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

XU Yiling[1]　SHAN Junjie[1]　安尾 萌[2]　西原 陽子[1]

1 立命館大学　情報理工学部

2 立命館グローバル・イノベーション研究機構

研究の背景と目的

- 背景:
  - 課題: 視覚に不自由がある方の画像情報アクセス
  - 現状技術: 「一画像一文」方式
    - **情報の圧縮**:
    - 物体の位置関係・色彩といった視覚情報の欠落
    - 人物間の関係性などの文脈情報の欠落
  - 本研究のアプローチ: **サブ画像記述**
    - 画像を複数領域に分割し、それぞれを個別に記述
- 目的:
  - 視覚に不自由がある方が、より 階層的で詳細なイメージを構築できるよう支援する

①		部屋に机と椅子とベッドがある
②		ホームベースにバッター、キャッチャー、審判が集まっている
③		会議に参加している人々

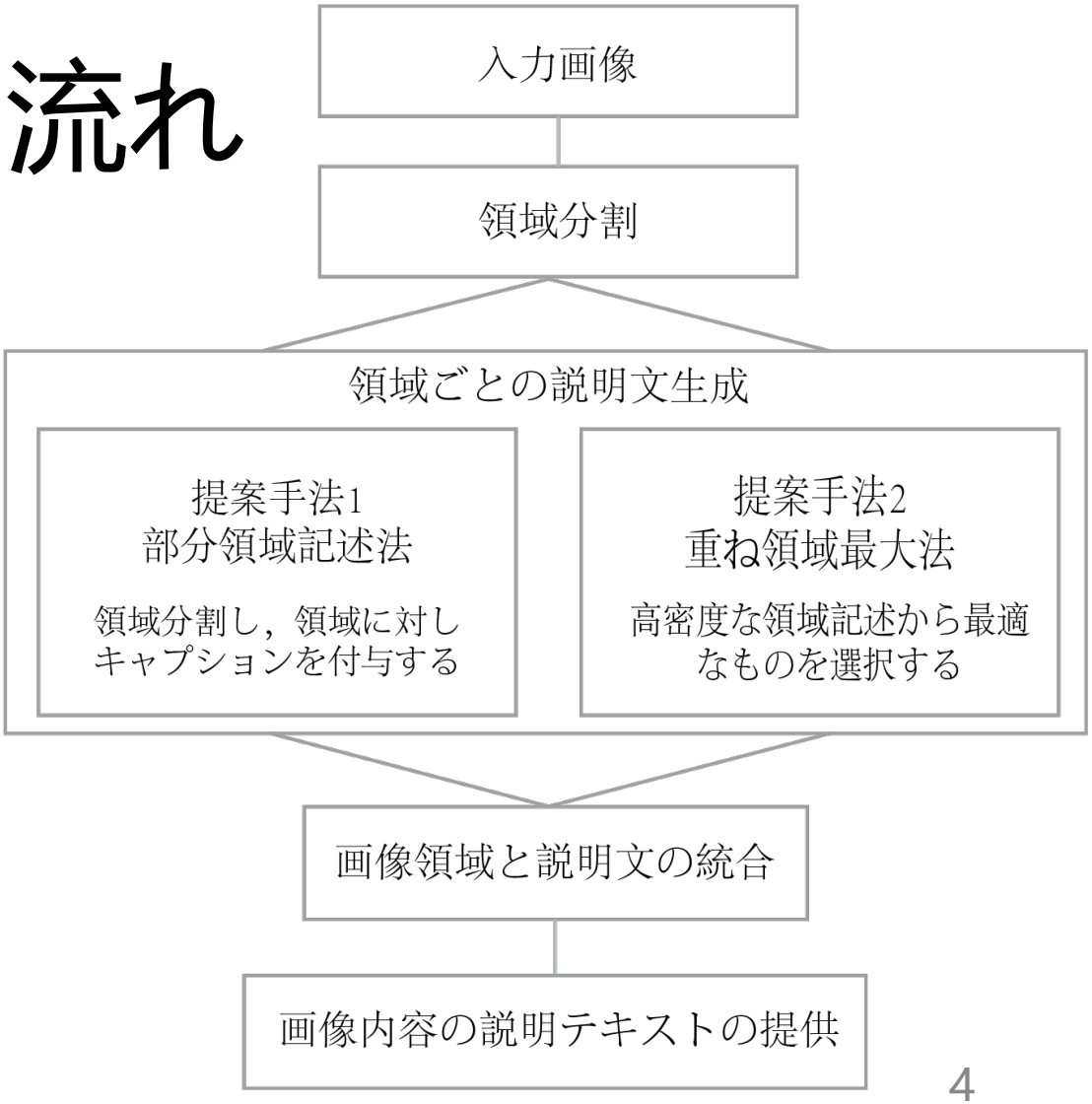
本研究で実現したこと

- 本研究は，単一の画像説明文では視覚障害を持つ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対し，画像を複数領域に分割して説明する「サブ画像記述」手法を提案し，その有効性を検証した．

空	山	山
湖	湖	湖
いす	机	いす

提案システムの流れ

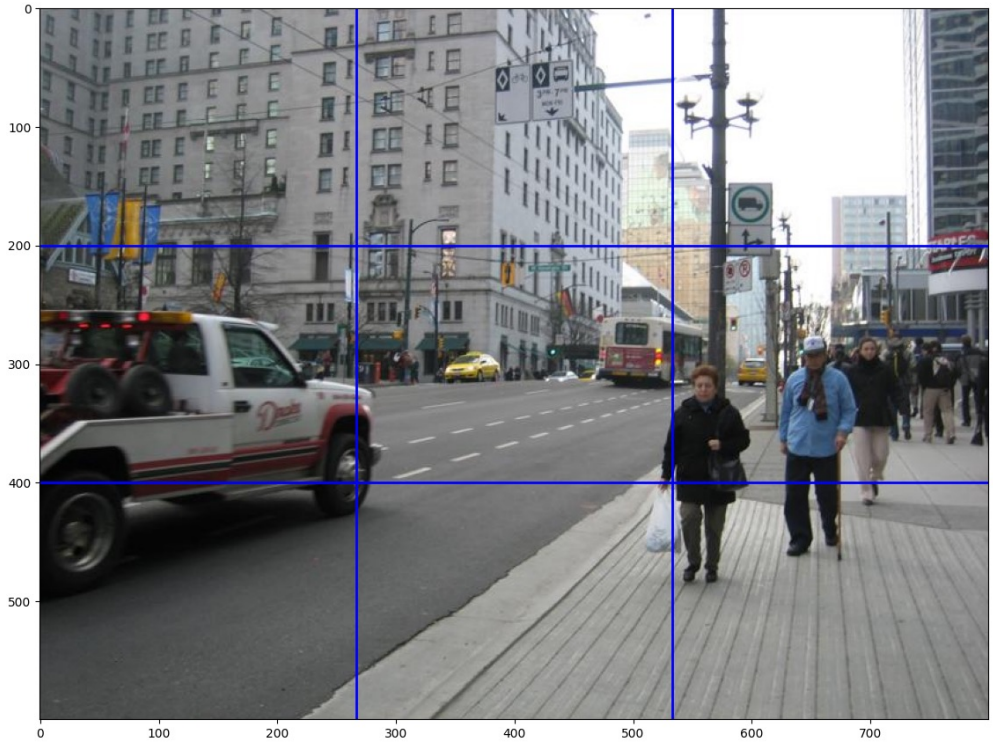
- 提案手法1：部分領域記述法 (simple sub-image captioning)
- 提案手法2：重ね領域最大法 (max cover dense captioning)



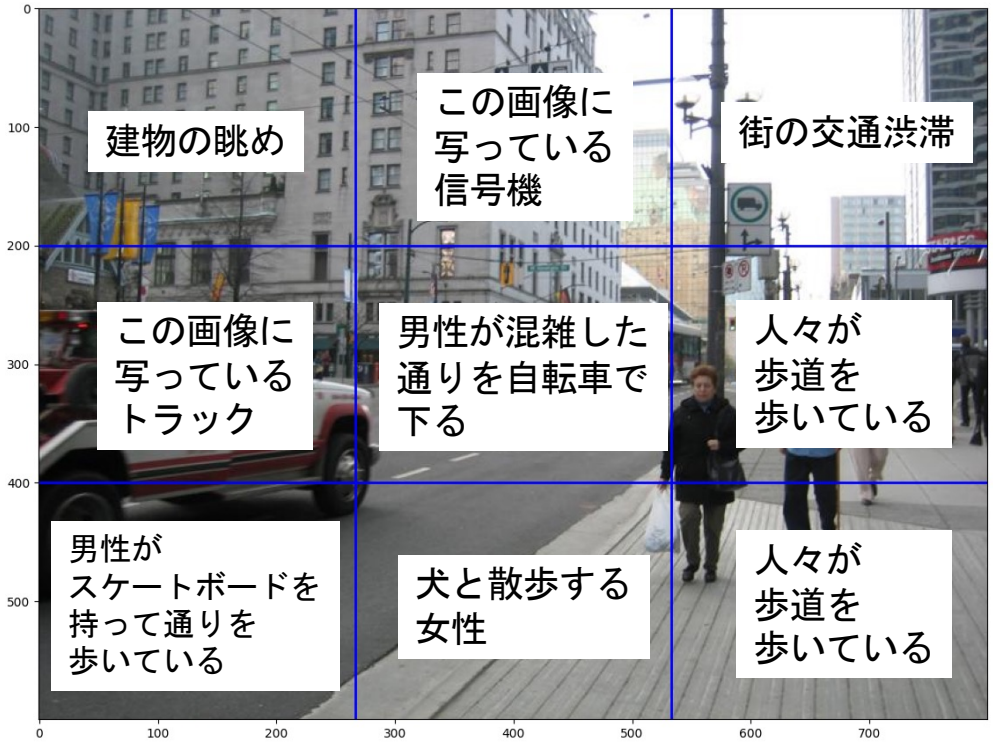
提案手法1： 部分領域記述法

- 入力画像を均等なグリッドに分割し，各領域に個別の説明文を生成
  - ClipCapモデル1

グリッドに分割



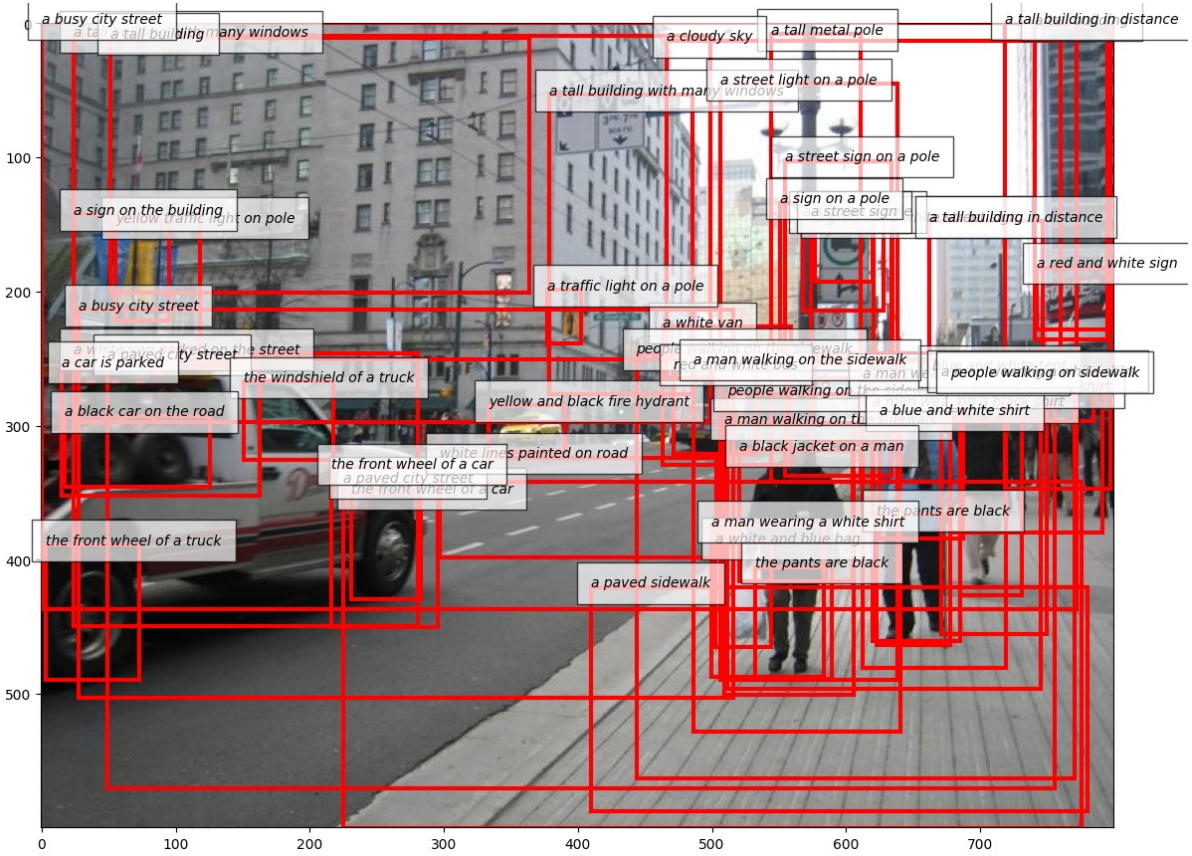
個別の説明文生成



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

提案手法2：重ね領域最大法-1

- (1)画像から多数の詳細な領域記述候補を網羅的に生成する段階
  - densecapモデル[1]

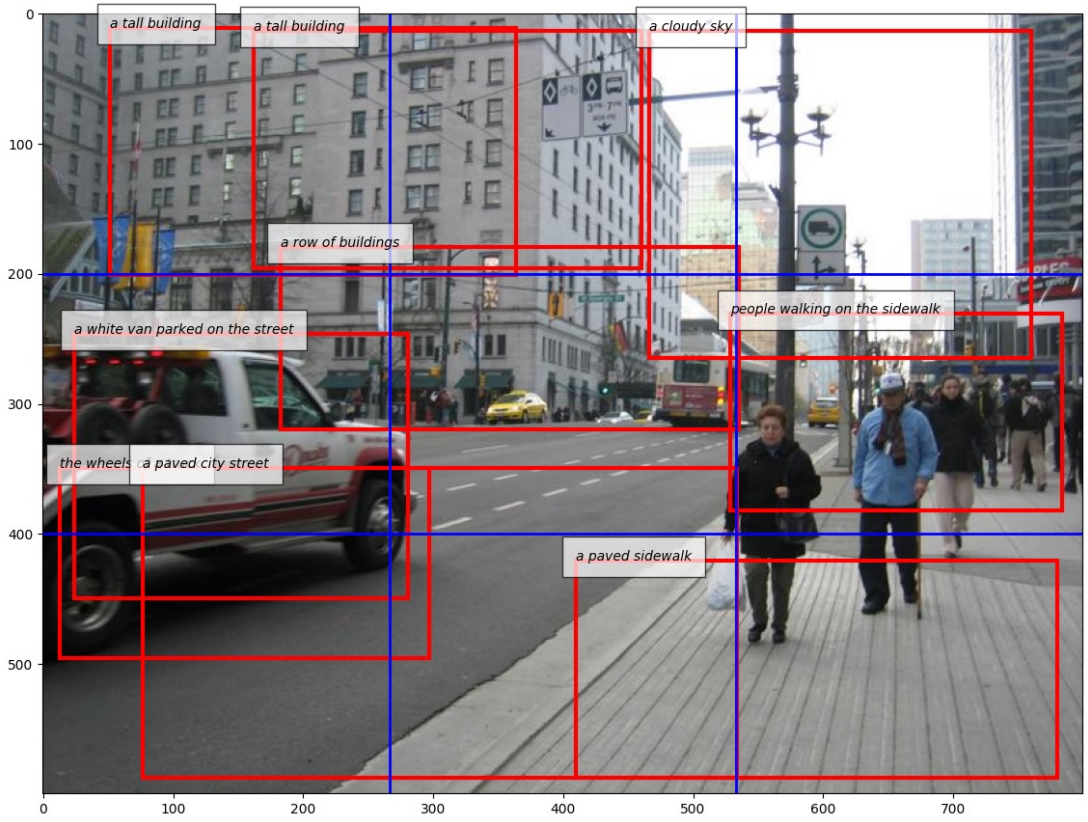


[1]Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4565–4574, 2016.

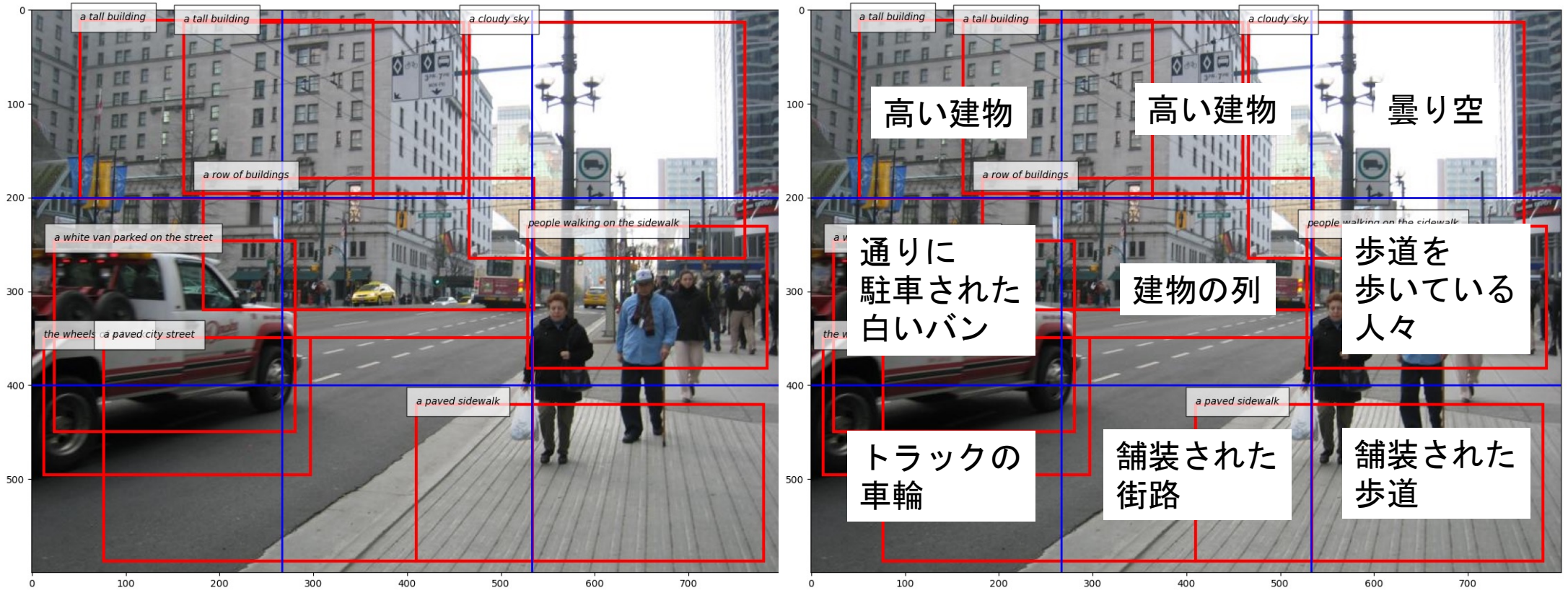
提案手法2：重ね領域最大法-2

- (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
  - 空間的な重なり具合をIoU（Intersection over Union）
  - 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

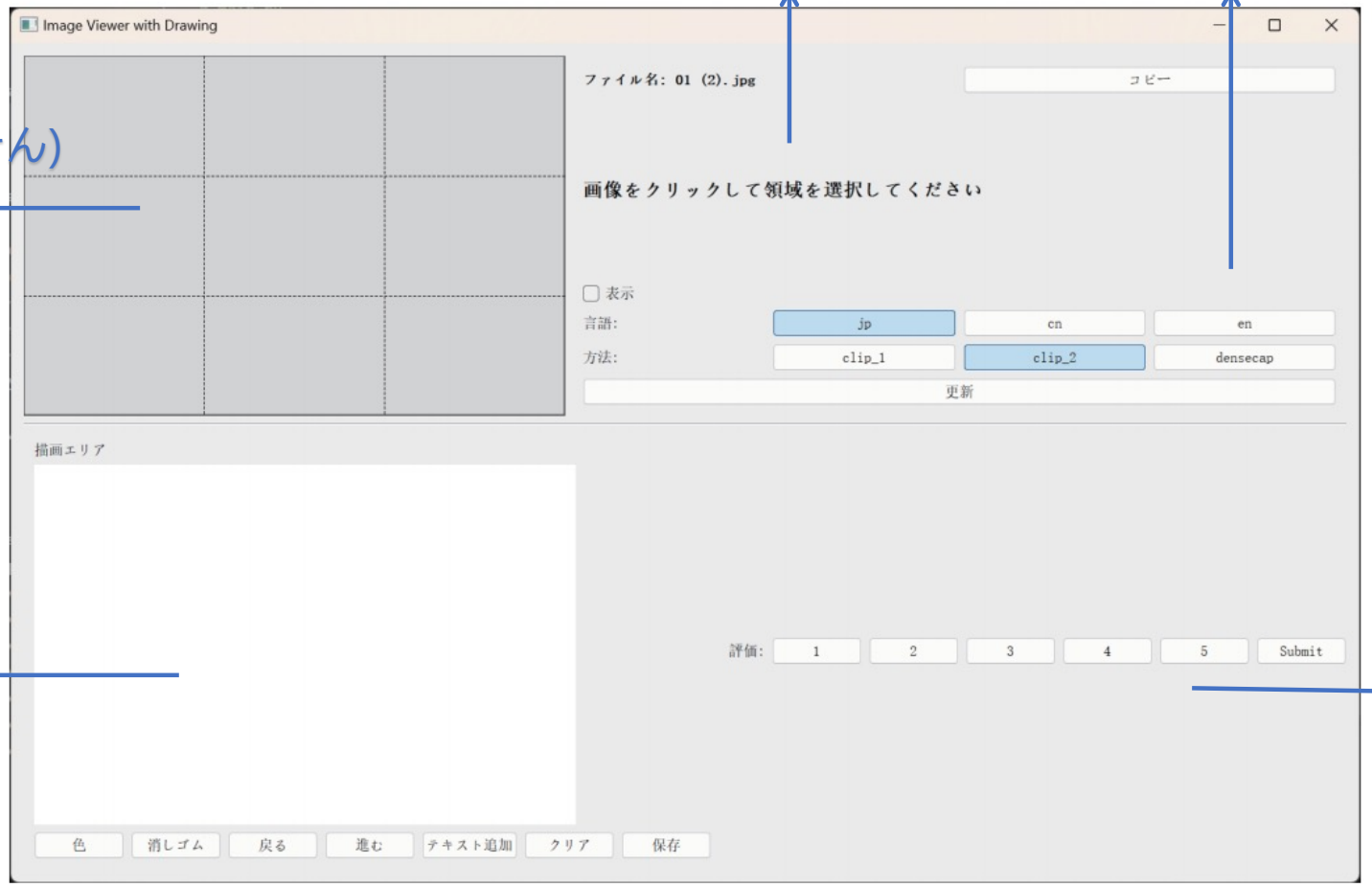
$$\mathrm{IoU} = \frac{\mathrm{Area}(B_{\mathrm{grid}} \cap B_{\mathrm{candidate}})}{\mathrm{Area}(B_{\mathrm{grid}} \cup B_{\mathrm{candidate}})}$$



提案手法2：重ね領域最大法-3



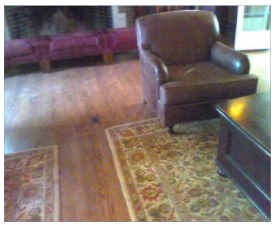
評価実験のUI


説明文
手法の選択
元画像(見せません)
描画エリア
主観評価
Image Viewer with Drawing
ファイル名: 01 (2).jpg
コピー
画像をクリックして領域を選択してください
表示
言語:
jp
cn
en
方法:
clip_1
clip_2
densecap
更新
描画エリア
評価:
1
2
3
4
5
Submit
色
消しゴム
戻る
進む
テキスト追加
クリア
保存

評価実験の説明

- 目的：2つの手法について，有効性を検証する
- 比較手法：一枚の画像に一つの説明文を生成する
- 実験に用いるデータ：各カテゴリの画像各10枚，計100枚
- 手順：
  - 20名参加者に対し，15枚の画像の説明文を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．
  - 加えて，参加者に対し「説明文から想像しやすいかどうか」について5段階で主観的に評価してもらった．

ベースラインの出力例：

①		部屋に机と椅子とベッドがある
②		ホームベースにバッター、キャッチャー、審判が集まっている
③		会議に参加している人々

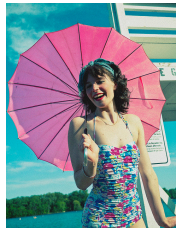
入力画像の選択

- 性質の異なる10種類のカテゴリから画像を抽出した









カテゴリ	画像内容の概要
01_ski	スキーをする人物と雪山
02_baseball player	試合中の野球選手
03_traffic light	街中の交通信号機
04_zebra	草原や動物園のシマウマ
05_street sign	街中の道路標識
06_StanfordCars	自動車（特定車種）
07_Country211	各国風景やランドマーク
08_Food101	皿に盛られた各種料理
09_room	家具が配置された室内
10_coco	物・人がいる日常風景

評価項目-主観評価

- 想像のしやすさ：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

評価項目-客観評価

- 大規模言語モデルで，参加者が描いたスケッチと，その元となったオリジナルの画像との類似度を算出させる
- 20回繰り返して平均値を取る

使用したLLMモデル:gemini-2.5-flash
プロンプト:

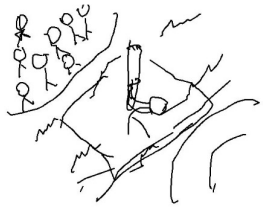
プロンプトの例

以下の 2 つの画像の類似度を比較し，0 から 100 の間のスコアで評価してください．1 枚目の画像は元の参照画像であり，2 枚目の画像は，その参照画像に基づいて描かれた線画です．評価にあたり，以下の点を考慮してください．

- **内容の一致性**： 線画は参照画像の主要な要素を捉えられていますか？ 欠落や誤解はありませんか？ 描画の技術レベルは考慮せず，同じ種類の物体に対する描画の簡略化の度合いも考慮せず，物体の内容の理解が正しいかどうかのみを評価してください．(40 点)
- **構図の一致性**： 線画は，参照画像の主要な要素が画像内のどの位置にあるかを正確に表現していますか？(40 点)
- **正確性と完全性**： 線画は，参照画像内の重要な要素を見落としていませんか？ 内容や位置が不正確，または誤って描かれている部分はありませんか？(20 点)

上記の各項目を総合的に加算し，最終的な類似度スコアを算出してください．描かれた線画の画力やスタイルは考慮せず，画像内の物体とその位置関係における類似度のみを考慮してください．

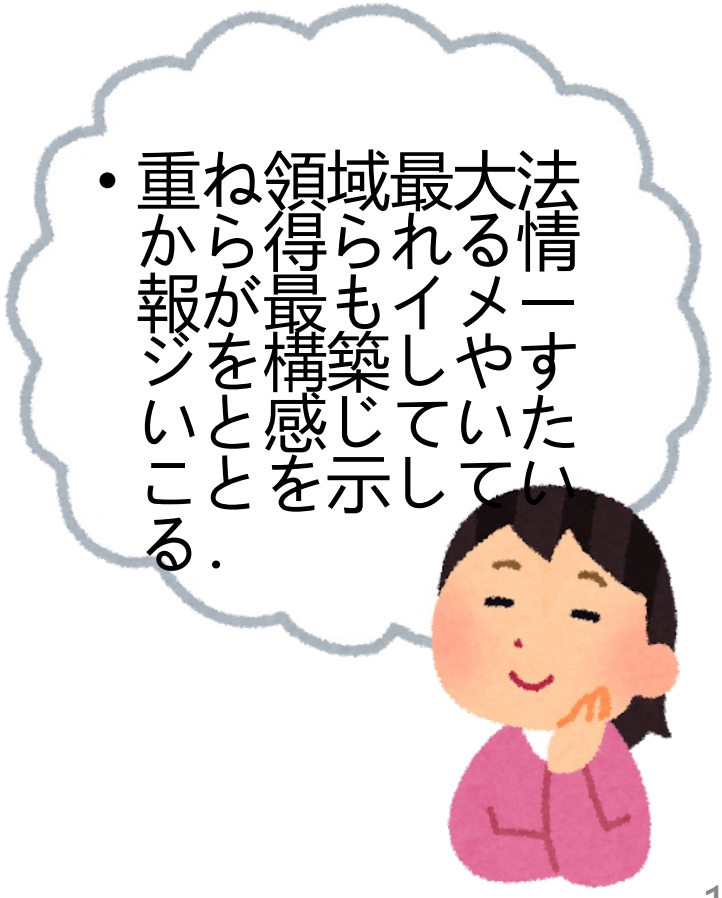
最後の行に，最終スコアを【】で囲んで単独で記載してください．例：【100】

参照画像	描画されたスケッチ	スコア
		88
		15

結果1-主観評価（想像しやすさ）

手法	平均評点
ベースライン	3.19
手法1	3.27
手法2	**3.88**

全カテゴリにおける各手法の主観評価



結果2-客観評価

- 提案手法である手法2は、10カテゴリ中8つで最も高い平均スコアを記録した。
- 全体の平均スコアにおいても、提案手法2つがベースラインを上回り、特に手法2（30.83）が最も高い評価を得た。
- ベースラインと比較し、提案手法2によるスコアの向上が最も顕著だったのは「05 標識」カテゴリであった。
- 例外として、「08 料理」カテゴリではベースラインが、「09 部屋」カテゴリでは手法1がそれぞれ最高スコアとなった。

カテゴリ	ベースライン	手法1	手法2	分散分析
01 スキー	32.97	32.23	**36.61**	p=0.863
02 野球	25.14	33.29	**37.26**	p=0.517
03 信号機	27.85	34.97	**35.19**	p=0.663
04 シマウマ	34.18	15.13	**34.33**	p=**0.022**
05 標識	14.03	32.55	**37.51**	p=**0.004**
06 車	21.36	15.57	**37.88**	p=0.097
07 風景	17.44	18.96	**26.82**	p=0.401
08 料理	**21.39**	14.49	7.34	p=0.214
09 部屋	20.70	**30.87**	23.75	p=0.579
10 日常	29.80	23.17	**31.67**	p=0.641
全体平均	24.48	25.12	**30.83**	
分散分析	p=0.052			

カテゴリごとの各手法の平均スコアと分散分析

考察: 最適な記述法は画像の特性に依存である

- **複雑なシーン（例：野球の試合）**：手法2が有効である．複数のアクター（動作主体）や，それらの空間的な関係性を捉えることができるため．【多くの画像がこれに該当】

- **単一被写体（例：料理の一皿）**：ベースラインが有効である．分割を行うと，不必要な複雑さ（情報の断片化）が生じる可能性があるため．

- **構造的なシーン（例：部屋）**：手法1が有効である．空間的なレイアウトを直感的に伝えることができるため．

結論

- 本研究では、画像を複数領域に分割して説明する「サブ画像記述」システムを提案した。
- 参加者が説明文から画像をスケッチする実験を行い、客観的指標（LLM類似度スコア）と主観的指標（想像のしやすさ）の両面から有効性を検証した。
- 実験の結果、提案システムが、客観・主観評価共に最も高い評価を得る傾向が見られた.
- 今後の改良点: より多くの参加者、領域分割の最適化と説明の個別化