Multimodal language models show syntax-dependent first-mention biases distinct from the human "like-me" effect

Bei Xiao[1], Xufeng Duan(xufengduan@cuhk.edu.hk)[2], Zhenguang Cai(zhenguangcai@cuhk.edu.hk)[1,2]

[1]Departement of Linguistics and Modern Languages, The Chinese University of Hong Kong,
[2]Brain and Mind Institute, The Chinese University of Hong Kong

Background. Humans preferentially mention same-gender or same-race targets first when describing depicted events, a "like-me" effect that is syntax-invariant across active, conjoined, and passive constructions [1]. Such first-mention preferences are generally linked to increased referential accessibility in humans (e.g., [2]). We use this phenomenon to test whether multimodal large language models (MLLMs) exhibit human-like first-mention biases or qualitatively different, potentially distributional, patterns. We further ask whether such biases can be modulated by demographic persona prompting.

Method. We replicated Brough et al.'s [1] paradigm using three MLLMs (Qwen2.5-Omni-7B, Qwen3-Omni-30B, GPT-4.1; Fig. 1a–b). Trials presented a group image (4 women/4 men; 4 Black/4 white) followed by a two-frame interaction to be described using a fixed sentence frame. Stimuli crossed 10 symmetric verbs with 3 syntactic conditions. Study 1 collected 640 generations per item for the Qwen models and 320 for GPT-4.1. Study 2 added a 2 (persona gender) $\times$ 2 (persona race) demographic persona prefix (Fig. 1c). Responses were analyzed using binomial GLMMs with random intercepts for Item.

Study 1: a syntax-specific first-mention bias in Qwen2.5. Qwen2.5-Omni-7B showed reliable woman-first ($P(\text{woman}) = 51.4\%$, $p = .009$) and white-first ($P(\text{white}) = 51.4\%$, $p = .047$) biases. Crucially, both were *syntactically specific*. The gender bias emerged only in the Conjoined condition (57.5%, $p = .014$; Condition $\chi^2(2) = 14.13$, $p < .001$), whereas the race bias appeared only in the Active condition (51.9%, $p = .014$). This contrasts with the syntax-invariant human like-me effect and suggests frame-specific distributional regularities rather than accessibility-driven production biases. Qwen3-Omni-30B and GPT-4.1 showed no reliable baseline bias.

Study 2: the bias is persona-stable. The woman-first bias in Qwen2.5-Omni-7B persisted under demographic persona prompting ($P(\text{woman}) \approx 52.2\%$, $p < .001$) and was unaffected by persona gender ($\beta = 0.003$, $p = .94$) or race ($\beta = 0.007$, $p = .86$). Persona-match rates remained at chance, and persona $\times$ Condition interactions were null.

Discussion. Qwen2.5-Omni-7B exhibits demographic first-mention biases, but these dissociate from the human like-me effect in two critical respects: they are syntactically specific rather than syntax-invariant, and they remain stable under demographic persona prompting. Together, these signatures suggest that the observed biases are primarily *distributional* in origin, reflecting frame-specific co-occurrence statistics in training data rather than a self-referential accessibility mechanism that can be redirected by surface-level persona cues in prompts.

References.

[1] Brough, J., Harris, L. T., Wu, S. H., Branigan, H. P., & Rabagliati, H. (2024). Cognitive causes of "like me" race and gender biases in human language production. *Nature Human Behaviour*, 8, 1706–1715.

[2] Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, 27(6), 699–717.

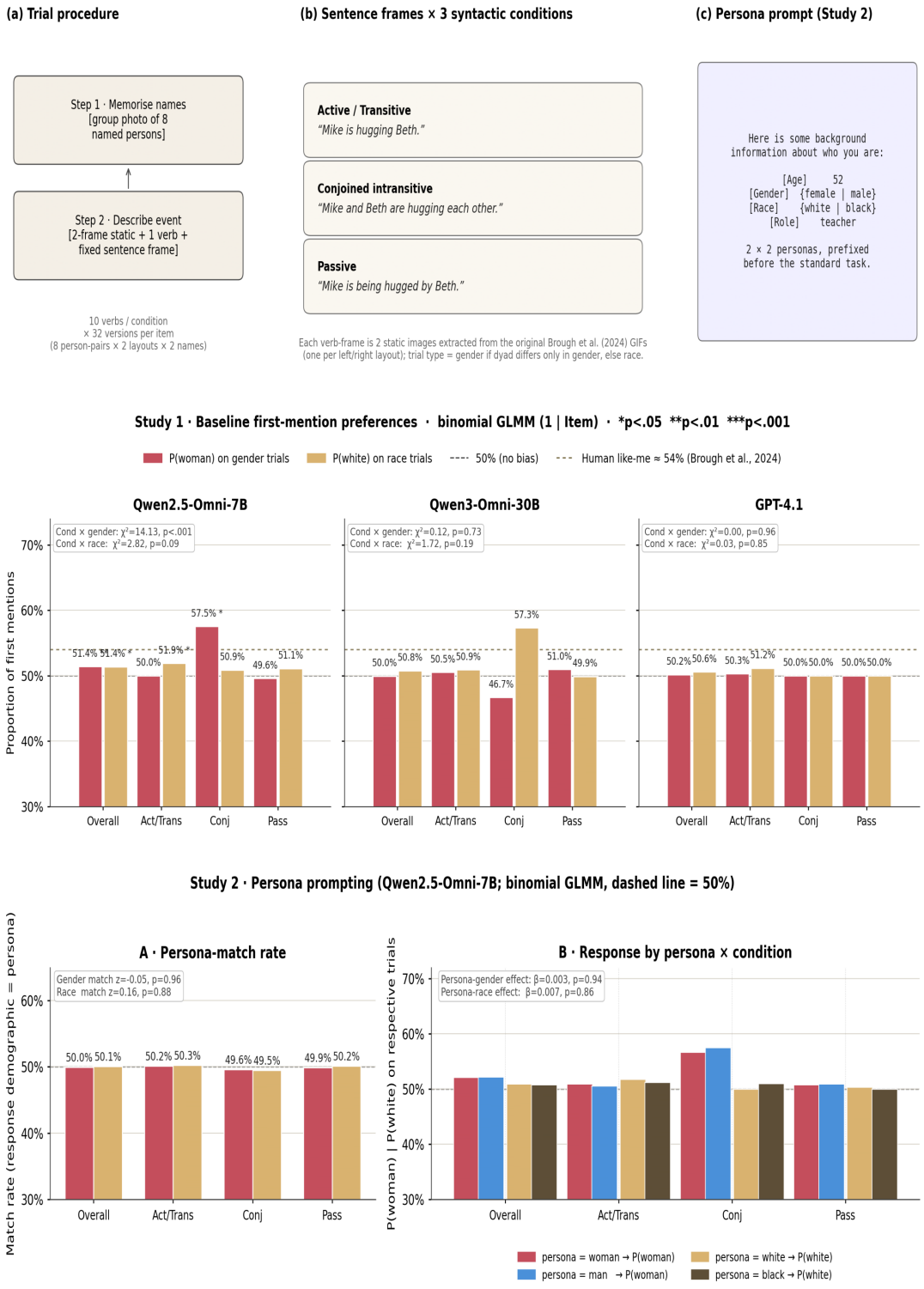

Figure 1. Replicating the Brough et al. (2024) like-me paradigm with three multimodal LLMs. (a) Trial procedure. (b) Three syntactic frames. (c) Persona prompt for Study 2. Middle: Study 1 baseline first-mention preferences across Qwen2.5-Omni-7B, Qwen3-Omni-30B, and GPT-4.1. Bottom: Study 2 persona prompting in Qwen2.5-Omni-7B. Binomial GLMMs with random intercept for Item; * $p < .05$, ** $p < .01$, *** $p < .001$; dashed line = 50% chance.