Beyond Words: Unlocking Compositionality of Vision-Language Models with Visualization Insights

Tong Li, Wang Xia, Jingwei Tang, Qi Jiang, Yunchao Wang and Guodao Sun*

01 INTRODUCTION

In vision-language research, **"compositional understanding"** involves the ability to process text and image — managing not only words, phrases and their combinations but also recognizing independent elements in images (such as objects, actions, or scenes), understanding how these elements are interrelated, and how they collectively function within a given context.

For instance, the model should be able to recognize each component such as *"lawn"*, *"girl"*, *"white dress"*, *"yellow ball"* as well as their combination the entire scene of *"a girl in a white dress playing a yellow ball on the lawn"* …

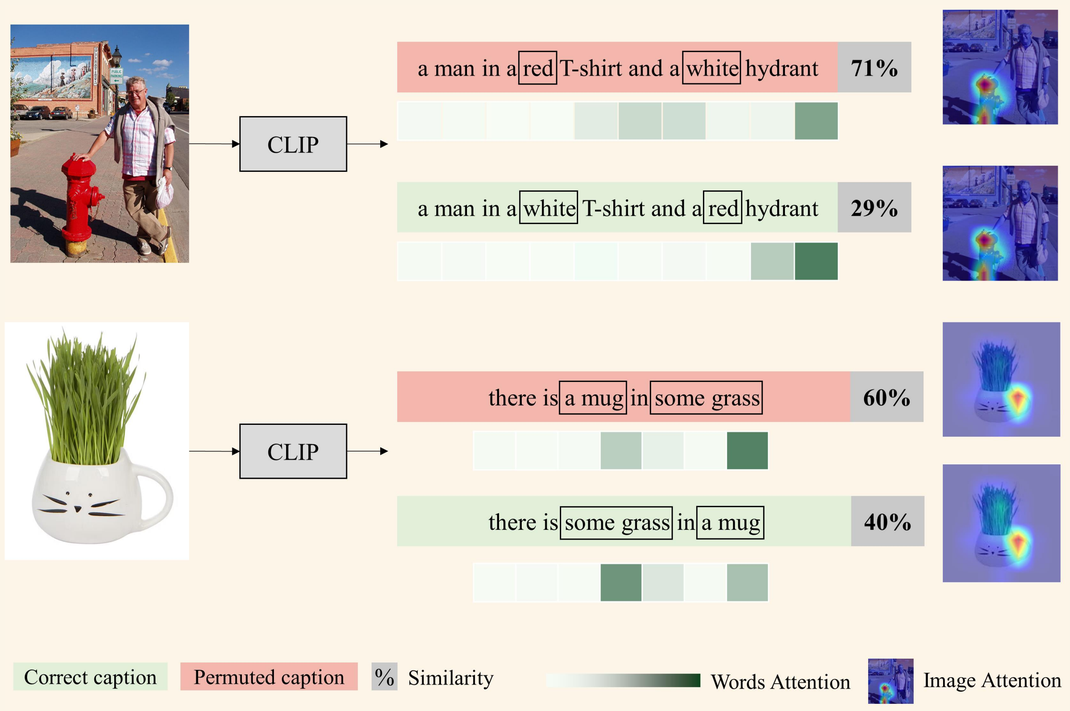
vision-linguistic compositionality

Though vision-language models (VLMs) show high performance on numerous established benchmarks, however, their effectiveness in compositional understanding remains a matter of debate.

Humans can easily perceive the vision differences between images depicting "there is a mug in some grass" and "there is some grass in a mug". It's still unclear how well VLMs grasp the complexity of such vision-linguistic compositionality.

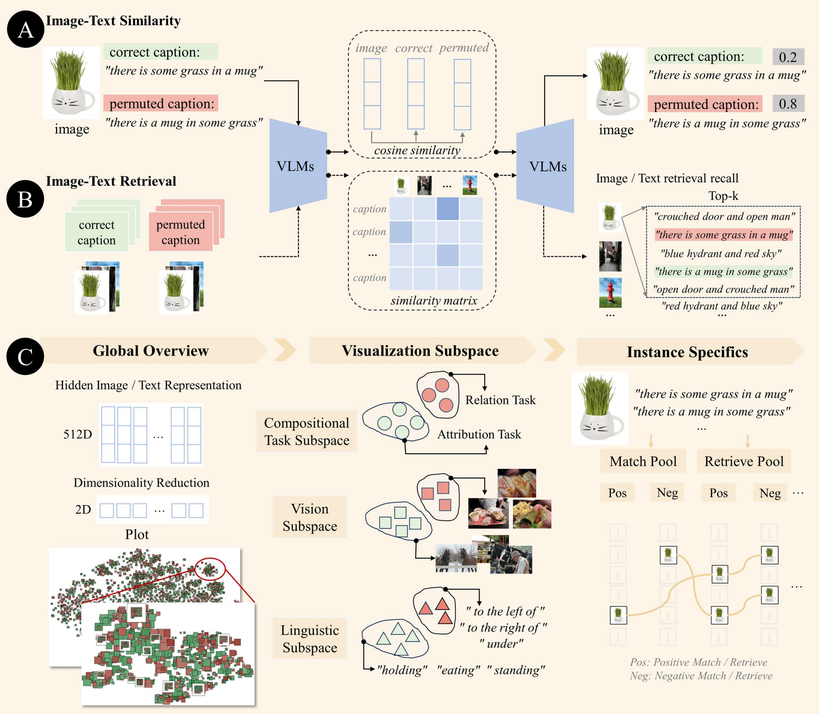
Recent studies have begun to investigate such information. **Even advanced VLMs struggle with challenge of integrating vision and linguistic information**, especially when dealing with fine-grained linguistic phenomena.

As shown in the figure, we permuted the word order of correct image captions to evaluate the response of the vision-language model CLIP. The results indicate that CLIP has a higher acceptance of the permuted captions than the correct ones, contrary to our expectations.

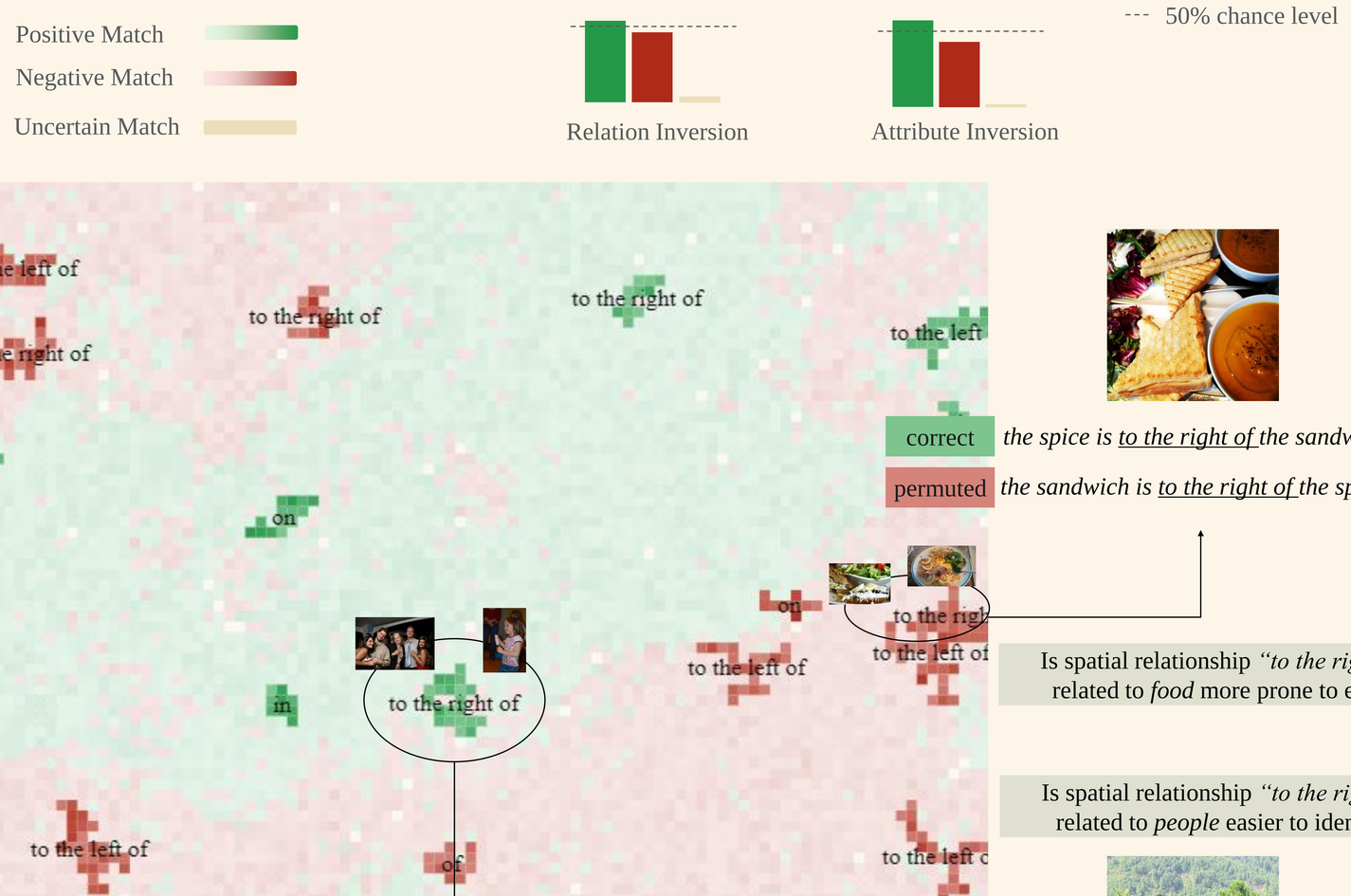


02 PIPELINE

Firstly, we evaluate VLMs' reactions to specific correct and permuted captions by image-text similarity matching. (b) Further, we add permuted captions to "pollute" the data pool and perform a fine-tuned image-text/text-image retrieval. (c) Lastly, based on above data insights, we employ a global-subspace-instance visual analysis approach and develop publicly available tools. It enables users to dynamically explore VLMs' behavioral patterns when processing different vision-linguistic structures.



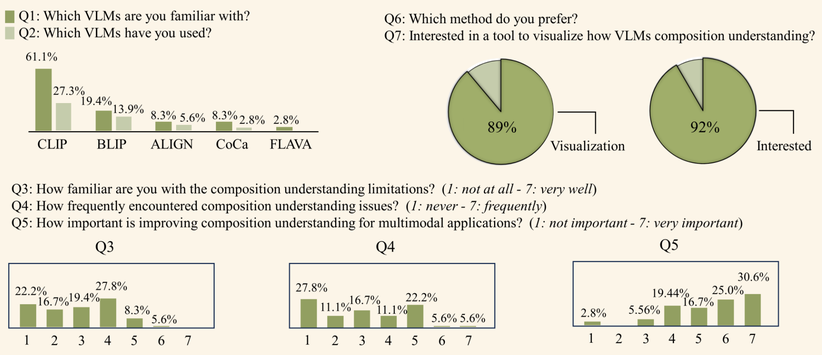
03 VISUALIZATION INSIGHTS



04 USER SURVEY

Our user survey (20 and 16) found that many participants experienced poor performance in composition understanding when using popular models like CLIP, BLIP, and ALIGN. Notably, **5.6% participants frequently encountered these issues**. 89% participants prefer visualization methods to analyze this issue, and **92% participants look forward to the introduction of related visual analysis tool**.

Also, the survey reveals that while many researchers use VLMs, only 13.9% of participants recognize their limitations in integrating vision and linguistic information. 90% participants believe that enhancing this capability could advance related technologies and impact multi-modal applications.



Tong Li (李童)

litong@zjut.edu.cn

TongLi97.github.io

Zhejiang University of Technology

Biography

I am a PH.D. student in ZJUT VIS Lab. My research focus is developing interactive visual analytics approaches for analyzing vision-language models and images / video data (associated with computer vision).

Guodao Sun (孙国道)

guodao@zjut.edu.cn

https://godoorsun.org/

Zhejiang University of Technology

ChinaVis 2024

Hong Kong, China
July 22-25, 2024