CompoVis: Probing the Compositional Understanding in VLMs with Visualization Representation and Insights

Tong Li
ZJUT Vis
litong@zjut.edu.cn

Guodao Sun
ZJUT Vis
guodao@zjut.edu.cn

Xueqian Zheng
ZJUT Vis
xqzheng@zjut.edu.cn

Haixia Wang
ZJUT Vis
hxwang@zjut.edu.cn

Ronghua Liang
ZJUT Vis
hxwang@zjut.edu.cn

2025 PacificVis

01 Introduction

In vision-language research, "compositional understanding" refers to the ability of a model to jointly handle images and texts.

The model should be able to recognize, comprehend and align each component in vision and textual modality (such as *"red"*, *"hydrant"*, *"man"*) and their combination (the scene semantics of *"a man in white T-shirt leans on a red hydrant"*).

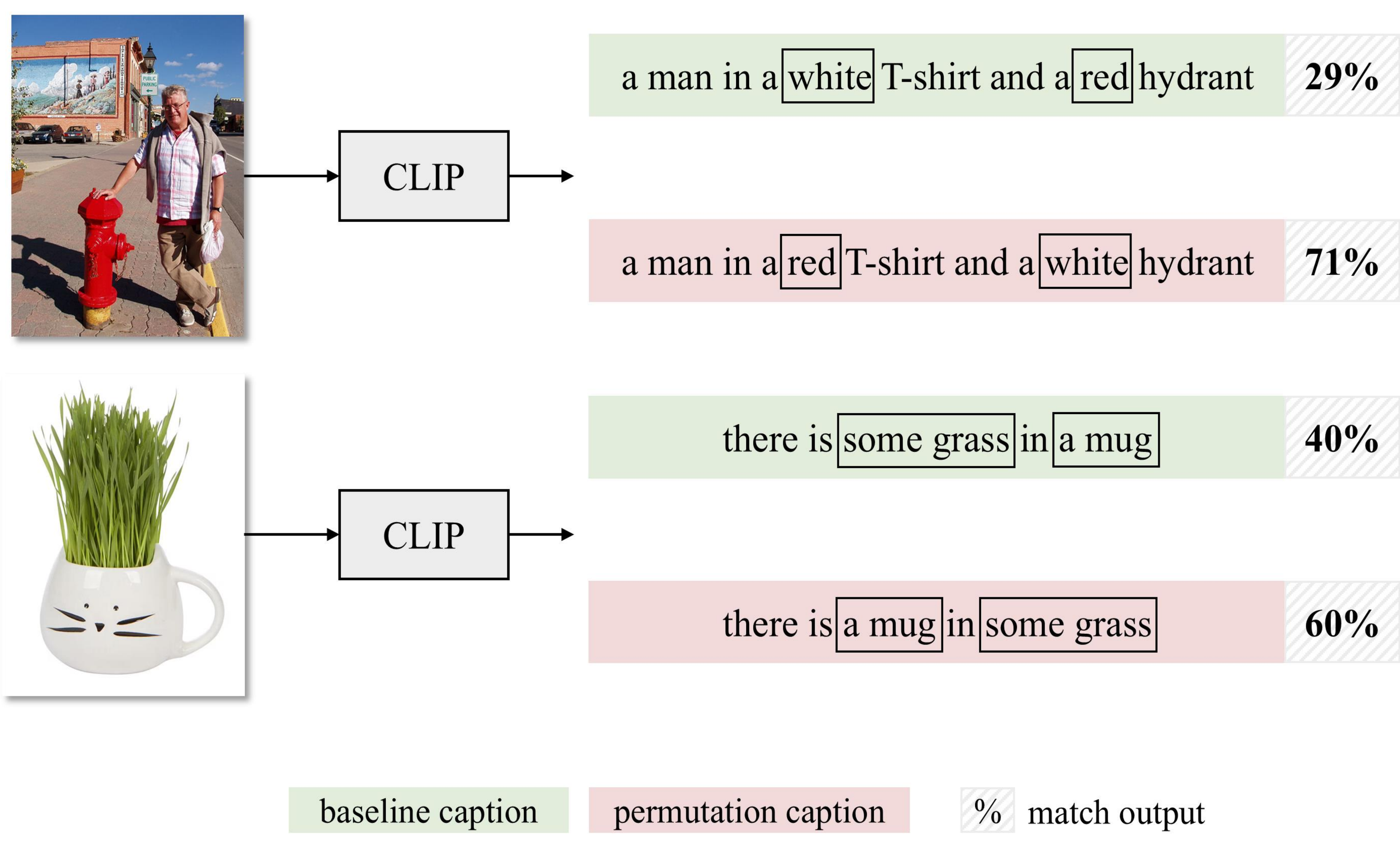
Recent studies reveal even advanced vision-language models (VLMs) struggle with "compositional understanding", particularly when dealing with fine-grained linguistic phenomena.

vision-linguistic compositionality

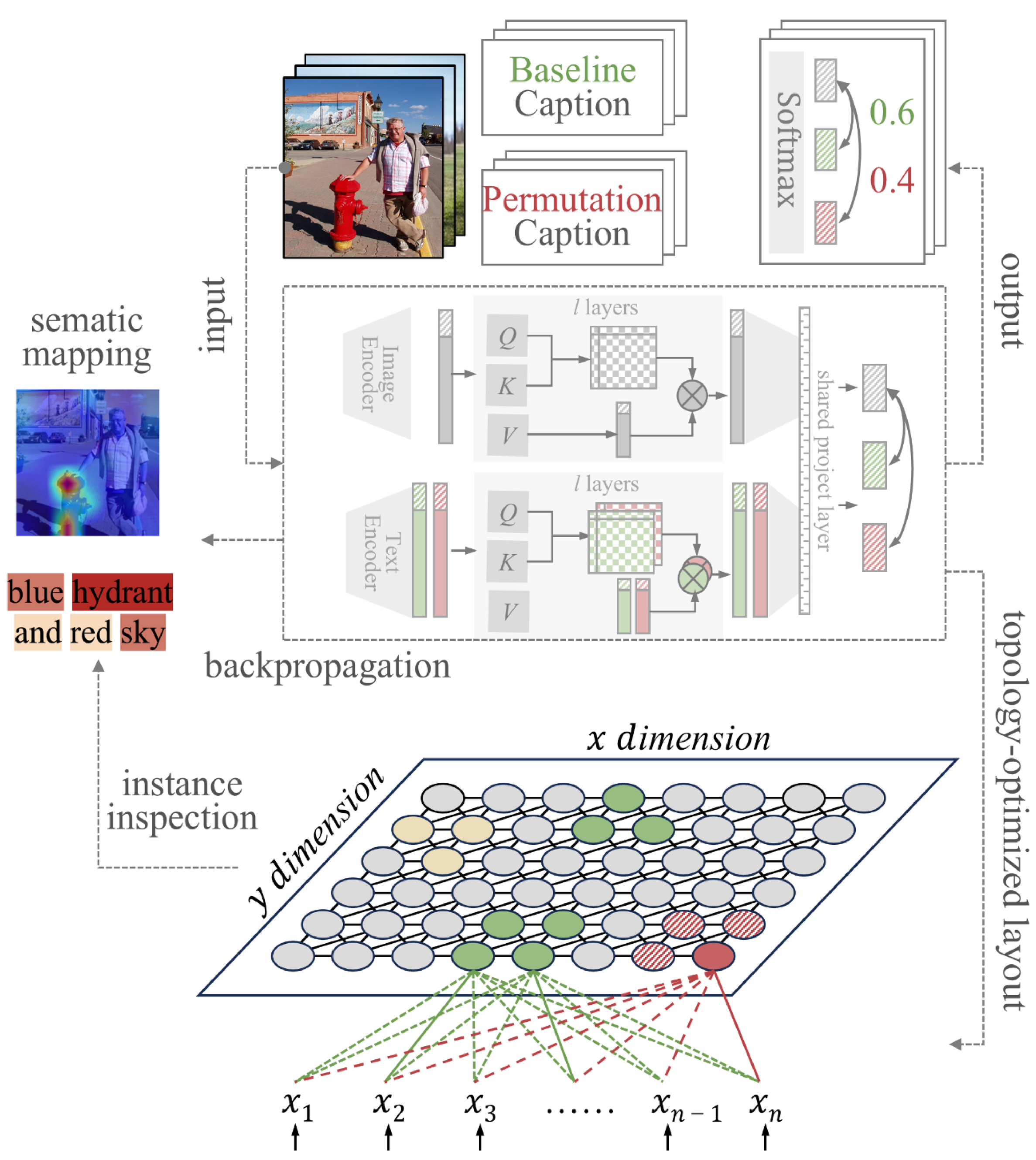
As shown in the figure, CLIP assigns higher scores to the permutation caption. Computer vision methods focus on quantitative metrics and model architectures, constrained by closed datasets and predefined parameters.

To our knowledge, we are the first to elucidate the "bag-of-objects" behavior of VLMs from a visualization perspective. In summary, the contributions of our paper are as follows:

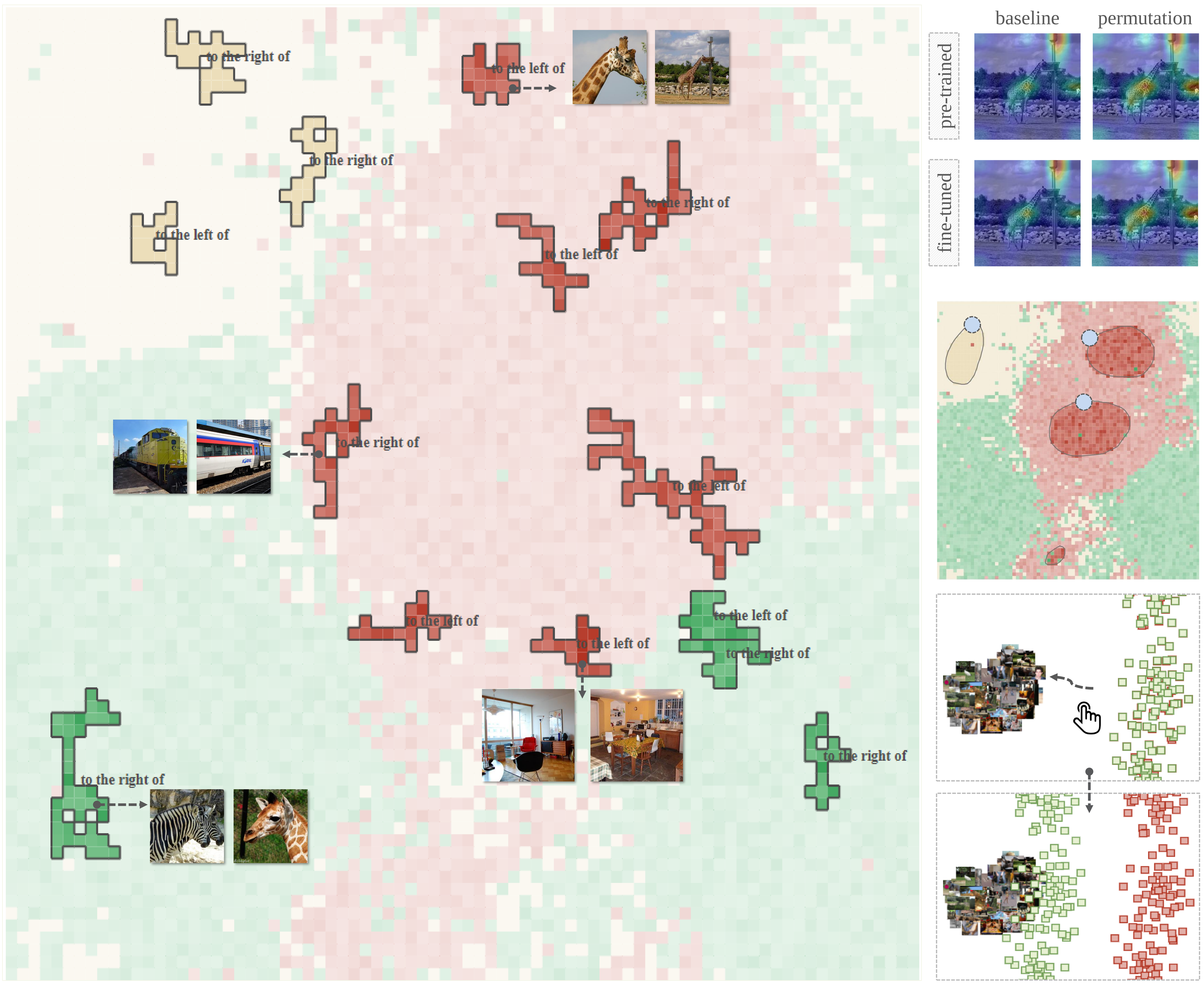
- We introduce a visual analysis tool that supports non-expert users in exploring comprehension defects in VLMs, promoting user-guided fine-tuning evaluation.
- We optimize the topological layout to enhance the visual proximity between semantically consistent samples in large-scale data, providing a reference for grid layout in the visualization community.
- We introduce a permutation dataset, supporting the visual probing of cross-modal alignment ability, and facilitating human-in-the-loop alignment evaluation.

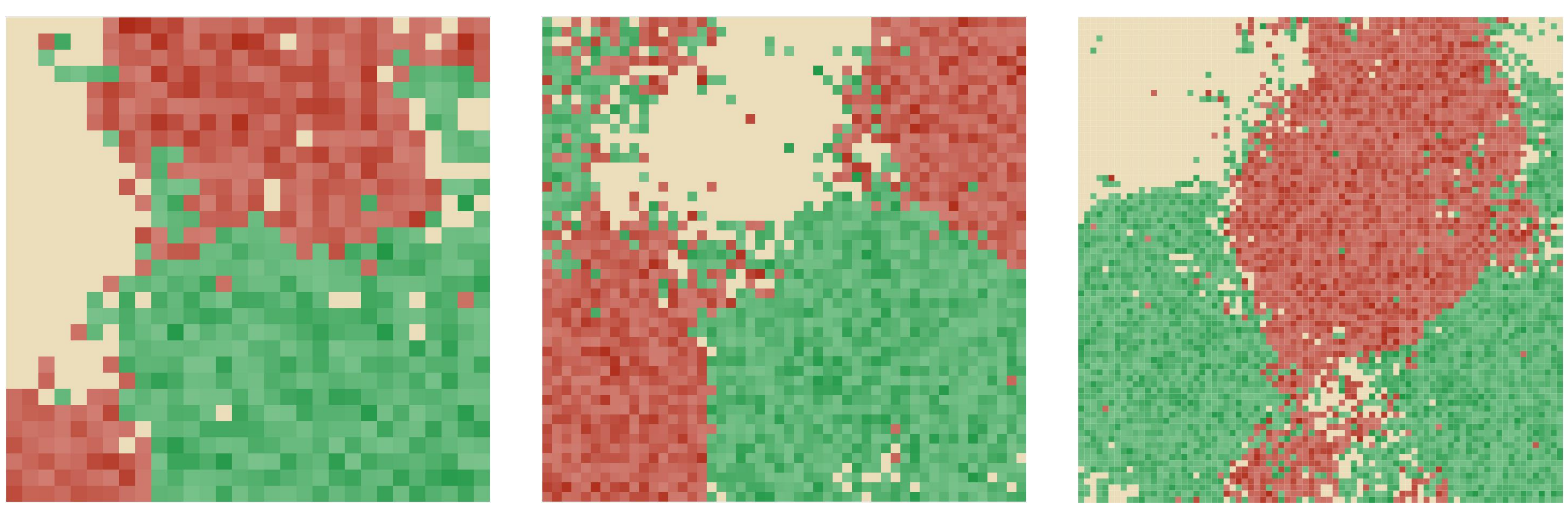


02 Pipeline



03 Visualization Insights





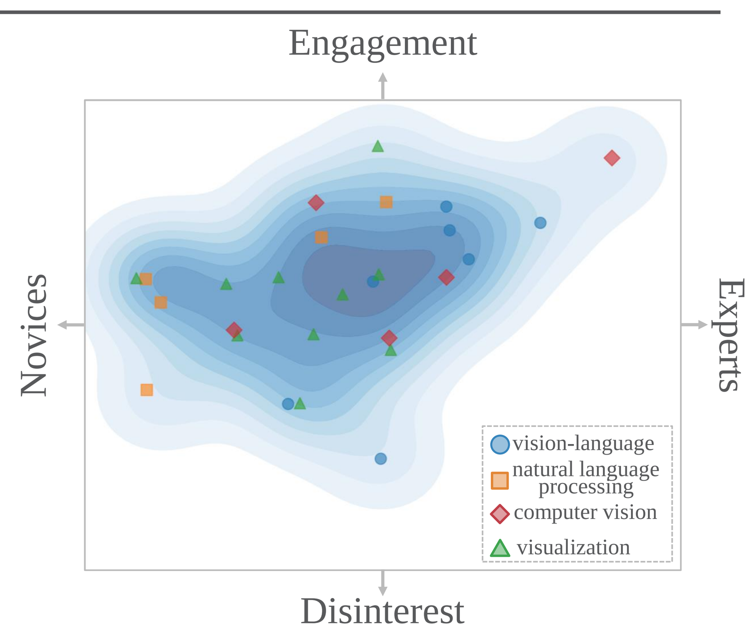
data size: 0.9k data size: 2.5k data size: 3.6k

04 Evaluation

The table compares self-organizing map (SOM), resource-controlled self-organizing map (RC-SOM), and HERC-SOM (ours) in terms of quantization error (QE), topological error (TE), boundary entropy (BE), and neuron activation ratio (NA).

Methods	Indicators			
	QE.	TE.	BE.	NA.
SOM	3.899	0.032	1.113	0.310
RC-SOM	3.969	0.040	1.256	1.0
HERC-SOM(ours)	4.027	0.059	0.918	1.0

Our work sparks discussions among participants from diverse backgrounds. Approximately 90% of participants (24 and 3) stated that, compared to methods relying solely on statistics metrics and closed dataset, CompoVis offers a more innovative and effective approach for investigating modality gaps in VLMs.



We would like to express our sincere gratitude to the members of ZJUT Vis for their valuable contributions. Additionally, we thank all the participants who took part in the evaluation of this research.

TongLi97.github.io litong@zjut.edu.cn ZJUT Vis, Zhejiang University of Technology

浙江工業大學 ZHEJIANG UNIVERSITY OF TECHNOLOGY