The *Visual Iconicity Challenge*: Evaluating Vision-Language Models on Sign Language Form–Meaning Mapping

Onur Keleş[1,2], **Aslı Özyürek**[1,3], **Gerardo Ortega**[4], **Kadir Gökgöz**[2], **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
Correspondence: onur.keles@mpi.nl and esam.ghaleb@mpi.nl

Abstract

Iconicity, the resemblance between linguistic form and meaning, is pervasive in signed languages, offering a natural testbed for visual grounding. For vision–language models (VLMs), the challenge is to recover such essential mappings from dynamic human motion rather than static context. We introduce the *Visual Iconicity Challenge*, a novel video-based benchmark that adapts psycholinguistic measures to evaluate VLMs on three tasks: (i) phonological sign-form prediction (e.g., handshape, location), (ii) transparency (inferring meaning from visual form), and (iii) graded iconicity ratings. We assess 13 state-of-the-art VLMs in zero- and few-shot settings on Sign Language of the Netherlands and compare them to human baselines. On *phonological form prediction*, VLMs recover some handshape and location detail but remain below human performance; on *transparency*, they are far from human baselines; and only top models correlate moderately with human *iconicity ratings*. Interestingly, *models with stronger phonological form prediction correlate better with human iconicity judgment*, indicating shared sensitivity to visually grounded structure. Our findings validate these diagnostic tasks and motivate human-centric signals and embodied learning methods for modelling iconicity and improving visual grounding in multimodal models.

1 Introduction

Language is inherently multimodal: besides speech and text, it includes co-speech gesture and signed languages. Across these modalities, iconicity is the non-arbitrary link between form and meaning. Iconicity can be visual (e.g., speakers use iconic gestures through drawing shapes or trajectories in the air, adding depictive content alongside speech) or even vocal, as in onomatopoeia like "knock knock", showing that form can transparently reflect meaning (Perlman and Lupyan, 2018). Within

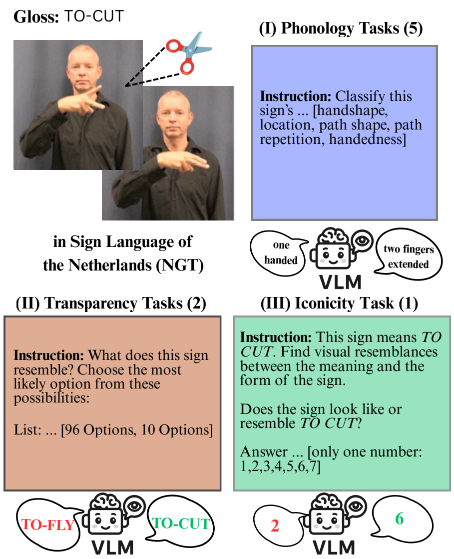


Figure 1: Overview of the *Visual Iconicity Challenge*: evaluation pipeline of the sign TO-CUT in NGT for phonological form prediction, (top right), transparency (bottom left), and iconicity (bottom right) tasks.

signed languages, iconicity is widespread. Estimates suggest that at least a third of lexical signs are iconic (Boyes-Braem, 1986; Campbell et al., 2025) and that between 50–60% of signs' structure can be directly linked to the physical features of their referents (Ortega, 2017; Pietrandrea, 2002). They depict actions or shapes, providing a natural laboratory for studying the symbol grounding problem: how concepts connect to the physical world (Campbell et al., 2025; Taub, 2001).

For vision–language models (VLMs), sensitivity to form–meaning mapping is a core test of grounding in human-centric signals (Bisk et al., 2020). This is especially relevant for applications in sign language understanding and translation, as well as gesture and action recognition. A capable VLM should attend to *dynamic* bodily movements and hand configurations– not just static objects or text– when interpreting a sign or gesture. Following

Yin et al. (2021), signed languages offer a natural testbed for developing and evaluating models that must perceive temporally extended, simultaneous, visuospatial structure, rather than relying on static context alone. However, modern VLMs may exhibit static biases: they over-rely on contextual objects or background features and under-attend to dynamic human actions (Nishida et al., 2025; Yu et al., 2025). *Testing VLMs on iconicity thus offers a proof of concept for machine interpretation of visual-bodily form-meaning mappings, while revealing concrete directions for improvement*.

To address these questions, we build the **Visual Iconicity Challenge**[1]: a sign dataset of Sign Language of the Netherlands (NGT), manually annotated with ground-truth phonological features, iconicity types, and iconicity ratings by non-signers, based on Ortega et al. (2019). The dataset distinguishes between *iconic* signs (with clear visual links to meaning) and *arbitrary* signs (with no visual resemblance).

We evaluate whether models capture different layers of sign–meaning structure, introducing three complementary tasks. Because iconicity links visual form to meaning, it depends on both *phonological form competence* and *analogical reasoning*, requiring models to map structured movement onto conceptual meaning through perceptuomotor analogy (Thompson and Do, 2019). First, we test whether models can recognise the ***phonological form of signs***, including handshape, location, and movement features. Second, we examine ***transparency***: if a model can infer a sign's intended *meaning* from visual form alone (Hoemann, 1975), as non-signers often do (Sehyr and Emmorey, 2019). Finally, we test their sensitivity to ***iconicity*** itself, i.e., whether they can approximate human judgments of graded iconicity. Because previous work shows that VLMs may sometimes rely more on textual or contextual cues than on visual evidence (Nishida et al., 2025), the first two tasks also serve as checks that the models genuinely attend to the visual signal of the sign. Figure 1 shows an overview of the three components. In summary, our contributions are:

- introducing the *Visual Iconicity Challenge*, a benchmark of NGT signs with ground-truth sign phonological annotations and iconicity ratings;
- collecting human baselines for phonology and transparency from a deaf signer and hearing sign-naïve participants;
- conducting the first large-scale zero- and few-shot assessment of state-of-the-art VLMs on sign language iconicity, analysing models' biases for object-based iconicity and failures of form–meaning transparency;
- releasing evaluation code, annotations, and human baselines via a repository for reproducibility and reuse.[2]

2 Related Work

Iconicity in language and computational models. Iconicity has long been analysed as structure mapping between form and meaning in signed languages (e.g., depiction of shape or action) (Taub, 2001; Ortega, 2017; Pietrandrea, 2002). Psycholinguistic and lexical studies report substantial iconicity in signed lexicons and roles for iconicity in acquisition, processing, and L2 learning (Boyes-Braem, 1986; Campbell et al., 2025; Karadöller et al., 2024; Caselli and Pyers, 2020).

Recent NLP research has explored analogous patterns in spoken language models. Large language models can capture sound symbolism effects. For example, GPT-4 can generate iconic pseudowords whose meanings humans and models guess above chance (Marklová et al., 2025). Furthermore, larger language models align with human iconicity ratings, indicating some sensitivity to sound symbolism (Loakman et al., 2024). Metaphor understanding, like iconicity, depends on analogical mapping between domains (Lakoff and Johnson, 1980). Tong et al. (2024) introduce the *Metaphor Understanding Challenge*, which tests whether LLMs can interpret metaphors by distinguishing target-domain paraphrases from literal source-domain alternatives. Their findings show that even advanced models often rely on surface similarity rather than analogy.

In the visual modality, sound symbolism studies report weak or dataset-driven effects in CLIP/Stable Diffusion (Alper and Averbuch-Elor, 2023) and mixed evidence for shape/magnitude symbolism in VLMs (Loakman et al., 2024). Extending this understanding to the visual-manual modality of signed languages requires VLMs, which we address in this work.

[1] The name is inspired by the "vocal iconicity challenge" of Perlman and Lupyan (2018).

[2] https://github.com/kelesonur/Visual_Iconicity_Challenge

General multimodal benchmarks. Large-scale multimodal benchmarks have assessed VLM capabilities on image captioning, VQA, and social signals. For example, Zhang et al. (2025) introduce MMLA, a suite of 61K multimodal utterances with labels for intent, emotion, style, etc., and report that even fine-tuned state-of-the-art models plateau around 60–70% accuracy. Furthermore, Li et al. (2025) introduce a Multimodal Causal Reasoning benchmark testing whether multimodal models can infer causal relations when crucial evidence appears in visual details. Their results show that models with strong textual reasoning still struggle with visual–conceptual integration. These resources and findings evaluate general multimodal capabilities but do not measure whether models map signed visual form onto meaning or assess graded iconicity relative to human judgments.

Gesture and sign understanding with VLMs. VLMs underperform on indexical/iconic gestures, especially with visuals-only input, indicating reliance on textual priors (Nishida et al., 2025). Systems like GIRAF mitigate this by injecting structured descriptors (pose skeletons, segmentations, depth) before LLM reasoning, achieving 75% on deictic and 50% on iconic gestures (Lin et al., 2023). Similarly, Zhang et al. (2024) introduce Pose-enhanced VLM, which integrates a skeletal pose modality into a CLIP-like model: one module uses the 2D pose to guide the visual attention to body joints, and another enriches the pose representation with visual context. This integration yields fine-grained action recognition by encouraging the model to focus on human motion cues. In sign language specifically, recent systems fuse additional signals. For example, SignLLM leverages human poses to generate sign language poses for digital human or avatar generation (Fang et al., 2025).

To our knowledge, no prior work has systematically probed VLMs on iconicity in sign languages. Our study is the first to do so at scale, evaluating how well off-the-shelf VLMs perceive the form–meaning transparency that signers exploit.

3 Dataset: The Visual Iconicity Challenge

We present a dataset built on the Sign Language of the Netherlands (NGT) from Karadöller et al. (2024) and Ortega et al. (2019). It contains 96 sign videos (64 iconic signs and 32 arbitrary signs), each with an English gloss (meaning) and human iconicity ratings ranging from 1 to 7 (see the full dataset in Appendix A). This categorisation is based on human iconicity ratings. Signs with low ratings ($M = 2.10$, $SD = 0.50$) were classed as arbitrary, and signs with high ratings ($M = 5.13$, $SD = 1.02$) were classed as iconic.

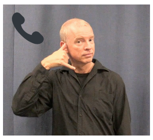
Item	Ortega et al.	Ours
Phonology form features (based on Klomp and Pfau 2020)		
Handshape	✗	✓
Location	✗	✓
Path shape	✗	✓
Path repetition	✗	✓
Handedness	✗	✓
Transparency labels (N=96)	✓	✓
Iconicity		
Ratings (1–7)	✓	✓
Labels (Iconic vs. arbitrary)	✓	✓
Types (e.g., Object or action based)	✗	✓
Human baselines		
Phon. form prediction	✗	✓
Transparency	✗	✓
Iconicity ratings	✓	✓

Table 1: Comparison of the original NGT sign videos dataset (Ortega et al., 2019; Karadöller et al., 2024) and our extensions for the visual iconicity challenge.

Our evaluation operationalises iconicity through three complementary tasks targeting different form-meaning mapping levels. *(i)* ***Phonological form prediction*** examines whether models perceive the *articulatory structure* of a sign (handshape, location, movement). *(ii)* ***Transparency*** asks models to recover a sign's *lexical meaning* from visual form alone, indexing analogical mapping from form to concept while minimising reliance on linguistic priors. *(iii)* ***Graded iconicity rating*** evaluates whether models are sensitive to the *degree* of resemblance between form and meaning by correlating model ratings with human judgments.

Hypothesis. Models that better predict phonological features (e.g., handshape, location, path) should better capture iconicity, since both require grounding in structured bodily properties.

Motivated by the view that iconicity relies on both *phonological form competence* and *analogical reasoning* (Thompson and Do, 2019), we extend the original resource with: (i) detailed phonological annotations for each sign, (ii) iconicity-type labels, and (iii) human baselines for phonology and transparency. These additions support evaluation of VLMs from sub-lexical perception to graded iconicity. See Table 1 for a comparison between



Gloss: TELEPHONE
Type: Iconic
Handshape: Two fingers extended
Location: Head
Path Shape: Hold
Path Repetition: No
Handedness: One-handed

Gloss: SUGAR
Type: Arbitrary
Handshape: All fingers extended
Location: Head
Path Shape: Straight
Path Repetition: No
Handedness: One-handed

Figure 2: Examples of an iconic vs. an arbitrary sign, with their annotated phonological form features. The sign TELEPHONE is iconic as its form resembles a telephone's shape, whereas SUGAR is arbitrary with no clear visual link to its meaning.

the original dataset and our extensions.

3.1 Sign Phonological Form Features

We annotate phonological form features of each sign using a standard NGT phonology framework (Klomp and Pfau, 2020). These are discrete, visual descriptors of articulation (elaboration on the annotation criteria is in Appendix B). In summary, we use five phonological parameters:

- *Handshape*: 7 categories (e.g., fist, flat hand, one finger extended, etc.). Figure 2 illustrates a few categories of the annotated handshapes.
- *Location*: 5 categories of where on the signer's body or space the sign is articulated, i.e., face/head, torso, arm/shoulder, the opposite hand, or neutral space.
- *Path Shape*: 4 categories of movement trajectory shape, i.e., no movement/hold, straight line, arched curve, and circular motion.
- *Path Repetition*: 2 categories (whether the movement is repeated or only single).
- *Handedness*: 3 categories (one-handed sign, two-handed symmetrical, or two-handed asymmetrical).

A deaf signer and a hearing non-signing researcher performed the annotations. To assess reliability, inter-annotator agreement ranged from 77.9% ($\kappa = 0.73$) for handshape to 98.9% ($\kappa = 0.98$) across parameters. All disagreements were discussed and resolved. These reliable annotations serve as a gold-standard reference or "ceiling" for assessing how well models can recognise sign form.

Task. Given a sign video, models perform multi-class prediction for each parameter: handshape, location, path shape, path repetition, and handedness. For this task, we report the accuracy of the model predictions per parameter and the overall average accuracy. This task checks whether the model is capable of extracting form information from the video: where and how signs were articulated.

Human baseline. We gather baseline results from human participants for all the phonological parameters. The 96 stimulus signs were divided into four lists of 24 signs each. Four sign-naive undergraduate participants (i.e., without prior knowledge of sign language) were recruited and randomly assigned to lists in a counterbalanced design. Each participant judged all 24 signs in their list on both the phonological feature tasks and the transparency (open-set meaning identification) task. This provided a sign-naive baseline for both tasks. The human baseline mean phonological accuracy was 0.79 (highest for handedness, lowest for handshape).

3.2 Sign Transparency

Task. Transparency tests whether meaning can be inferred from visual form alone. We use the gloss list (i.e., meaning) of each sign, which is provided in the original dataset. We evaluate two settings: **Transparency$_1$** (open-set identification among all 96 glosses) and **Transparency$_2$** (multiple choice with 10 candidates: the target gloss plus 9 distractors). We use accuracy as the primary metric (proportion of signs correctly identified).

Human baseline. The deaf signer (who annotated the phonological form features and is not a native NGT signer) identified $57/96$ glosses in the open-set setting; the sign-naïve group identified $40/96$ (same participants and lists as in the phonology baseline). These provide upper- and lower-bound human references for Transparency$_1$.

3.3 Sign Iconicity Ratings

Task. This task probes whether models capture the degree to which a sign's form resembles its meaning. We use the original crowdsourced iconicity ratings as **human baselines** (see Appendix A). Each sign has an average iconicity rating on a 1–7 scale (with 7 = "looks exactly like its meaning", 1 = "not iconic at all"). Models are prompted to produce the iconicity rating for each sign (i.e., the degree of the sign's resemblance to its meaning). We compute Spearman's rank correlation ρ between the model's ratings and the average human iconicity ratings for the signs.

Iconicity types. Iconicity type influences how signers perceive, process, and acquire signs (Ortega et al., 2014, 2017). We annotate each sign for its iconicity type to probe how well models align with these distinctions. These include *object-based* signs ($N = 16$), where the handshape visually resembles a property of the referent (e.g., the wings of a butterfly), and *action-based* signs ($N = 30$), where the hand depicts an action performed on or by the referent (e.g., brushing teeth). The remaining 16 signs belonged to a third category named "combined", where both strategies were employed for the same sign. The descriptions of these types can be found in Appendix D. These label types enable us to analyse how different iconic strategies affect model predictions and human perception.

4 Models and Inference

Models. We evaluate a representative and diverse set of 10 open-source VLMs and 3 proprietary models. Open models include: Qwen2.5-VL (72B/32B/7B) (Bai et al., 2025), VideoLLaMA2 (72B/7B) (Cheng et al., 2024), LLaVA-Video-Qwen2 (72B/7B) (Li et al., 2024), LLaVA-Onevision-Qwen2 (72B/7B) (Liu et al., 2024), and Gemma-3 (27B) (Team et al., 2025). We evaluate three *proprietary (closed-source) large multimodal models*: GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025), and Gemini 2.5 Pro (Team, 2025).

For inference, smaller models (7B) were run efficiently on a single NVIDIA A100 GPU, while larger models (27B >) were distributed across up to four A100 GPUs. Closed-source models were queried via API calls.

Zero-shot setup. All models are evaluated in a zero-shot manner first. We craft a prompt/instruction template for each task that is standardised across models. The prompts explicitly describe the task and the expected answer format, and we ensure the output format is constrained (e.g. just a single number for ratings, or a one-word answer for glosses). For example, for the iconicity rating task, the prompt to the model is:

> *This sign means: <MEANING>. Some signs are iconic, and some are arbitrary. Find visual resemblances between the meaning and the form of the sign. How much does the sign look like "<MEANING>"? Answer with only one number: 1,2,3,4,5,6,7 (1=not at all, 7=exactly).*

The prompts for all tasks and features can be found in Appendix C. We do not use chain-of-thought prompting or specialised prompting tools, as initial trials with those did not show clear benefits. Our aim is to first establish baseline performance; more sophisticated prompting or fine-tuning can be explored in future work.

Few-shot setup. To examine whether a few examples can improve models' performance, we conduct 4-shot experiments with four selected models. We choose the *open model families* that perform best in zero-shot: Qwen2.5-VL-72B, Qwen2.5-VL-32B, Qwen2.5-VL-7B and Gemma-3-27B. We omit closed models in these settings since few-shot probing suggests that GPT-5 and Gemini are already comparatively well-calibrated in zero-shot settings, showing only marginal benefit.

We provide 4 example QA pairs (two iconic signs and two arbitrary signs) before the test query, using the same instruction format and showing the correct outputs for those examples. This few-shot setup gives the model a better understanding of the task with visual examples of iconic & arbitrary signs, which effectively offer a window into the learnability of the tasks and how to improve the model's performance. For Qwen2.5-VL, exemplars and test items were provided as short video clips paired with gold answers. For Gemma-3, which does not natively support video, we extracted up to eight evenly spaced frames per clip and supplied both exemplars and test inputs as frame sequences.

5 Results & Discussions

5.1 Phonological Form Prediction

Zero-shot. We observe clear differences across VLMs in phonological form features. Figure 3 shows the accuracy of each model on the five phonological sub-tasks. For example, the location of the sign (where on the body the sign is articulated) and handedness (one vs. two hands) were the easiest. For location, $6/12$ models reach ≥ 0.70 (best 0.86), while for handedness, $9/12$ exceed 0.70. In contrast, the handshape and path shape are the hardest, whereas only GPT-5 and Gemini 2.5 Pro exceed 0.50. Most models exceed the random baseline but remain well below the human mean of 0.794. For example, Gemini 2.5 Pro leads ($M = 0.706$), closely followed by GPT-5 (0.698). The best open-source models are Qwen2.5-VL-72B ($M = 0.598$) and Gemma-3-27B (0.535).

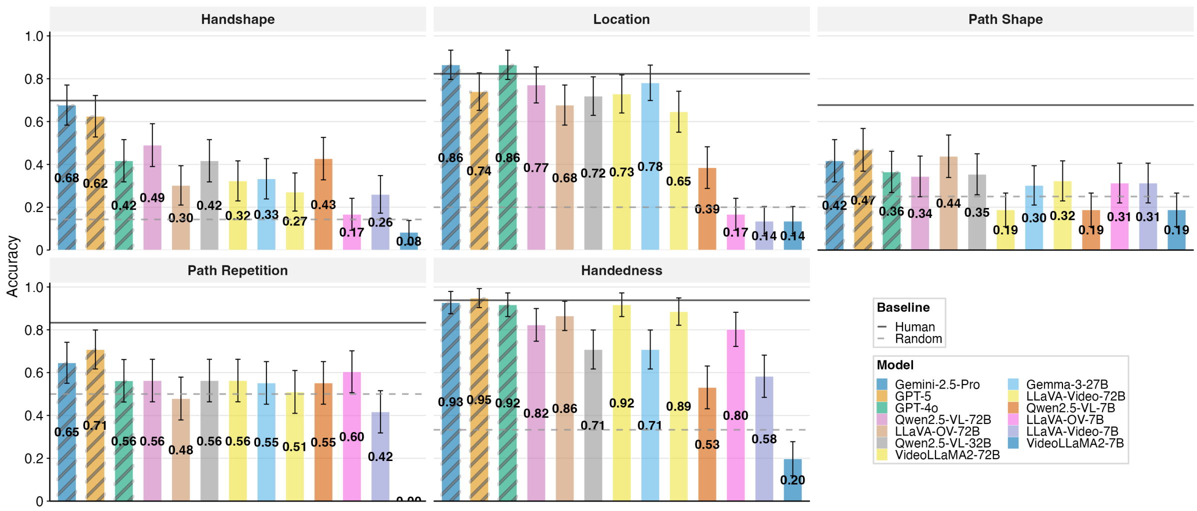


Figure 3: Zero-shot accuracy per form feature. Solid black lines indicate the human baseline, and dashed grey lines refer to random. Bars show VLMs. Across models, location and handedness are comparatively easy; handshape and path shape are hardest; path repetition is intermediate. Numbers on bars are mean accuracies.

Overall, while we see large models encode phonologically relevant structure, still, the absolute gap remains large: human accuracy is far higher for most features. A radar plot in Figure 8 in Appendix F visualises the performances of humans and representative models.

Interestingly, the performance patterns of models for each form feature in Figure 3 mirror well-established acquisitional asymmetries in sign language. Like deaf children and adults (Keleş et al., 2022; Sandler and Lillo-Martin, 2006; Morgan et al., 2007; Marentette and Mayberry, 1999), models find *location* easier than *handshape*.

Few-shot Few-shot prompting yields only modest, model-dependent accuracy gains for the selected open-source VLMs. (Table 2). Qwen2.5-VL-32B and Gemma-3-27B improve slightly, while Qwen2.5-VL-72B changes little, indicating that few-shot prompting mainly benefits models with lower zero-shot performance (possibly by helping them interpret the task format).[3]

5.2 Sign Transparency

Zero-shot. Open-set gloss identification is highly challenging (Table 3). Even the strongest closed-source models perform poorly: best model (i.e., Gemini 2.5 Pro) identified only 17 of the 96 glosses (≈17.7%). This is well below human baselines (57/96 for the deaf expert, 40/96 for hearing non-experts). Restricting the task to a 10-multiple-choice format improves scores (e.g., 42/96 for GPT-5 and 41/96 for Gemini 2.5 Pro) Open-source VLMs perform even worse, with the best achieving only 5/96 correct identifications (Qwen2.5-VL-32B). The consistent advantage of the closed models over all open-source systems indicates that high-capacity proprietary VLMs are better at leveraging visual and linguistic cues.

Across models, correct predictions often cluster on visually obvious signs such as TELEPHONE, TO-WRING, TO-CUT, and PISTOL with 9 out of 13 models correctly guessing them (see Figure 7 in Appendix F). Interestingly, some arbitrary but cross-linguistically shared signs (e.g., TO-ORDER, PERSON, TO-ARGUE, and TO-DIE) were successfully guessed by a handful of VLMs.[4] Identification of such arbitrary signs suggests that their forms

Model	Mean Accuracy	
	0-shot	4-shot
Qwen2.5-VL-72B	0.598	0.600
Qwen2.5-VL-32B	0.552	0.620
Gemma-3-27B	0.535	0.572
Qwen2.5-VL-7B	0.417	0.550

Table 2: Comparison of zero-shot and 4-shot performance on the phonological form prediction task.

[3]A closer look at task-level breakdowns reveals uneven effects: the largest gains occur for path shape and handedness, while location remains unstable, handshape shows minimal improvement, and path repetition is largely unaffected. This suggests that few-shot prompting primarily helps models disambiguate structural features like path shape and handedness.

[4]Most of these arbitrary signs were guessed correctly by our human participants too.

Model	96 options	10 options
Human baselines		
Deaf signer	0.594	–
Hearing non-signer	0.417	–
Models		
GPT-5	0.156	0.438
Gemini-2.5-Pro	0.177	0.427
GPT-4o	0.073	0.354
Qwen2.5-VL-32B	0.052	0.177
LLaVA-OV-Qwen2-72B	0.031	0.156
VideoLLaMA2-72B	0.031	0.156
Qwen2.5-VL-72B	0.021	0.167
Gemma3-27B	0.021	0.125
LLaVA-Video-72B-Qwen2	0.021	0.125
Qwen2.5-VL-7B	0.021	0.115
LLaVA-OV-Qwen2-7B	0.021	0.073
LLaVA-Video-7B-Qwen2	0.010	0.146
VideoLLaMA2-7B	0.010	0.125
Chance (random)	0.010	0.100

Table 3: Transparency task accuracy in 96-option and 10-option conditions.

still contain strong visual cues, possibly through conventional metaphorical mappings shared across sign languages (Meir and Cohen, 2018).

Few-shot. Four-shot prompting yields no meaningful gains. In the 96-way setting, Qwen2.5-VL (72B & 32B) and Gemma-3-27B each identify 2 of 92 items, and Qwen2.5-VL-7B identifies 1 of 92. In the 10-choice format, the same models score 15–16 of 92, matching their zero-shot levels. These results suggest the bottleneck is not just understanding the task format, but a fundamental limitation in the models' visual–semantic grounding.

5.3 Sign Iconicity

Zero-shot. For iconicity ratings (Table 4), we observe that some models show positive moderate correlation with human iconicity judgment ($\rho \geq 0.40$, $p < .001$). For instance, GPT-5 reaches the highest correlation with human ratings ($\rho \approx 0.61$). Among open VLMs, Qwen2.5-VL-72B achieves the strongest correlation with human judgments ($\rho = 0.501$), while Gemma-3-27B shows the best categorical separation (Cohen's $d = 1.216$) of iconics vs arbitrary signs.

Despite these promising results, most models compress the scale around the midpoint and systematically over-rate arbitrary signs, thereby reducing contrast between iconic and arbitrary categories compared to human ratings.

Few-shot. Few-shot prompting yields clear but uneven effects (Table 5). The larger Qwen2.5 models benefit most, with Qwen2.5-VL-32B showing a sharp gain in correlation (ρ: 0.344 $\rightarrow$ 0.510) and Qwen2.5-VL-72B improving moderately on both correlation and separation. For Gemma-3-27B, few-shot examples produce only small or inconsistent changes, while Qwen2.5-VL-7B declines substantially. Overall, these mixed results indicate that few-shot cues are most helpful for large open models that underperform relative to their capacity, but provide little advantage for the strongest models already close to human-like calibration.

Model	ρ	d
GPT-5	0.607^{***}	1.382
Gemini-2.5-Pro	0.577^{***}	1.435
GPT-4o	0.248^{*}	0.432
Qwen2.5-VL-72B	0.501^{***}	0.800
Gemma-3-27B	0.452^{***}	1.216
Qwen2.5-VL-7B	0.456^{***}	0.693
VideoLLaMA2-72B	0.400^{***}	0.790
Qwen2.5-VL-32B	0.344^{***}	0.519
LLaVA-OV-Qwen2-72B	0.223^{*}	0.278
LLaVA-OV-Qwen2-7B	0.119^{ns}	0.204
LLaVA-Video-7B-Qwen2	0.109^{ns}	0.195
LLaVA-Video-72B-Qwen2	0.102^{ns}	0.133
VideoLLaMA2-7B	0.101^{ns}	0.136

Table 4: Graded iconicity rating results in terms of Spearman ρ and Cohen's d. Significance codes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, ns $p \geq .05$.

Model	ρ		d	
	0-shot	4-shot	0-shot	4-shot
Gemma-3-27B	0.452	0.484	1.216	1.021
Qwen2.5-VL-72B	0.501	0.521	0.800	1.021
Qwen2.5-VL-32B	0.344	0.510	0.693	0.941
Qwen2.5-VL-7B	0.456	0.321	0.519	0.418

Table 5: Comparison of zero-shot and 4-shot performance on the graded iconicity rating.

Type of iconicity. We perform a post-hoc analysis to examine whether models differ in how they handle different kinds of iconic signs. Iconicity is commonly classified by whether a sign depicts an object's shape or a human action (Ortega et al., 2019). Human raters show a robust preference for *action*-based signs over *object*-based ones, consistent with findings that action signs are acquired earlier and processed more easily due to their transparent "hand-represents-hand" mappings (Ortega et al., 2017; Sümer and Özyürek, 2025). As illustrated in Figure 4, both humans and large models clearly distinguished arbitrary from iconic signs, indicating that models can broadly recognise iconic

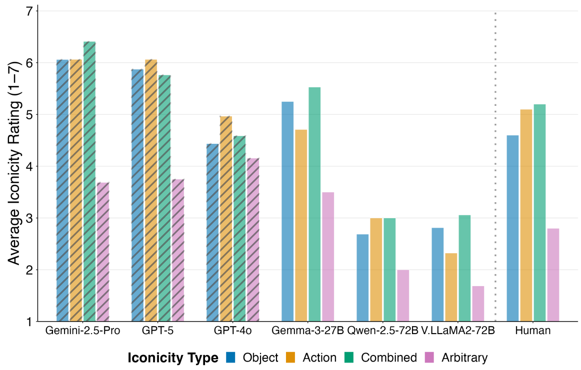


Figure 4: Average iconicity ratings by iconicity type (higher = more iconic).

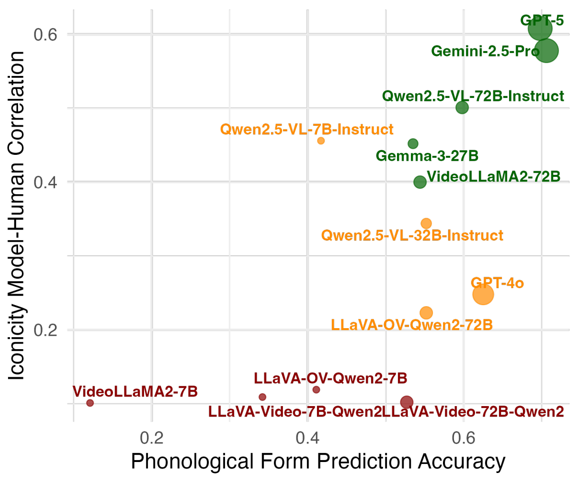


Figure 5: Overall model landscape by zero-shot phonological form prediction accuracy and iconicity scores. Top-right are best; dot size encodes model size.

structure. However, within iconic signs, differences emerged. Humans show a consistent *action* bias, whereas most open-source models displayed the reverse pattern, favouring *object*-based signs that depict visual features rather than actions. Closed-source models such as Gemini and GPT-5 showed little to no preference between the two types. This inversion suggests that while models might identify iconicity, they rely more on static visual resemblance than on dynamic mappings.

We argue that humans ground iconicity in embodied experience, mapping hand actions onto conceptual structure, while models—lacking such bodily grounding—depend on surface correlations between form and referent. As a result, they might tend to overvalue object-based resemblance and underestimate dynamic agency, highlighting the gap between visual pattern recognition and embodied understanding of sign meaning.

6 Interaction of Iconicity and Phonology

We hypothesise that models with stronger phonological form predictions are better at rating graded iconicity, as both require grounding in structured bodily properties. Indeed, as shown in Figure 5, models with higher phonological form accuracies, such as Gemma-3, GPT-5, and Gemini 2.5, also achieve closer alignment with human iconicity ratings. Conversely, models with weaker phonological representations (e.g., smaller Qwen variants) show both lower accuracy on phonological features and less consistent treatment of iconicity. This pattern suggests that sensitivity to phonological form and to form–meaning mappings are not independent, but may partly co-develop (Emmorey, 2014).

From a cognitive perspective, this link mirrors the human case: iconicity is tied to phonological awareness in sign language because mapping form features onto conceptual structure requires attending to both form and meaning simultaneously. Yet, unlike humans, models systematically overrate object-based iconicity (see Figure 4), showing that their phonological sensitivity alone does not reproduce embodied biases. In other words, while better phonology helps models approximate human iconicity ratings in general, it does not prevent them from favouring visually simpler object correspondences over dynamic action mappings.

7 Conclusion

We introduced the *Visual Iconicity Challenge*, a diagnostic evaluation that probes phonological form prediction, meaning prediction from form, and iconicity ratings in the Sign Language of the Netherlands. Our evaluations suggest that, compared to human baselines, larger vision large language models partly mirror human phonological difficulty patterns (e.g., handshape is more difficult than location), can distinguish iconic from arbitrary signs, and correlate moderately with graded human iconcity ratings. Yet they fail to infer lexical meaning and show a different iconicity-type bias. Bridging this gap requires richer gesture/sign pretraining and dynamic pose encoding. For future work, we suggest integrating structured pose information via tools such as MediaPipe (Lugaresi et al., 2019) or VideoPrism (Cheng et al., 2024), and fine-tuning with auto-generated phonological descriptors (e.g., "fist moves upward near head") to provide geometric grounding that raw video lacks.

8 Limitations

Our evaluation has several constraints. The dataset is small (96 isolated NGT signs) with citation-style clips that may not generalise to other sign languages or continuous discourse. Phonological annotations cover five major parameters but omit finer-grained features (orientation, aperture changes, non-manual markers), and the mixed lexical classes (verbs vs. nouns) may affect transparency and iconicity patterns.

Furthermore, we evaluated models only in zero-shot and few-shot settings without fine-tuning, which establishes a diagnostic baseline but likely underestimates potential performance with sign-specific training. Future work may explore fine-tuning, examine model-specific factors (parameter count, memory footprint, mixture-of-experts activation patterns during inference to examine the processing of signs with different levels of iconicity), test robustness under visual perturbations (noise, motion blur), conduct stratified analyses by iconicity type (action-based vs. object-based) and sign difficulty levels, and perform qualitative error analysis to identify whether failures stem from visual perception, analogical reasoning, or lexical-semantic grounding deficits.

References

Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. *Advances in Neural Information Processing Systems*, 36:78347–78359.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, and 1 others. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Penny Boyes-Braem. 1986. Two aspects of psycholinguistic research: iconicity and temporal structure. In *Proceedings of the Second European Congress on Sign Language Research; Signs of Life*, Amsterdam. University of Amsterdam.

Erin E Campbell, Zed Sevcikova Sehyr, Elana Pontecorvo, Ariel Cohen-Goldberg, Karen Emmorey, and Naomi Caselli. 2025. Iconicity as an organizing principle of the lexicon. *Proceedings of the National Academy of Sciences*, 122(16):e2401041122.

Naomi K Caselli and Jennie E Pyers. 2020. Degree and not type of iconicity affects sign language vocabulary acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(1):127.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Karen Emmorey. 2014. Iconicity as structure mapping. *Philosophical transactions of the Royal Society B: Biological sciences*, 369(1651):20130301.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. 2025. Signllm: Sign language production large language models. In *The International Conference on Computer Vision*. IEEE/CVF.

Harry W Hoemann. 1975. The transparency of meaning of sign language gestures. *Sign language studies*, 7(1):151–161.

Dilay Z Karadöller, David Peeters, Francie Manhardt, Aslı Özyürek, and Gerardo Ortega. 2024. Iconicity and gesture jointly facilitate learning of second language signs at first exposure in hearing nonsigners. *Language Learning*, 74(4):781–813.

Onur Keleş, Furkan Atmaca, and Kadir Gökgöz. 2022. Effects of age of acquisition and category size on signed verbal fluency. *Language Acquisition*, 29(4):361–383.

Ulrika Klomp and Roland Pfau, editors. 2020. *A Grammar of Sign Language of the Netherlands (NGT)*, 1 edition. SIGN-HUB Sign Language Grammar Series. SIGN-HUB. Accessed 31-10-2021.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. 2025. Multimodal causal reasoning benchmark: Challenging multimodal large language models to discern causal links across modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5509–5533, Vienna, Austria. Association for Computational Linguistics.

Li-Heng Lin, Yuchen Cui, Yilun Hao, Fei Xia, and Dorsa Sadigh. 2023. Gesture-informed robot assistance via foundation models. In *7th Annual Conference on Robot Learning*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge. *arXiv preprint arXiv:2404.04060*.

Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. *arXiv preprint arXiv:2409.14917*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and 1 others. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Paula F. Marentette and Rachel I. Mayberry. 1999. Principles for an emerging phonological system: A case study of early asl acquisition. In Charlene Chamberlain, Jill P. Morford, and Rachel I. Mayberry, editors, *Language Acquisition by Eye*, pages 71–90. Psychology Press, Mahwah, NJ.

Anna Marklová, Jiří Milička, Leonid Ryvkin, L'udmila Lacková Bennet, and Libuše Kormaníková. 2025. Iconicity in large language models. *Digital Scholarship in the Humanities*, page fqaf095.

Irit Meir and Ariel Cohen. 2018. Metaphor in sign languages. *Frontiers in psychology*, 9:1025.

Gary Morgan, Sarah Barrett-Jones, and Helen Stoneham. 2007. The first signs of language: Phonological development in british sign language. *Applied Psycholinguistics*, 28(1):3–22.

Noriki Nishida, Koji Inoue, Hideki Nakayama, Mayumi Bono, and Katsuya Takanashi. 2025. Do multimodal large language models truly see what we point at? investigating indexical, iconic, and symbolic gesture comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 514–524, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2025. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`.

Gerardo Ortega. 2017. Iconicity and sign lexical acquisition: A review. *Frontiers in Psychology*, 8:1280.

Gerardo Ortega, Annika Schiefner, and Aslı Özyürek. 2019. Hearing non-signers use their gestures to predict iconic form-meaning mappings at first exposure to signs. *Cognition*, 191:103996.

Gerardo Ortega, Beyza Sumer, and Asl Ozyurek. 2014. Type of iconicity matters: Bias for action-based signs in sign language acquisition. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Gerardo Ortega, Beyza Sümer, and Aslı Özyürek. 2017. Type of iconicity matters in the vocabulary development of signing children. *Developmental psychology*, 53(1):89.

Marcus Perlman and Gary Lupyan. 2018. People can create iconic vocalizations to communicate various meanings to naïve listeners. *Scientific reports*, 8(1):2634.

Paola Pietrandrea. 2002. Iconicity and arbitrariness in italian sign language. *Sign Language Studies*, 2(3):296–321.

Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Zed Sevcikova Sehyr and Karen Emmorey. 2019. The perceived mapping between form and meaning in american sign language depends on linguistic knowledge and task: Evidence from iconicity and transparency judgments. *Language and Cognition*, 11(2):208–234.

Beyza Sümer and Aslı Özyürek. 2025. Action bias in describing object locations by signing children. *Sign Language & Linguistics*.

Sarah F Taub. 2001. *Language from the body: Iconicity and metaphor in American Sign Language*. Cambridge University Press.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Arthur Lewis Thompson and Youngah Do. 2019. Defining iconicity: An articulation-based methodology for explaining the phonological structure of ideophones. *Glossa: a journal of general linguistics*.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *arXiv preprint arXiv:2403.11810*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360.

Yating Yu, Congqi Cao, Yifan Zhang, and Yanning Zhang. 2025. Learning to generalize without bias for open-vocabulary action recognition. *arXiv preprint arXiv:2502.20158*.

Hanlei Zhang, Zhuohang Li, Yeshuang Zhu, Hua Xu, Peiwu Wang, Haige Zhu, Jie Zhou, and Jinchao Zhang. 2025. Can large language models help multimodal language analysis? mmla: A comprehensive benchmark. *arXiv preprint arXiv:2504.16427*.

Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. 2024. Pevl: Pose-enhanced vision-language model for fine-grained human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18867.

A Stimuli

Iconic Signs (n = 64)

Sign	Iconicity rating	Sign	Iconicity rating
TO-BREAK	6.79	TO-INJECT	5.0
TO-CRY	6.74	TO-STAPLE	4.89
WINDSCREEN WIPER	6.63	CALCULATOR	4.8
TO-CUT	6.61	PENGUIN	4.78
ELEPHANT	6.53	RATTLE	4.75
BICYCLE	6.44	CAR	4.7
BIRD	6.42	CURTAINS	4.7
BABY	6.39	BRIDGE	4.6
KEY	6.26	DEER	4.6
TELEPHONE	6.22	HELICOPTER	4.6
TO-WRING	6.12	MONKEY	4.5
TO-SWIM	6.11	SPIDER	4.5
TO-SLAP	6.11	ZIMMER	4.42
TO-PUMP	6.11	TO-ERASE	4.2
PIANO	6.05	TO-SMS	4.2
TO-KNOCK	6.05	PLANE	4.11
BUTTERFLY	5.94	BALL	4.05
TO-CRASH	5.79	DOOR	4.0
SNAKE	5.74	WHEELCHAIR	4.0
TO-FLY	5.74	CHICKEN	3.83
TABLE	5.7	BLANKET	3.8
PISTOL	5.61	CELL	3.8
EAGLE	5.53	DRILL	3.8
TO-CUT	5.51	TO-PLAY-CARDS	3.8
LAPTOP	5.44	BOTTLE	3.68
UMBRELLA	5.42	CAT	3.61
TO-JUGGLE	5.42	SUITCASE	3.6
CAMEL	5.4	LOBSTER	3.5
SPOON	5.3	TO-PUT-CLOTHES-ON	3.32
TO-STEAL	5.11	BED	3.21
TOWEL	5.1	RESTAURANT	3.21
BOX	5.05	RABBIT	3.16

Arbitrary Signs (n = 32)

Sign	Iconicity rating	Sign	Iconicity rating
AMBULANCE	3.11	MUMMY	2.06
TO-ARGUE	2.94	KIWI	2.06
BEAR	2.89	TO-GOSSIP	2.0
TO-SHOUT	2.79	TO-GO-OUT	1.89
INTERPRETER	2.74	TOILET	1.79
DOG	2.69	ELECTRICITY	1.79
TO-DIE	2.58	PRAM	1.74
PERSON	2.53	DOCTOR	1.74
TO-ORDER	2.44	BUS	1.74
TREE	2.26	HORSE	1.67
TO-LAUGH	2.26	WATER	1.63
SOFA	2.26	BUILDING	1.63
ROOM	2.22	PUPPET	1.53
SHEEP	2.16	FRUIT	1.47
FIRE	2.11	SUGAR	1.37
TO-COOK	2.11	LIGHTBULB	1.22

B Criteria for Phonological Feature Annotation

The following guidelines summarize the decision criteria we applied when annotating the five phonological features of each NGT sign. Our annotations were mainly based on the descriptions drawn from the phonology chapters of *A Grammar of Sign Language of the Netherlands (NGT)* (Klomp and Pfau, 2020). We follow the general phonological descriptions in the NGT grammar but use our own simplified label set for annotation and model evaluation.

Handshape: Handshapes were coded using seven discrete labels:

- All fingers closed to a fist
- All fingers extended
- All fingers curved or clawed
- One (selected) finger extended
- One (selected) finger curved or clawed
- Two or more (selected) fingers extended
- Two or more (selected) fingers curved or clawed

These categories are drawn from the NGT phonological inventory, but we simplify them by collapsing sub-types and by omitting features such as orientation or aperture change. However, our labels treat each sign as having a single static handshape; they therefore do not fully capture signs in which the handshape itself changes over time. For example, signs where a fist closes or opens during the articulation. For such dynamically changing signs we accepted *multiple answers as correct*, so that both start and end configurations are treated as valid.

Location: Each sign was assigned to one of five major location categories:

- Hands touching head/face
- Hands touching torso
- Hands touching arm
- Hands touching weak/passive hand
- Hands in front of the body or face (neutral space)

If a sign involved contact with multiple regions, the primary lexical target location was coded.

Path Shape: Primary path movement was classified using four labels:

- Hold: no path or directional movement
- Straight: linear horizontal, vertical, or diagonal trajectory
- Arched: curved or semicircular trajectory
- Circular: full or near-full circular path

Path Repetition: Repetition of the movement was coded as:

- Single: one primary stroke
- Repeated: movement is repeated

Handedness: Handedness was coded according to the two-handed typology:

- One-handed
- Two-handed symmetrical: both hands share the same handshape and movement
- Two-handed asymmetrical: hands differ in handshape and/or movement

C Used Prompts

Phonological Form Prediction Instructions:

> *Handshape?* H1=all fingers closed to a fist, H2=all fingers extended, H3=all fingers curved or clawed, H4=one (selected) finger extended, H5=one (selected) finger curved or clawed, H6= two or more (selected) fingers extended, H7=two or more(selected) fingers curved or clawed)
>
> *Location?* Major sign location? Answer with only one: L1, L2, L3, L4, L5 (L1=hands touching head/face, L2=hands touching torso, L3=hands touching arm, L4=hands touching weak/passive hand, L5=hands in front of the body or face)
>
> *Path Shape?* Movement path shape? Answer with only one: Hold, Straight, Arched, Circular. (Hold=no path or direction, Straight=move in a straight line, Arched=move in an arched line, Circular=move in a circular path)
>
> *Path Repetition?* Answer with only one: Single, Repeated. (Single=one movement, Repeated=multiple or repeated movements)
>
> *Handedness?* Answer with only one: One-handed, Two-handed symmetrical, Two-handed asymmetrical. (One-handed=only one hand is used in the sign, Two-handed symmetrical=two hands are used but the hands move together and have the same handshape, Two-handed asymmetrical=two hands are visible, but one hand does not move and the hands have different handshapes)

Transparency (open-set over 96 glosses) and Transparency$_2$ (10-choice Instructions)

> *What does this sign resemble? Look at the form and movement of the sign. Choose the most likely option from these possibilities: <OPTIONS>. Answer with only the exact word from the list.*

Iconicity Rating Instructions:

> *This sign means: <MEANING>. Some signs are iconic and some are arbitrary. Find visual resemblances between the meaning and the form of the sign. How much does the sign look like "<MEANING>"? Answer with only one number: 1,2,3,4,5,6,7 (1=not at all, 7=exactly).*

D Iconicity Type Examples

Combined iconic
SPIDER: The wiggling motion of the hands conveys the spider's movement, while the curved fingers depict its legs.

Object-based iconic
BUTTERFLY: two hands mirror the referent's *wings*.

Action-based iconic
TO-SMS: thumb and fingers enact the *typing/texting* action.

Arbitrary
ELECTRICITY: hand configuration and path show no transparent visual resemblance; the form–meaning link is purely conventional.

Figure 6: Representative frames illustrating the four iconicity categories. Each pair of frames shows how the sign fits its category.

E Full Results Tables for Phonology

Model	Handshape	Location	Path Shape	Path Repetition	Handedness	Mean
Human baseline (hearing non-expert)	0.698	0.823	0.677	0.833	0.938	0.794
Gemini-2.5-Pro	0.677	0.865	0.417	0.646	0.927	0.706
GPT-5	0.625	0.740	0.468	0.708	0.948	0.698
GPT-4o	0.417	0.865	0.365	0.562	0.917	0.625
Qwen2.5-VL-72B	0.490	0.771	0.344	0.563	0.823	0.598
LLaVA-OV-Qwen2-72B	0.302	0.677	0.438	0.479	0.865	0.552
Qwen2.5-VL-32B	0.417	0.719	0.354	0.563	0.708	0.552
VideoLLaMA2-72B	0.323	0.729	0.188	0.563	0.917	0.544
Gemma-3-27B	0.333	0.781	0.302	0.552	0.708	0.535
LLaVA-Video-72B-Qwen2	0.271	0.646	0.323	0.510	0.885	0.527
Qwen2.5-VL-7B	0.427	0.385	0.188	0.552	0.531	0.417
LLaVA-OV-Qwen2-7B	0.167	0.167	0.313	0.604	0.802	0.411
LLaVA-Video-7B-Qwen2	0.260	0.135	0.313	0.417	0.583	0.342
VideoLLaMA2-7B	0.083	0.135	0.188	0.000	0.198	0.121
Random baseline	0.143	0.200	0.250	0.500	0.333	0.285

Table 8: Phonological form prediction accuracy by model and phonological subtasks, with random and human baselines, and mean accuracy across all subtasks.

F Additional Figures

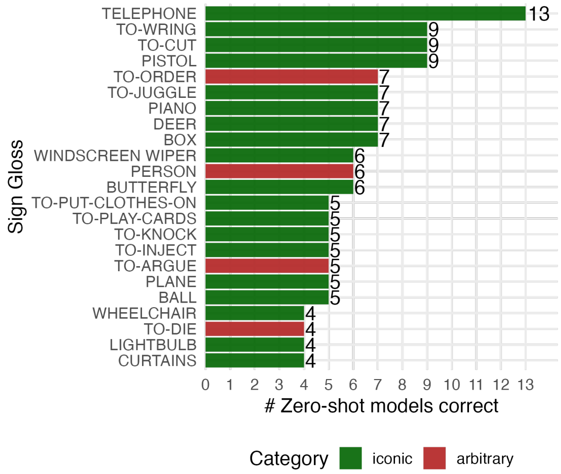


Figure 7: Correctly guessed signs from sign video only (>3 models) in the Transparency$_2$ Task.

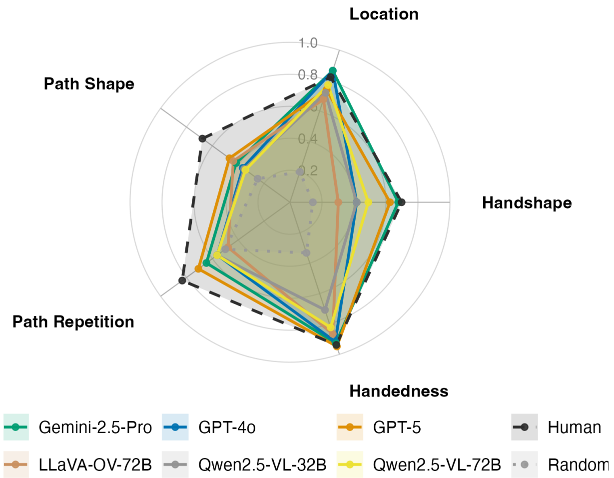


Figure 8: Zero-shot accuracy of the top 6 performing models (3 open, 3 closed) across five phonological features, averaged over 96 signs. Solid colored lines are models; black dashed and grey dotted are the human non-expert and random baselines, respectively.