分散分析（ANOVA）を実施した．全体では有意差は観察されなかった．カテゴリ別では,「04 シマウマ」と「05 標識」において，p 値がそれぞれ 0.022 と 0.004 と有意水準（$p < 0.05$）を下回り，統計的に有意な差が認められた（表 3 参照）．

表 2: カテゴリごとの各手法の平均スコア

カテゴリ	ベースライン	手法 1	手法 2
01 スキー	32.965	32.230	**36.610**
02 野球	25.135	33.285	**37.260**
03 信号機	27.850	34.965	**35.185**
04 シマウマ	34.180	15.130	**34.325**
05 標識	14.030	32.550	**37.510**
06 車	21.360	15.565	**37.875**
07 風景	17.435	18.960	**26.820**
08 料理	**21.390**	14.485	7.335
09 部屋	20.695	**30.865**	23.750
10 日常	29.795	23.165	**31.670**

表 3: 各カテゴリにおける 3 手法間の客観評価スコアの一元配置分散分析（ANOVA）結果

カテゴリ	F 値 (F-statistic)	p 値 (p-value)
全データ	2.99	0.052
01 スキー	0.15	0.863
02 野球	0.68	0.517
03 信号機	0.42	0.663
04 シマウマ	**4.44**	**0.022**
05 標識	**6.84**	**0.004**
06 車	2.55	0.097
07 風景	0.95	0.401
08 料理	1.63	0.214
09 部屋	0.56	0.579
10 日常	0.45	0.641

表 4: カテゴリ別の主観評価スコア，および全体の平均と分散

カテゴリ	ベースライン	手法 1	手法 2
01 スキー	3.3	3.6	4.0
02 野球	3.0	2.9	4.1
03 信号機	2.8	3.3	3.8
04 シマウマ	4.0	2.8	4.5
05 標識	3.0	3.1	3.7
06 車	3.6	3.6	3.9
07 風景	2.8	3.6	3.6
08 料理	2.7	2.2	4.1
09 部屋	2.6	3.7	3.3
10 日常	4.1	3.9	3.8
平均スコア	**3.19**	**3.27**	**3.88**
分散	**1.27**	**1.18**	**0.89**

4.5.3 主観評価（想像しやすさ）の集計結果

次に,「想像のしやすさ」に関する主観評価の集計結果について述べる．表 4 は，この評価の詳細な結果を示す．各カテゴリの平均スコアと共に，各手法の全体的な平均スコアと分散が示されている．全体として，手法 2（重ね領域最大法）が最も高い平均スコア（3.88）と最も低い分散（0.89）を達成した．カテゴリ別に見ると,「04 シマウマ」や「08 料理」で手法 2 の評価が特に高く，一方で「09 部屋」では手法 1 が最も高い評価を受けた．

5 実験結果に対する考察

本章では，実験結果に基づき，各手法の有効性と画像カテゴリの特性が与える影響について多角的な考察を行う．

5.1 手法ごとの全体的な有効性に関する考察

まず，全カテゴリを総合した際の各手法の有効性について考察する．客観評価において，提案手法，特に重ね領域最大法（手法 2）が優れている傾向が観察された．表 1 が示す通り，重ね領域最大法の平均スコアは 30.83 であり，部分領域記述法（手法 1）の 25.12，ベースラインの 24.48 を明確に上回った．この差について一元配置分散分析（ANOVA）で検定したところ，有意差はなかった．本実験の参加者数 (N=20) が，全体差を検出する上での統計的検出力に影響した可能性が考えられる．

一方で，この高い平均スコアが示す潜在的な優位性の傾向は，主観評価の結果によっても裏付けられた．表 4 を見ると，重ね領域最大法の「想像のしやすさ」に関する平均スコアは 3.88 であり，ベースライン（3.19）および部分領域記述法（3.27）を大幅に上回った．この結果は，ユーザが重ね領域最大法から得た情報を，心的イメージの構築に最も容易だと感じたことを示している．

客観・主観両方の評価結果を総合すると，単一の記述よりも複数の構造化された記述を提供する方がユーザの画像理解を効果的に支援するという傾向が観察された．その中でも，意味のある領域を的確に捉えて記述する重ね領域最大法が最も有望なアプローチとして際立っており，本提案システムの有効性を示している．

5.2 画像カテゴリの特性による影響の考察

次に，結果をカテゴリ別に詳細に分析し，最適な記述法が画像の特性に強く依存していることを明らかにする．第一に，単一の明確な被写体が画像の大部分を占める場合，ベースライン手法が有効な場合がある．例えば，表 2 の「08 料理」カテゴリでは，ベースラインの平均スコアが 21.390 と最も高かった．ただし，この差は統計的に有意ではなかった．この傾向は，画像が単一の被写体と単純な情報を含む場合，画像を分割するアプローチが不必要に情報を断片化させ，単一の包括的な記述文の方が効率的である可能性を示唆している．

第二に，単一被写体であっても，機械的なグリッド分割は有効に機能しないことが示された．「04 シマウマ」カテゴリでは，手法間に統計的に有意な差が検出された（$p = 0.022 < 0.05$）．表 2 を見ると，部分領域記述法のスコア（15.130）は，ベースライン（34.180）および重ね領域最大法（34.325）より著しく低い．これは，グリッド分割がゼブラという被写体を不自然に分断したためと考えられる．対照的に，ゼブラ全体を単一の意味のある領域として捉えた重ね領域最大法と，画像全体を記述したベースラインは，共に高い評価を受けた．

第三に，画像内に比較的小さくとも重要な要素が含まれる場合，重ね領域最大法が圧倒的に優位であった．「05 標識」カテゴリでは，手法間に有意な差が観察され（$p = 0.004 < 0.05$），重ね領域最大法のスコア（37.510）はベースラインの（14.030）を遥かに上回った（表 2）．これは，DenseCap ベースの手法が，道路標識という特定の重要領域を正確に検出し，その内容を詳細に記述したことが，ユーザの理解に直接貢献したためと考えられる．

最後に，構造的なシーンでは部分領域記述法が有効な場合もあった．「09 部屋」カテゴリでは，部分領域記述法の平均スコアが 30.865 と最も高かった（表 2）．この差は統計的に有意ではなかったが，この傾向は，部屋のような人工的な環境では，「左上・中央・右下」といった単純なグリッド構造が，空間全体のレイアウトを直感的に伝える上で有効であった可能性を示唆している．

これらのカテゴリ別の詳細な考察は，単一の画像記述アプローチですべてのカテゴリに最適に対応することが困難であることを示している．したがって，画像の構成や内容に応じて最適な記述戦略を動的に選択できる，柔軟な画像記述システムが有効であろうことが確認された．

6 結論

本研究は，単一の包括的な記述文では目の不自由なユーザが複雑な画像を理解するには不十分であるという課題に取り組んだ．画像を複数領域に分割して説明する「サブ画像記述」アプローチを提案し，その有効性を検証した．具体的には，ベースラインである単一記述法，部分領域記述法，および重ね領域最大法の 3 つの手法を，参加者が説明に基づいてスケッチを描くという独自の実験手法を用いて比較した．客観・主観評価による実験結果は，複数の記述を提供する提案手法，特に重ね領域最大法がベースラインを上回る傾向を示した．同時に，単一の被写体が中心の画像ではベースラインがより効果的な場合があるなど，最適な記述法が画像の特性に強く依存することも明らかになった．本研究の貢献は，「サブ画像記述」アプローチの有効性を実証し，ユーザの理解度を定量化する新たな評価方法を提示した点にある．

今後の改良点として，領域分割の最適化と説明の個別化が挙げられる．具体的には，画像の複雑さに応じて領域を動的に決定する適応的セグメンテーションや，ユーザが関心領域を指定する機能の導入を計画している．また，ユーザの知識レベルや好みに応じて説明を調整し，リアルタイムのフィードバックに基づいて内容を動的に変更する機能の開発も目指す．これらの機

能強化は，言語モデルのパラメータ調整や，より表現豊かなデータセットでのモデルの再訓練によって実現可能である．さらに，本実験 (N=20) では $p = 0.052$ という有意傾向に留まった全体差を明確に検証するため，提案システムの有効性をより強固に検証するため，今後はより多くの参加者による評価を実施する予定である．

参考文献

[1] GBD 2019 Blindness and Vision Impairment Collaborators: Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study, *The Lancet Global Health*, Vol. 9, No. 2, pp. e130–e143 (2021)

[2] Vision Loss Expert Group of the Global Burden of Disease Study.: Publisher Correction: Global estimates on the number of people blind or visually impaired by cataract: a meta-analysis from 2000 to 2020, *Eye*, Vol. 38, No. 11, pp. 2229 (2024)

[3] Hou, W., Riccò, D.: Accessible Design for Museums: A Systematic Review on Multisensory Experience Based on Digital Technology, *Advances in Design and Digital Communication V*, Vol. 51, pp. 282–298 (2025)

[4] Cavazos Quero, L., Iranzo Bartolomé, J., Cho, J.: Accessible visual artworks for blind and visually impaired people: comparing a multimodal approach with tactile graphics, *Electronics*, Vol. 10, No. 3, pp. 297 (2021)

[5] Petrie, H.: Crowdsourcing descriptions of visual works of art for blind and partially sighted people, *Diss. York* (2023)

[6] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, *arXiv preprint arXiv:2301.12597* (2023)

[7] Liu, H., Li, C., Wu, Q., Lee, Y. J.: Visual Instruction Tuning, *arXiv preprint arXiv:2304.08485* (2023)

[8] Fernando, S., Ndukwe, C., Virdee, B., Djemai, R.: Image recognition tools for blind and visually impaired users: An emphasis on the design considerations, *ACM Transactions on Accessible Computing*, Vol. 18, No. 1, pp. 1–21 (2025)

[9] Li, F. M., Zhang, L., Bandukda, M., Stangl, A., Shinohara, K., Findlater, L., Carrington, P.: Understanding visual arts experiences of blind people, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023)

[10] Doore, S. A., Istrati, D., Xu, C., Qiu, Y., Sarrazin, A., Giudice, N. A.: Images, Words, and Imagination: Accessible Descriptions to Support Blind and Low Vision Art Exploration and Engagement, *Journal of Imaging*, Vol. 10, No. 1, pp. 26 (2024)

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision, *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, pp. 8748–8763 (2021)

[12] Mokady, R., Hertz, A., Bermano, A. H.: Clipcap: Clip prefix for image captioning, *arXiv preprint arXiv:2111.09734* (2021)

[13] Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning, *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016)

[14] Delloul, K., Larabi, S.: Towards Real Time Egocentric Segment Captioning for The Blind and Visually Impaired in RGB-D Theatre Images, *arXiv preprint arXiv:2308.13892* (2023)

[15] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., Li, F.-F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73 (2017)

[16] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565 (2018)