CVPR
JUNE 17-21, 2024
SEATTLE, WA

GROUNDHOG: Grounding Large Language Models to Holistic Segmentation

Yichi Zhang[1], Ziqiao Ma[1], Xiaofeng Gao[2], Suhaila Shakiah[2], Qiaozi Gao[2], Joyce Chai[1]
[1]University of Michigan; [2]Amazon AGI

TL;DR

We present GROUNDHOG, a multimodal large language model that:

- Enables pixel-level language grounding to segmentation masks of diverse semantic granularity, including objects, background stuff, parts, text, and groups of instances;
- Integrates seamlessly with any segmentation foundation model (e.g., SAM) to support arbitrary spatial prompts as input;
- Achieves state-of-the-art performance on various vision-language grounding tasks, reduces hallucination, and enhances interpretability and diagnosability.

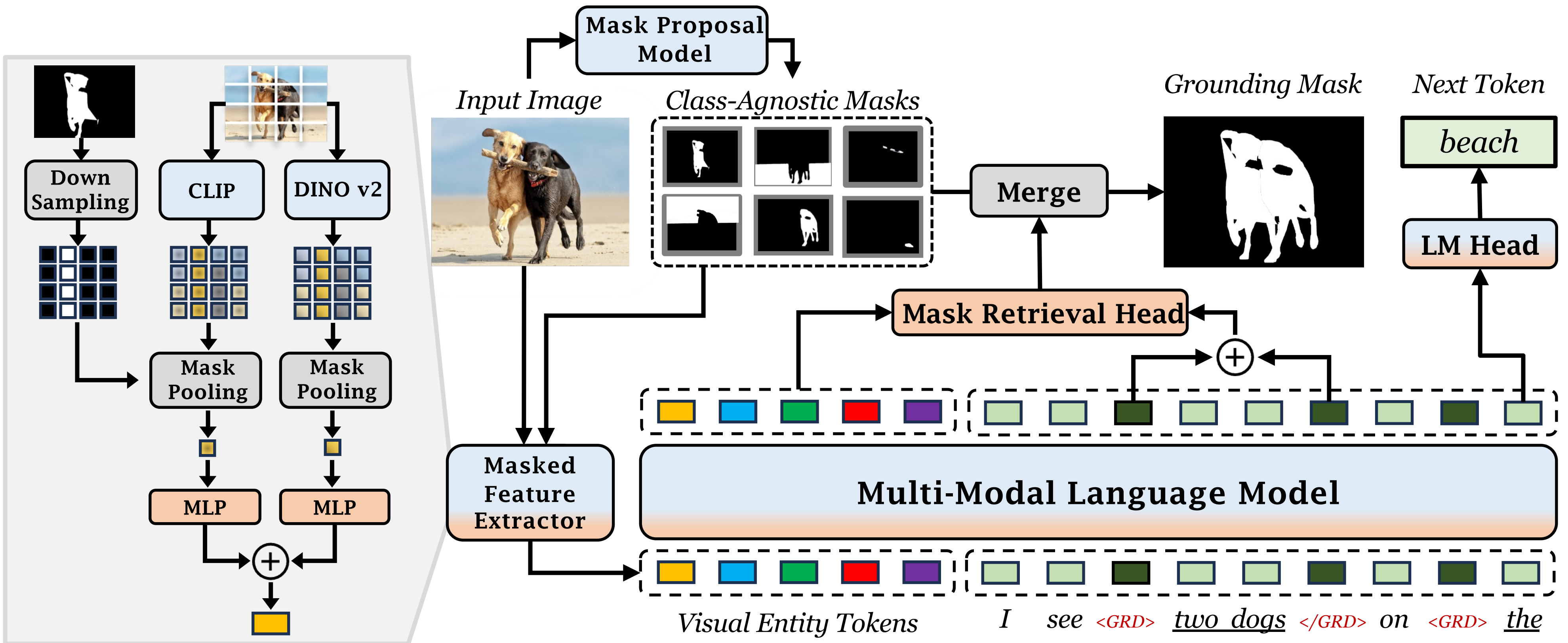
MODEL ARCHITECTURE

GROUNDHOG formulates the grounding process as an entity segment selection problem:

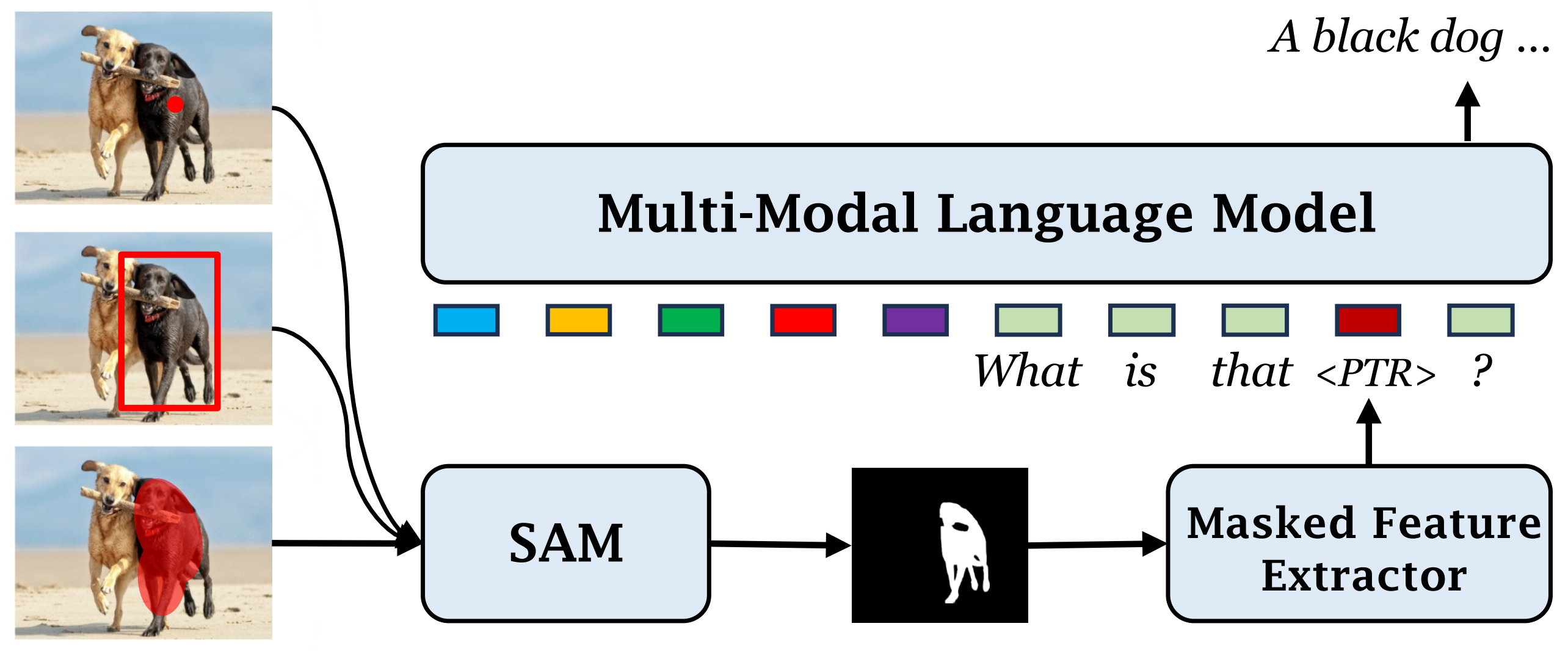
- Propose entity segmentations where the masks capture regions with discernible semantic content;
- Recognize the retrieved entities through the understanding of both visual and language context.

Given a set of class-agnostic entity masks generated by a mask proposal model, GROUNDHOG first leverages a masked feature extractor to extract features from each mask, serving them as entity-centric visual tokens for the LLM backbone. To ground the generated text to these visual entities, we introduce special grounding tokens `<GRD>` and `</GRD>` and use their output embeddings to retrieve the entities. The retrieved masks are then merged to form the final grounding mask.

To enable holistic entity mask proposals, we train an enhanced Mask2Former model on a combination of existing segmentation datasets encompassing various semantic granularities, with 50 additional queries each for segmenting parts and text regions alongside the original 200 entity queries.

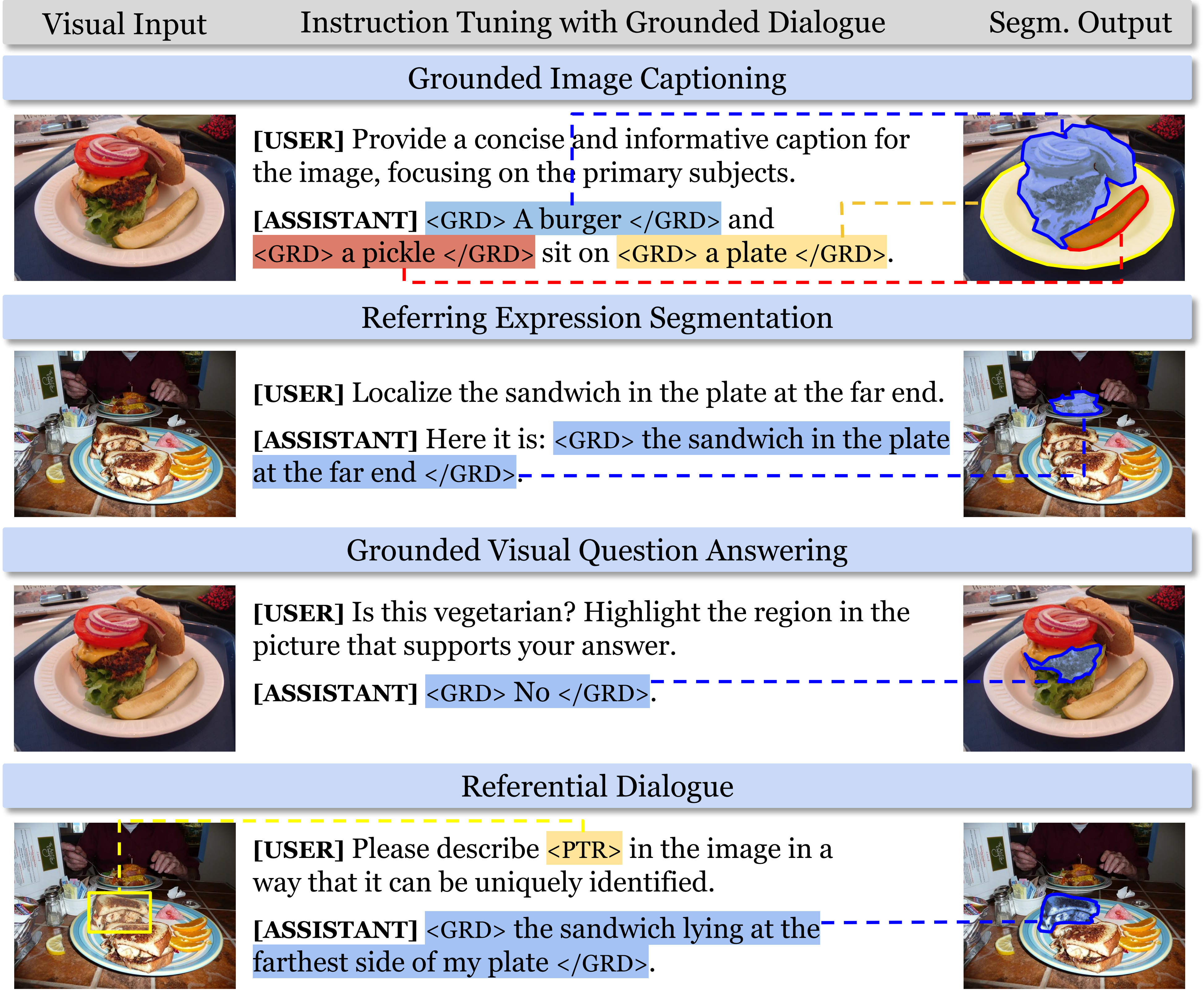


We use a pointer token `<PTR>` to refer to a specific point or region in the image input. To form its input embedding, we leverage an existing interactive segmentation model (e.g. SAM) to obtain the mask from the prompt and extract the mask feature using the same mask feature extractor.



DATASET

We introduce the M3G2 dataset for Multi-Modal Multi-Grained Grounding. M3G2 is a comprehensive dataset consisting of 2.5M text-image pairs with 36 sub-problems, derived and augmented from 27 existing datasets with grounded vision-language annotations. The dataset is categorized into four main types: (1) Grounded Image Captioning (GCAP), (2) Referential Expression Segmentation (RES), (3) Grounded Visual Question Answering (GVQA), and (4) Referential Dialogue (RD).



Grounded Image Captioning

[USER] Provide a concise and informative caption for the image, focusing on the primary subjects.

[ASSISTANT] <GRD> A burger </GRD> and <GRD> a pickle </GRD> sit on <GRD> a plate </GRD>.

Referring Expression Segmentation

[USER] Localize the sandwich in the plate at the far end.

[ASSISTANT] Here it is: <GRD> the sandwich in the plate at the far end </GRD>.

Grounded Visual Question Answering

[USER] Is this vegetarian? Highlight the region in the picture that supports your answer.

[ASSISTANT] <GRD> No </GRD>.

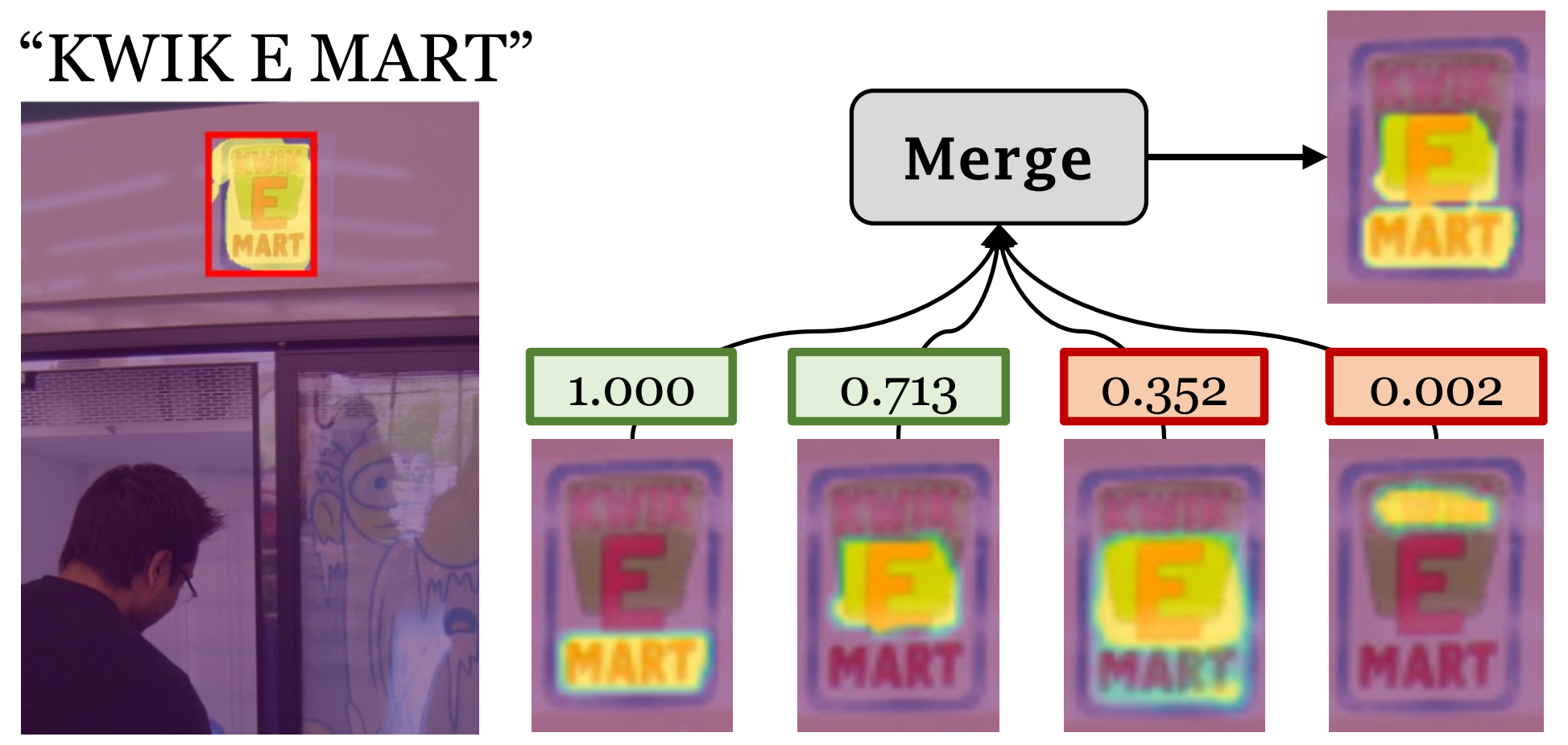
Referential Dialogue

[USER] Please describe <PTR> in the image in a way that it can be uniquely identified.

[ASSISTANT] <GRD> the sandwich lying at the farthest side of my plate </GRD>.

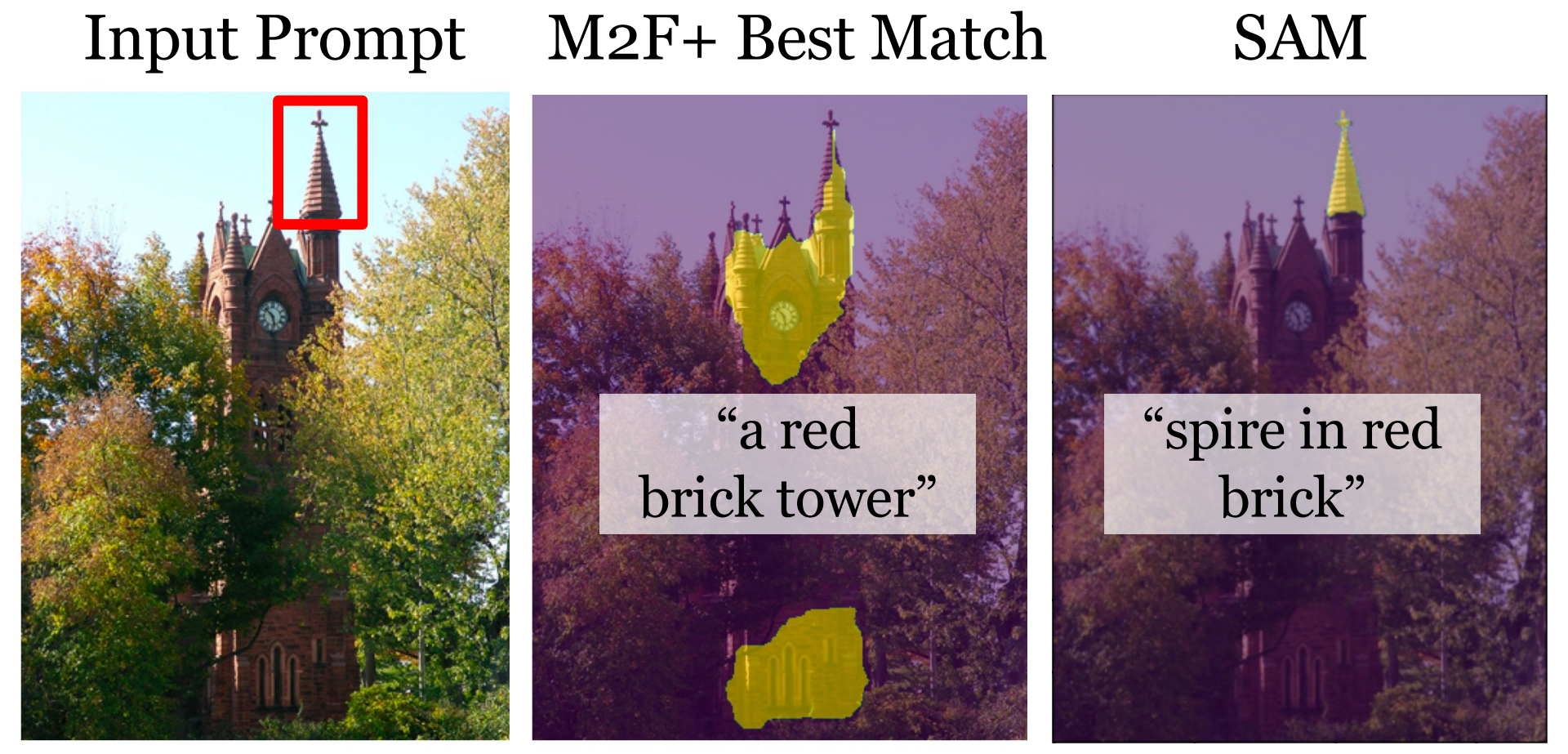
DIAGNOSABILITY

Illustration of GROUNDHOG's entity proposal scoring and grounding mask merging process.

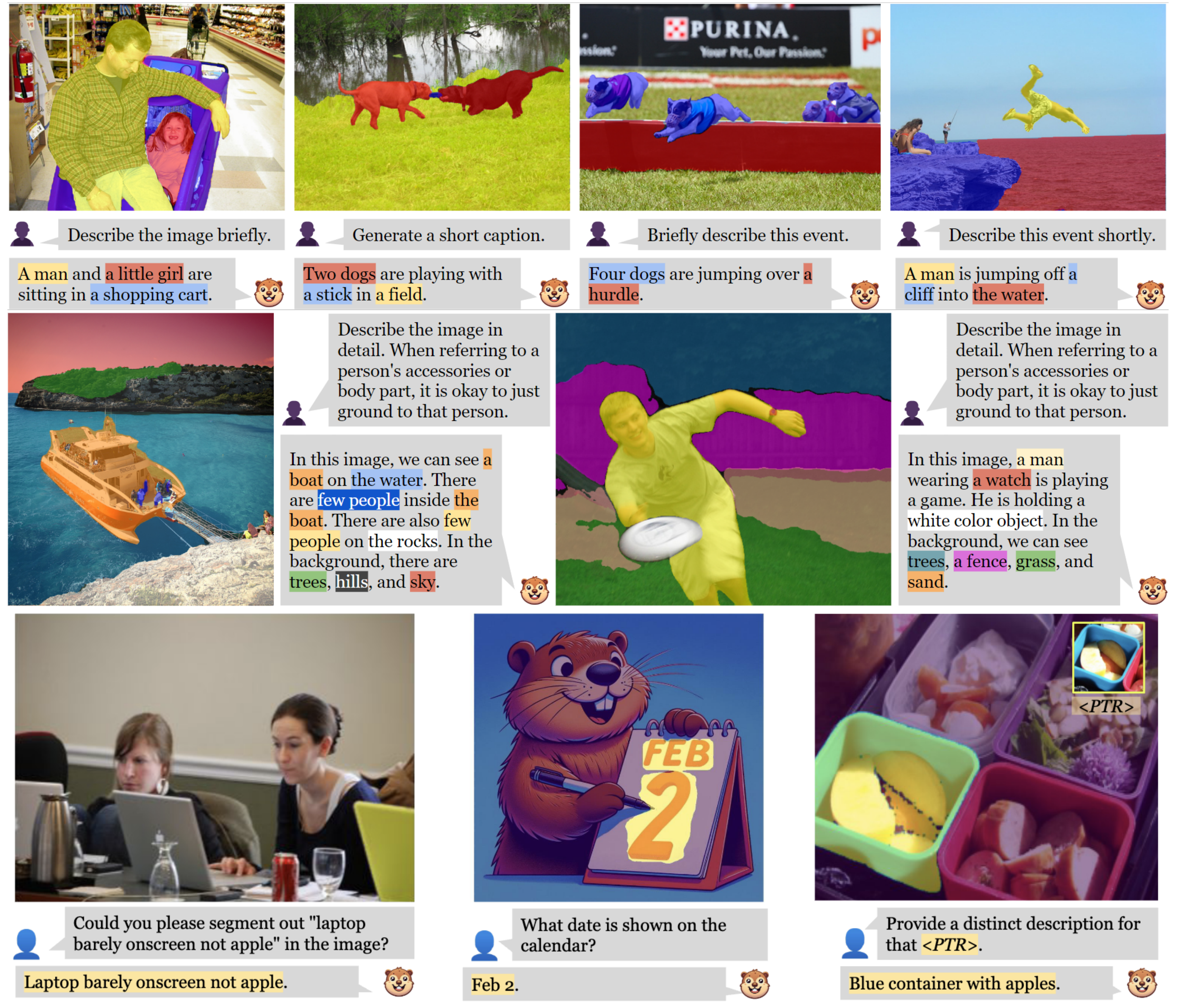


PLUG-IN-AND-PLAY

Leverage SAM to produce a more accurate mask than selecting from the mask proposals.



RESULTS



Describe the image briefly.

A man and a little girl are sitting in a shopping cart.

Generate a short caption.

Two dogs are playing with a stick in a field.

Briefly describe this event.

Four dogs are jumping over a hurdle.

Describe this event shortly.

A man is jumping off a cliff into the water.

Describe the image in detail. When referring to a person's accessories or body part, it is okay to just ground to that person.

In this image, we can see a boat on the water. There are few people inside the boat. There are also few people on the rocks. In the background, there are trees, hills, and sky.

Describe the image in detail. When referring to a person's accessories or body part, it is okay to just ground to that person.

In this image, a man wearing a watch is playing a game. He is holding a white color object. In the background, we can see trees, a fence, grass, and sand.

Could you please segment out "laptop barely onscreen not apple" in the image?

Laptop barely onscreen not apple.

What date is shown on the calendar?

Feb 2.

Provide a distinct description for that <PTR>.

Blue container with apples.

- Achieves state-of-the-art performance on 13 benchmarks, including referring expression segmentation, phrase grounding, grounded caption generation, grounded VQA, and pointer-based QA.
- Reduces object hallucination: obtains 86.33% accuracy on the POPE-adversarial split, surpassing Shikra-13B's 83.10%.
- Ablation study insights: (1) Mask2Former+ excels in non-COCO domains, (2) Combining CLIP and DINOv2 features is superior, (3) Effective mask retrieval representation, (4) Decoupling the mask proposal model and MLLM speeds up training at lower resolution (320px) and boosts performance at higher evaluation resolution.

Setups	RefCOCO+	Flickr30K	TextVQA-X
Mask Proposal Models			
Mask2Former	**67.1**	69.0	9.8
Mask2Former+	66.6	**77.2**	**34.0**
Entity Features			
CLIP	59.8	75.0	32.0
DINOv2	62.3	76.3	28.4
CLIP+DINOv2	**66.6**	**77.2**	**34.0**
Grounding Query			
`<GRD>` only	64.4	67.5	**34.2**
`</GRD>` only	64.4	**77.2**	33.5
Sum	**66.6**	**77.2**	34.0
Eval Input Resolution			
224–480	54.7	67.2	27.6
480–640	65.5	76.7	27.6
800–1024	**66.6**	**77.2**	**34.0**

Table 10. Ablation study on model design choices and evaluation setups. Models are trained on RefCOCO+, Flickr30K, TextVQA-X and tested on corresponding validation sets.