

Image Understanding Support Method for Visually Impaired Users via Multi-Region Caption Generation

Yiling XU, Junjie SHAN, Megumi Yasuo, and Yoko Nishihara

Ritsumeikan University
Ritsumeikan Global Innovation Research Organization




Introduction & Motivation

- **Problem:**

- OrCam MyEye, Envision Glasses, Seeing AI.....
- A single descriptive sentence is often insufficient for visually impaired users to understand complex images.

- **Our Goal:**

- To develop an interactive system that provides structured, multi-regional descriptions to help users build a richer, more detailed image picture.

①		There's a desk, a chair and a bed in the room.
②		Batters, catchers and referees gather at home base.
③		The people in the meeting.

Our Contributions

- Validated the "sub-image captioning approach" (Method 1 and 2) for enhancing image accessibility.
- Proposed and demonstrated the superiority of a semantically-aware method (Method 2).
- Introduced a novel evaluation method using user sketches and LLM similarity scores to quantify comprehension.



Proposed System: Overview

"Multi-Region Caption Generation"

- Intuitive description
 - > **more detail**
- Construct spatial relationships
 - > **easier to imagine**

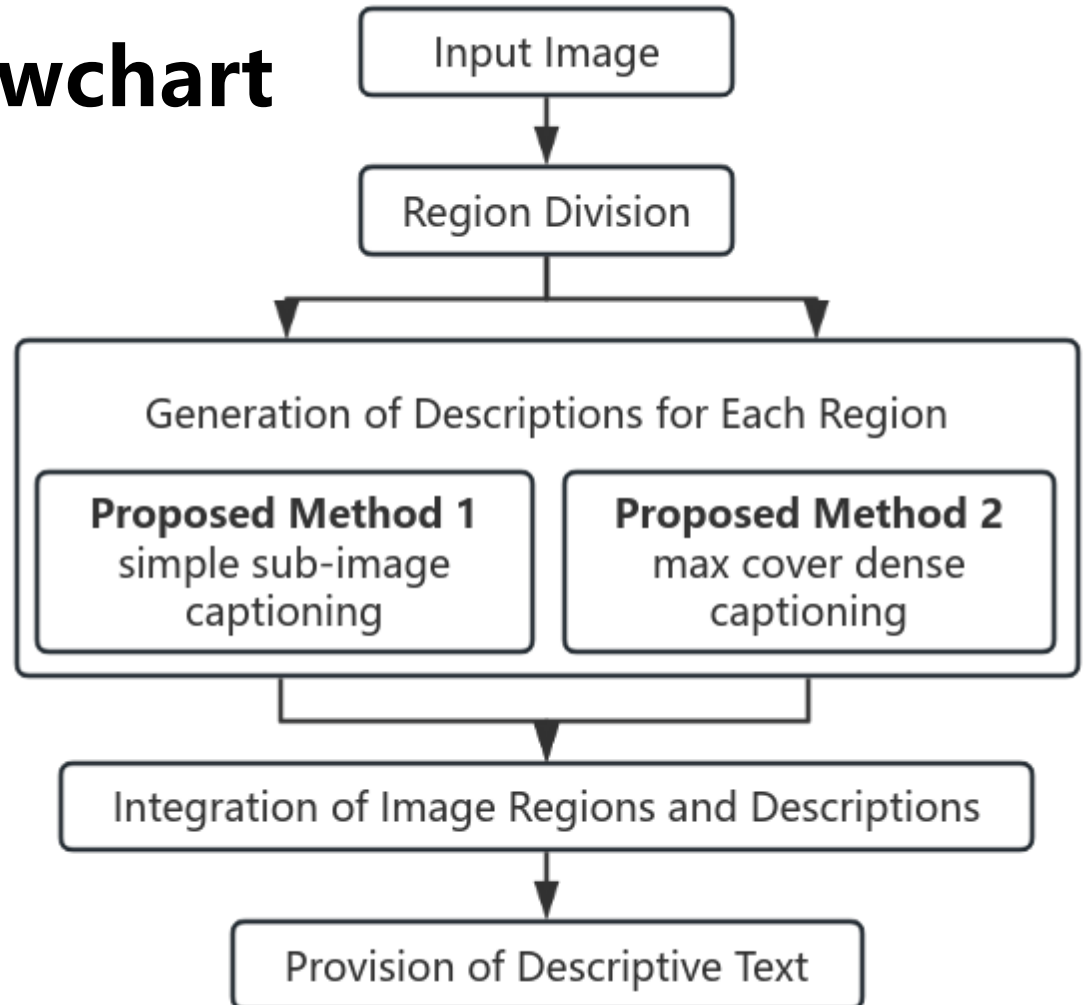
sky	hill	hill
lake	lake	lake
chair	table	chair



Proposed System: Flowchart

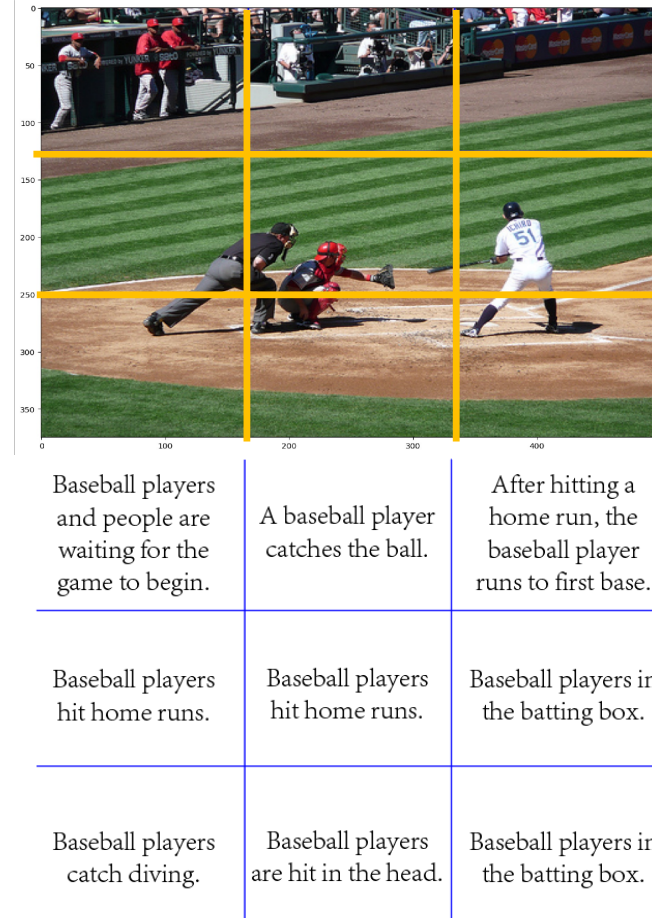
- **Method 1:**
(simple sub-image captioning)

- **Method 2:**
(max cover dense captioning)



Proposed Method 1: Simple Sub-image Captioning

- (1) divide the input image into equal grids
- (2) generate a separate description for each area
 - using ClipCap model¹



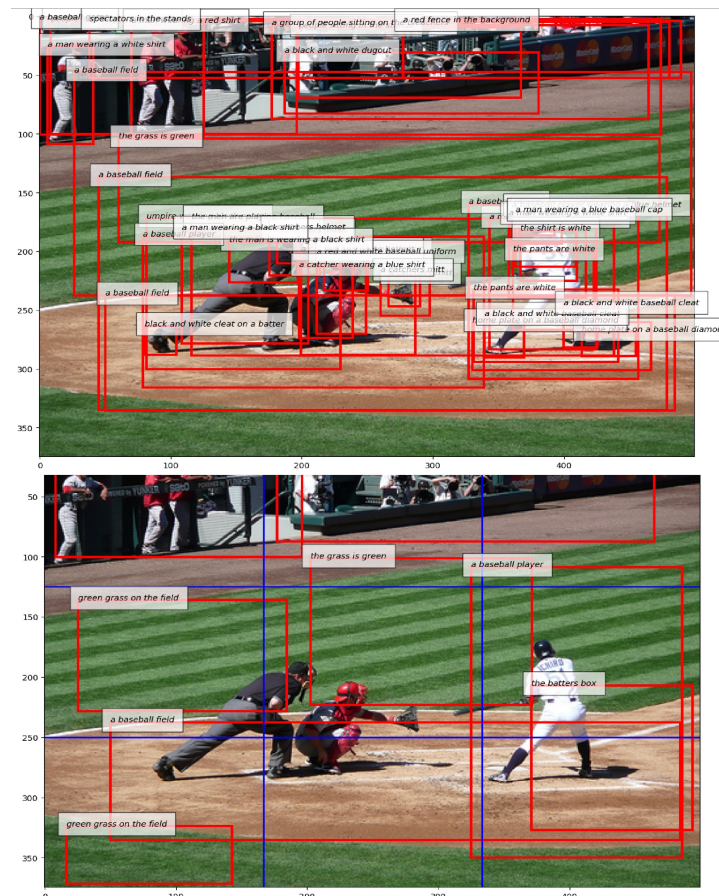
¹Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734, 2021.

Proposed Method 2: Max Cover Dense Captioning

- (1) Generate a large number of detailed candidate regions from the image
 - using densecap model¹

- (2) Selecting the best description according to the grid structure
 - IoU (Intersection over Union)
 - B_{grid} : blue grid
 - $B_{\text{candidate}}$: red boxes

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



¹Justin 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.

Proposed Method 2: Max Cover Dense Captioning

- (1) Generate a large number of detailed candidate regions from the image

- using densecap model¹

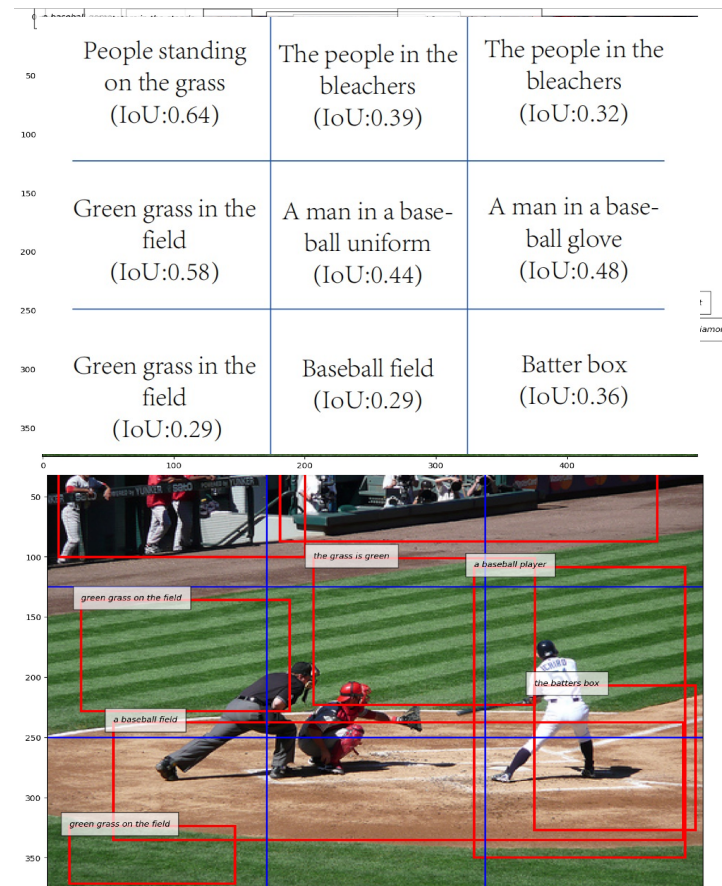
- (2) Selecting the best description according to the grid structure

- IoU (Intersection over Union)

- B_{grid} : blue grid

$B_{\text{candidate}}$: red boxes

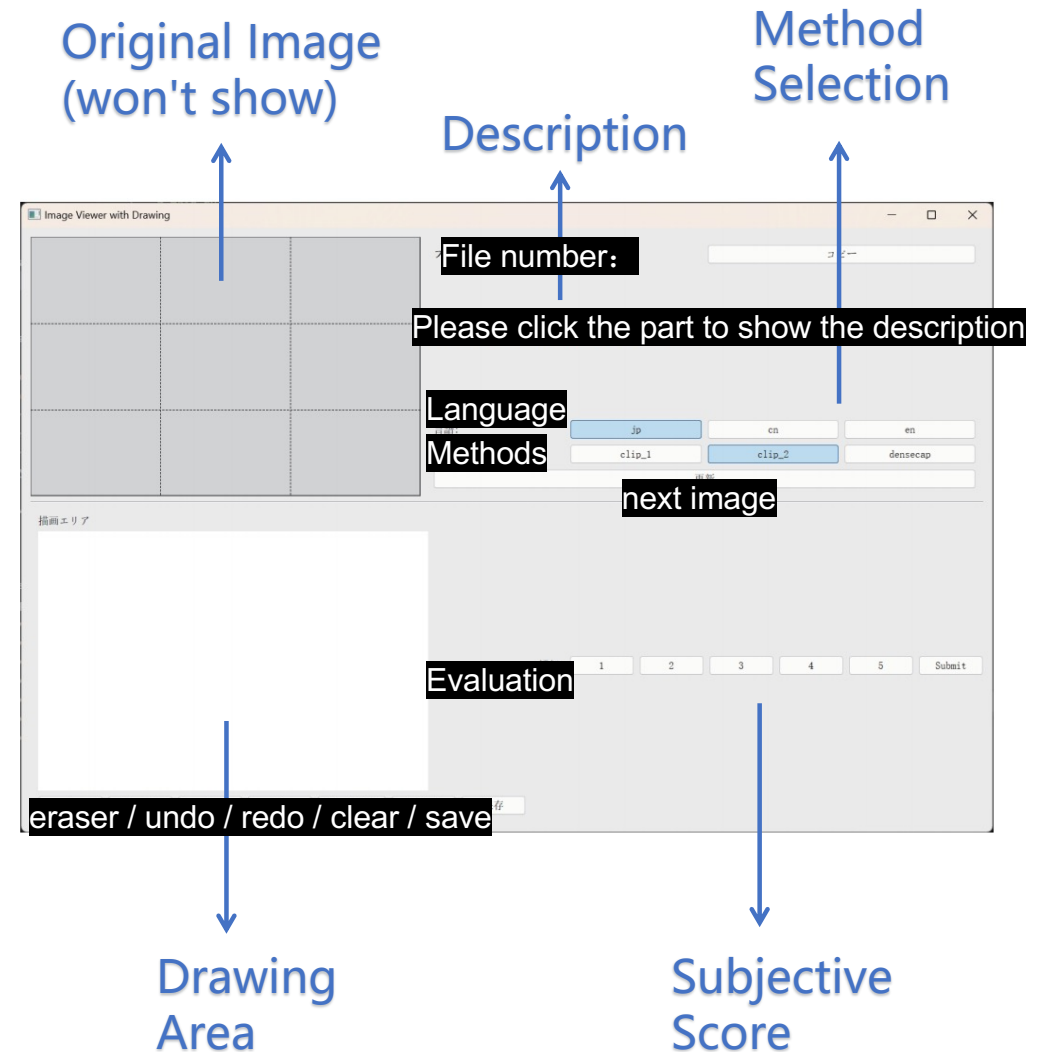
$$\text{IoU} = \frac{\text{Area}(B_{\text{grid}} \cap B_{\text{candidate}})}{\text{Area}(B_{\text{grid}} \cup B_{\text{candidate}})}$$



¹Justin 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.

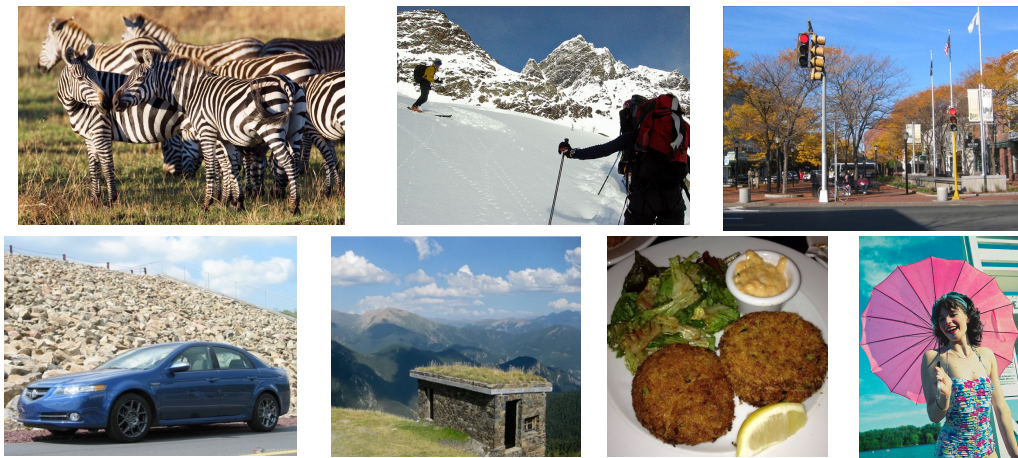
Evaluation Experiment

- **Methods:**
 - baseline:
 - simply using ClipCap model
 - method1
 - method2
- **Task:**
 - 20 participants drew sketches based only on the generated text descriptions.



Experiment Data

- extract images from 10 different categories



categories	image content
01_ski	Skiing people and mountains
02_baseball	Baseball players in the game
03_traffic light	Traffic lights in the city
04_zebra	Zebras in grasslands and zoos
05_street sign	Street signs
06_StanfordCar	Cars (specific types)
07_Country211	Scenery of various countries
08_Food101	Various dishes on a plate
09_room	Furnished room
10_coco	Everyday scenery

Evaluation Experiment-Metrics

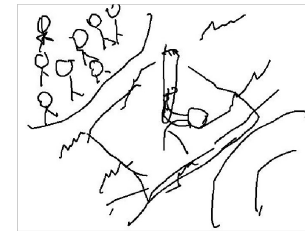
- **Subjective Score**

- **"Imaginability"**: Participants rated how easily they could imagine the scene on a 5-point scale.

reference image

sketch

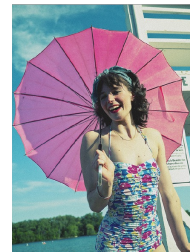
score



88

- **Objective Score**

- "Perception & Comprehension"
- An LLM (gemini-2.5-flash) calculated a similarity score between the user's sketch and the original image.
- Repeat 20 times to get an average.



15

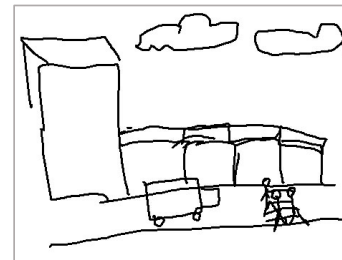
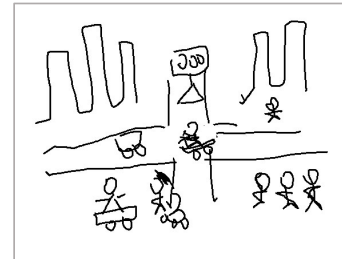
Objective Evaluation

Prompt

.....Instruction.....

- Content Consistency:(40 points)
- Compositional Consistency:(40 points)
- Accuracy and Completeness: (20 points)

.....Instruction.....



Mean score
of 20 times

39.35

56

60.95

Result1: Subjective Score

Method	Mean Score
Baseline	3.19
Method 1	3.27
Method 2	3.88

Subjective evaluation of each method in all categories

Information from
the Method 2 was
the easiest to
construct an image.



Result2: Objective Score

- Method 2 achieved the highest score in 8 out of 10 categories.
- Method 2 consistently outperformed the others, especially for complex scenes with small but important objects (like a 'street sign').
- As exceptions, baseline was the highest score in the "08_Food101" category, and Method 1 was the highest score in the "09_room" category.

categories	Baseline	Method 1	Method 2
01_ski	32.96	32.23	36.61
02_baseball player	25.13	33.28	37.26
03_traffic light	27.85	34.96	35.18
04_zebra	34.18	15.13	34.32
05_street sign	14.03	32.55	37.51
06_StanfordCars	21.36	15.56	37.87
07_Country211	17.43	18.96	26.82
08_Food101	21.39	14.48	7.33
09_room	20.69	30.86	23.75
10_coco	29.79	23.16	31.67
Mean	24.48	25.12	30.83
ANOVA	p=0.052		

Mean score for each method by category 14

Discussion: One Size Does Not Fit All

- **Complex Scene** (e.g., Baseball game): Method 2 is best. It captures multiple actors and their spatial relationships. **【most images】**
- **Single Subject** (e.g., a plate of food): A single sentence (Baseline) can sometimes be more effective. Dividing it can add unnecessary complexity.
- **Structured Scene** (e.g., a room): The simple grid (Method 1) can be effective at conveying the spatial layout intuitively.



Conclusion

- **Conclusion:**
 - Providing multiple, structured descriptions is more effective than a single sentence, and a semantically-aware approach (Method 2) is the most promising.
 - The optimal description strategy is content-dependent.

