Qeios

Peer Review

Review of: "Are Vision-Language Models Truly Understanding Multi-vision Sensor?"

Dong Huo[1]

1. University of Alberta, Canada

I have two concerns:

1. The training set is generated with ChatGPT. The authors mentioned that existing VLM models do not work well on multi-sensor data; then how can we trust ChatGPT?
2. Does the model trained for multi-sensor VLM perform worse on RGB images? In the tables of the paper, it never shows the performance on the RGB images. I wonder if there is a performance drop.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/39KNZ3