

Supplementary Materials

Anonymous Authors

Model	Flickr8k-Expert		Pascal-50S	FOIL
	Kendall τ_b	Kendall τ_c	Acc.	Acc.(1 Ref)
YOLOv8	53.9	55.2	84.5	92.0
MobileSAMv2	55.9	56.4	86.1	93.1

Table 1: Ablation studies of semantic visual regions extractor.

1 MORE IMPLEMENTATION DETAILS

In this section, we introduce more implementation details about each module of our proposed HICE-S.

Image region generation. A single image typically contains numerous visual concepts, distributed throughout various local regions of the image. To extract local visual representation, previous metrics [2, 3] utilize an off-the-shelf detector to acquire objects bounding boxes in the image. However, the rectangular bounding boxes include undesired background information in that most objects have irregular boundaries. To better represent local visual concepts, we leverage a segmentation model to obtain exact semantic regions. Recently, the advent of the segment-anything (SAM) model has been a prominent milestone for general image segmentation owing to powerful zero-shot capabilities. SAM addresses two practical yet challenging segmentation tasks: segment anything (SegAny), which utilizes a certain point to predict the mask for a single object of interest and segment everything (SegEvery), which predicts the masks for all objects on the image. To conduct fast and comprehensive image segmentation, we choose MobileSAMv2 [5] which is a lightweight SAM model designed for fast SegEvery mode. To extract semantic visual regions, we leverage MobileSAMv2 to extract region masks $\{A_n\}_{n=1}^N$ to represent different local visual concepts. For better model relationships between objects in the image, we utilize a full-one mask \tilde{A} to represent the whole image. Finally, we get a region mask set $(\{A_n\}_{n=1}^N, \tilde{A})$ to hierarchically represent the original image. To compare mobileSAMv2 with the conventional detector, we also tried another alternative region mask generator YOLOv8 [1]. Comparison results are shown in Table. 1. It is observed that mobileSAMv2 is superior than YOLOv8, indicating the effectiveness of the current design.

Text phrase generation. A complete caption sentence usually contains multiple entities and relationships between them, where each entity represents an instance in the image. Thus we use a frozen TextGraphParsor [4] to extract a textual scene graph from captions and references as shown in Fig.2 in the main paper. The textual scene graph contains several subject-predicate-object triplets, where the subject and object are entities existing in the image. Each triplet usually represents a specific local image region. Then we transform the triplet into a short phrase in the form of a “subject predicate object”. For example, we transform the “(man, with, bike)” into a short description “man with bike.” To conduct local evaluation, we can calculate the IITC between image regions and caption

phrases. Meanwhile, ITTC are computed between caption phrases and reference phrases.

Fig. 1 has shown more qualitative examples for image region generation and text phrase generation.

2 HUMAN EVALUATION

We invite 5 human experts to assess the detailed captions from two perspectives: correctness and completeness. Specifically, we give human experts an image and a candidate LLaVa caption each time and require them to provide an evaluation score. The evaluation scores range from 1 to 4. Concrete instructions are as follows:

Correctness. Correctness is a metric to evaluate whether the captions contain mistakes or unrelated hallucinations. Specific instruction is:

We would like to evaluate the correctness of the captions. The evaluation scores range from 1 to 4, where 1 indicates all the content of the candidate caption is unrelated to the image and 4 means the candidate description is accurate and has no mistakes.

Completeness. Completeness is a metric to evaluate whether all concepts that appear in the image are included in the description. Specific instruction is:

We would like to evaluate the completeness of the descriptions. The evaluation scores range from 1 to 4, where 1 indicates the caption contains no visual concepts from the image, and 4 means that the caption includes all critical visual details and is comprehensive enough to represent all image content.

The whole human evaluation involves 5000 images from the MSCOCO Karpathy-split test subset with each image having 4 candidate LLaVa captions. The sentence lengths of 4 LLaVa captions are 10, 25, 60, and 75. Generally, longer captions contain more visual details. HICE’s correlation with human evaluation results is shown in Sec. 4.5 system-level correlation.

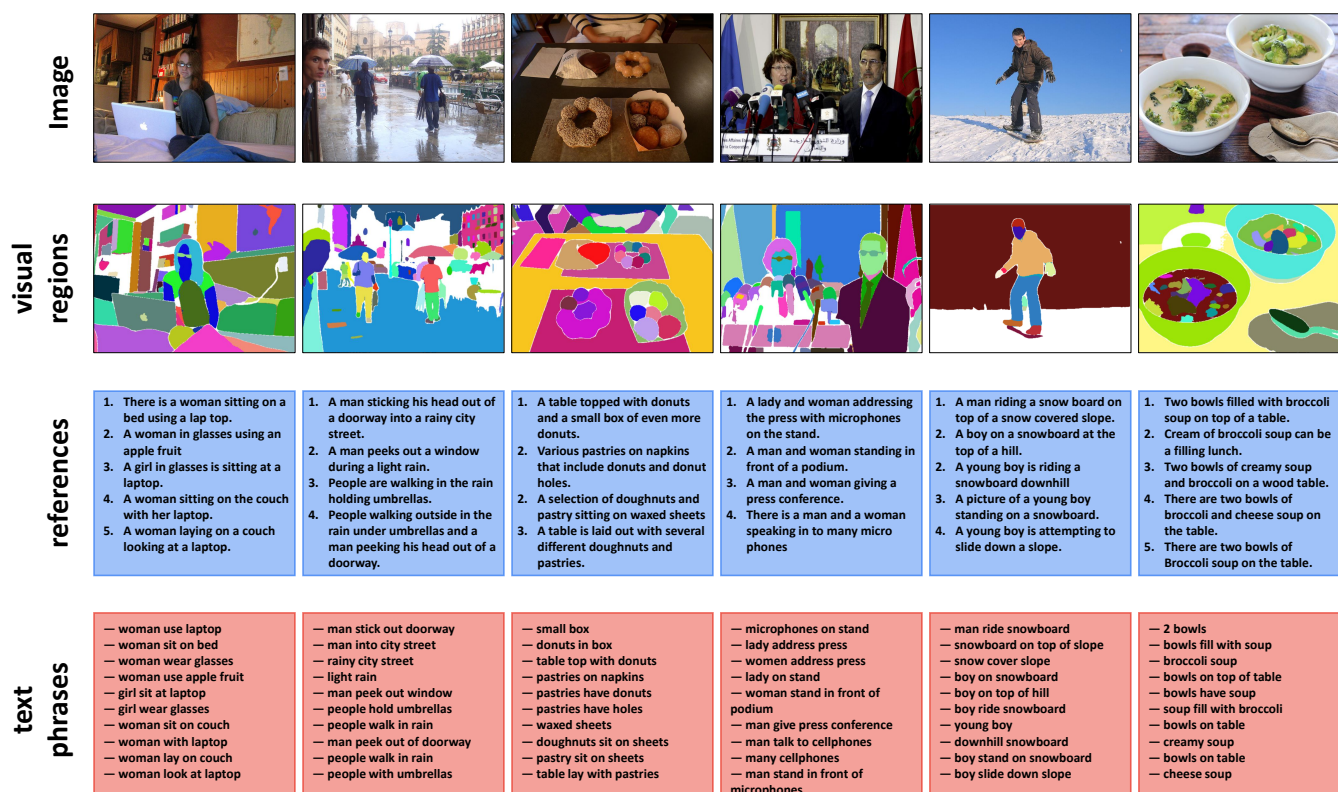


Figure 1: More qualitative examples.

REFERENCES

[1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. 2022. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. *2022 IEEE International Conference on Image Processing (ICIP)* (2022), 966–970. <https://doi.org/10.1109/ICIP46576.2022.9897990>

[2] Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. *arXiv preprint arXiv:2305.06002* (2023).

[3] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 34–39.

[4] Zhuang Li, Yuyang Chai, Terry Zhuo Yue, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023. FACTUAL: A Benchmark for Faithful and Consistent Textual Scene Graph Parsing. *arXiv preprint arXiv:2305.17497* (2023).

[5] Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. 2023. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579* (2023).