CLIP-Adapted Region-to-Text Learning for Generative Open-Vocabulary Semantic Segmentation

Supplementary Material

User: If an image contains [cap1], [cap2]... [capk]. Are/Is/Does/Do the [target] in this image? Answer the Yes-No question with only the single word with nothing else.
ASSISTANT: Yes | No

========== Example 1 ==========
User: If an image contains a dog on the beach, green grass on the ground, a large rock in the background, a dog is standing in the dirt. Is the grass green in this image? Answer the Yes-No question with only the single word with nothing else.
ASSISTANT: Yes.

========== Example 2 ==========
User: If an image contains a red and white striped shorts, a man wearing a white shirt, a boy with a red shirt, a blue and white bag, a little boy sitting on the ground with a frisbee. Does the child stand in this image? Answer the Yes-No question with only the single word with nothing else.
ASSISTANT: No

========== Example 3 ==========
User: If an image contains a man wearing a white shirt, a section of green grass, a man wearing a blue shirt, a baseball player, a green grassy field, two men playing soccer. Are there three men in this image? Answer the Yes-No question with only the single word with nothing else.
ASSISTANT: No

Figure 1. We utilize captions to assist Llama2 in understanding vision-related questions, facilitating the evaluation of our model's semantic expressiveness. Blue indicates the captions extracted by the model, orange highlights the targets of interest for evaluation, and red represents Llama2's answers.

In the supplementary material, we provide details and additional results of the evaluation for the model's expressiveness in Sec. 1. Sec. 2 provides a detailed description of the parsing and filtering techniques in data processing, while Sec. 3 includes additional visualizations.

1. Evaluation of Semantic Expressiveness

We adopt evaluation types from AMBER [3], including coverage, existence, and attribute, to better assess our model's expressiveness. For coverage, we use the Cover metric to reflect the comprehensiveness of category coverage, while Precision, Recall, and F1 are used for the other types. Specifically, existence assesses the model's ability to perceive category presence, and attribute evaluates its ability to express features like color, action, and quantity. Following AMBER, the Cover metric is obtained by extracting key nouns from captions. In contrast, evaluating existence and attribute requires Yes-No questions to query a multimodal large language model [1, 4]. Since our model lacks conversational capabilities, we instead save region-level captions and use them to construct prompts for Llama2 [2], allowing it to answer Yes-No questions for these metrics. As shown in Figure 1, for each image, we formulate questions based on a set of extracted captions $[cap_1], [cap_2], \ldots, [cap_k]$ and a target $[target]$, and use the instruction "Answer the Yes-No question with only a single word and nothing else." to constrain Llama2's output. This allows the purely textual Llama2 model to indirectly interpret visual content and determine the existence of the $[target]$ in the image. For example, in Example 2 of Figure 1, the model can infer that "child stand" does not exist based on the caption "a little boy sitting on the ground with a frisbee" and thus responds with "No".

In Table 9 of the main paper, we present the results of our method on the evaluation types of coverage, existence, and attribute. Notably, the attribute evaluation encompasses three aspects: **state**, which reflects properties such as color and shape; **number**, which captures the quantity of target objects; and **action**, which assesses the target's actions. The detailed results for these three aspects are shown in Table 1. Overall, our method demonstrates strong performance in state and number compared to most multimodal large language models. This can be attributed to the effective model design and the integration of data processing strategies, enabling our approach to provide detailed descriptions of individual regions. However, our method shows relatively weaker performance in action recognition. This limitation

Table 1. Detailed evaluation results for attribute expression. State refers to attributes such as color and shape, number indicates the quantity of the target, and action focuses on the target's movements.

Methods	Attribute (Att)			Att-State			Att-Number			Att-Action		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
mPLUG-Owl	87.5	13.2	22.9	86.4	17.1	28.5	92.3	1.2	2.4	92.5	21.7	35.2
LLaVA	78.9	35.1	48.6	76.2	43.5	55.4	91.6	15.7	26.8	87.1	35.9	50.8
CogVLM	79.7	44.9	57.4	76.7	42.1	54.4	79.8	39.0	52.4	91.4	77.5	83.9
LLaVA-1.5	88.0	51.0	64.6	86.6	43.8	58.2	87.7	59.1	70.6	93.8	73.0	82.1
mPLUG-Owl2	88.1	61.5	72.4	87.4	59.4	70.5	89.1	60.6	72.1	94.0	76.3	84.1
InstructBLIP	75.9	76.7	76.3	73.5	81.4	77.2	78.9	61.2	68.9	85.4	88.6	87.0
Ours+Llama2	66.3	83.9	74.1	65.2	86.4	74.3	72.4	73.7	73.0	61.4	95.2	74.6

Figure 2. Examples of parsing and filtering techniques for processing image-text pair data.

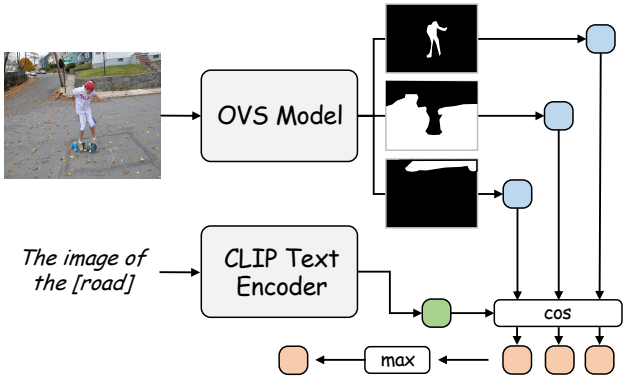


Figure 3. Details of the filtering process. Blue represents region-level proposal visual features, green denotes CLIP text embeddings, and orange indicates similarity scores.

arises because action understanding often requires reasoning across multiple regions. While our model excels at independently describing isolated semantic regions, it struggles with cross-region reasoning.

It is important to emphasize that our work primarily focuses on developing a recognition model with strong capabilities in identifying and describing independent semantic regions, including attributes such as color, shape, and quantity. Moving forward, we will try to explore approaches for modeling relationships across multiple regions, thereby improving the model's ability to understand action-related attributes.

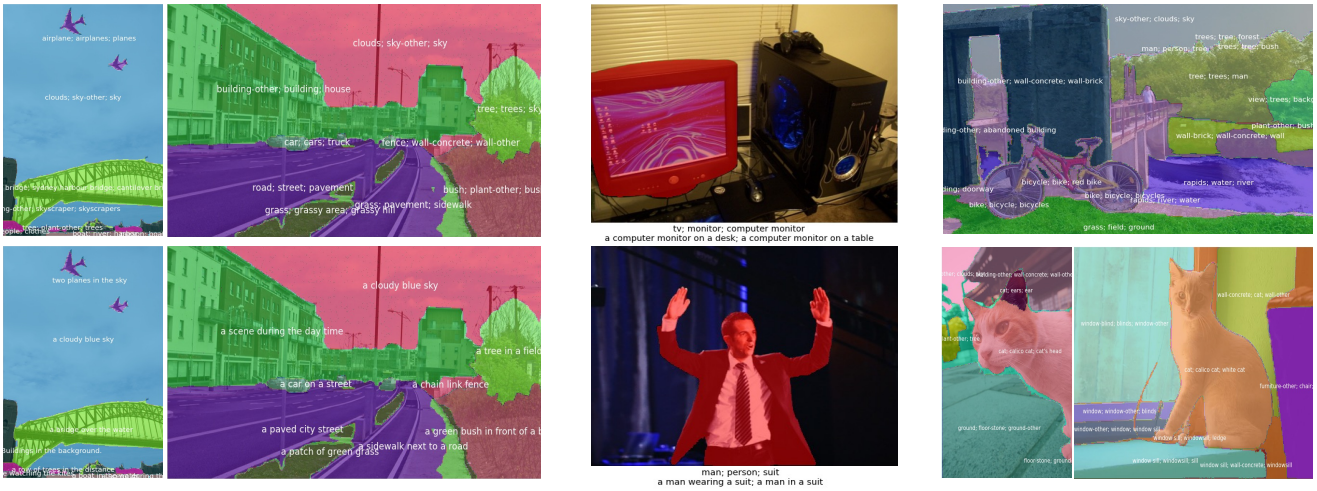
2. Details of Parsing and Filtering in Image-Text Data Processing

Enabling a language model to comprehend visual content typically requires a large amount of data, but existing segmentation datasets provide only limited resources. To address this, we incorporated object detection data and image-text pairs as weakly-supervised data into the training process. While object detection data, which includes class labels, can be directly utilized, image-text pairs only provide captions, making them unsuitable for direct use. To overcome this limitation, we designed a **parsing** and **filtering** pipeline to process the image-text pairs, as illustrated in Figure 2. Specifically, we used spacy to extract noun phrases from the captions. However, this extraction introduces a significant number of invalid nouns, such as "history", which do not have corresponding visual entities in the image.

To address this limitation, we implemented a filtering step to eliminate irrelevant phrases. As shown in Figure 3, we first employed a pre-trained open-vocabulary segmentation (OVS) model and calculated the similarity between each noun phrase and the extracted regions using CLIP Text Encoder. The maximum similarity score across all regions was assigned as the phrase's similarity to the image. Second, after processing all image-text pairs data, we ranked the similarity scores of all extracted phrases and discarded the bottom fraction $R \in [0, 1]$ with the lowest confidence. In this way, we have completed the processing of image-text pair data for our method.

3. More visualization results

In Figure 4, we present additional visualization results. In Figure 4(a), when using our original MaskFormer-like mask extractor, we rely on semantic-agnostic masks. Although these masks provide less refined boundaries, they still achieve satisfactory semantic segmentation results. Notably, our model generates diverse categories and captures higher-level semantics, such as "*two planes in the sky*".



(a) Semantic Segmentation with categories and captions

(b) Visual understanding based on a single mask through integration with SAM

(c) Visual understanding based on an image through integration with SAM

Figure 4. More visualization results.

Furthermore, the model's design, which processes binarized masks, allows seamless integration with tools like Segment Anything Model (SAM). Leveraging SAM's capabilities, we enable interactive understanding, as illustrated in Figure 4(b), where specific semantic targets can be selected through clicks, as well as holistic scene understanding, shown in Figure 4(c). It is also worth noting that SAM can extract more localized targets, such as "*ear*". This is because SAM uniformly samples across the entire image, and adjusting the sampling distance influences the granularity of the extracted masks. Therefore, integrating with SAM enables some control over the granularity of the objects understood by the model, further broadening the application scenarios of our approach.

References

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[3] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 1

[4] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1