Who Did What to Whom? Visually Grounded Role Assignment in Humans but Not in Vision–Language Models

Enjie Xu (xuenjiek15@gmail.com)[1], **Ning Tang**[2], **Alvin Tang**[2], **Joshua K Hartshorne**[3], **Tao Gao (taogao@ucla.edu)**[1,2]

[1]Department of Communication, University of California, Los Angeles
[2]Department of Statistics & Data Science, University of California, Los Angeles
[3]Communication Sciences and Disorders, MGH Institute of Health Professions

Abstract

Event role assignment is central to event understanding in both language and vision. Crucially, the foundation of this semantic structure is likely rooted in visual experience. However, it remains unclear whether recent vision–language models (VLMs) can attain this fundamental human cognitive capability. To compare humans and VLMs, we conducted two studies using Heider–Simmel–style animations that minimize object and scene semantics. In Study 1, humans identified roles near ceiling (∼97%), whereas VLMs were less accurate and less stable across actions (GPT-5 84%; GPT-4o 47%). In Study 2, we introduced a Stroop-inspired visual–linguistic mismatch by pairing animations with occasionally incongruent role statements. Humans' role judgments remained highly vision-consistent (92.7%), but VLMs shifted away from the visual event under conflict (GPT-5: 46.9%; GPT-4o: 29.7%), indicating heavier reliance on linguistic cues. In Study 3, a mismatch-shape baseline left both models perfectly vision-consistent (100.0%), ruling out a generic object-recognition failure explanation for Study 2. Together, these results demonstrate that VLMs do not ground event-role understanding in vision as reliably as humans do, indicating a substantial human–VLM gap in integrating visual versus linguistic information.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*