# Agentive Iconicity in Signs: Visual Context Helps Form-Meaning Mapping in Minds & Machines

**Onur Keleş[1,2], Aslı Özyürek[1,3], Gerardo Ortega[4], Kadir Gökgöz[2], and Esam Ghaleb[1,3]**
[1]Multimodal Language Department, Max Planck Institute for Psycholinguistics
[2]Department of Linguistics, Boğaziçi University
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University
[4]Department of Linguistics and Communication, University of Birmingham

**Background.** This paper investigates whether Vision-Language Models (VLMs) can capture iconicity in sign languages, which refers to the resemblance between visual form and meaning [1, 2], and if such mappings require embodied experience beyond multimodal distributions. This question extends recent psycholinguistic work using neural models to test the sufficiency of statistical learning for human-like linguistic generalizations [3, 4]. However, while previous research examined text-based iconicity in LLMs [e.g., 5], multimodal computational approaches to sign language iconicity remain underexplored. Our benchmarking work (in progress), investigated 13 VLMs, and found that closed-source and large open-source VLMs (≥32B) correlate moderately with human iconicity judgments (ρ≈0.50), but exhibit an agentive iconicity gap: humans perceive action-based iconic signs (dynamic agent-object manipulation as in the sign BABY, Figure 1) as highly iconic, while open-source VLMs favor signs with static properties. As a follow-up, we test whether providing a visual context depicting agent-object relations can scaffold pretrained VLM ratings. For this, we compare open-source VLMs against sign-naïve hearing speakers on transparency and iconicity rating tasks.

**Methodology.** We tested 102 sign-naïve Turkish speakers ($M_{Age}$ = 20 years) and the two largest variants of Qwen 2.5-VL model (72B and 32B; responses averaged) [6] as the top-performing model family in our benchmark. We selected 18 nouns in Sign Language of the Netherlands (NGT) that exhibit action-based iconicity, plus 78 control items (32 arbitrary, 46 other iconic), which were citation-form short sign videos taken from [7]. Participants and models completed a *Transparency* (open-set meaning identification) and an *Iconicity* Rating (1–7) task (Figure 1). For the 18 agentive items, meanings were presented in three conditions: (a) Text (written gloss only); (b) Visual Congruent (VC, image depicting manipulation); (c) Visual Incongruent (VIC, static object image without agent). Control items used two conditions: Text and Visual only. Items were distributed across three lists, with conditions counterbalanced across participants. We recorded accuracy for *Transparency*; ratings & response/inference time (R/ITs) for *Iconicity*.

**Results.** *Transparency*: Mixed-effects models were fit to Accuracy with Type (Iconic, Arbitrary) as fixed effect, and participant and item as random effects. Speakers outperformed VLMs (33.79% vs. 6% accuracy on the 96-choice task). Interestingly, while sign iconicity significantly predicted guess accuracy for humans ($\beta$=4.16, $p$<.001), this effect was absent in VLMs ($\beta$=0.30, $p$=.93). *Iconicity*: Mixed-effects models were fit to Ratings and R/ITs with Gloss Presentation (Text, VC, VIC) as fixed effect, and participant and item as random effects. The results are given in Figure 2 and Tables 1-2. Here, VC significantly increased ratings for humans and VLMs ($p$'s<.05). This effect was larger for the VLMs ($p$<.001). VIC also increased ratings but did not reach significance for speakers ($p$=.34), but was significant for the VLMs ($p$=.041). VLM ratings increased with congruent context, partially closing the agentive iconicity gap, though remaining below human levels (3.75 vs. 5.40). As for RTs, speakers showed slower response times for VC/VIC vs. Text. This result patterns with [8], which suggests effortful analogical reasoning for novice hearing signers. In contrast, VLMs exhibited increased IT only for VIC ($p$<.001).

**Conclusion.** We conclude that the agentive iconicity gap in these VLMs is representational, likely stemming from insufficient dynamic visual grounding in training. While congruent visual context brings model ratings closer to human judgments, processing times vary systematically across conditions. At a descriptive level, this pattern of congruent facilitation and incongruent conflict in VLMs resembles processing effects reported for native signers [8]. These findings highlight possible distinct processing mechanisms in humans versus models, and suggest that training corpora emphasizing dynamic agent-object interactions may be necessary to improve embodied semantics.

Figure 1. Experimental Paradigm (Example NGT Sign: BABY).

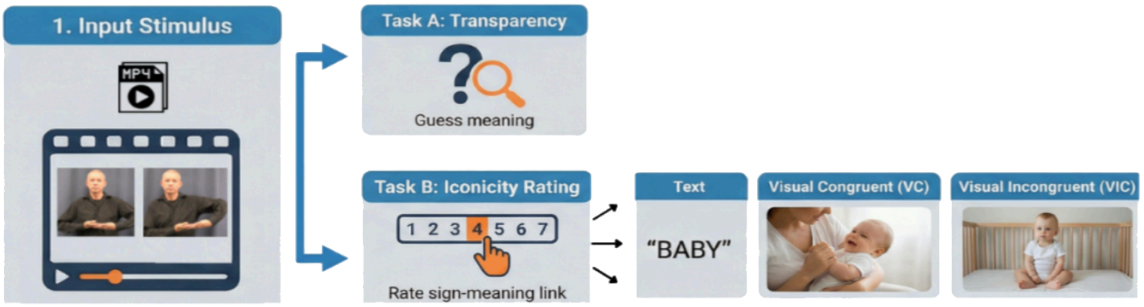


Figure 2. Impact of Visual Context on Iconicity Ratings (A, B) and R/ITs (C, D) in the Experimental Items (*N* = 18).

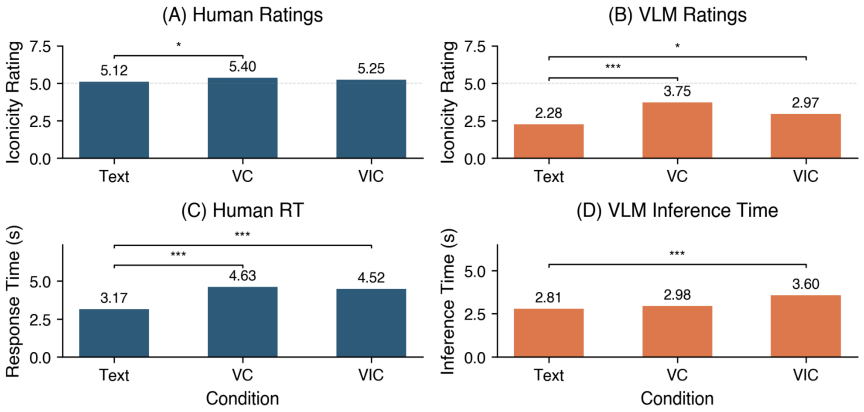


Table 1. Mean Ratings and R/ITs by Stimulus Type and Condition

| | | Human | | VLM | |
|---|---|---|---|---|---|
| Stimulus Type | Condition | Rating | Time (s) | Rating | Time (s) |
| Experimental | Text | 5.12 | 3.17 | 2.28 | 2.81 |
| | VC | 5.40 | 4.63 | 3.75 | 2.98 |
| | VIC | 5.25 | 4.52 | 2.97 | 3.60 |
| $Control_{Iconic}$ | Text | 5.45 | 3.07 | 3.00 | 3.32 |
| | Visual | 5.48 | 5.12 | 3.34 | 3.84 |
| $Control_{Arbitrary}$ | Text | 2.39 | 3.47 | 2.14 | 1.93 |
| | Visual | 2.61 | 6.17 | 2.83 | 2.91 |

*Note.* All experimental items are iconic.

**References.** [1] Campbell et al. (2025). Iconicity as an organizing principle of the lexicon. PNAS. [2] Taub (2001). Language from the body: Iconicity and metaphor in ASL. CUP. [3] Wilcox et al. (2023). Testing the predictions of surprisal theory in 11 languages. TACL. [4] Goldstein et al. (2022). Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. [5] Marklová et al. (2025). Iconicity in large language models. Digit. Scholarsh. Humanit. [6] Bai et al. (2025). Qwen2.5-VL technical report. arXiv. [7] Ortega et al. (2019). Gestures predict iconic form–meaning mappings at first exposure to signs. Cognition. [8] Thompson et al. (2009). Form–meaning links in ASL lexical processing. JEP:LMC.

Table 2. Regression Model Results for Humans and VLMs

| | | | Human | | VLM | |
|---|---|---|---|---|---|---|
| Stimulus Type | Outcome | Comparison | $\beta$ | $p$ | $\beta$ | $p$ |
| Experimental | Rating | VC vs. Text | 0.275 | .041* | 1.472 | <.001*** |
| | | VIC vs. Text | 0.127 | .342 | 0.694 | .041* |
| | Time | VC vs. Text | 1.461 | <.001*** | 0.171 | .333 |
| | | VIC vs. Text | 1.354 | <.001*** | 0.795 | <.001*** |
| $Control_{Iconic}$ | Rating | Visual vs. Text | 0.024 | .708 | 0.337 | .190 |
| | Time | Visual vs. Text | 2.051 | <.001*** | 0.520 | <.001*** |
| $Control_{Arbitrary}$ | Rating | Visual vs. Text | 0.217 | .002** | 0.687 | <.001*** |
| | Time | Visual vs. Text | 2.707 | <.001*** | 0.976 | <.001*** |

*Note.* $*p < .05$. $**p < .01$. $***p < .001$.