MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

Iconicity in Visual Communication: From Silent Gestures and Signs to Vision-Language Models

Onur Keleş

Boğaziçi University, Dept. of Linguistics

Introduction

Introduction

PhD Candidate in Linguistics, currently at Boğaziçi University, Istanbul

MA Thesis: Discourse Cohesion and Phonetics in Turkish Sign Language (2024; supervised by Kadir Gökgöz and Nazik Dinçtopal Deniz)

Planning a PhD with MPI collaboration:

Supervisors: Dr. Kadir Gökgöz, Prof. Aslı Özyürek, Dr. Esam Ghaleb

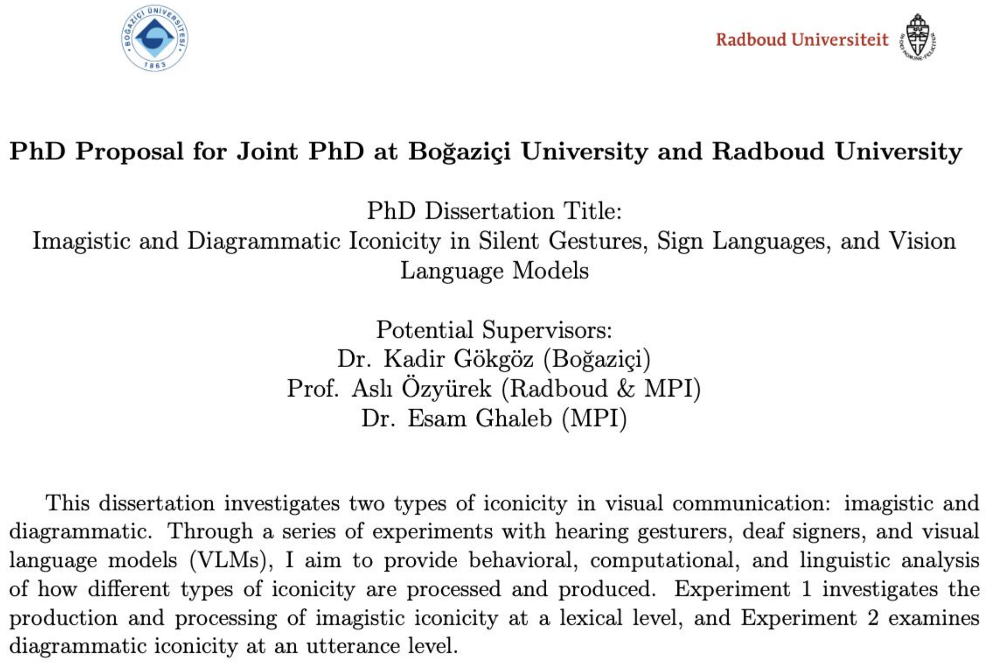
PhD Plans

What is going to my PhD focus?

Iconicity in silent gestures, sign languages, and AI models (VLMs)

Why this matters:

Bridges cognitive science, linguistics, and engineering for better understanding of visual language processing



Radboud Universiteit

PhD Proposal for Joint PhD at Boğaziçi University and Radboud University

PhD Dissertation Title:
Imagistic and Diagrammatic Iconicity in Silent Gestures, Sign Languages, and Vision Language Models

Potential Supervisors:
Dr. Kadir Gökgöz (Boğaziçi)
Prof. Aslı Özyürek (Radboud & MPI)
Dr. Esam Ghaleb (MPI)

This dissertation investigates two types of iconicity in visual communication: imagistic and diagrammatic. Through a series of experiments with hearing gesturers, deaf signers, and visual language models (VLMs), I aim to provide behavioral, computational, and linguistic analysis of how different types of iconicity are processed and produced. Experiment 1 investigates the production and processing of imagistic iconicity at a lexical level, and Experiment 2 examines diagrammatic iconicity at an utterance level.

Defining Iconicity

A "structure-preserving mapping between mental models of linguistic form and meaning" (Taub, 2001, p. 23).

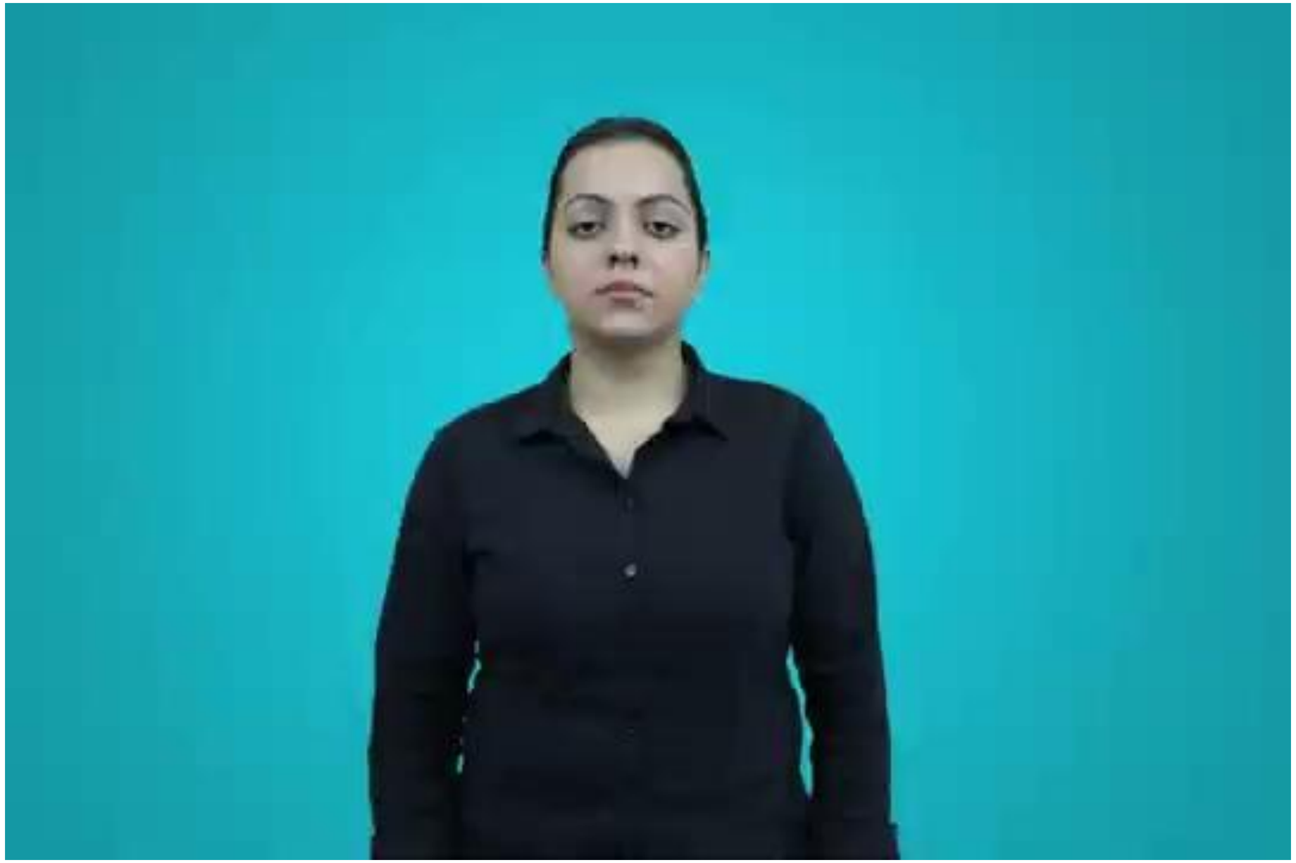
The form of a sign visually resembles its concept, unlike arbitrary forms.

Iconicity is not an absolute property but may be conventionalized within each sign language through an analogue-building model (Emmorey, 2014).

(Examples are from http://tidsozluk.aile.gov.tr; Makaroğlu and Dikyuva, 2017)

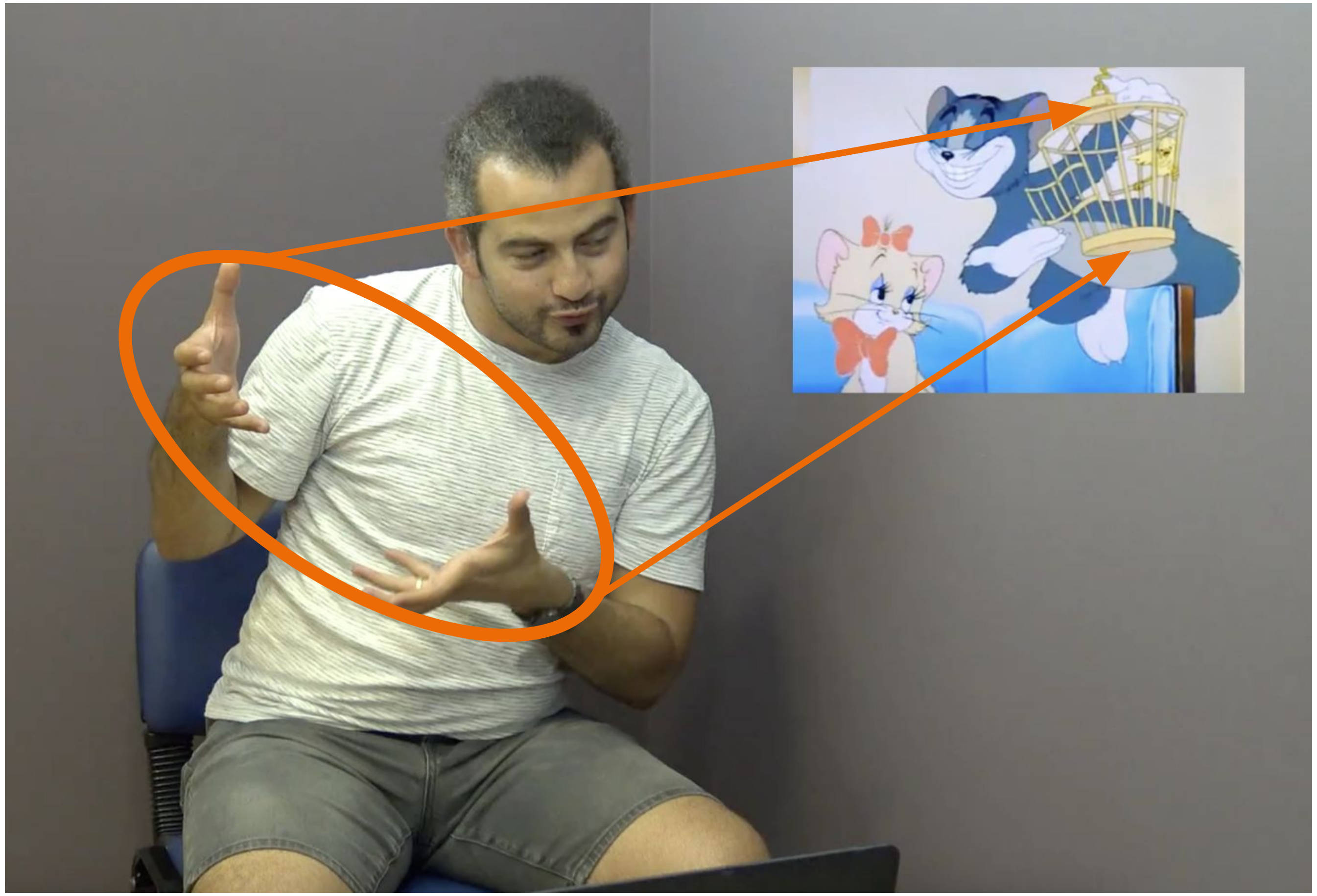
Imagistic Iconicity

Based on perceptual resemblance between linguistic form and meaning elements



Diagrammatic Iconicity

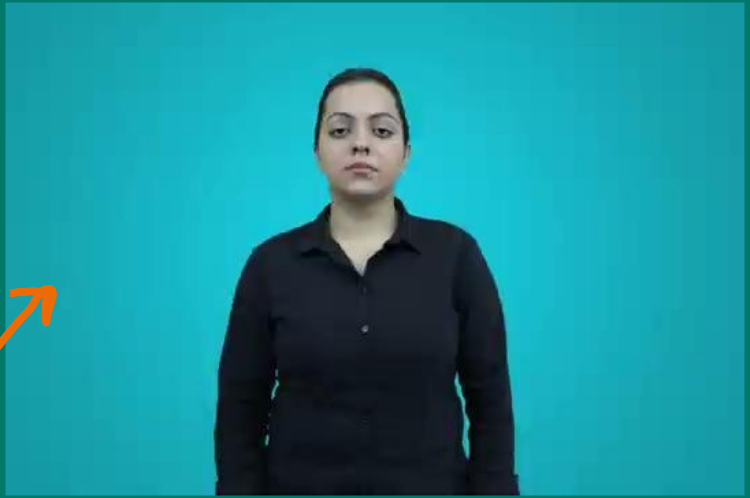
Based on structural resemblance between meaning elements and the relationship between articulators



See Ortega, Sümer, & Özyürek (2017)

Action-based Iconicity:

Perception-based Iconicity:



PhD Proposal

Proposal (2025-2028)

Aim

- Investigate how imagistic and diagrammatic iconicity are produced and processed in silent gestures, sign languages, and vision–language models. Combine behavioral, computational, and linguistic perspectives.

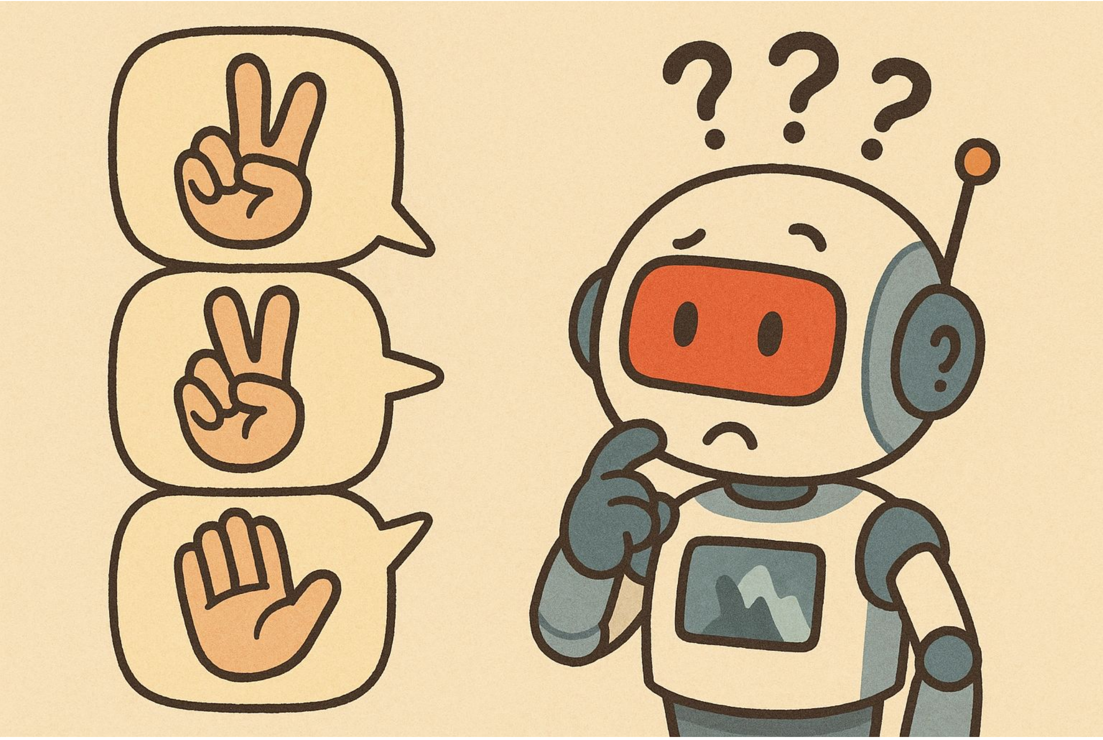
Research Questions

1. How do hearing gesturers and deaf signers differ in producing and processing imagistic iconicity at the lexical level?
2. How is diagrammatic iconicity expressed and comprehended in complex utterances by signers compared to gesturers?
3. To what extent can state-of-the-art vision–language models learn and interpret iconicity?

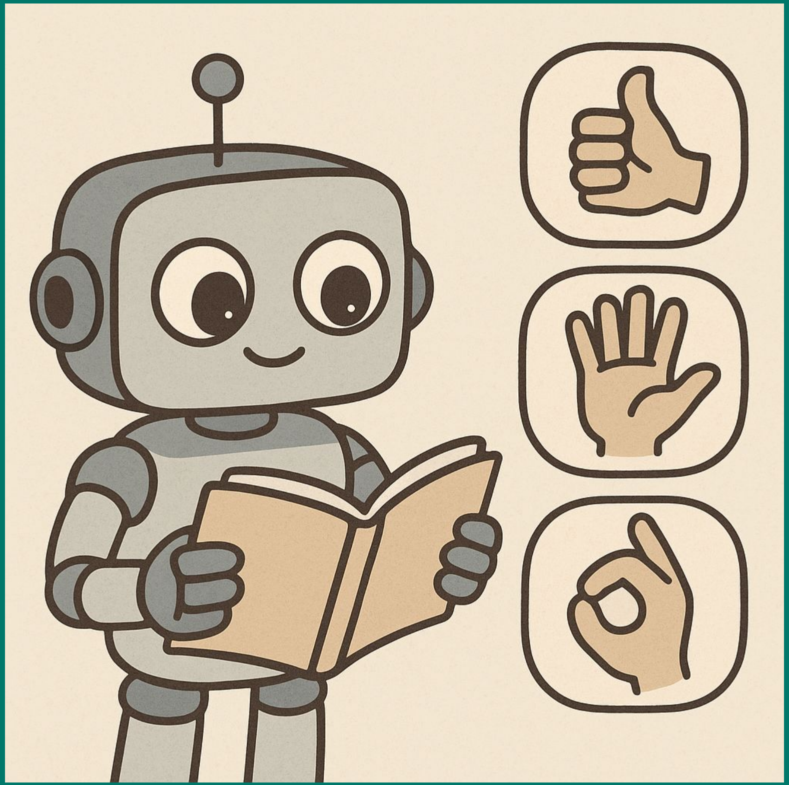
Why Is This Important?

Research Question to Focus

To what extent can state-of-the-art vision–language models learn and interpret iconicity?



VLM Benchmarking Project with Esam Ghaleb

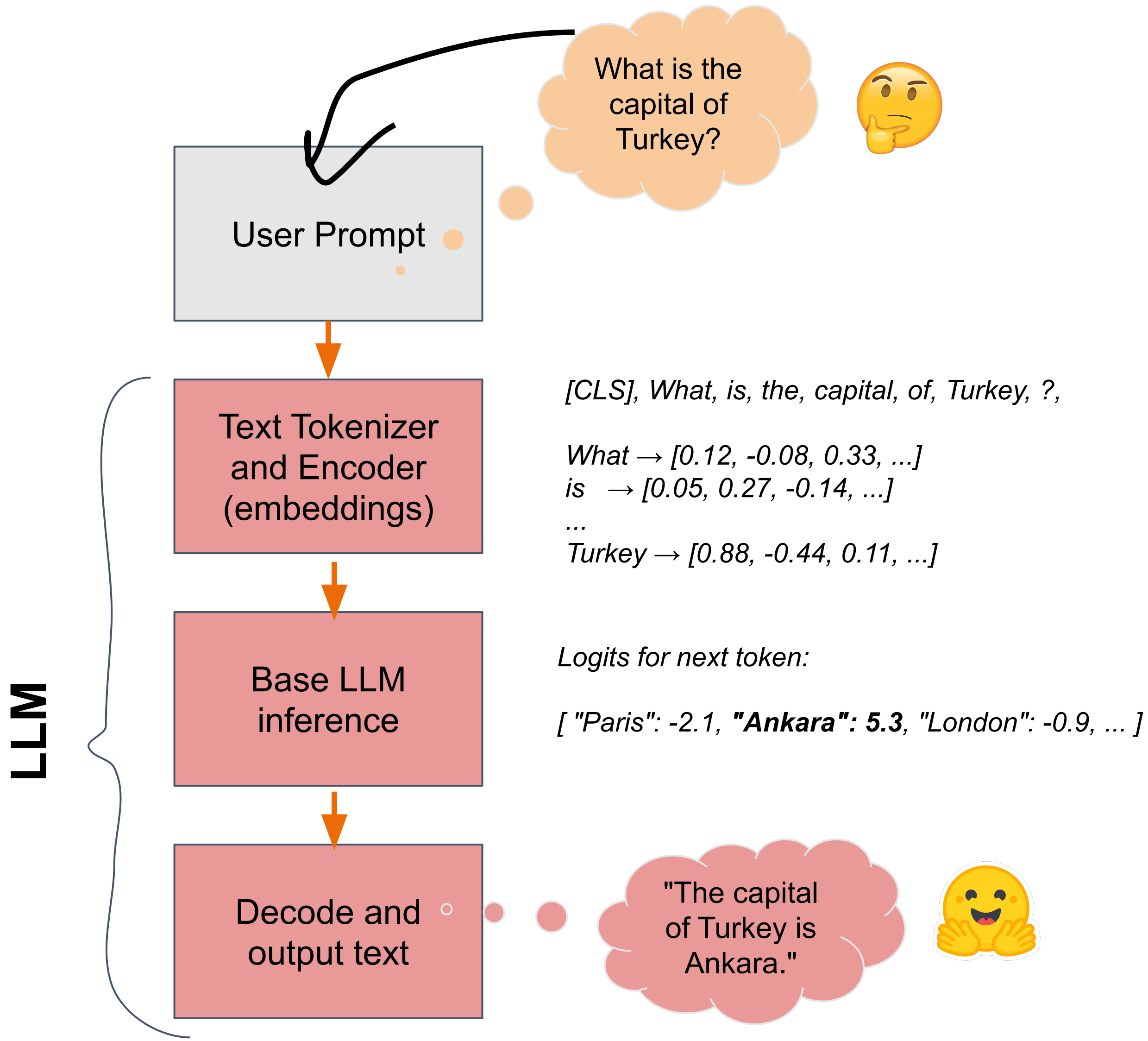


A (VERY) short tutorial on LMs for linguists

What are LLMs?

Definition: LLMs are AI systems trained on massive text datasets to generate and understand human-like language. Text-based LLMs are restricted to text input and output.

- Learn statistical patterns in billions of words
- Use deep neural networks (transformers) to model relationships between words.
- Predict the next word, sentence, or answer in context.



What is the capital of Turkey?

User Prompt

LLM

Text Tokenizer and Encoder (embeddings)

[CLS], What, is, the, capital, of, Turkey, ?,

What → [0.12, -0.08, 0.33, ...]
is → [0.05, 0.27, -0.14, ...]
...
Turkey → [0.88, -0.44, 0.11, ...]

Base LLM inference

Logits for next token:

*["Paris": -2.1, **"Ankara": 5.3**, "London": -0.9, ...]*

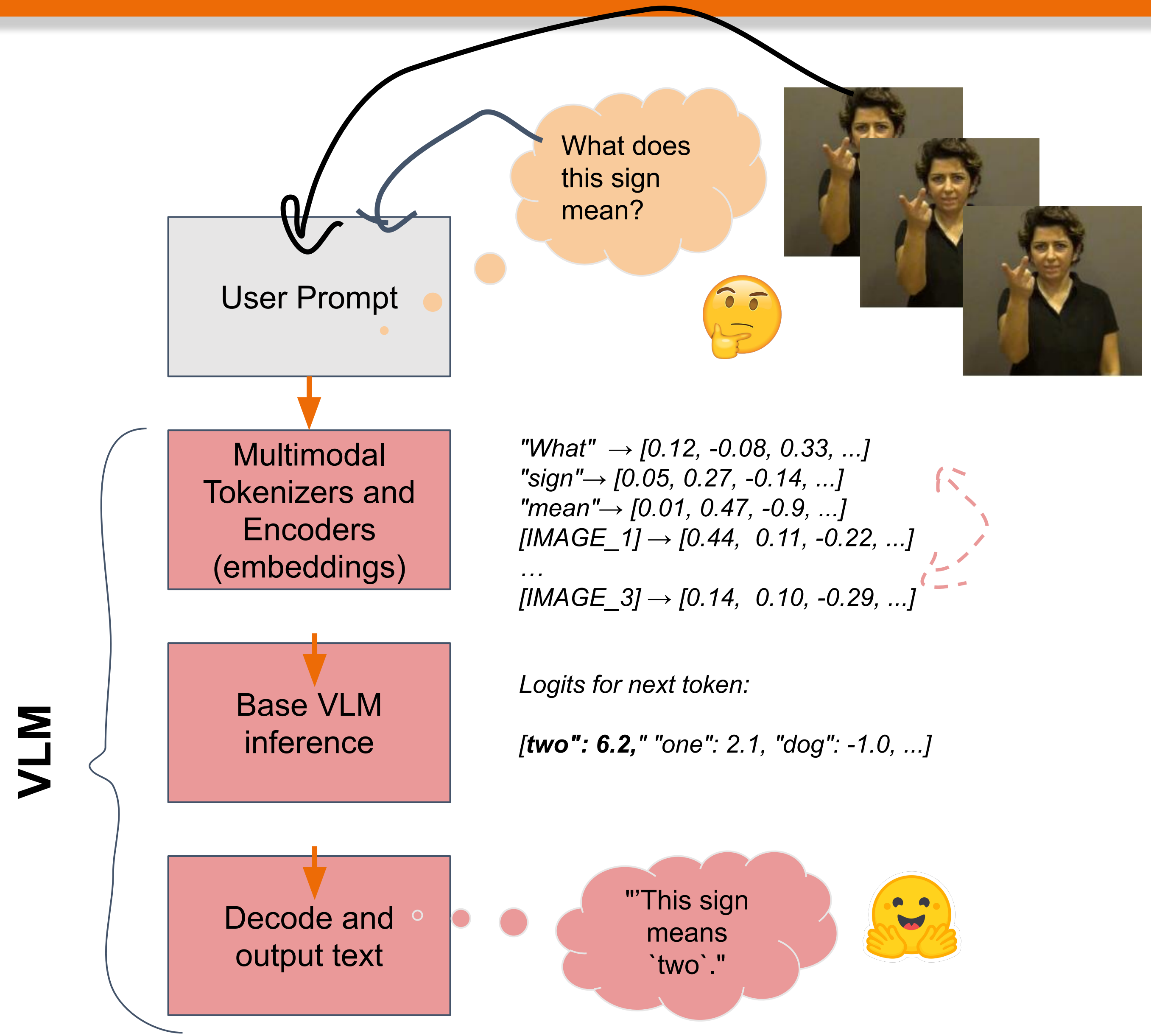
Decode and output text

"The capital of Turkey is Ankara."

What are VLMs?

Definition: VLMs (a type of multimodal LM) extend LLMs by integrating visual data (images, video) with language.

- Encode visual features (objects, motion, scenes).
- Align them with language embeddings.
- Generate text that describes, interprets, or reasons about visuals.



What does this sign mean?

User Prompt

VLM

Multimodal Tokenizers and Encoders (embeddings)

"What" → [0.12, -0.08, 0.33, ...]
"sign"→ [0.05, 0.27, -0.14, ...]
"mean"→ [0.01, 0.47, -0.9, ...]
[IMAGE_1] → [0.44, 0.11, -0.22, ...]
…
[IMAGE_3] → [0.14, 0.10, -0.29, ...]

Base VLM inference

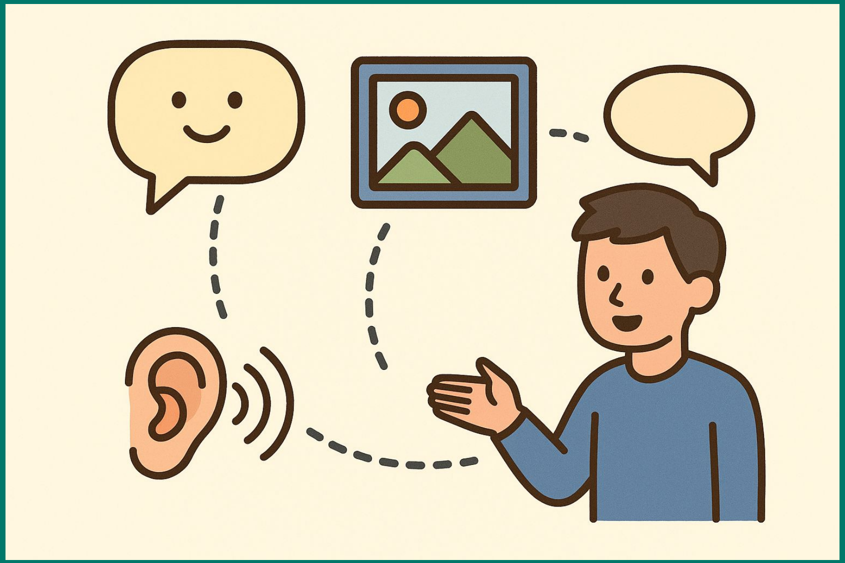
Logits for next token:

*[**two": 6.2,**" "one": 2.1, "dog": -1.0, ...]*

Decode and output text

"’This sign means `two`."

Multimodal in NLP
≠
Multimodal in Lingustics



What do we know so far?

LLMs (text-only):

- *Marklová (2025):* GPT-4 generates iconic pseudowords in text → humans and models can guess meanings.
- *Loakman (2024):* Larger models' text iconicity ratings align more closely with humans.

VLMs (multimodal, general):

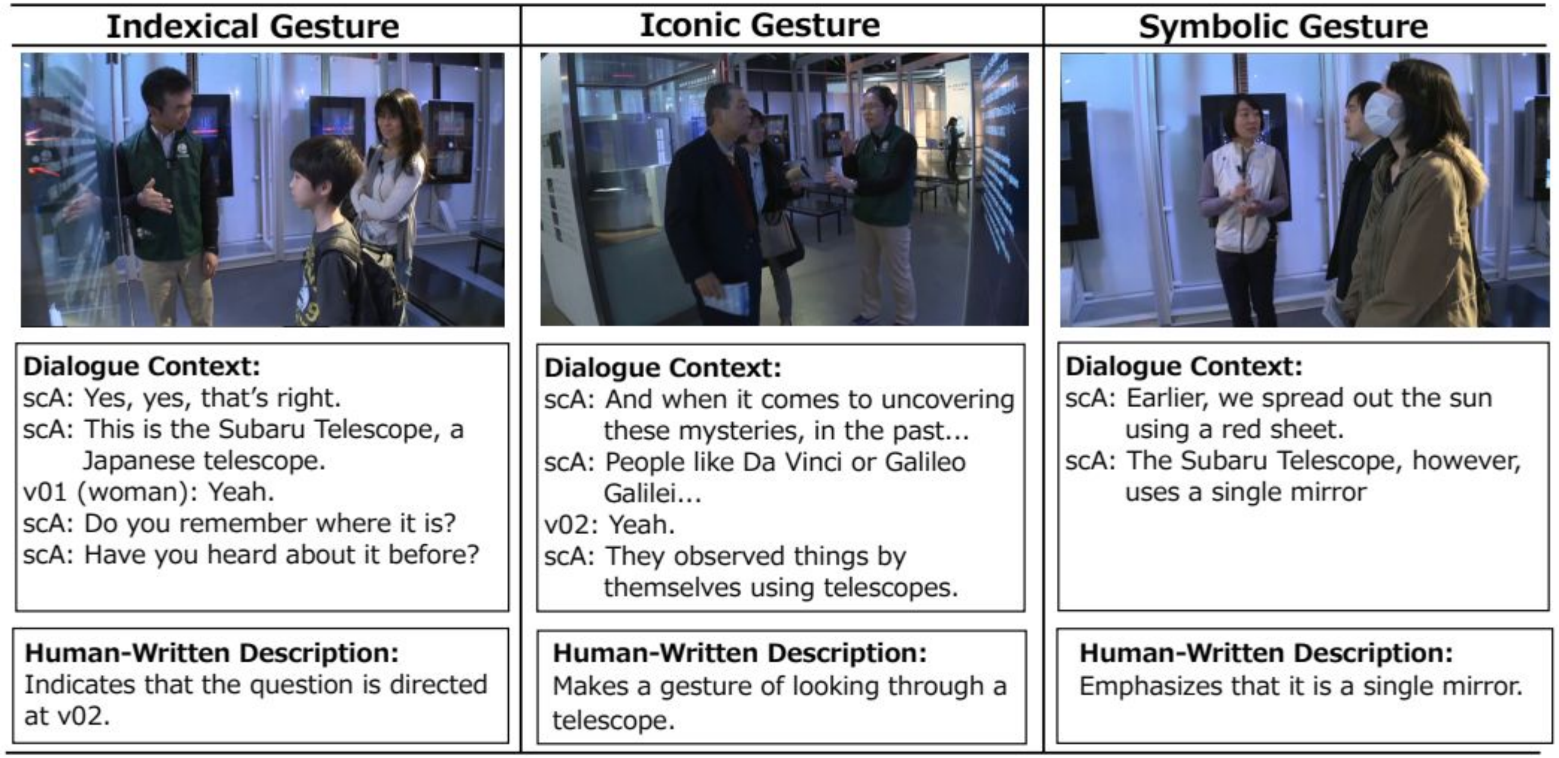
- *Alper & Averbuch-Elor (2023):* CLIP/Stable Diffusion show weak kiki–bouba alignment → likely dataset co-occurrence.

(Nishida et al., 2025)

What do we know so far?

Gesture & Sign:

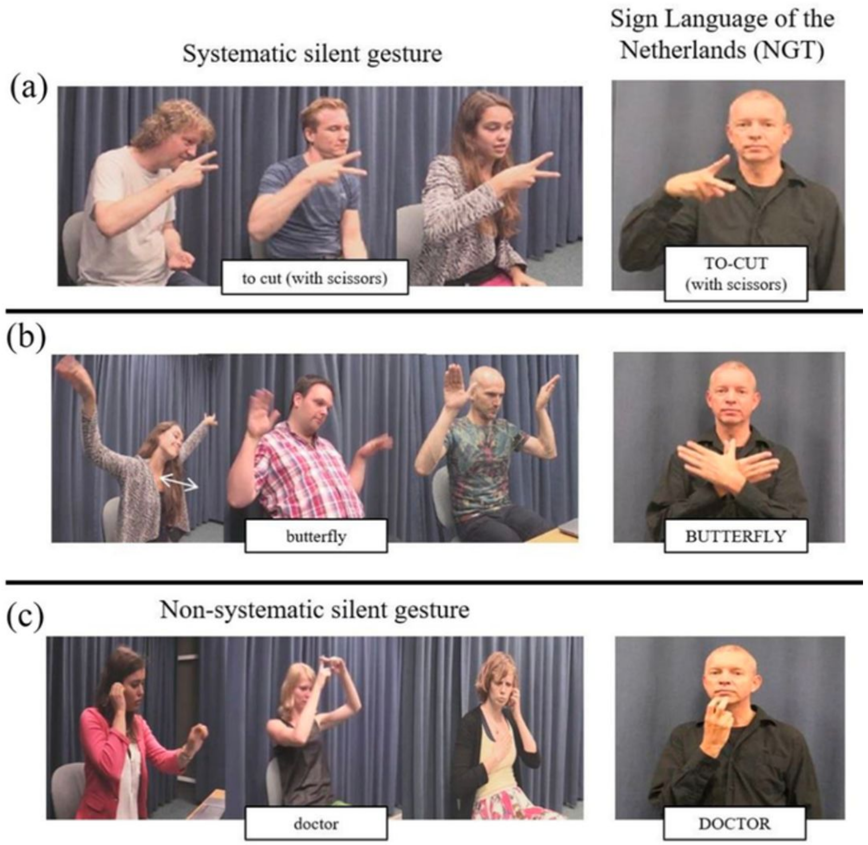
Nishida et al. (2025): VLMs underperform on indexical/iconic gestures, especially with visual-only input → heavy bias toward text cues.



Indexical Gesture	Iconic Gesture	Symbolic Gesture
Dialogue Context: scA: Yes, yes, that's right. scA: This is the Subaru Telescope, a Japanese telescope. v01 (woman): Yeah. scA: Do you remember where it is? scA: Have you heard about it before?	**Dialogue Context:** scA: And when it comes to uncovering these mysteries, in the past... scA: People like Da Vinci or Galileo Galilei... v02: Yeah. scA: They observed things by themselves using telescopes.	**Dialogue Context:** scA: Earlier, we spread out the sun using a red sheet. scA: The Subaru Telescope, however, uses a single mirror
Human-Written Description: Indicates that the question is directed at v02.	**Human-Written Description:** Makes a gesture of looking through a telescope.	**Human-Written Description:** Emphasizes that it is a single mirror.

Gap: No systematic benchmark of VLMs on **imagistic iconicity in signed languages and gestures**.

96-Item Stimuli in Karadöller et al. (2024)
Gesture Database: Ortega et al. (2020)


Systematic silent gesture
Sign Language of the Netherlands (NGT)
(a)
to cut (with scissors)
TO-CUT (with scissors)
(b)
butterfly
BUTTERFLY
(c)
Non-systematic silent gesture
doctor
DOCTOR

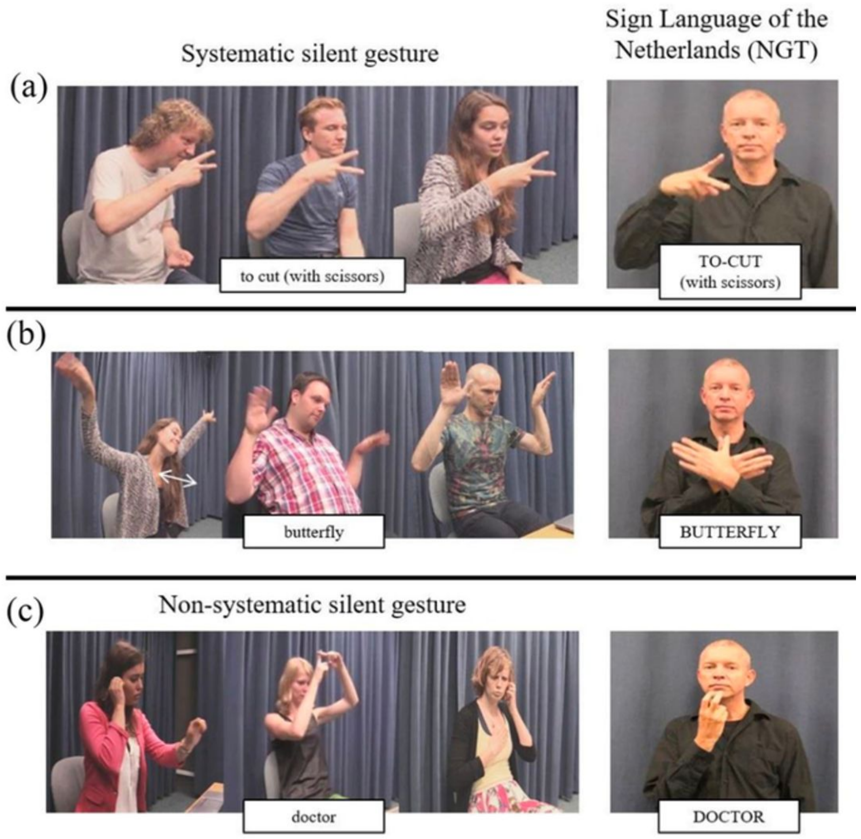
Categories:

(a): Iconic Signs with High Overlap with Gestures (N = 32)

(b): Iconic Signs with Low Overlap with Gestures (N = 32)

(c): Arbitrary Signs with No Overlap with Gestures (N = 32)

96-Item Stimuli in Karadöller et al. (2024)
Gesture Database: Ortega et al. (2020)


Systematic silent gesture
Sign Language of the Netherlands (NGT)
(a)
to cut (with scissors)
TO-CUT (with scissors)
(b)
butterfly
BUTTERFLY
(c)
Non-systematic silent gesture
doctor
DOCTOR

Stimulus selection:

- Based on gesture database (Ortega et al., 2020).
- Classified by overlap in **handshape, location, movement, orientation**.

Iconicity ratings (7-point scale):

- High-overlap iconic: **M = 5.13** (SD = 1.02)
- Low-overlap iconic: **M = 4.42** (SD = 1.08)
- Arbitrary: **M = 2.10** (SD = 0.50)

(Emmorey, 2014)

Aims and RQs

Test whether VLMs capture **structured form–meaning mappings (iconicity)** given 96 NGT signs

RQ1: Can VLMs produce reliable iconicity judgments?
RQ2: Do VLMs recognize key phonological features (handshape, location, path shape, repetition, handedness)?

- **From theory:** Iconicity as structure mapping between phonological form and meaning (Emmorey, 2014).
- **Expectation:** Stronger phonological competence ⇒ better-calibrated iconicity and fewer text-biased errors.

(Klomp & Pfau, eds., 2020)

Why Test Phonological Competence?

Pilot attempt: Directly queried models for iconicity ratings. 🤔

Issue observed: Some models hallucinated and showed bias toward textual prompts rather than visual evidence.

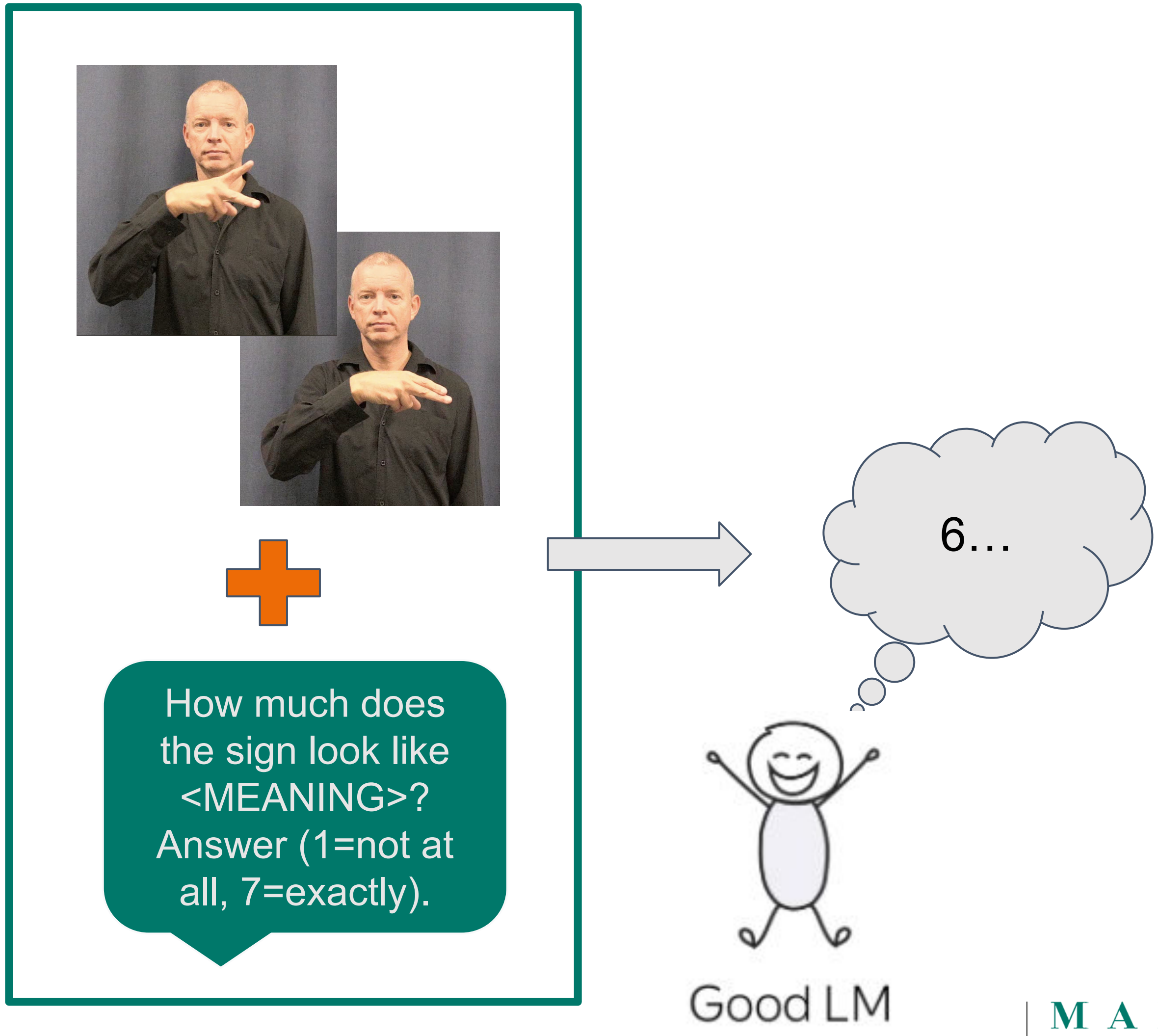
Our response: Add phonological competence tasks that force attention to sublexical form. 🤗

- **Labels & data:** Phonology labels adapted from NGT (Klomp & Pfau, eds., 2020) to standardize feature definitions.
- **Takeaway:** Benchmark both iconicity judgments and phonological feature recognition to disentangle text-bias from genuine visual understanding.

Our Benchmarking Project

We present the **first benchmark** of state-of-the-art VLMs on sign iconicity:

- New evaluation pipeline with multiple tasks
- Dataset: 96 NGT signs with videos, phonological annotations, human iconicity ratings
- Models: 12 recent VLMs **(zero-shot)**
- Tasks: phonological competence, transparency, binary iconicity, graded rating



How much does the sign look like <MEANING>? Answer (1=not at all, 7=exactly).

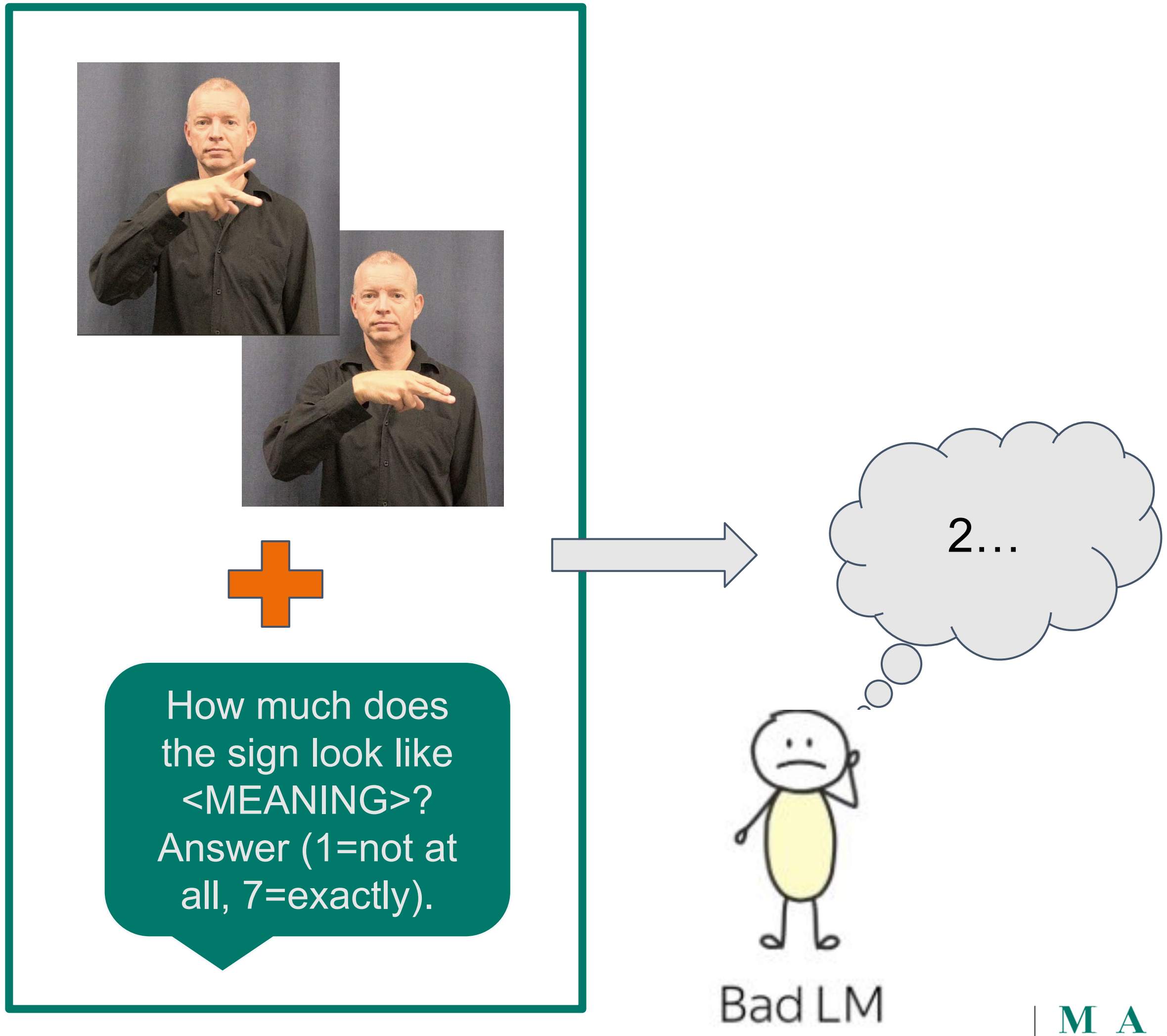
6…

Good LM

Our Benchmarking Project

We present the **first benchmark** of state-of-the-art VLMs on sign iconicity:

- New evaluation pipeline with multiple tasks
- Dataset: 96 NGT signs with videos, phonological annotations, human iconicity ratings
- Models: 12 recent VLMs **(zero-shot)**
- Tasks: phonological competence, transparency, binary iconicity, graded rating



How much does the sign look like <MEANING>? Answer (1=not at all, 7=exactly).

Phonology Tasks (manually annotated by me for this project)

1

Major sign handshape? **Answer with only one: H1,H2,H3,H4,H5,H6,H7"**

(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)

2

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)

Phonology Tasks (manually annotated by me for this project)

3

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)

4

Movement repetition? Answer with only one: Single, Repeated.

(Single=one movement, Repeated=multiple or repeated movements)

5

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 Tasks

6

Transparency-OpenSet (96): What does this sign resemble?

"Choose the most likely option from these possibilities: {gloss_options}.

"Answer with only the exact word from the list that best matches what the sign looks like."

"If the sign does not resemble any of the above, answer 'UNKNOWN'"

7

Transparency-Small Set (10): What does this sign resemble?

"Choose the most likely option from these possibilities: {gloss_options}.

"Answer with only the exact word from the list that best matches what the sign looks like."

"If the sign does not resemble any of the above, answer 'UNKNOWN'"

Iconicity Tasks

8

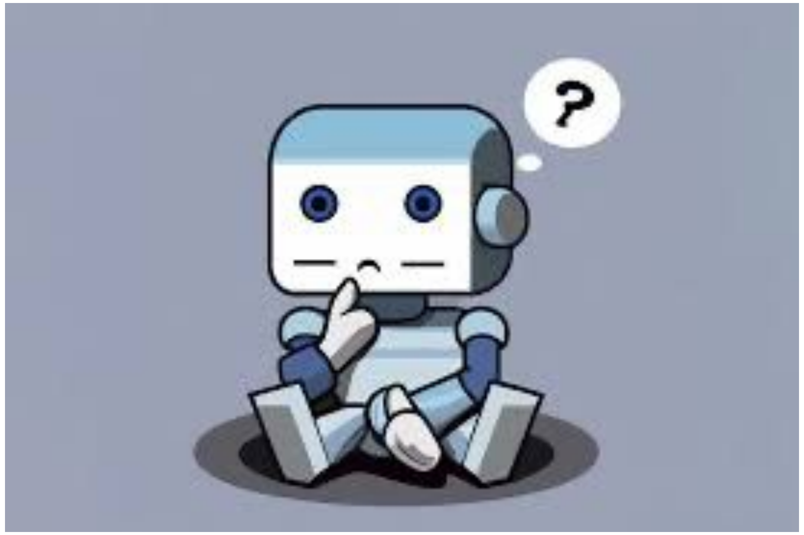
Iconicity Binary: Meaning: {meaning}.
Some signs are iconic and some are arbitrary.
Find visual resemblances between the meaning and the form of the sign.
Does the sign look like or resemble '{meaning}'? Answer only one word: yes or no

9

Iconicity Ratings: Meaning: {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).

Tested Models

(but top 5 will be reported for each task)



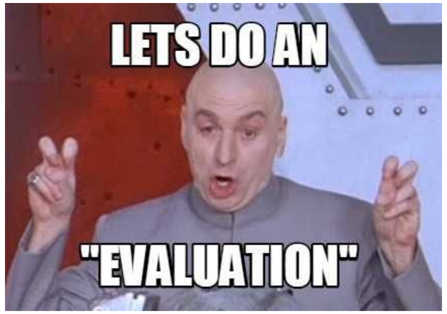
- Gemma-3-27B
- Qwen2.5-VL-72B
- Qwen2.5-VL-32B
- Qwen2.5-VL-7B
- VideoLLaMA2-72B
- VideoLLaMA2-7B
- LLaVA-Video-Qwen2-72B
- LLaVA-Video-Qwen2-7B
- LLaVA-Onevision-Qwen2-72B
- LLaVA-Onevision-Qwen2-7B
- MiniCPM-V-4-4B
- MiniCPM-V-2_6-7B

Evaluation Metrics

- **Phonology & Transparency (categorical):**
  - *Accuracy* = overall correctness
  - *F1 Score* = unweighted average across classes, penalizes bias

- **Binary Iconicity (yes/no):**
  - *Balanced Accuracy* = equal weight to iconic vs. arbitrary classes
  - *Matthews Corr. Coef. (MCC)* = correlation-like score, −1 to +1

- **Graded Iconicity Ratings (1–7 scale):**
  - *Spearman's ρ* = rank correlation with human ratings
  - *AUC* = sensitivity to separating iconic vs. arbitrary categories
  - *Normalized Cohen's d* = effect size for category separation

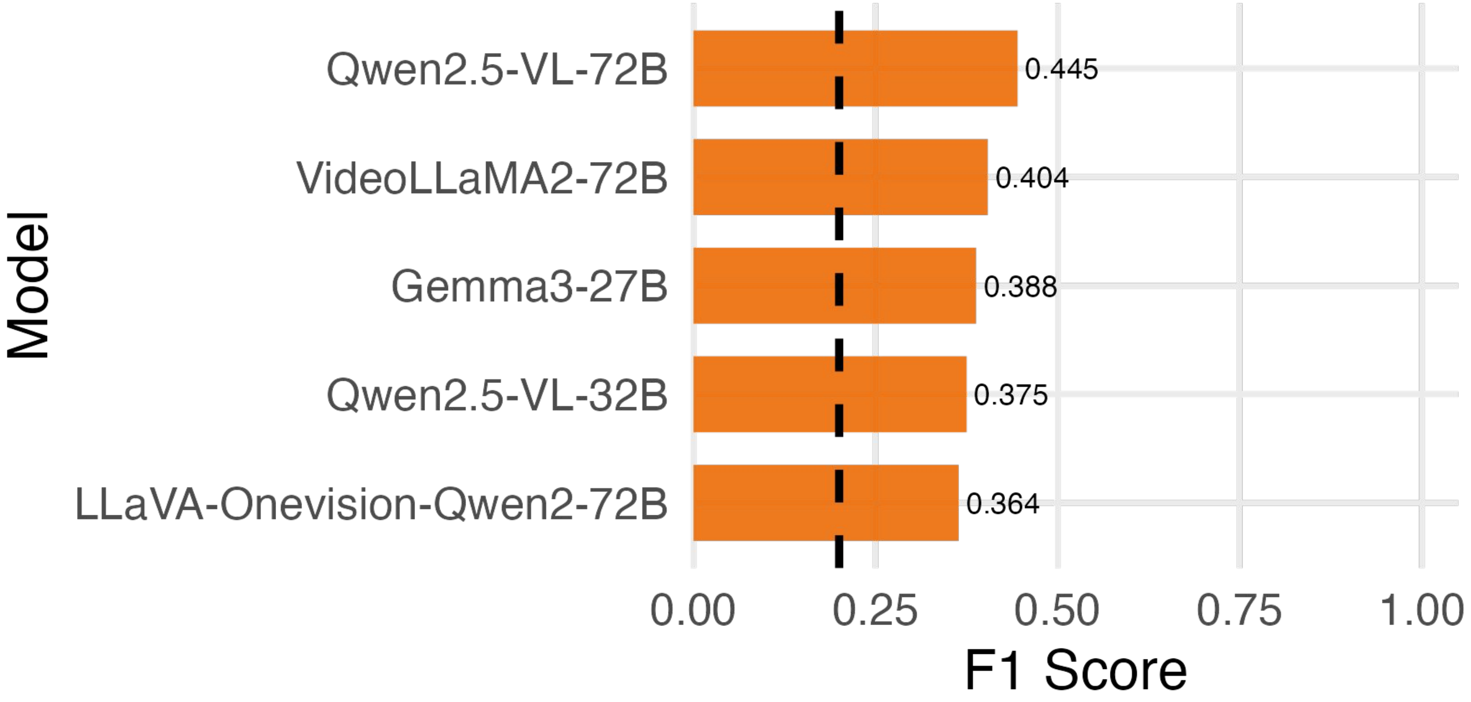


Results

Phonological Competence

Overall Phonological Competence Results (F1 only)

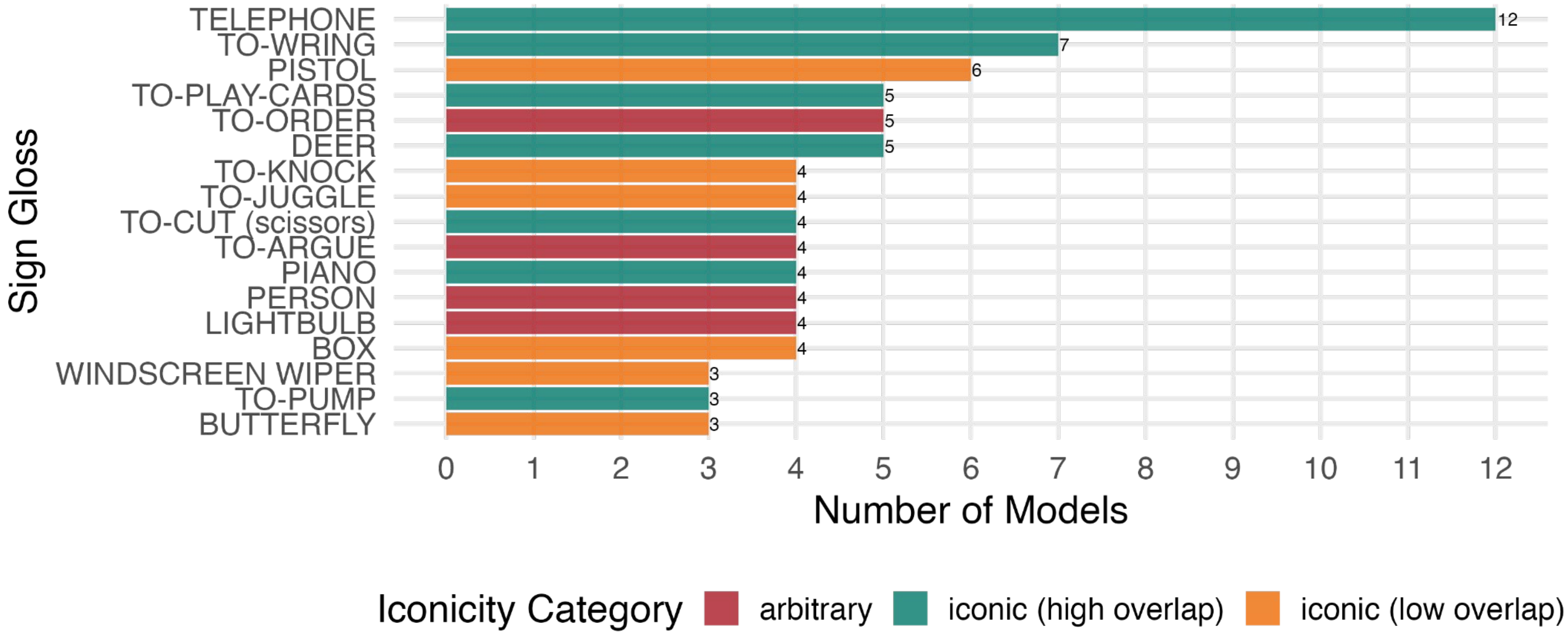
Random baseline is the dashed vertical black line.



- Models exceed baselines but remain modest
- Best: Qwen2.5-VL-72B and VideoLLaMA2-72B
- Strongest features: **location, handedness**
- Hardest: **handshape, path shape**

Transparency

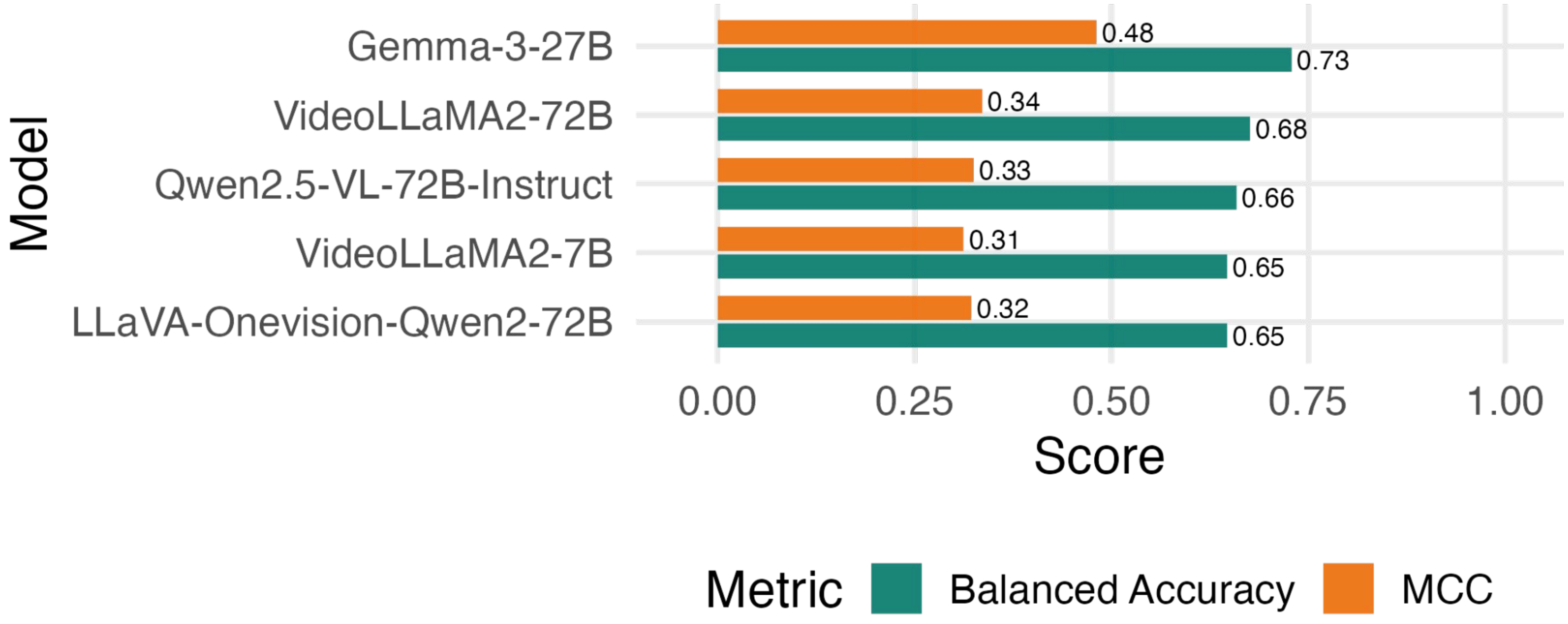
Signs correctly guessed by ≥3 Models



- Correct guesses cluster on visually salient items (TELEPHONE, PISTOL)
- Some “arbitrary” but cross-linguistically common signs guessed correctly (PERSON, TO-ORDER) → likely training-data redundancy

Iconicity

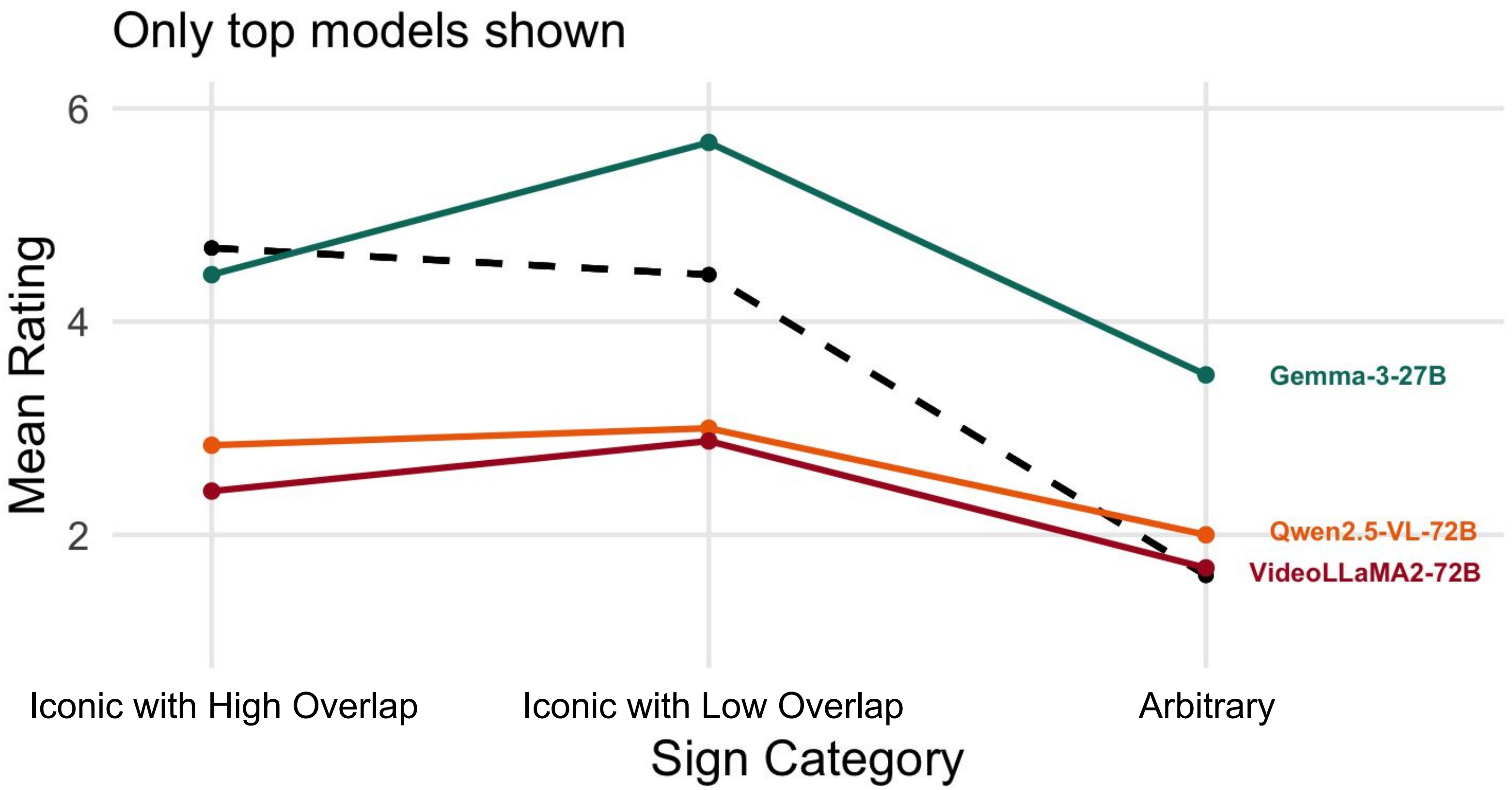
Balanced Accuracy and MCC Results for Binary Iconicity



- Best: **Gemma-3-27B** (Balanced Acc 0.73, MCC 0.48)
- Next tier: VideoLLaMA2-72B, Qwen2.5-VL-72B
- Smaller models ~ chance, often over-predict iconicity

Mean Iconicity Ratings by Model and Sign Category.

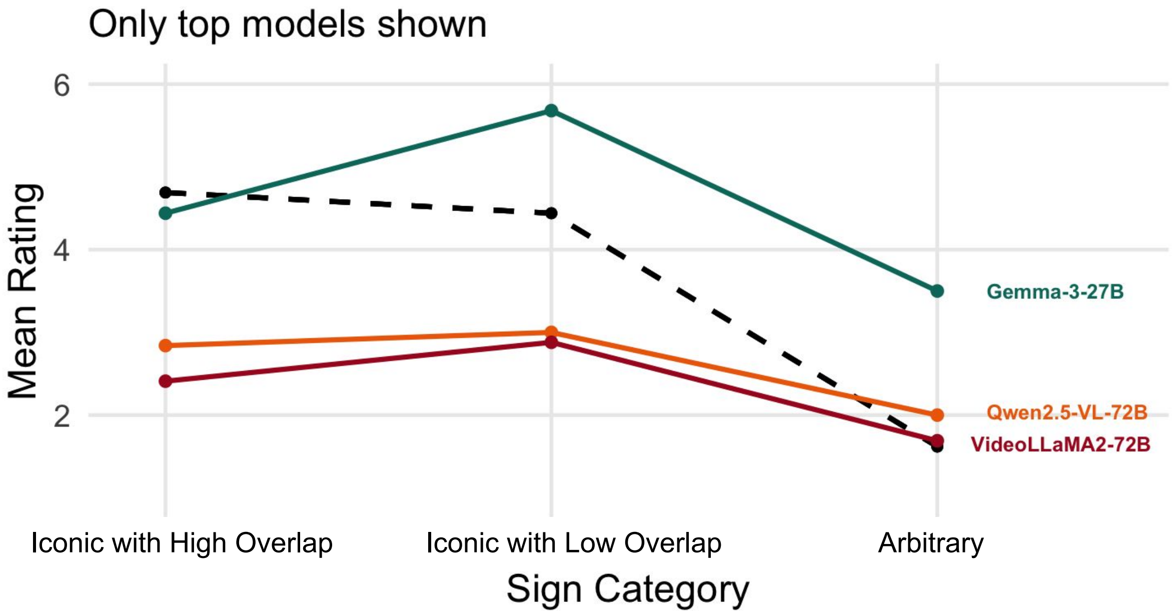
Black dashed line indicates average human ratings



- Models compress or distort human scale
- Best: **Gemma-3-27B** (ρ = 0.43, d = 1.03, Overall = 0.63)
- Qwen2.5-VL-72B: higher correlation but weaker separation
- Smaller & LLaVA/MiniCPM: collapse distinctions entirely

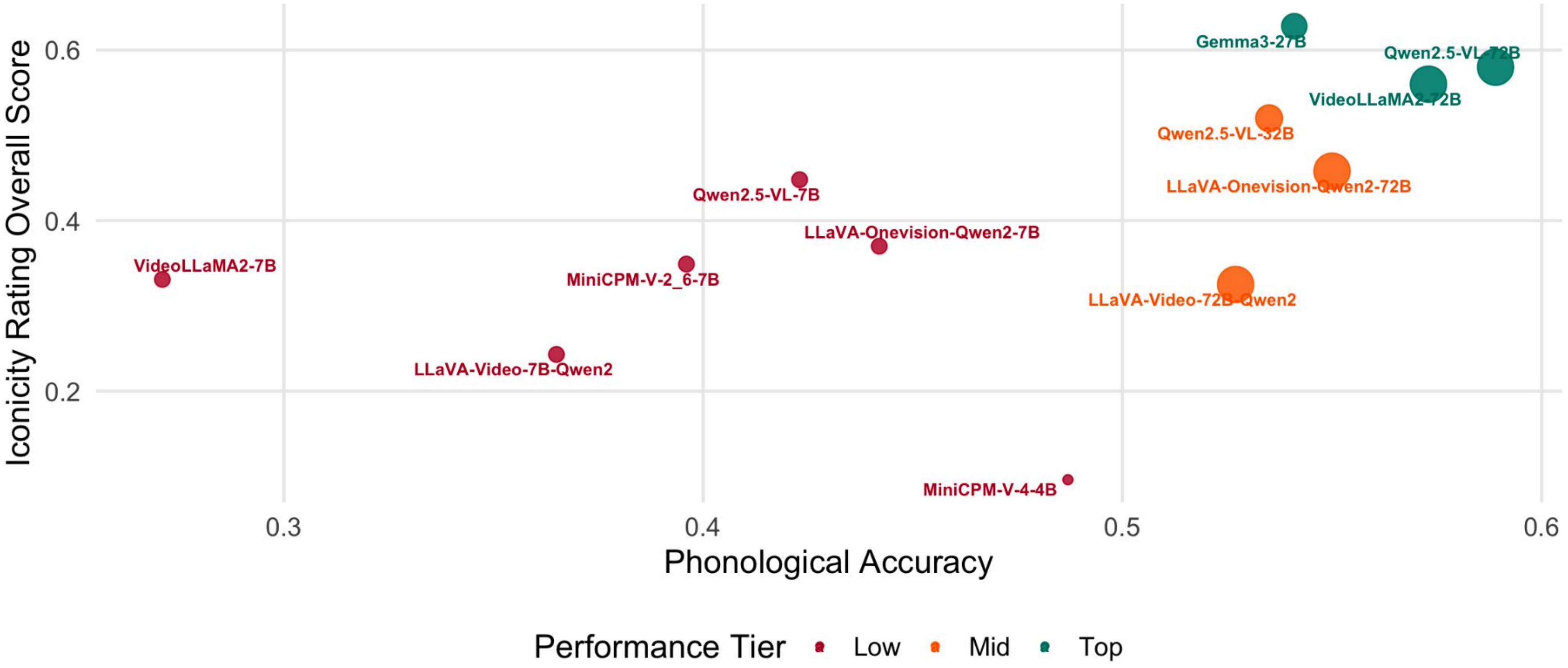
Mean Iconicity Ratings by Model and Sign Category.

Black dashed line indicates average human ratings



Model	H-L	Iconic-Arbitrary
Gemma-3-27B	** p = .001	*** p < .001
Qwen2.5-VL-72B	ns	*** p < .001
VideoLLaMA2-72B	ns	*** p < .001

Overall Model Performance



As a summary…

Only a handful of models:

- Endorse iconic signs **and**
- Reject arbitrary signs **and**
- Reproduce graded human ratings

Gemma-3-27B emerges most reliable

Qwen2.5-VL-72B, VideoLLaMA2-72B: second tier
Most smaller/LLaVA/MiniCPM: fail to distinguish categories

Conclusions

Partial Sensitivity: VLMs show some awareness of form–meaning resemblance but success clusters on *visually salient* signs (e.g., TELEPHONE, PISTOL) or *cross-linguistic redundancies*.

Systematic Failures: Iconic signs with low gesture overlap and arbitrary signs expose weaknesses. Models often over-predict iconicity or compress rating scales to midpoints.

Similar Mechanisms: Phonological description accuracy and iconicity alignment **do** correlate.

Implications: Current zero-shot VLMs rely on shortcuts rather than structured iconic reasoning and they require scaffolding to improve.

Ideas for Future

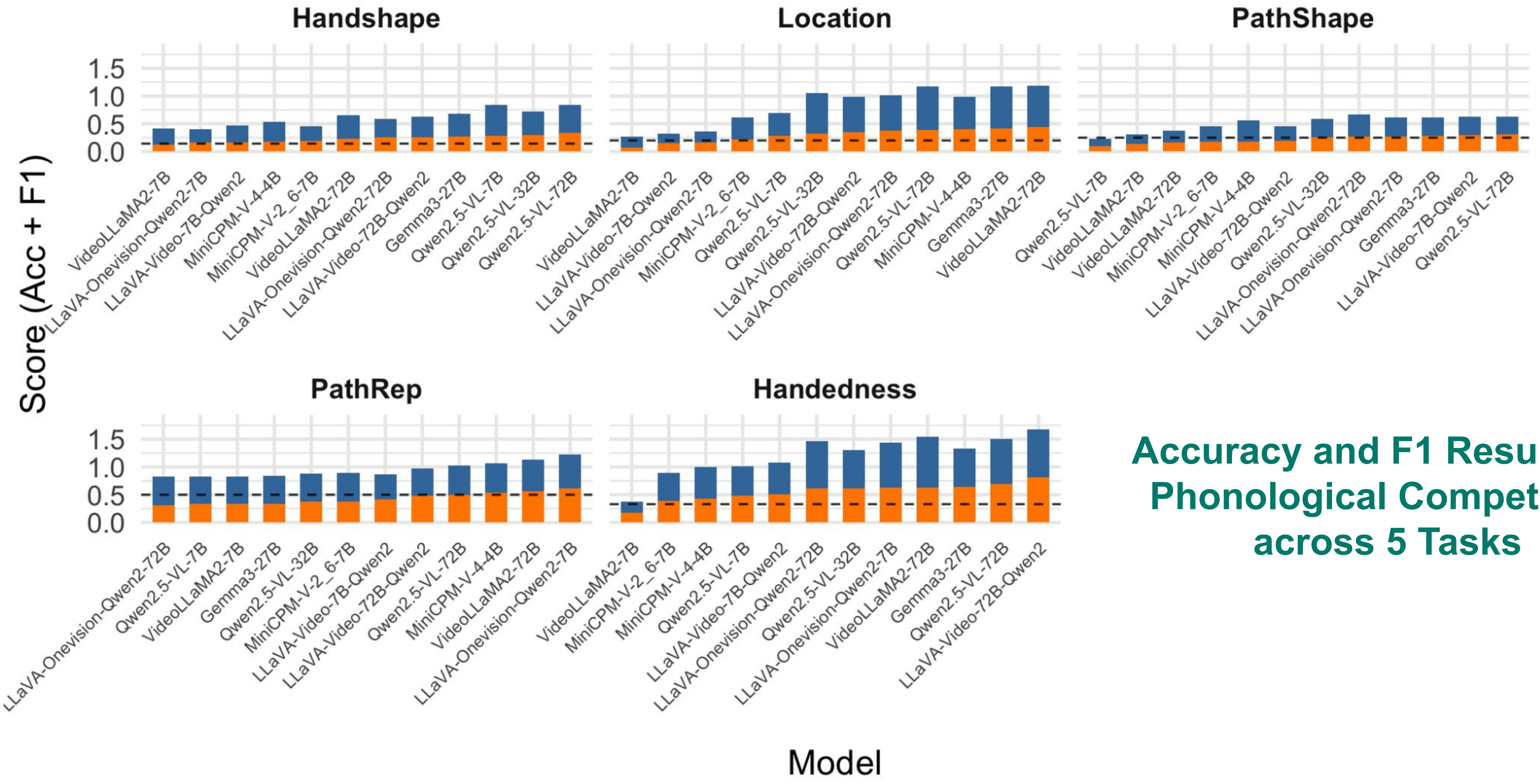
- **Smarter Prompts (Few-shot & Chain-of-thought):** Give models examples and try with step-by-step thinking.
- **Improving Models (Instruction-tuning, Fusion with pose/motion encoders):** Train them further with mixed text-and-video tasks, and add extra input from body and hand movements.
- **Taking Away Clues (Ablation Studies):** Blur or remove parts of the sign (handshape, location, movement) to see which features matter most compared to humans.
- **Clear Descriptions (Mid-fidelity gesture descriptors):** Provide short, simple written descriptions of gestures (e.g., “a fist moves up and down near the head”) as a bridge between video and meaning.

THANK YOU!

Any questions or comments?

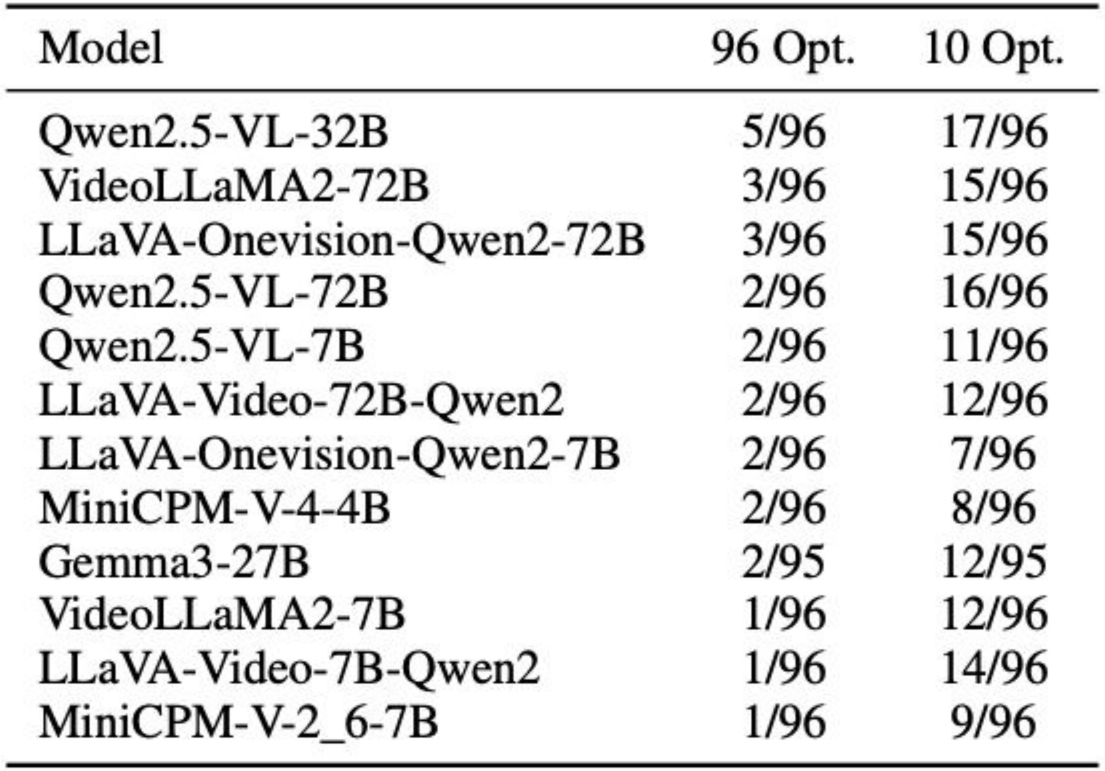
Appendix

(Taub, 2001; Emmorey, 2014)



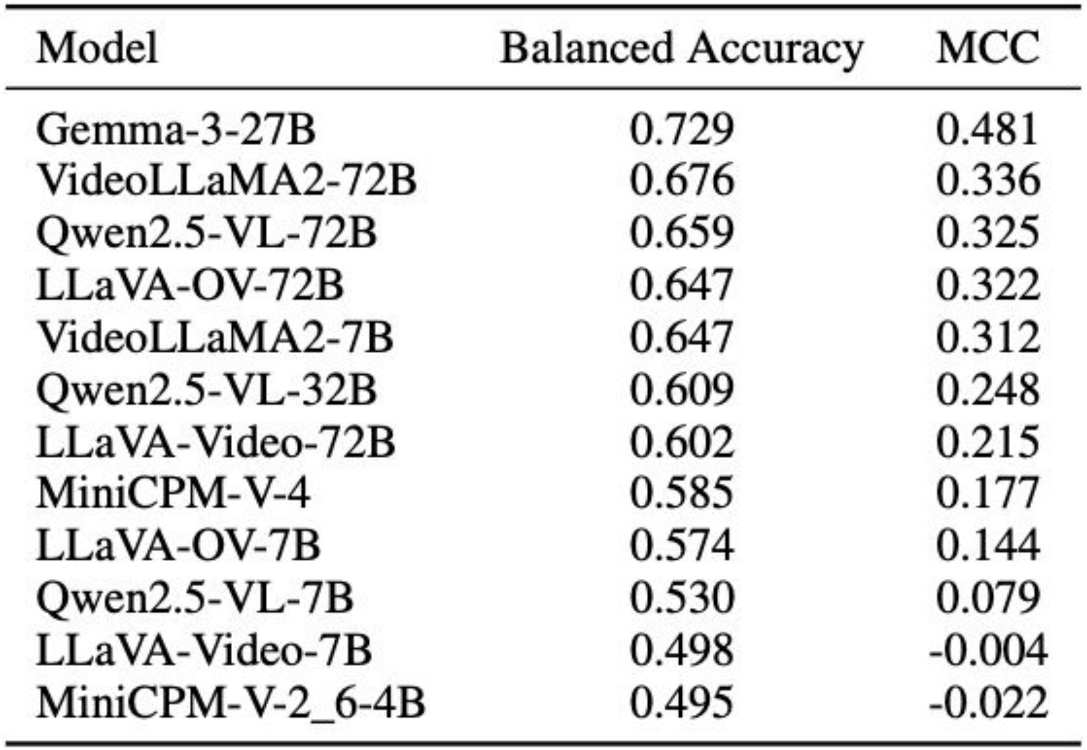
Accuracy and F1 Results for Phonological Competence across 5 Tasks

Transparency Results (96 vs. 10 Options). Number of correctly guessed words



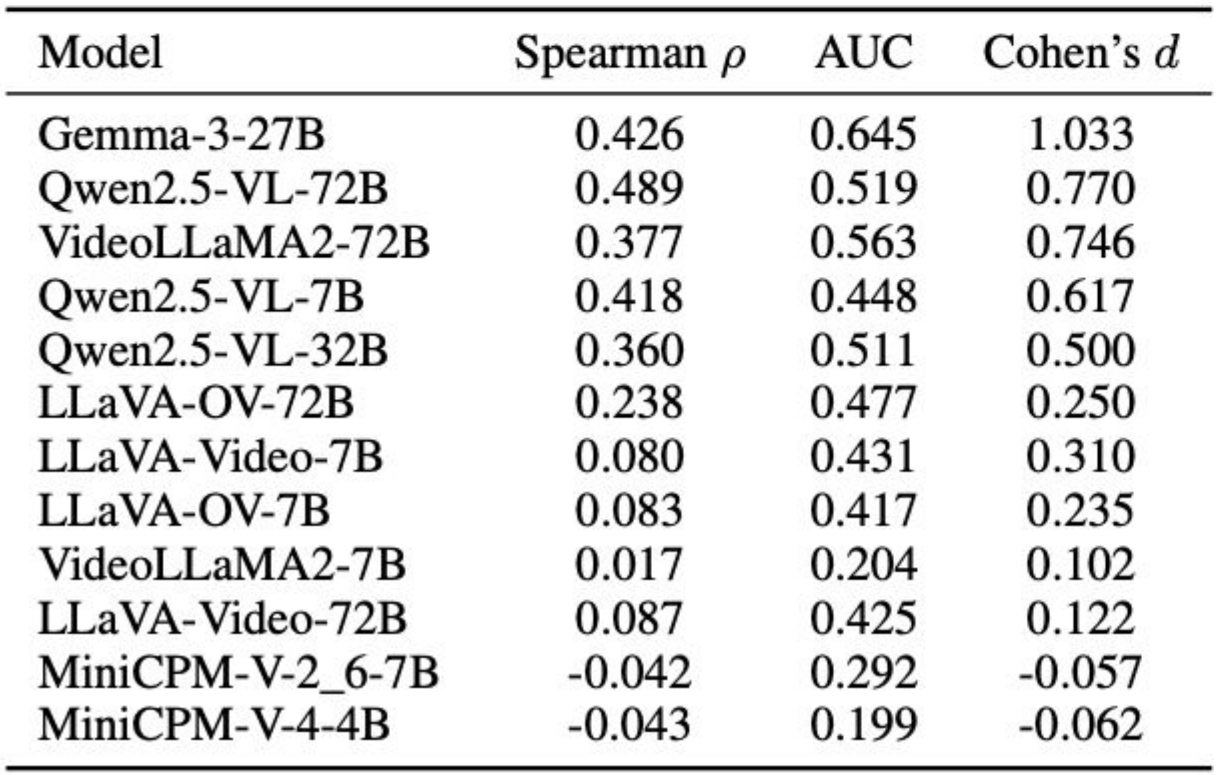
Model	96 Opt.	10 Opt.
Qwen2.5-VL-32B	5/96	17/96
VideoLLaMA2-72B	3/96	15/96
LLaVA-Onevision-Qwen2-72B	3/96	15/96
Qwen2.5-VL-72B	2/96	16/96
Qwen2.5-VL-7B	2/96	11/96
LLaVA-Video-72B-Qwen2	2/96	12/96
LLaVA-Onevision-Qwen2-7B	2/96	7/96
MiniCPM-V-4-4B	2/96	8/96
Gemma3-27B	2/95	12/95
VideoLLaMA2-7B	1/96	12/96
LLaVA-Video-7B-Qwen2	1/96	14/96
MiniCPM-V-2_6-7B	1/96	9/96

Binary iconicity classification performance. Balanced Accuracy averages sensitivity and specificity across iconic and arbitrary classes; Matthews Correlation Coefficient (MCC) provides a correlation-based measure accounting for all confusion matrix elements (range: -1 to +1).



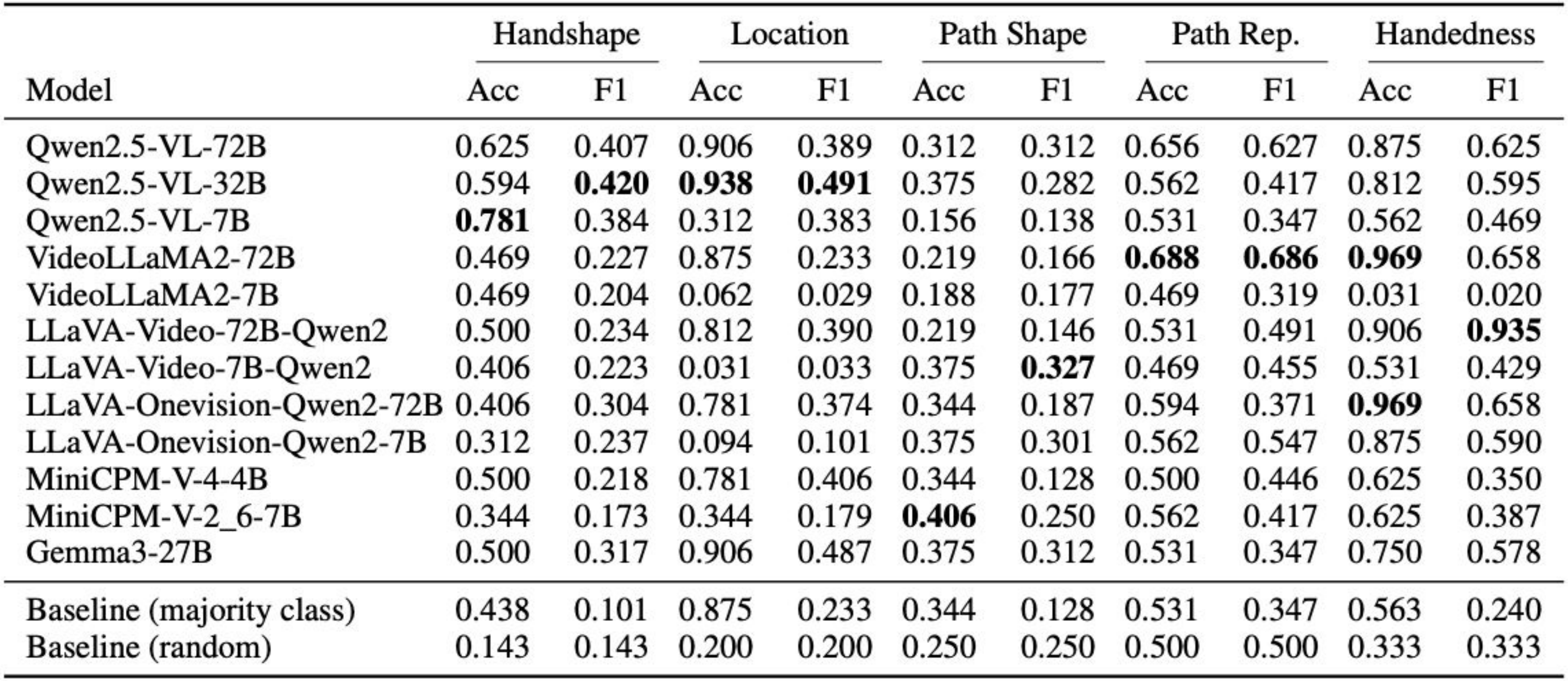
Model	Balanced Accuracy	MCC
Gemma-3-27B	0.729	0.481
VideoLLaMA2-72B	0.676	0.336
Qwen2.5-VL-72B	0.659	0.325
LLaVA-OV-72B	0.647	0.322
VideoLLaMA2-7B	0.647	0.312
Qwen2.5-VL-32B	0.609	0.248
LLaVA-Video-72B	0.602	0.215
MiniCPM-V-4	0.585	0.177
LLaVA-OV-7B	0.574	0.144
Qwen2.5-VL-7B	0.530	0.079
LLaVA-Video-7B	0.498	-0.004
MiniCPM-V-2_6-4B	0.495	-0.022

Graded iconicity rating performance. Spearman ρ measures rank correlation with human ratings; AUC evaluates binary iconic vs. arbitrary discrimination; Cohen’s d quantifies effect size between iconic and arbitrary rating distributions



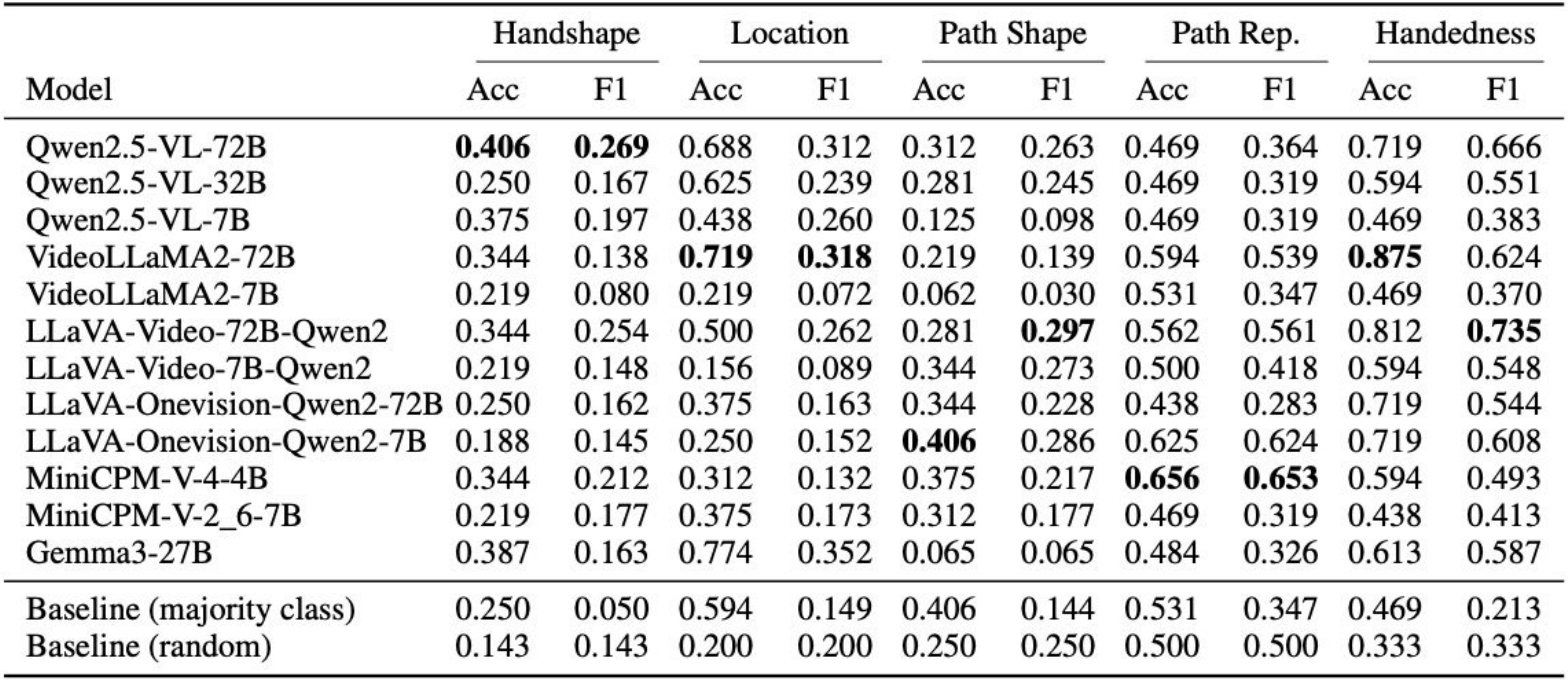
Model	Spearman ρ	AUC	Cohen’s d
Gemma-3-27B	0.426	0.645	1.033
Qwen2.5-VL-72B	0.489	0.519	0.770
VideoLLaMA2-72B	0.377	0.563	0.746
Qwen2.5-VL-7B	0.418	0.448	0.617
Qwen2.5-VL-32B	0.360	0.511	0.500
LLaVA-OV-72B	0.238	0.477	0.250
LLaVA-Video-7B	0.080	0.431	0.310
LLaVA-OV-7B	0.083	0.417	0.235
VideoLLaMA2-7B	0.017	0.204	0.102
LLaVA-Video-72B	0.087	0.425	0.122
MiniCPM-V-2_6-7B	-0.042	0.292	-0.057
MiniCPM-V-4-4B	-0.043	0.199	-0.062

Phonological Competence Results for Iconic (high overlap) Signs (n = 32)



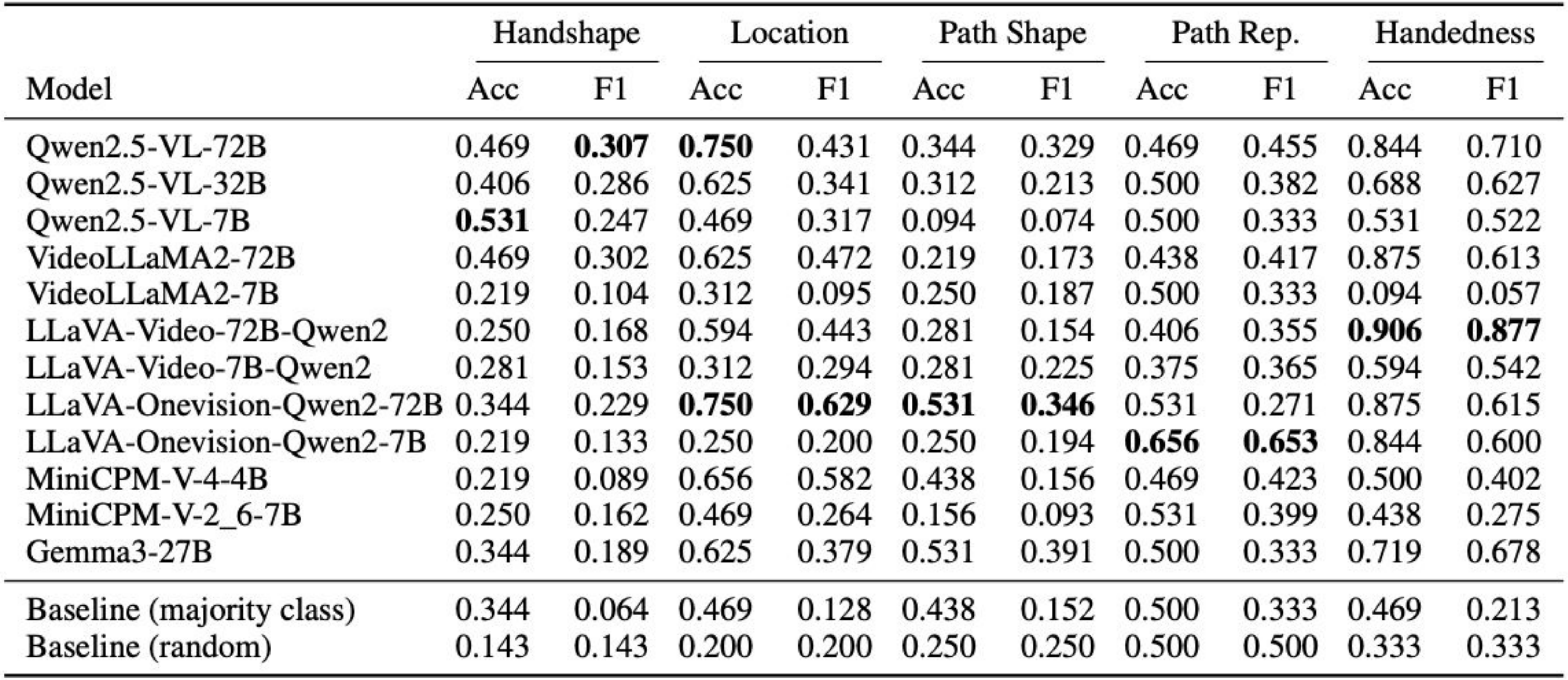
	Handshape		Location		Path Shape		Path Rep.		Handedness	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-VL-72B	0.625	0.407	0.906	0.389	0.312	0.312	0.656	0.627	0.875	0.625
Qwen2.5-VL-32B	0.594	**0.420**	**0.938**	**0.491**	0.375	0.282	0.562	0.417	0.812	0.595
Qwen2.5-VL-7B	**0.781**	0.384	0.312	0.383	0.156	0.138	0.531	0.347	0.562	0.469
VideoLLaMA2-72B	0.469	0.227	0.875	0.233	0.219	0.166	**0.688**	**0.686**	**0.969**	0.658
VideoLLaMA2-7B	0.469	0.204	0.062	0.029	0.188	0.177	0.469	0.319	0.031	0.020
LLaVA-Video-72B-Qwen2	0.500	0.234	0.812	0.390	0.219	0.146	0.531	0.491	0.906	**0.935**
LLaVA-Video-7B-Qwen2	0.406	0.223	0.031	0.033	0.375	**0.327**	0.469	0.455	0.531	0.429
LLaVA-Onevision-Qwen2-72B	0.406	0.304	0.781	0.374	0.344	0.187	0.594	0.371	**0.969**	0.658
LLaVA-Onevision-Qwen2-7B	0.312	0.237	0.094	0.101	0.375	0.301	0.562	0.547	0.875	0.590
MiniCPM-V-4-4B	0.500	0.218	0.781	0.406	0.344	0.128	0.500	0.446	0.625	0.350
MiniCPM-V-2_6-7B	0.344	0.173	0.344	0.179	**0.406**	0.250	0.562	0.417	0.625	0.387
Gemma3-27B	0.500	0.317	0.906	0.487	0.375	0.312	0.531	0.347	0.750	0.578
Baseline (majority class)	0.438	0.101	0.875	0.233	0.344	0.128	0.531	0.347	0.563	0.240
Baseline (random)	0.143	0.143	0.200	0.200	0.250	0.250	0.500	0.500	0.333	0.333

Phonological Competence Results for Iconic (low overlap) Signs (n = 32)



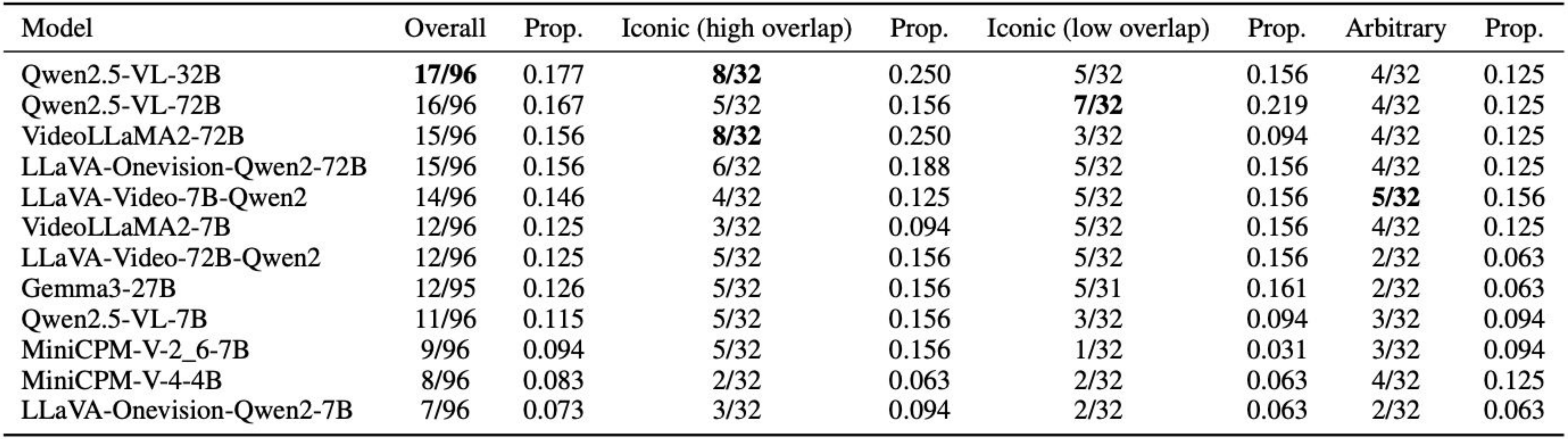
	Handshape		Location		Path Shape		Path Rep.		Handedness	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-VL-72B	**0.406**	**0.269**	0.688	0.312	0.312	0.263	0.469	0.364	0.719	0.666
Qwen2.5-VL-32B	0.250	0.167	0.625	0.239	0.281	0.245	0.469	0.319	0.594	0.551
Qwen2.5-VL-7B	0.375	0.197	0.438	0.260	0.125	0.098	0.469	0.319	0.469	0.383
VideoLLaMA2-72B	0.344	0.138	**0.719**	**0.318**	0.219	0.139	0.594	0.539	**0.875**	0.624
VideoLLaMA2-7B	0.219	0.080	0.219	0.072	0.062	0.030	0.531	0.347	0.469	0.370
LLaVA-Video-72B-Qwen2	0.344	0.254	0.500	0.262	0.281	**0.297**	0.562	0.561	0.812	**0.735**
LLaVA-Video-7B-Qwen2	0.219	0.148	0.156	0.089	0.344	0.273	0.500	0.418	0.594	0.548
LLaVA-Onevision-Qwen2-72B	0.250	0.162	0.375	0.163	0.344	0.228	0.438	0.283	0.719	0.544
LLaVA-Onevision-Qwen2-7B	0.188	0.145	0.250	0.152	**0.406**	0.286	0.625	0.624	0.719	0.608
MiniCPM-V-4-4B	0.344	0.212	0.312	0.132	0.375	0.217	**0.656**	**0.653**	0.594	0.493
MiniCPM-V-2_6-7B	0.219	0.177	0.375	0.173	0.312	0.177	0.469	0.319	0.438	0.413
Gemma3-27B	0.387	0.163	0.774	0.352	0.065	0.065	0.484	0.326	0.613	0.587
Baseline (majority class)	0.250	0.050	0.594	0.149	0.406	0.144	0.531	0.347	0.469	0.213
Baseline (random)	0.143	0.143	0.200	0.200	0.250	0.250	0.500	0.500	0.333	0.333

Phonological Competence Results for Arbitrary Signs (n = 32)



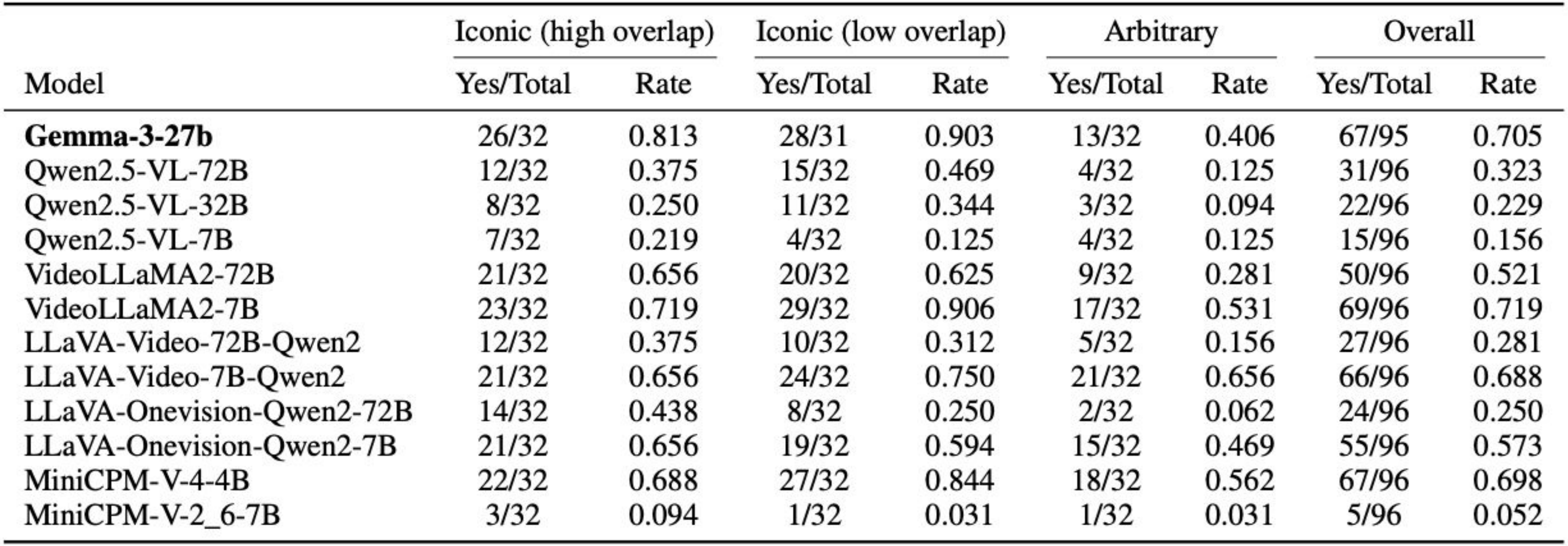
	Handshape		Location		Path Shape		Path Rep.		Handedness	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-VL-72B	0.469	**0.307**	**0.750**	0.431	0.344	0.329	0.469	0.455	0.844	0.710
Qwen2.5-VL-32B	0.406	0.286	0.625	0.341	0.312	0.213	0.500	0.382	0.688	0.627
Qwen2.5-VL-7B	**0.531**	0.247	0.469	0.317	0.094	0.074	0.500	0.333	0.531	0.522
VideoLLaMA2-72B	0.469	0.302	0.625	0.472	0.219	0.173	0.438	0.417	0.875	0.613
VideoLLaMA2-7B	0.219	0.104	0.312	0.095	0.250	0.187	0.500	0.333	0.094	0.057
LLaVA-Video-72B-Qwen2	0.250	0.168	0.594	0.443	0.281	0.154	0.406	0.355	**0.906**	**0.877**
LLaVA-Video-7B-Qwen2	0.281	0.153	0.312	0.294	0.281	0.225	0.375	0.365	0.594	0.542
LLaVA-Onevision-Qwen2-72B	0.344	0.229	**0.750**	**0.629**	**0.531**	**0.346**	0.531	0.271	0.875	0.615
LLaVA-Onevision-Qwen2-7B	0.219	0.133	0.250	0.200	0.250	0.194	**0.656**	**0.653**	0.844	0.600
MiniCPM-V-4-4B	0.219	0.089	0.656	0.582	0.438	0.156	0.469	0.423	0.500	0.402
MiniCPM-V-2_6-7B	0.250	0.162	0.469	0.264	0.156	0.093	0.531	0.399	0.438	0.275
Gemma3-27B	0.344	0.189	0.625	0.379	0.531	0.391	0.500	0.333	0.719	0.678
Baseline (majority class)	0.344	0.064	0.469	0.128	0.438	0.152	0.500	0.333	0.469	0.213
Baseline (random)	0.143	0.143	0.200	0.200	0.250	0.250	0.500	0.500	0.333	0.333

Transparency$_2$ Results (10 Options Per Trial)



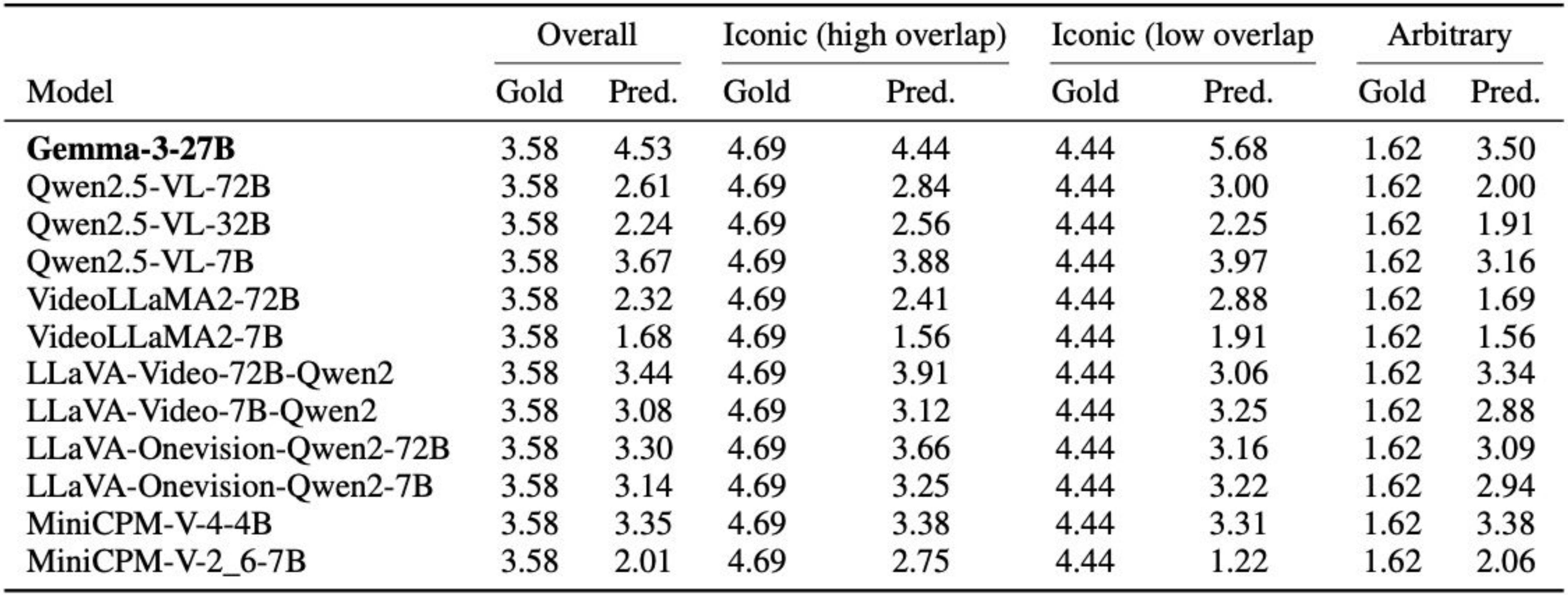
Model	Overall	Prop.	Iconic (high overlap)	Prop.	Iconic (low overlap)	Prop.	Arbitrary	Prop.
Qwen2.5-VL-32B	**17/96**	0.177	**8/32**	0.250	5/32	0.156	4/32	0.125
Qwen2.5-VL-72B	16/96	0.167	5/32	0.156	**7/32**	0.219	4/32	0.125
VideoLLaMA2-72B	15/96	0.156	**8/32**	0.250	3/32	0.094	4/32	0.125
LLaVA-Onevision-Qwen2-72B	15/96	0.156	6/32	0.188	5/32	0.156	4/32	0.125
LLaVA-Video-7B-Qwen2	14/96	0.146	4/32	0.125	5/32	0.156	**5/32**	0.156
VideoLLaMA2-7B	12/96	0.125	3/32	0.094	5/32	0.156	4/32	0.125
LLaVA-Video-72B-Qwen2	12/96	0.125	5/32	0.156	5/32	0.156	2/32	0.063
Gemma3-27B	12/95	0.126	5/32	0.156	5/31	0.161	2/32	0.063
Qwen2.5-VL-7B	11/96	0.115	5/32	0.156	3/32	0.094	3/32	0.094
MiniCPM-V-2_6-7B	9/96	0.094	5/32	0.156	1/32	0.031	3/32	0.094
MiniCPM-V-4-4B	8/96	0.083	2/32	0.063	2/32	0.063	4/32	0.125
LLaVA-Onevision-Qwen2-7B	7/96	0.073	3/32	0.094	2/32	0.063	2/32	0.063

Binary Iconicity Classification: "Yes" (Iconic) Response Rates by Sign Category



	Iconic (high overlap)		Iconic (low overlap)		Arbitrary		Overall	
Model	Yes/Total	Rate	Yes/Total	Rate	Yes/Total	Rate	Yes/Total	Rate
Gemma-3-27b	26/32	0.813	28/31	0.903	13/32	0.406	67/95	0.705
Qwen2.5-VL-72B	12/32	0.375	15/32	0.469	4/32	0.125	31/96	0.323
Qwen2.5-VL-32B	8/32	0.250	11/32	0.344	3/32	0.094	22/96	0.229
Qwen2.5-VL-7B	7/32	0.219	4/32	0.125	4/32	0.125	15/96	0.156
VideoLLaMA2-72B	21/32	0.656	20/32	0.625	9/32	0.281	50/96	0.521
VideoLLaMA2-7B	23/32	0.719	29/32	0.906	17/32	0.531	69/96	0.719
LLaVA-Video-72B-Qwen2	12/32	0.375	10/32	0.312	5/32	0.156	27/96	0.281
LLaVA-Video-7B-Qwen2	21/32	0.656	24/32	0.750	21/32	0.656	66/96	0.688
LLaVA-Onevision-Qwen2-72B	14/32	0.438	8/32	0.250	2/32	0.062	24/96	0.250
LLaVA-Onevision-Qwen2-7B	21/32	0.656	19/32	0.594	15/32	0.469	55/96	0.573
MiniCPM-V-4-4B	22/32	0.688	27/32	0.844	18/32	0.562	67/96	0.698
MiniCPM-V-2_6-7B	3/32	0.094	1/32	0.031	1/32	0.031	5/96	0.052

Mean Iconicity Ratings by Model and Sign Category



	Overall		Iconic (high overlap)		Iconic (low overlap		Arbitrary	
Model	Gold	Pred.	Gold	Pred.	Gold	Pred.	Gold	Pred.
Gemma-3-27B	3.58	4.53	4.69	4.44	4.44	5.68	1.62	3.50
Qwen2.5-VL-72B	3.58	2.61	4.69	2.84	4.44	3.00	1.62	2.00
Qwen2.5-VL-32B	3.58	2.24	4.69	2.56	4.44	2.25	1.62	1.91
Qwen2.5-VL-7B	3.58	3.67	4.69	3.88	4.44	3.97	1.62	3.16
VideoLLaMA2-72B	3.58	2.32	4.69	2.41	4.44	2.88	1.62	1.69
VideoLLaMA2-7B	3.58	1.68	4.69	1.56	4.44	1.91	1.62	1.56
LLaVA-Video-72B-Qwen2	3.58	3.44	4.69	3.91	4.44	3.06	1.62	3.34
LLaVA-Video-7B-Qwen2	3.58	3.08	4.69	3.12	4.44	3.25	1.62	2.88
LLaVA-Onevision-Qwen2-72B	3.58	3.30	4.69	3.66	4.44	3.16	1.62	3.09
LLaVA-Onevision-Qwen2-7B	3.58	3.14	4.69	3.25	4.44	3.22	1.62	2.94
MiniCPM-V-4-4B	3.58	3.35	4.69	3.38	4.44	3.31	1.62	3.38
MiniCPM-V-2_6-7B	3.58	2.01	4.69	2.75	4.44	1.22	1.62	2.06