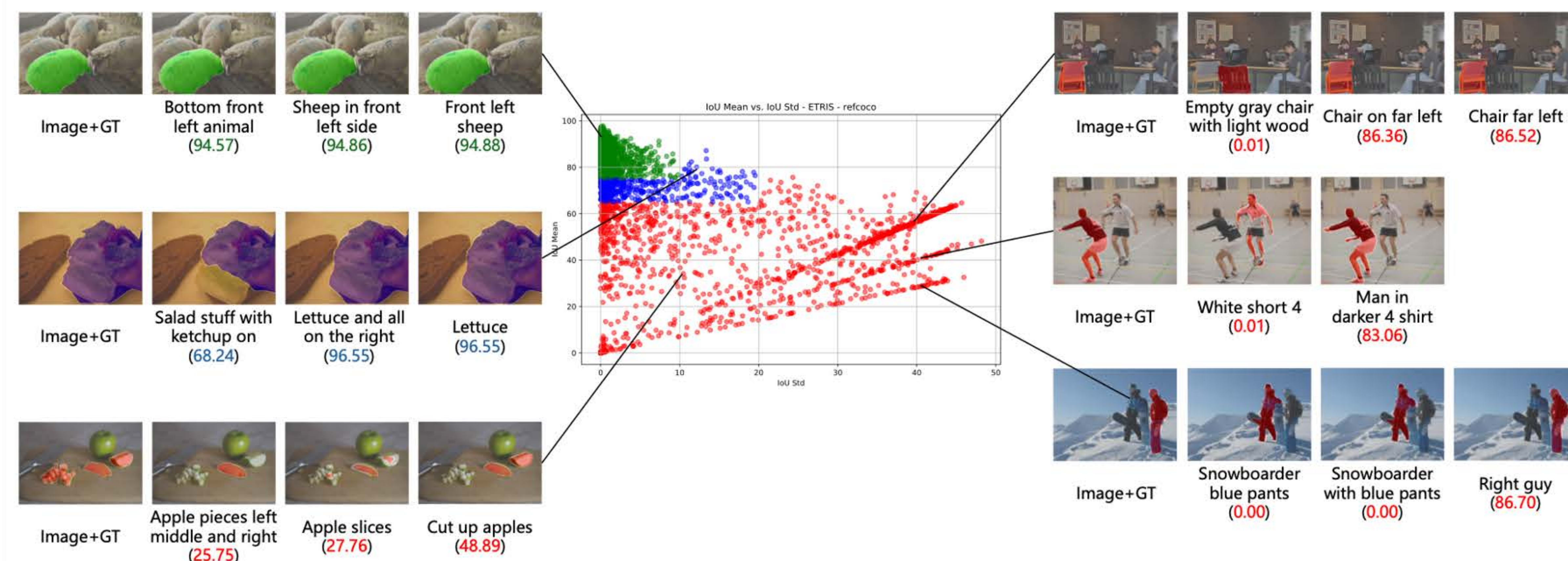




Challenge

- Different sentences, but with similar description of an image.
- Open-Set (free-form image reference expressions).
- Heavy computational costs on training / fine-tuning VLM/MLLM

Setup: Reference Image Segmentation

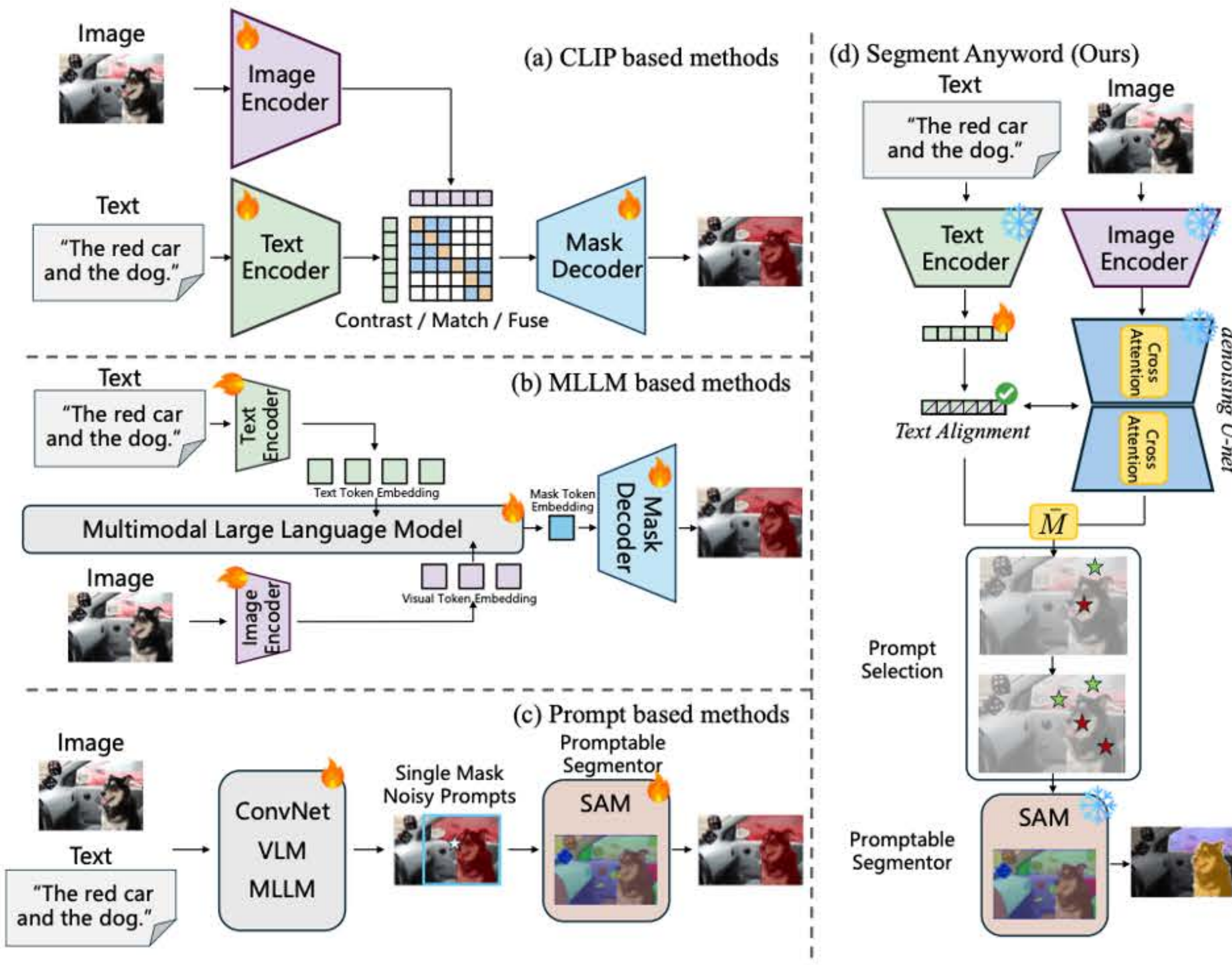


Observation

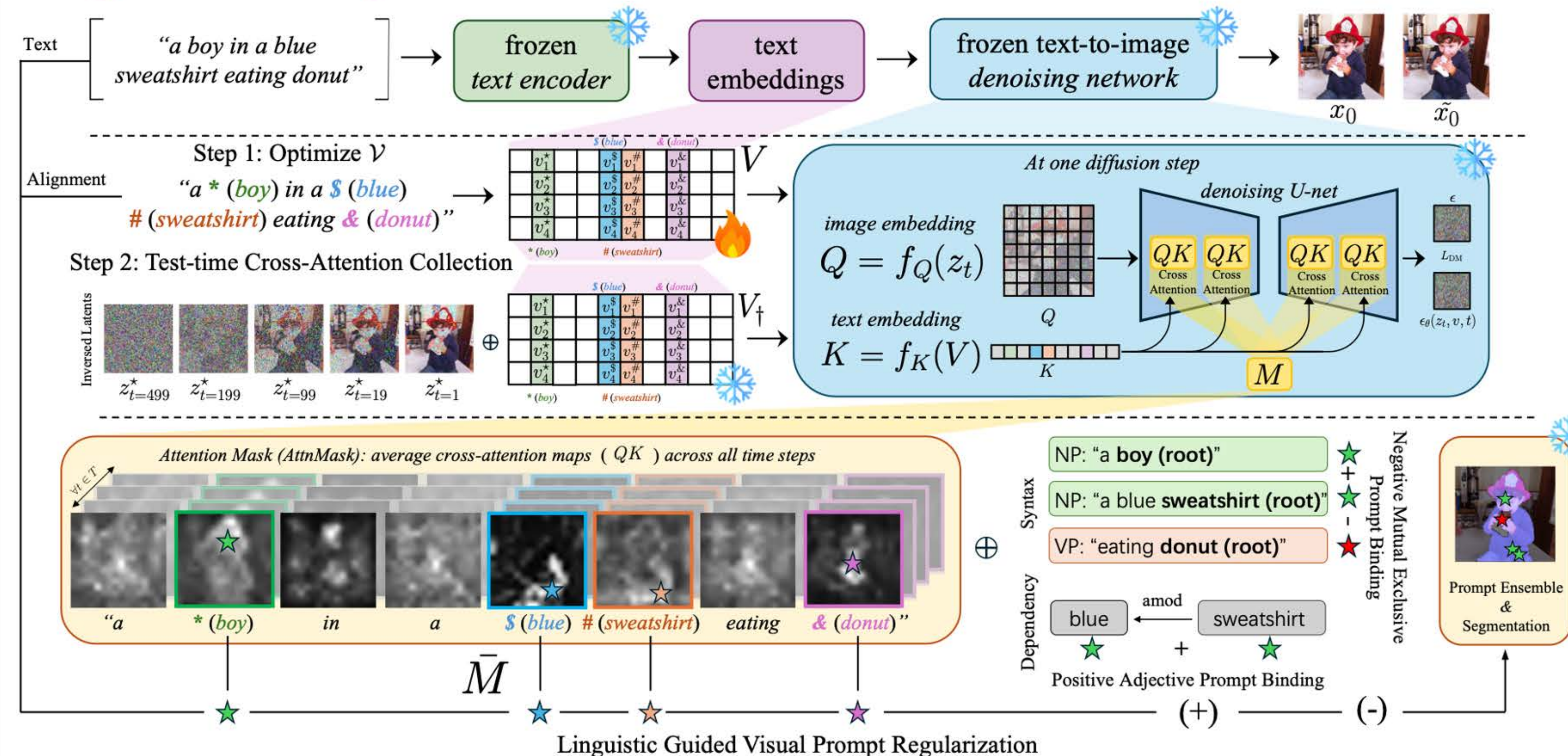
- Strong VLM/MLLM **NOT** stable (sensitive) to expression variation.
- Confounded from expression variation and image complexity.
- A lightweight, test-time adaptation method is desired.

Contribution

- Segment Anyword: A training-free test-time prompt-learning method encodes token-level visual concept.
- Linguistic guided regularization techniques for prompt binding and clustering.
- Superior performance among training free methods across several metrics.



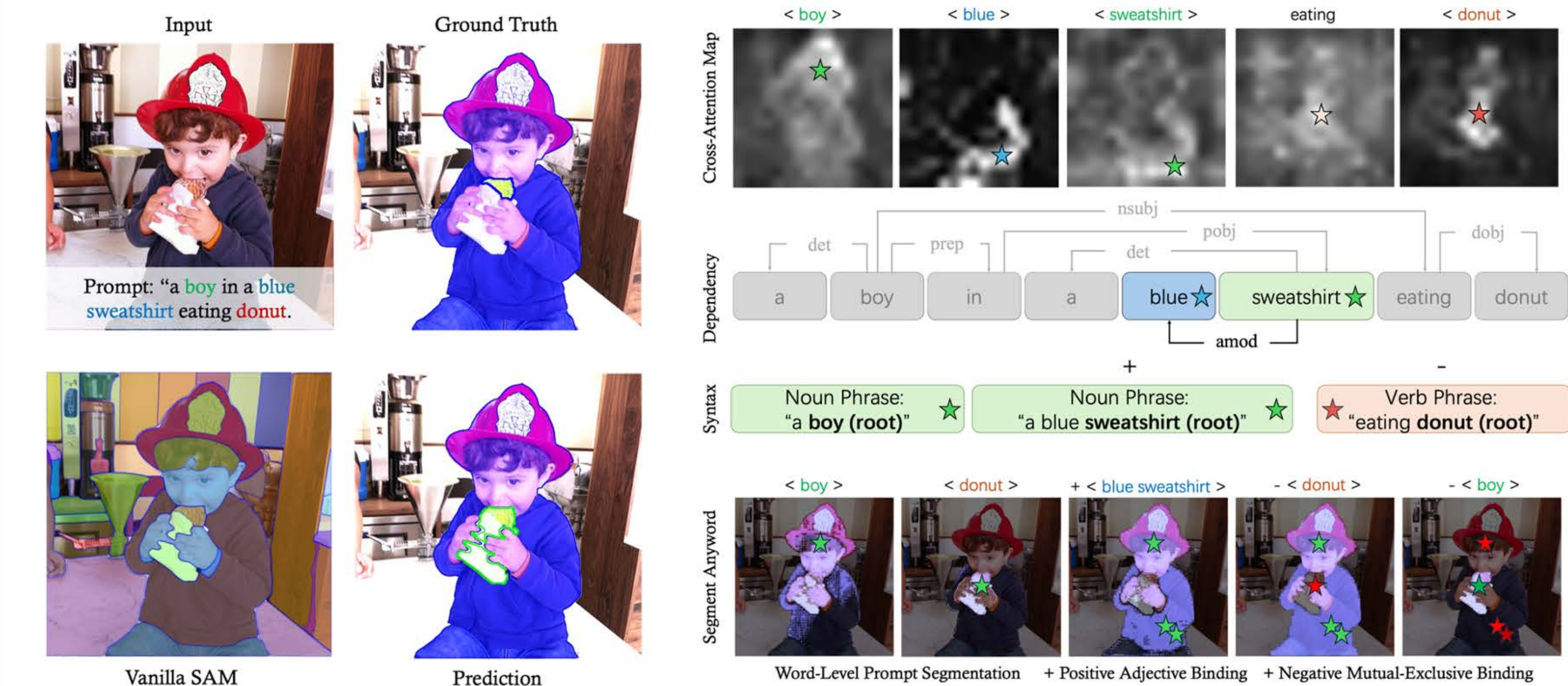
Segment Anyword: Framework



Plain Cross-Attention Cannot Segment Well



Linguistic Guided Visual Prompt Regularization



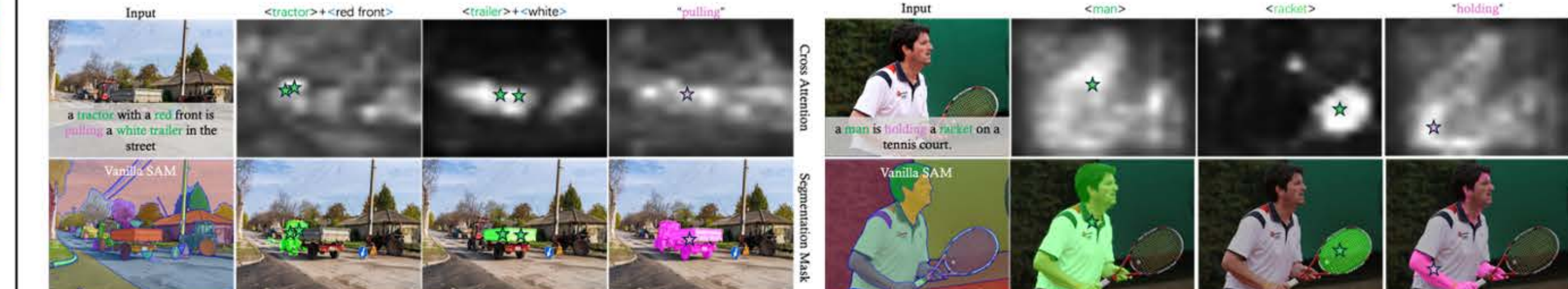
Quantitative Results

Table 2. Open-Set Language Grounded Segmentation Performance on GranDf Dataset validation and test split. “#Images” indicates the number of training images. “-” indicates the result is not reported from original paper.

Method	#Images	GranDf Val			GranDf Test		
		AP50	mIoU	Recall	AP50	mIoU	Recall
<i>Training End-to-End MLLM</i>							
1. BuboGPT	130M	19.1	54.0	29.4	17.3	54.1	27.0
2. Kosmos-2	9M	17.1	55.6	28.3	17.2	56.8	29.0
3. LISA	0.12M	25.2	62.0	36.3	24.8	61.7	35.5
4. GLaMM	11M	28.9	65.8	39.6	27.2	64.6	38.0
5. OMG-LLaVA	74K	26.9	64.6	-	26.1	62.8	-
<i>Fine Tuning MLLM</i>							
6. GLaMM _f	200K	30.8	66.3	41.8	29.2	65.6	40.8
7. OMG-LLaVA _f	200K	29.9	65.5	-	28.6	64.7	-
<i>Training-Free Prompt Learning</i>							
8. Segment Anyword	0	31.3	67.4	40.7	26.6	63.4	34.7
9. Segment Anyword _f w/ GT Text	500	30.2	65.9	42.4	31.1	64.1	33.9

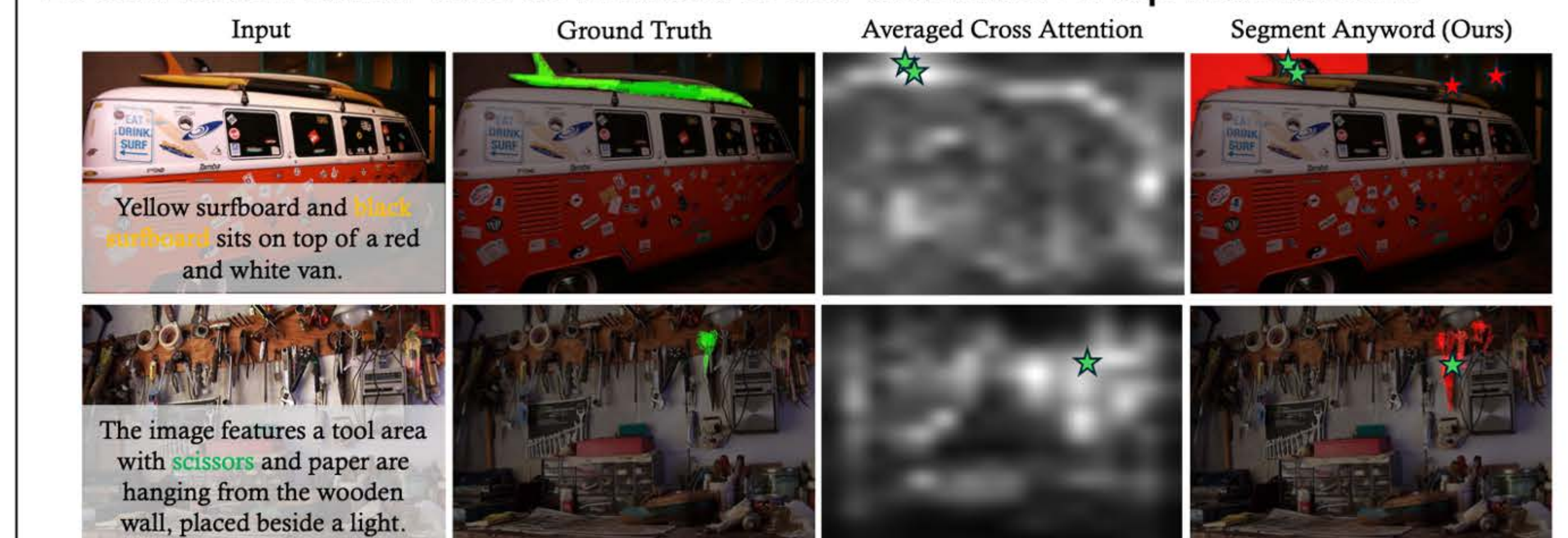
Predicate Segmentation

Predicate words without explicit semantic meaning, our method can be prompted to identify the predicate masks such as **(Left)** relation between object entities and **(Right)** human-object interaction.



Failure Case Analysis

Localization error due to limited cross-attention map resolution.



Acknowledgement: Work was done when Zhihua interned at AZ. Experiments conducted on Sulis HPC funded by EPSRC Grant EP/T022108/1 and the HPC Midlands+ consortium. Poster Template from Bolei Zhou (github.com/zhoubolei/bolei_awesome_posters).