Similar Choices, Different Attention: Cross-Modal Associations in Humans and Vision–Language Models

Anonymous authors
Paper under double-blind review

Abstract

Cross-modal associations are systematic pairings of features across modalities, such as the association of 'bouba' with round shapes and 'kiki' with sharp shapes. Prior work has compared humans and vision–language models (VLMs) on such associations, but often using different stimuli or tasks between humans and models. Here, we ask whether VLMs align with humans not only in choices, but also in where they look when making those choices. We study both VLMs and humans ($N = 53$), presenting them with the same stimuli, a pseudo-word and two images, and record participants' choices and eye movements, which we release[1]. We find choice alignment in a few larger VLMs, but their saliency matches human gaze less closely than a center-bias baseline, a fixed Gaussian at the center of each image. Fine-tuning small VLMs on human choices brings their choice alignment to the level of a human majority-vote reference on unseen words and images, yet their attention still matches human gaze less closely than this baseline. Training model attention on human gaze raises attention–gaze correlation without improving choice alignment, and a single average gaze map per image position raises it by a similar amount. Matching human choices, or even human gaze patterns, is therefore not sufficient evidence of human-aligned cross-modal processing.

1 Introduction

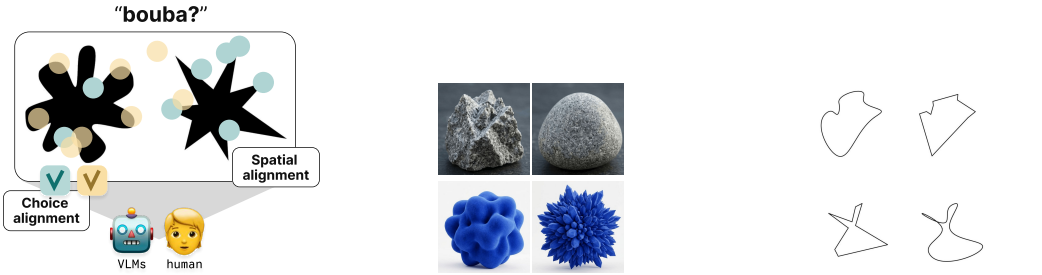
Humans systematically map information across sensory modalities. A common example is the bouba–kiki effect: When asked to match the pseudo-words *bouba* and *kiki* to rounded and spiky shapes, participants overwhelmingly pair *bouba* with the rounded shape and *kiki* with the spiky shape (Ramachandran & Hubbard, 2001). This sound–shape correspondence has been replicated across languages and writing systems, and appears to reflect cross-modal correspondences between speech sounds and visual properties (Ćwiek et al., 2022). It emerges even in pre-literate children (Maurer et al., 2006) and infants (Ozturk et al., 2013), making it a standard testbed for studying human cross-modal association.

The rise of vision–language models (VLMs) raises a natural question: Do models trained on visual and linguistic data develop cross-modal associations comparable to those of humans? The bouba–kiki effect is a useful test case because it challenges the assumption that how a word sounds is arbitrary with respect to what it refers to. If VLMs align with humans on novel pseudo-words, this would suggest that their representations capture not only lexical semantics, but also non-arbitrary associations between word form and perceived meaning. Aligning with humans in their choices, however, is not enough to answer this question. The similar choices can rest on different attention: a model may select the round image because of its shape, the property the association concerns, or because of other properties that co-vary with roundness in the stimuli. Choice agreement alone cannot tell these apart. We therefore also ask where models attend, by comparing their saliency with human gaze on the same trials.

Prior work has reached conflicting conclusions. Alper & Averbuch-Elor (2023) evaluated CLIP (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022) and claimed to find cross-modal association in their representations. Kouwenhoven et al. (2025), in contrast, reported that such associations were absent from CLIP models. Critically, however, prior studies often stop short of a

[1] Dataset: https://anonymous.4open.science/r/Sound-Symbolism-Dataset/README.md

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Overview (b) Gemini-generated pairs (c) Line-based minimal pairs

Figure 1: (a) Humans and VLMs receive the same stimulus, a pseudo-word (*bouba*) and a round–sharp image pair; we compare which shape they choose (checks) and which regions they attend to (circles). Here, both choose the round shape but attend to different regions. (b, c) Example image pairs: abstract pairs generated with Gemini (Gemini Team et al., 2023) and line-based minimal pairs following Nielsen & Rendall (2013).

direct, matched comparison: they evaluate humans and VLMs on different stimuli or different tasks, or analyze them at different levels.

We address this gap by comparing humans and VLMs on identical stimuli (Fig. 1a), using choices and eye-tracking from 53 participants (8,208 trials), which we release. We consider three questions:

- **Do VLMs choose as humans do?** *Choice alignment* asks whether, on trials matched to the human study, a model selects the same image. We find that it does mainly in a subset of larger VLMs; smaller models match adjectives but are near chance and side-locked on pseudo-words.
- **Do they attend where humans look?** *Spatial alignment* compares VLM saliency (attention maps or Grad-CAM (Selvaraju et al., 2017)) with the same participant's gaze on the same trial, using gaze as a reference for where people look when choosing (Pani & Yang, 2025; Lai et al., 2020; Das et al., 2017). We find that a simple center-bias baseline, a fixed Gaussian at the center of each image that uses no model and no information about the word or trial, aligns with human gaze better than the saliency of even the models whose choices track human judgments.
- **Are the two levels linked?** We fine-tune three small VLMs on human choices and ask whether, once their choices are human-aligned, their attention becomes more gaze-aligned, and conversely whether, once attention is more gaze-aligned, their choices become more human-aligned. Neither holds. Training on ~2k choices reaches human-level choice agreement that transfers to unseen words and images, but the center-bias baseline still aligns with human gaze better than the models' attention does. Adding human gaze as a training target for attention brings attention closer to human gaze but has no detectable improvement on agreement with human choices.

2 Related work

2.1 Cross-modal associations in humans

Word–meaning mappings are largely arbitrary (Hockett, 1960), but *sound symbolism* is a well-documented exception (Köhler, 1929; 1947): most people match *bouba* to a round shape and *kiki* to a sharp one (Ramachandran & Hubbard, 2001; Maurer et al., 2006; Ćwiek et al., 2022), and specific phonological properties correlate with roundness and sharpness (Nielsen & Rendall, 2013; McCormick et al., 2015). Related associations exist for magnitude (Sapir, 1929; Shinohara & Kawahara, 2010; Haynie et al., 2014; Winter & Perlman, 2021) and texture (Greenberg & Jenkins, 1966; Sakamoto & Watanabe, 2018; Kumagai, 2020; Sidhu et al., 2022; Winter et al., 2022).

2.2 Cross-modal associations in VLMs

Prior work on VLMs has focused on round–sharp symbolism, with mixed conclusions. Alper & Averbuch-Elor (2023) defined a round–sharp direction in CLIP (Radford et al., 2021) embedding space from adjective embeddings and found that pseudo-words and images of the same category

(round/sharp) project to the same side of this axis; Iida & Funakura (2024) could not replicate this in Japanese VLMs, suggesting dependence on the English-based setup, including the adjective sets. Verhoef et al. (2024) and Kouwenhoven et al. (2025) instead compared pseudo-word probabilities per image, the latter adding multiple prompts, an adjective baseline, and Grad-CAM (Selvaraju et al., 2017) analysis, and found no consistent evidence of the association in CLIP models despite successful adjective matching. Beyond CLIP, studies of multimodal LLMs (Loakman et al., 2024; Jeong et al., 2026) and audio–visual models (Tseng et al., 2024) extend the models and semantic dimensions tested, with mixed findings. Across these studies, humans and models are typically evaluated on different stimuli or tasks, and compared only at the output level; whether VLMs match humans under matched conditions, and whether they attend to the same image regions, remains open.

3 Dataset and human study

We construct a set of pseudo-words paired with abstract round–sharp image pairs and evaluate humans and VLMs under the same forced-choice paradigm, recording each participant's choice and gaze on every trial. Gaze data serves two purposes: it extends the comparison beyond choices to the allocation of attention (Section 4), and provides a per-trial supervision signal (Section 5). Throughout, labels and images have an *intended* round or sharp category, based on tendencies in prior human studies rather than objective ground truth.

3.1 Stimuli

Linguistic stimuli. Linguistic inputs come from three sources: the canonical pairs *takete–maluma* and *bouba–kiki* (Köhler, 1929; Ramachandran & Hubbard, 2001); the 20 English round–sharp adjectives of Alper & Averbuch-Elor (2023) as a semantic baseline (Appendix A); and 162 constructed two-syllable pseudo-words. Following prior human studies, in which sonorant consonants and rounded vowels are associated with roundness, and unvoiced plosive consonants and unrounded vowels with sharpness (Nielsen & Rendall, 2013; McCormick et al., 2015), we use $V_{\text{sharp}} = \{\text{EE}, \text{AY}, \text{UH}\}$, $V_{\text{round}} = \{\text{OO}, \text{OH}, \text{AH}\}$, $C_{\text{sharp}} = \{\text{T}, \text{K}, \text{P}\}$, and $C_{\text{round}} = \{\text{M}, \text{N}, \text{L}\}$. These patterns have also informed the stimuli used in prior VLM evaluations (Alper & Averbuch-Elor, 2023; Kouwenhoven et al., 2025; Verhoef et al., 2024). We combine vowels and consonants only within the same intended category, yielding 18 one-syllable words (9 round and 9 sharp); combining them into two-syllable words yields 162 pseudo-words (e.g., KEEKEE, MOHMAH, PUHTAY), for 186 labels in total. All labels are presented in English orthography to standardize intended pronunciation, following the convention established in Nielsen & Rendall (2013). Participants read each label aloud before choosing (Section 3.2), whereas models receive only the written form, so we treat the correspondence as one between written word form and shape, and we do not separate orthographic from phonological contributions, which are known to be intertwined in literate participants (Cuskley et al., 2017).

Visual stimuli. We use 184 round–sharp image pairs: 17 from prior work (Köhler, 1929; 1947; Maurer et al., 2006; Ramachandran & Hubbard, 2001; Westbury, 2005; Kouwenhoven et al., 2025) (Appendix D); 75 line-based minimal pairs generated by connecting random points with curved or straight lines (Nielsen & Rendall, 2013), also used in Verhoef et al. (2024) and Kouwenhoven et al. (2025); and 92 more realistic abstract pairs generated with Gemini (Gemini Team et al., 2023) (prompt in Appendix B), which better resemble images VLMs encounter in training (Fig. 1b,c). Every pair is prepared in both left–right orientations, and label–pair assignments and orientations are randomized per participant (Appendix E), so no label is tied to an image pair or side, except for the canonical labels which keep their original images.

3.2 Human study procedure

Following the human study structure of Ćwiek et al. (2022), participants saw a written label followed by a pair of images, read the label aloud, and verbally selected the better-matching image by saying *A* (left) or *B* (right); no audio was played. Gaze data was recorded with a Tobii Pro Fusion monitor-mounted eye-tracker, and screen/audio recordings are used for analysis. The study had three sessions, each preceded by eye-tracker calibration; after each session, participants reported pseudo-words that reminded them of real words or names, and those items were excluded (Appendix E). We recruited 55 participants, and after retaining trials with a complete image presentation, a valid choice, and at least

two in-image fixations with the gaze recording covering ≥80% of the time the images were displayed, the cleaned dataset contains 8,208 trials from 53 participants (7,141 constructed pseudo-word, 910 adjective, 157 canonical); cleaning changed agreement rates by less than one point (Appendix E.5).

Human choices. *Intended-category agreement (ICA)* is the rate at which a response matches the label's intended round–sharp category, which is based on prior human studies (e.g., LOONOO with the round image or KEETUH with the sharp image). ICA is highest for English adjectives (94.84%), while canonical pseudo-words reach 84.71%, replicating the standard bouba–kiki effect. Constructed pseudo-words reach 73.98% agreement, substantially above chance, indicating reliable sound–shape associations in our pseudo-word set. Roughly one response in four departs from the intended category, with stronger consensus for round-associated phonemes (Fig. 9). In pseudo-words, word-level round-choice rates are reproducible across participants (mean split-half Pearson $r = .892$; Spearman–Brown-corrected reliability = .943 over 500 random within-word splits), so we compare models both with the intended category and with the actual response of the participant on the same trial.

Gaze–choice coupling. The image a participant looked at more (by gaze samples or summed fixation duration) is the one they chose on 77–80% of trials (80.58% for constructed pseudo-words, by gaze samples), whereas the first fixation agrees at chance (51%; Appendix E.4): gaze accumulates on the eventual choice (Shimojo et al., 2003; Krajbich et al., 2010). This association, not a causal relation, is what makes per-trial gaze informative as a supervision signal in Section 5.

3.3 Splits and behavioral references

Fixed splits. Fine-tuning and evaluation use only the 7,141 constructed pseudo-word trials. Adjectives are in-vocabulary words whose shape associations are encoded in their lexical meaning, and the canonical pairs are discussed throughout the pretraining corpora of current VLMs, so neither can test a sound–shape association on novel forms. We partition the pseudo-word trials by word and by image pair into a training set (Train; 1,932 trials), a validation set (Val; 721 trials), and a held-out test set (Test; 1,030 trials, 62 words, 70 image pairs, 53 participants) in which both the word and the image pair are unseen during training. Participants are shared across splits, and participant identity is never a model input; the evaluation therefore measures generalization to new stimuli, not to new people.

Human references. To provide a reference point for the fine-tuning results, we report two predictors on Test; these values serve as references, not ceilings. Always choosing the intended category agrees with participants' actual choices on 72.82% of Test trials, and choosing each word's majority response among the other participants (leave-one-out) agrees on 72.86%; the best attainable agreement cannot be estimated, because 915 of the 972 word×pair×orientation conditions in Test were seen by only one participant. For within-image alignment, a *center-bias baseline* places an isotropic Gaussian ($\sigma = 0.125$ in normalized composite-image coordinates) at the center of each image, and a *population-mean map* averages the gaze of the other Test trials separately for the left and right image positions, excluding the evaluated trial. The first asks whether saliency beats a generic center-bias; the second asks how much alignment a shared gaze pattern reaches without trial-specific information. Both are model-free maps that we integrate over each model's spatial cells and compare with the participant's gaze exactly as model saliency is compared (Section 4.1), so their values depend on the model's grid. The center-bias baseline is also computed on all pseudo-word trials for the zero-shot comparison (Table 2); the population-mean map is used on Test only (Section 5).

4 Cross-modal associations in pretrained VLMs

4.1 Evaluation protocol

We evaluate six embedding models (CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and SigLIP2 (Tschannen et al., 2025)) and six generative VLMs (Qwen3-VL (Bai et al., 2025), Gemma4 (Gemma Team et al., 2026), and Molmo2 (Clark et al., 2026)) on the human forced-choice task. Embedding models choose the image with the higher image–text cosine similarity, averaged over ten fixed caption templates (Appendix F). Generative models receive the original forced-choice prompt (Appendix C), with the composite image panel resized to 512×512 before each model's standard image preprocessing (Appendix H.2). Their choices and attention come from one eager forward

pass per input, with A/B probabilities from an FP32 projection of the final hidden state (Gemma's output softcap retained) and exact ties counted as half. The original prompt names the label type and asks for "visual/sound-symbolic fit," making it a favorable test of whether the model can express the association. Because the cue could inflate agreement, we rerun with a neutral prompt that mirrors the participants' instruction ("Choose the image that better matches this word") only for Gemma4-26B and Qwen3-VL-30B, the two models whose choices track human judgments; the smaller models are near chance and side-locked even with the cue.

For generative VLMs, we read out decision-token attention, attention from the position that predicts the A/B answer to the image tokens, averaged over heads and over the last half of full-attention decoder layers (Liu et al., 2026) (Gemma: global-attention layers only; Molmo: global and high-resolution tokens). For embedding models, we follow Kouwenhoven et al. (2025) and compute Grad-CAM maps (Selvaraju et al., 2017), using the fixed caption "The label for this image is {word}" on the 512×512 composed image input, keeping each model's native spatial cells without upsampling. Note that this map explains the label's similarity to the whole panel, not the choice between the two images. We then compare each model's map with the participant's gaze from the same trial. We sum the participant's smoothed gaze within each of the model's cells and compute Spearman ρ between gaze and model saliency separately for image A and image B; we then average across the two images and all trials. Because ρ is computed within each image, it does not depend on how attention is split between the two images; allocation between images is analyzed separately (Section 5). As a reference, we apply the same computation to the center-bias baseline of Section 3.3; because Grad-CAM and attention differ in resolution and smoothness, we compare each model with its own center reference rather than ranking models by raw ρ. Fine-tuning is evaluated with the same computation.

4.2 Zero-shot results

Choice alignment. Table 1 shows that embedding models match adjectives well but generalize little to pseudo-words. CLIP RN50 and ViT-B/32 reach ICA of 0.75 and 0.76 on adjectives, suggesting that CLIP representations encode lexical semantics, but pseudo-word ICA is near chance for most models. Only SigLIP-so400m shows a clear signal (ICA = 0.57, Maj. = 0.72, Spearman = 0.56).

Most generative VLMs are near ceiling on adjectives, but pseudo-word alignment varies with size. Gemma4-26B achieves the highest pseudo-word ICA (0.62), followed by Qwen3-VL-30B (0.57). Human distribution metrics show that both models partially capture human judgments on pseudo-words, rather than merely achieving above-chance ICA: majority match 0.73 and 0.68, Pearson 0.58 and 0.57, and Spearman 0.58 and 0.56, respectively. Molmo2-8B shows weaker positive signals, while Gemma4-E4B, Qwen3-VL-8B, and Molmo2-4B models show limited correlation with human judgments despite high adjective ICA. Their zero-shot answers are largely side-locked (on Test, Qwen3-VL-8B answers B on 96% of trials; Appendix L), so chance-level ICA in the small models reflects a positional default rather than a label-dependent choice.

Instruction robustness. Removing the sound-symbolic framing does not remove the association. Correlation with human pseudo-word round-choice rates stays positive (Gemma4-26B: $r = .580 \rightarrow .534$; Qwen3-VL-30B: $.571 \rightarrow .535$; 95% intervals above zero; paired changes include zero), and agreement with the participant's choice on the same trial does not decrease (57.30→57.74%; 55.73→57.72%, a gain of +1.99 percentage points (pp) [0.76, 3.22] for Qwen3-VL-30B). Individual decisions do depend on the instruction: 33.23% and 20.55% of inputs receive a different choice, Gemma4-26B's ICA falls by 3.99 pp [−5.29, −2.74], and majority match decreases (0.73→0.70; 0.68→0.57; unadjusted intervals, Appendix G). The graded, human-like pattern is therefore not an artifact of the prompt, although individual decisions depend on the instruction.

Spatial alignment. Table 2 compares each model's spatial ρ with its center-bias baseline (Section 3.3), a fixed Gaussian at each image center scored against the same participant's gaze on the same grid. A model with lower ρ than this baseline predicts gaze worse than a map that ignores the input. Every evaluated model falls below its center-bias baseline. Among embedding models, pseudo-word Grad-CAM–gaze correlations range from −.091 to .588; even the highest, CLIP ViT-B/32 ($\rho = .588$), is below its center-bias baseline (.743). Among generative models, Gemma4-E4B and Gemma4-26B partly track gaze ($\rho = .340$ and .265; center = .637), whereas Qwen3-VL-8B, Qwen3-VL-30B (.022, −.003; center .637), Molmo2-4B, and Molmo2-8B (−.074, .004; center = .601) are near zero or

Table 1: VLM–human behavioral alignment on the 8,208 cleaned trials. Maj. = majority-choice match per label; r, ρ = Pearson and Spearman correlations between human and model label-level round-choice rates; ICA = intended-category agreement. Embedding rows pool ten captions per trial; generative rows use the joint forward pass of Section 4.1; neutral-prompt rows, Section 4.2. An asterisk (*) means that nominal, unadjusted label-level $p < .05$; 95% participant-cluster intervals are in Appendix G.

	Overall				Pseudo-word				Adjective			
Model	Maj.	r	ρ	ICA	Maj.	r	ρ	ICA	Maj.	r	ρ	ICA
CLIP RN50	0.62	0.45*	0.42*	0.55	0.58	0.29*	0.27*	0.52	0.95	0.88*	0.72*	0.75
CLIP ViT-B/32	0.60	0.38*	0.34*	0.53	0.55	0.12	0.13	0.51	**1.00**	0.92*	0.75*	0.76
SigLIP-base	0.61	0.38*	0.37*	0.55	0.57	0.21*	0.20*	0.52	0.95	0.92*	0.85*	0.76
SigLIP-so400m	0.73	0.53*	0.60*	0.57	0.72	0.56*	0.56*	0.57	0.90	0.82*	0.69*	0.66
SigLIP2-base	0.50	0.32*	0.16*	0.53	0.44	−0.10	−0.12	0.50	0.95	0.92*	0.76*	0.79
SigLIP2-so400m	0.48	0.02	−0.04	0.50	0.50	0.01	−0.02	0.50	0.35	−0.02	−0.03	0.50
Gemma4-E4B	0.57	0.45*	0.33*	0.55	0.51	0.15	0.09	0.51	**1.00**	0.99*	0.85*	0.91
Gemma4-26B	**0.76**	0.70*	**0.70***	**0.66**	**0.73**	**0.58***	**0.58***	**0.62**	**1.00**	**0.99***	0.78*	**0.97**
Qwen3-VL-8B	0.61	0.53*	0.46*	0.57	0.56	0.26*	0.24*	0.52	**1.00**	0.99*	0.83*	0.94
Qwen3-VL-30B	0.72	**0.71***	0.69*	0.61	0.68	0.57*	0.56*	0.57	0.95	0.99*	**0.91***	0.94
Molmo2-4B	0.57	0.42*	0.35*	0.53	0.53	0.13	0.14	0.51	0.95	0.87*	0.82*	0.72
Molmo2-8B	0.66	0.62*	0.58*	0.58	0.61	0.41*	0.39*	0.53	**1.00**	0.99*	0.83*	0.94
Neutral prompt												
Gemma4-26B	0.73	0.68*	0.66*	0.63	0.70	0.53*	0.52*	0.58	0.95	0.98*	0.77*	0.95
Qwen3-VL-30B	0.62	0.67*	0.64*	0.61	0.57	0.54*	0.52*	0.57	0.95	0.98*	0.87*	0.95

Table 2: Zero-shot choice and spatial alignment on the 7,141 pseudo-word trials. Choice r = pseudo-word Pearson r (Table 1). Spatial ρ = within-image Spearman correlation between model saliency and the same participant's gaze on the model's native grid; Center = the center-bias baseline (Section 3.3) on the same grid, the reference for each model. For grids and denominators, see Appendix H.2.

	Embedding (Grad-CAM)						Generative (decision-token attention)					
	CLIP RN50	CLIP ViT-B/32	SigLIP base	SigLIP so400m	SigLIP2 base	SigLIP2 so400m	Gemma4 E4B	Gemma4 26B	Qwen3-VL 8B	Qwen3-VL 30B	Molmo2 4B	Molmo2 8B
Choice r	0.29	0.12	0.21	0.56	−0.10	0.01	0.15	0.58	0.26	0.57	0.13	0.41
Spatial ρ	0.31	0.59	−0.05	−0.09	−0.06	−0.07	0.34	0.27	0.02	0.00	−0.07	0.00
Center	0.74	0.74	0.64	0.61	0.64	0.60	0.64	0.64	0.64	0.64	0.60	0.60

slightly negative. Because the two families require different saliency methods, this comparison cannot distinguish between model-level and readout-level differences. Section 5 therefore holds the model and the readout fixed and changes only the training signal.

5 Fine-tuning on human choices and gaze

The zero-shot results leave the link between choices and attention unclear. The small models' side-locked answers may hide an association or reflect its absence, the large models' choices are partly human-aligned, and saliency is read out differently for each model family. Fine-tuning helps address those gaps: we ask (i) whether the association is learnable from a small set and generalizes to unseen words and images, (ii) whether human-aligned choices are then accompanied by attention closer to human gaze, and (iii) conversely, whether adding human gaze as a training target for attention makes choices more human-aligned. All conditions train LoRA adapters on the Train trials with the observed participant choice as the label and select checkpoints by Val choice agreement, using the prompt and 512×512 input of Appendix C. The decision-token attention readout serves both as the gaze-loss target and as the attention measure.

5.1 Training conditions

Choice-only trains on the participant's A/B choice alone. **+Gaze** adds, for each training trial, the within-image gaze map of the participant who supplied that trial's choice; this is the main comparison against Choice-only, since it changes exactly one thing—one added loss term. **+AvgGaze** adds the same loss with a single fixed target: one average gaze map per image position (left or right), estimated only from Train gaze and applied to every training trial. It is the control for +Gaze: whatever alignment +AvgGaze reaches is what a target that carries no trial-specific information can produce. **+Gaze+Side** additionally supervises how much attention falls on each of the two images through a jointly normalized gaze loss (Appendix K.1). It is a robustness condition rather than the main comparison: it changes the normalization and image weighting at once, has no matched constant-target control, and its side target is coupled to the choice label in 80% of trials (Table 4), so a choice gain under it would not isolate gaze information (Appendix K.1).[2]

5.2 Training objective

All conditions minimize the same per-trial loss: the full-vocabulary cross-entropy of the participant's answer token y_t, plus a gaze term with weight λ that is the only element varied across conditions,

$$\mathcal{L}_t = \mathrm{CE}(y_t) + \lambda \frac{\sum_{i\in\{A,B\}} m_{t,i}\,\mathrm{KL}\big[H_{t,i} \parallel A_{t,i}\big]}{\sum_i m_{t,i}}, \tag{1}$$

where $m_{t,i}$ marks a valid human gaze image $i \in \{A, B\}$ on trial t, $H_{t,i}$ is the normalized gaze target defined below, and $A_{t,i}$ is the decision-token attention readout normalized within that image after adding 10^{-6} to model token masses. Choice-only uses $\lambda = 0$; +Gaze, +AvgGaze, and +Gaze+Side use $\lambda = 1$, and we do not sweep λ. The gaze term is zero if no image is valid; the choice loss is retained, and the training pool has at least one valid image per trial. KL uses natural logarithms during training, with zero human bins contributing zero. Because each image is normalized separately, Eq. (1) does not supervise the relative total attention mass on A versus B. Gaze targets are sample-density maps (Gaussian $\sigma = 40$ screen pixels) integrated over each model's image-token regions; for +AvgGaze the target is replaced by the fixed Train average for the same image position (Appendix K.1).

5.3 Alignment measures

On Test, we report human-choice agreement and mean human-response NLL, $-\log_2 p(y_{\text{human}})$ averaged over trials (uniform prediction: 1 bit); within-image ρ and gaze KL as in Section 4.1, with KL in bits summarized as

$$\Delta\,\mathrm{KL} = \mathrm{KL}(H \parallel H_{\text{train-mean}}) - \mathrm{KL}(H \parallel A), \tag{2}$$

so that positive values indicate lower divergence than the fixed training-mean map; and allocation between images, via between-image JS and two side-agreement rates (choice–attention side agreement and gaze–attention side agreement) reported as gaps over the agreement expected from their marginal A/B rates. We compare within-image ρ with the center-bias baseline and population-mean map of Section 3.3, which are .637/.637/.603 and .731/.731/.680 on the token grids of Gemma4-E4B/Qwen3-VL-8B/Molmo2-8B. Intervals are paired word×pair×participant cluster bootstraps conditional on the fitted seed and checkpoints; Holm correction is applied separately to the 18 +Gaze–+AvgGaze comparisons and the nine choice-NLL comparisons (Appendix K.1).

5.4 Results

We first assess whether human-choice alignment generalizes to unseen words and image pairs, then examine how adding human gaze as a training target for attention changes choices, within-image attention, and allocation between images. Table 3 reports all five conditions on the same Test trials, using selected checkpoints from a single training seed; values below follow the order Gemma4-E4B/Qwen3-VL-8B/Molmo2-8B.

[2]Because the image receiving more gaze is usually the chosen one, +Gaze+Side partly re-supplies the choice label through the gaze channel. A choice gain under it would therefore not isolate gaze information beyond the label; the absence of such a gain is the more informative result, and Section 5.4 reports +Gaze+Side in that role.

Table 3: Matched Test evaluation on 1,030 trials with unseen words and image pairs. *Choice alignment.* ICA: intended category agreement; Human agr.: agreement with the actual participant choice; NLL: average negative log-likelihood of the participant's choice (lower is better, and assigning 50% to each image gives 1 bit). *Within-image alignment.* ρ: Spearman correlation between model attention and the participant's gaze inside each image; ΔKL: how much closer attention is to the participant's gaze than the average training gaze map is (Eq. (2)); positive means closer; *: +Gaze has a larger ΔKL than +AvgGaze after Holm correction ($p < .05$; Appendix K.3). *Allocation between images.* Side JS: divergence between the shares of attention and of gaze falling on each image; lower is closer to humans. Choice–attn: how often the model chooses the side it attends to more. Gaze–attn: how often the model attends more to the side the participant looked at more. Gap: observed agreement minus that expected if the two were independent, given how often each favors image B; a positive Gap indicates association, not that attention causes the choice. Zero-shot rows are the models before fine-tuning, with choices read out as for the fine-tuned models.

		Choice alignment			Within-image alignment		Allocation between images				
								Choice–attn		Gaze–attn	
Model	Cond.	ICA (%)	Human agr. (%)	NLL ↓ (bits)	ρ ↑	ΔKL ↑ (bits)	Side JS ↓ (nats)	Agr. (%)	Gap (pp)	Agr. (%)	Gap (pp)
Gemma4-E4B	Zero-shot	53.69	52.91	1.310	0.349	−2.398	0.0285	78.54	+17.68	54.37	+2.45
	Choice-only	92.72	72.72	0.817	0.308	−2.655	0.0325	93.11	+43.12	64.27	+14.28
	+Gaze	92.04	72.82	0.858	0.733	+0.083*	0.0281	92.14	+42.11	64.37	+14.20
	+AvgGaze	94.37	73.01	0.848	0.733	+0.002	0.0312	53.20	+3.83	46.21	+1.32
	+Gaze+Side	94.56	72.62	0.814	0.730	+0.070	0.0244	91.07	+40.87	64.85	+14.31
Qwen3-VL-8B	Zero-shot	49.90	48.54	1.835	0.032	−3.776	0.0447	65.34	+2.45	46.50	−1.95
	Choice-only	94.08	72.91	0.823	0.065	−3.694	0.0469	59.32	+9.59	49.51	+0.77
	+Gaze	92.43	72.82	0.813	0.729	+0.078*	0.0253	88.64	+38.13	63.40	+12.77
	+AvgGaze	89.61	71.75	0.814	0.731	+0.004	0.0258	82.82	+32.51	61.36	+12.23
	+Gaze+Side	90.97	71.94	0.833	0.726	+0.068	0.0247	90.49	+40.36	64.37	+14.48
Molmo2-8B	Zero-shot	53.50	50.00	1.054	0.010	−5.548	0.0667	80.68	−1.86	48.93	+0.44
	Choice-only	92.52	71.94	0.883	0.062	−5.457	0.0703	47.48	+0.51	48.74	+0.43
	+Gaze	89.03	73.11	0.819	0.673	+0.070	0.0293	86.41	+35.43	64.37	+14.15
	+AvgGaze	89.71	72.43	0.819	0.679	−0.007	0.0293	87.57	+36.49	65.44	+15.18
	+Gaze+Side	89.22	72.72	0.815	0.672	+0.059	0.0285	89.61	+39.19	65.24	+15.14

Learning from human choices. Training on 1,932 observed human choices raises human-choice agreement from 52.91/48.54/50.00% under Zero-shot to 72.72/72.91/71.94% under Choice-only. Under Choice-only, ICA is 92.72/94.08/92.52%, indicating that the models choose the intended shape category more often than they match the participant's actual response. Mean human-response NLL also decreases, from 1.310/1.835/1.054 to 0.817/0.823/0.883 bits, below the uniform-choice reference of 1 bit. Alignment with human responses is thus learnable from this small training set and transfers to new words and image pairs, although the comparison does not establish whether the association was already present before fine-tuning. The observed human-choice agreement under Choice-only is similar to the 72.86% human majority-vote reference (Section 3.3). On the 280 trials where participants chose against the intended category, Choice-only agrees with them on only 13.21/11.07/12.14% of trials. The learned choices are also robust to image position, so the model chooses the same shape regardless of whether it is shown on the left or right (Appendix L).

Effects of adding human gaze as a training target for attention. Relative to Choice-only, adding trial-specific within-image gaze changes human-choice agreement by +0.10 [−3.87, +4.49], −0.10 [−5.19, +4.84], and +1.17 [−3.98, +6.62] pp (individual 95% intervals), and all nine NLL contrasts among Choice-only, +Gaze, and +AvgGaze have intervals containing zero (Holm-adjusted $p = 1.000$). The probability-based evaluation gives the same qualification. Although gaze accumulates on the chosen image in 80.58% of pseudo-word trials (Table 4), supervising allocation between images with +Gaze+Side improves between-image JS relative to Choice-only for all three models (Appendix K.4) without a detectable additional choice gain (72.62/71.94/72.72%; all paired confidence intervals against Choice-only and +Gaze include zero).

Agreement without gaze alignment. Despite the increase in human-choice agreement, Choice-only reaches within-image gaze correlations of only .308/.065/.062, below the center-bias base-

line (.637/.637/.603; Section 3.3). The changes from Zero-shot are also inconsistent in direction. Correlation decreases for Gemma4-E4B and increases modestly for Qwen3-VL-8B and Molmo2-8B. Choice training therefore does not produce large, consistent gaze alignment at the decision-token readout used here.

Average-gaze control. +Gaze raises within-image gaze correlation to .733/.729/.673, but +AvgGaze reaches .733/.731/.679 while using the same fixed training-average gaze target on every trial. High within-image correlation therefore does not by itself demonstrate learning of trial-specific gaze patterns. The +Gaze−+AvgGaze difference in ΔKL (Table 3) is an estimated KL reduction of .074–.081 bits, meeting the Holm-adjusted .05 threshold for Gemma4-E4B and Qwen3-VL-8B, but not Molmo2-8B (Appendix K.3).

Allocation between images. Similar choice scores also coexist with different links between choice and attention. Under Choice-only, human-choice agreement differs by less than one point across models, yet the gap between choice–attention side agreement and its value expected under independence is +43.12/+9.59/+0.51 pp (Table 3). A near-zero gap suggests that an agreement reflects shared side preferences rather than a link between attention and choice (Appendix K.4). This is an association between the attended side and the answer, not evidence that attention causes the choice.

6 Discussion

Our results show a gap between adjective matching, which most models handle, and pseudo-word sound–shape alignment, which emerges mainly in a subset of larger VLMs. Molmo2-8B illustrates the distinction: near-perfect adjective ICA (0.94) with near-chance pseudo-word ICA (0.53). Adjective performance can rest on learned associations between familiar word forms and their referents; the partial success of Gemma4-26B and Qwen3-VL-30B on pseudo-words cannot rest on such associations, since the words are novel; it more plausibly reflects statistical regularities linking written form, phonology, and visual features in training data. Choice alignment does not imply human-aligned visual attention. Gemma4-26B's pseudo-word choices correlate with human judgments at $r \approx .58$ while its attention correlates with gaze at only .27, and fine-tuning on human choices raises human-choice agreement by ∼20 pp while within-image gaze correlation stays below a center-bias baseline. These models likely reach human-aligned choices through holistic image representations rather than the local shape-informative features on which human fixations concentrate, consistent with findings that vision transformers integrate information globally from early layers onward (Raghu et al., 2021). Judged by choice agreement alone, our fine-tuned models would appear to match the human majority; spatial alignment is the more diagnostic test. Adding human gaze as a training target for attention raises within-image attention–gaze correlation from below the center-bias baseline to .67–.73, with no detectable change in human-choice agreement or NLL, and a single average gaze map per image position produces the same correlation gain (Section 5.4). Prior work reports that gaze regularization improves attention alignment (Pani & Yang, 2025; Yu et al., 2017; Sood et al., 2020); our results agree on attention, but the gain does not carry over to choices and is matched by a constant average gaze map, echoing VQA gains from human-attention supervision that random maps reproduce (Shrestha et al., 2020).

Limitations. Fine-tuning uses three models, one optimization seed, one gaze-loss weight, and the original prompt. Attention and Grad-CAM are layer-dependent correlational readouts, not causal explanations. Written labels confound phonology and orthography; participants were English-speaking members of one institution and shared across splits.

7 Conclusion

We presented a direct human–VLM comparison of sound–shape associations, evaluating both choice alignment and spatial alignment on the same stimuli. Many models matched English adjectives to corresponding visual shapes, but pseudo-word alignment appears mainly in a subset of generative VLMs, and even there, model saliency diverges from human gaze. Fine-tuning closes the choice gap on unseen words and images without closing the spatial gap. Choice agreement alone is therefore insufficient evidence of human-aligned cross-modal processing; both choices and their visual attention must be evaluated.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, as detailed in Section 3.1. Additionally, we used generative AI tools to polish writing to improve readability. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The protocol for human subject research in this work was approved by the IRB at our institution. Participants gave informed consent, including consent to release de-identified choice and gaze data. Audio and video recordings were used only to verify responses and will not be released; released data contain no names, audio, video, or other personal identifiable information.

REPRODUCIBILITY STATEMENT

We will provide all code after publication. We will release all datasets used, including the stimuli and gaze data.

REFERENCES

Morris Alper and Hadar Averbuch-Elor. Kiki or bouba? Sound symbolism in vision-and-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 78347–78359, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f74054328beeb0c21a9b8e99da557f5a-Abstract-Conference.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv:2601.10611*, 2026. URL https://arxiv.org/abs/2601.10611.

Christine Cuskley, Julia Simner, and Simon Kirby. Phonological and orthographic influences in the bouba–kiki effect. *Psychological Research*, 81(1):119–130, 2017.

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841), 2022. doi:10.1098/rstb.2020.0390.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. ISSN 1077-3142. doi:https://doi.org/10.1016/j.cviu.2017.10.001. URL https://www.sciencedirect.com/science/article/pii/S1077314217301649. Language in Vision.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Joseph H. Greenberg and James J Jenkins. Studies in the psychological correlates of the sound system of American English. *Word*, 22(1-3):207–242, 1966. doi:10.1080/00437956.1966.11435451.

Hannah Haynie, Claire Bowern, and Hannah LaPalombara. Sound symbolism in the languages of Australia. *PLoS One*, 9(4):e92852, 2014. doi:10.1371/journal.pone.0092852.

Charles F. Hockett. The origin of speech. *Scientific American*, 203(3):88–97, 1960. URL https://web.stanford.edu/class/linguist197a/hockett60sciam.pdf.

Hinano Iida and Hayate Funakura. Investigating iconicity in vision-and-language models: A case study of the bouba/kiki effect in Japanese models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024. URL https://escholarship.org/uc/item/0br4n9tn.

Jinhong Jeong, Sunghyun Lee, Jaeyoung Lee, Seonah Han, and Youngjae Yu. Do language models associate sound with meaning? A multimodal study of sound symbolism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31247–31255, 2026. URL https://arxiv.org/abs/2511.10045.

Wolfgang Köhler. *Gestalt psychology*. Horace Liveright, 1929.

Wolfgang Köhler. *Gestalt psychology*. Horace Liveright, 1947.

Tom Kouwenhoven, Kiana Shahrasbi, and Tessa Verhoef. Cross-modal associations in vision and language models: Revisiting the bouba-kiki effect. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/6ebb92aad3a4fe7aae230b0e63c2ef35-Abstract-Conference.html.

Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298, 2010. doi:10.1038/nn.2635.

Gakuji Kumagai. The pluripotentiality of bilabial consonants: The images of softness and cuteness in Japanese and English. *Open Linguistics*, 6(1):693–707, 2020. doi:10.1515/opli-2020-0040.

Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2020. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9133499.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JAI7afWA9e.

Tyler Loakman, Yucheng Li, and Chenghua Lin. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2849–2867, 2024. URL https://aclanthology.org/2024.emnlp-main.167/.

Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, 9(3):316–322, 2006. doi:10.1111/j.1467-7687.2006.00495.x.

Kelly McCormick, Jee Young Kim, Sara List, and Lynne C Nygaard. Sound to meaning mappings in the bouba-kiki effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37, 2015. URL https://escholarship.org/uc/item/3s632891.

Alan K.S. Nielsen and Drew Rendall. Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(2):153, 2013. doi:10.1037/a0030553.

Ozge Ozturk, Madelaine Krehm, and Athena Vouloumanos. Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of experimental child psychology*, 114(2):173–186, 2013. doi:10.1016/j.jecp.2012.05.004.

Anupam Pani and Yanchao Yang. Gaze-VLM: Bridging gaze and VLMs through attention regularization for egocentric understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/ef72fa6579401ffff9da246a5014f055-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 12116–12128, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf.

Vilayanur S. Ramachandran and Edward M. Hubbard. Synaesthesia – a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34, 2001. URL https://www.ingentaconnect.com/content/imp/jcs/2001/00000008/00000012/1244.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. doi:10.1109/CVPR52688.2022.01042.

Maki Sakamoto and Junji Watanabe. Bouba/kiki in touch: Associations between tactile perceptual qualities and Japanese phonemes. *Frontiers in Psychology*, 9:295, 2018. doi:10.3389/fpsyg.2018.00295.

Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA)*, pp. 71–78, 2000. doi:10.1145/355017.355028.

Edward Sapir. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225, 1929. doi:10.1037/h0070931.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi:10.1109/ICCV.2017.74.

Shinsuke Shimojo, Claudiu Simion, Eiko Shimojo, and Christian Scheier. Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322, 2003. doi:10.1038/nn1150.

Kazuko Shinohara and Shigeto Kawahara. A cross-linguistic study of sound symbolism: The images of size. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, pp. 396–410, 2010. URL https://journals.linguisticsociety.org/proceedings/index.php/BLS/article/view/3926.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for VQA. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8172–8181, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.727. URL https://aclanthology.org/2020.acl-main.727/.

David M. Sidhu, Gabriella Vigliocco, and Penny M. Pexman. Higher order factors of sound symbolism. *Journal of Memory and Language*, 125:104323, 2022. doi:10.1016/j.jml.2022.104323.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6327–6341, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/460191c72f67e90150a093b4585e7eb4-Paper.pdf.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Wei-Cheng Tseng, Yi-Jen Shih, David Harwath, and Raymond Mooney. Measuring sound symbolism in audio-visual models. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 1165–1172, 2024. doi:10.1109/SLT61566.2024.10832254.

Tessa Verhoef, Kiana Shahrasbi, and Tom Kouwenhoven. What does Kiki look like? Cross-modal associations between speech sounds and visual shapes in vision-and-language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 199–213, 2024. doi:10.18653/v1/2024.cmcl-1.17.

Chris Westbury. Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and language*, 93(1):10–19, 2005. URL https://www.sciencedirect.com/science/article/pii/S0093934X04002202.

Bodo Winter and Marcus Perlman. Size sound symbolism in the english lexicon. *Glossa*, 6(1):1–13, 2021. URL https://www.glossa-journal.org/article/id/5459/.

Bodo Winter, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse. Trilled /r/ is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, 12(1):1035, 2022. URL https://www.nature.com/articles/s41598-021-04311-7.

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6119–6127, 2017. URL https://api.semanticscholar.org/CorpusID:25392305.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.pdf.

A Stimuli

Below is the full list of adjectives from Alper & Averbuch-Elor (2023), used as a baseline:

- A_{sharp} = {sharp, spiky, angular, jagged, hard, edgy, pointed, prickly, rugged, uneven}
- A_{round} = {round, circular, soft, fat, chubby, curved, smooth, plush, plump, rotund}.

Note on AH. AH is not a rounded vowel in every English dialect, but we include it in the round-associated set following prior human studies (Nielsen & Rendall, 2013; Maurer et al., 2006), where labels such as *bouba/BOOBAH* have been used as round-associated stimuli.

B Prompt for Image Generation

The prompt template used to generate images using Google's Gemini (Gemini Team et al., 2023) in Section 3.1 is provided in Figure 2.

For all prompts, $adjective_1$ and $adjective_2$ are the 20 adjectives from Alper & Averbuch-Elor (2023), listed in Section 3.1. Each pair of adjectives is always from different sets (i.e., one adjective from A_{sharp} and one from A_{round}). Some prompts do not include $material$, while others specify materials such as *plastic-like* or *metal-like*.

Generate a distinct image pair of $\{material\}$, $\{adjective_1\}$ vs $\{adjective_2\}$ abstract objects, where both objects are $\{color\}$, do not include text, no frames, ensure it is different from all previous images, and use a white background with no separation.

Figure 2: Prompt used to generate images.

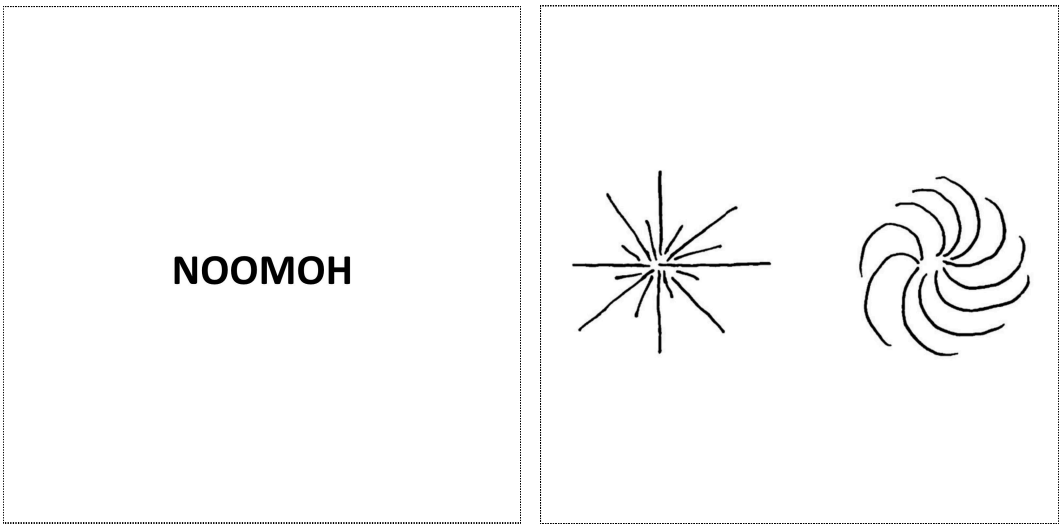


Figure 3: Stimuli pair in the human study. Word stimuli come first, and following image stimuli come second.

C Prompt for Forced Choice Analysis

The prompt used for forced choice analysis of generative VLMs is in Figure 4.

The neutral prompt used for the instruction-robustness check (Section 4.2) follows the participants' instruction, "Choose the image that better matches this word" and is shown in Figure 10. Information about the label type and sound symbolism is removed, as well as the rules, except for returning a JSON.

D Prior Visual Stimuli

Visual stimuli from prior works are in Figures 5, 6, 7, and 8.

E Human Study details

E.1 Recruitment

Participants were recruited through word of mouth and email lists, and included undergraduate and graduate students, and staff at our institution. Each participant was compensated with a $15 gift card for the full study (30 minutes), as approved by our institution's IRB.

E.2 Methodology

We use a slide deck comprised of pairs of labels and images (shown in Figure 3). We provide a word label for each image pair and ask participants to choose the image that they thought better matched with the label. Participants received the following verbal instructions prior to the task: "You will see a word followed by two images. Choose the image that best matches the word by saying A (left) or B

You are doing a forced-choice visual alignment task.

You will see one image containing two options:
- Image A is on the LEFT.
- Image B is on the RIGHT.

Label: "label"
Label type: label_type

Task: Choose which image, A or B, better aligns with the label.

Important rules:

1. You must choose exactly one option: A or B.
2. Do not answer "both", "neither", "unclear", or "tie".
3. If the label is a pseudoword, do not try to define it. Judge only by visual/sound-symbolic fit.
4. Base your answer only on the visual appearance of A and B.
5. Return only valid JSON with this exact schema:
"choice":"A"
or
"choice":"B"

Figure 4: Prompt used for forced choice analysis.

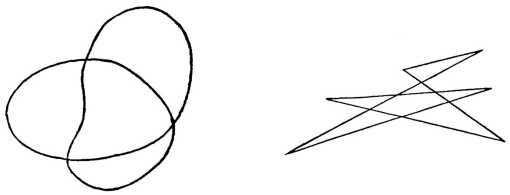


Figure 5: Image pairs from Köhler (1929; 1947)

(right)." No additional written instruction screen was shown beyond the procedure described above. For all labels, we ask them to read out loud and then verbally express their choice.

The study was divided into three equal-length sessions, with eye-tracker calibration before each session. At the end of each session, participants reported whether any pseudo-words in that session reminded them of real words or names; these items were excluded from analysis to minimize the influence of semantic associations. The study lasted approximately 30 minutes, participants were compensated with a $15 gift card, and the protocol was approved by our institutional IRB.

During the studies, two authors listened to participants' comments and observed no offensive content. After data collection, all personally identifying information was removed, as we processed only user choice and gaze data. Audio recordings were collected only from participants who provided explicit consent on the consent form; participants who declined audio recording were not recorded. Gaze data and choice responses were retained in anonymized form, linked only to participant identifiers.

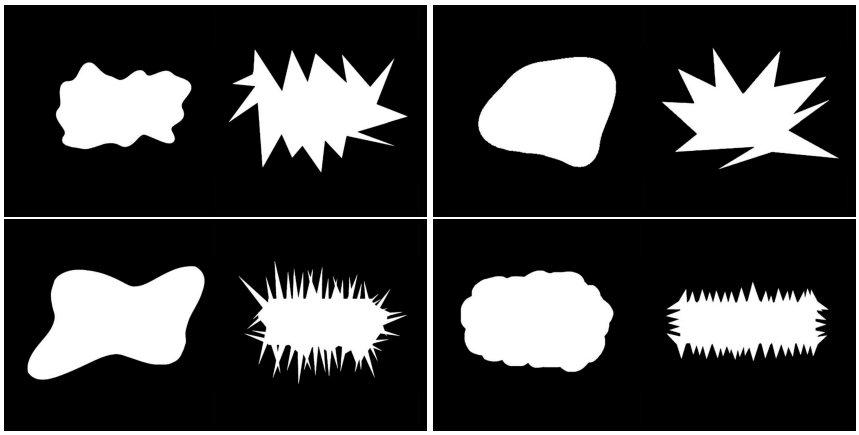


Figure 6: Image pairs from Westbury (2005)

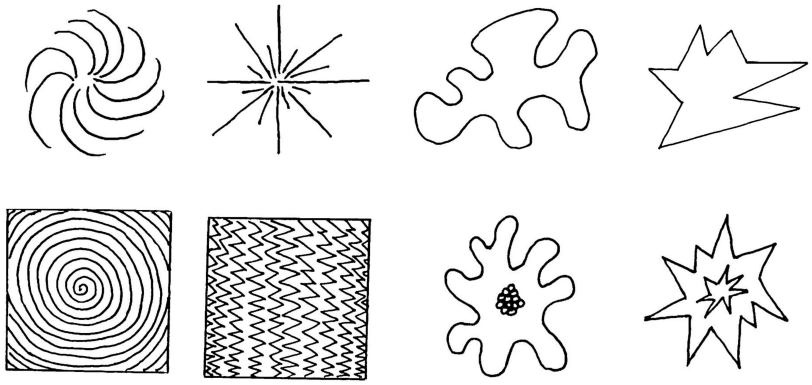


Figure 7: Image pairs from Maurer et al. (2006)

E.3 Stimuli

Each stimulus is presented via a slide-based format in which a word appears on one slide, followed immediately by the corresponding image pair on the next. As described in Section 3.1, the word set comprises 4 canonical pseudo-words, 162 phonemically constructed pseudo-words, and 20 adjectives. Stimuli are presented in the following order: pseudo-words first (canonical and phonemic combined, $n = 166$), followed by adjectives ($n = 20$). Within each block, the order of words is randomized across participants to counterbalance side bias. As described in Section 3.1, the image-pair set comprises 75 minimal line-drawing pairs, 92 language-model-generated pairs, and 17 pairs drawn from prior work. Two of these 17 original pairs were each used twice: one paired with *bouba* and *kiki*, and one with *takete* and *maluma*, thereby replicating the stimulus conditions of the original sound symbolism studies for our canonical word set. All image pairs were created in both left–right orientations (round shape on the left or right), yielding $186 \times 2 = 372$ total image instances, of which $184 \times 2 = 368$ are distinct. The four canonical words were fixed to their designated image pairs. The remaining 182 phonemically constructed pseudo-words and adjectives were then randomly mapped to the 364 non-canonical image pairs, with each word assigned to one instance of an image pair per participant. An example of stimuli is in Figure 3.

E.4 Human Gaze–Choice Coupling

Because gaze is recorded on the same trial as the choice, we can ask whether where a participant looked relates to what they chose. Raw gaze points are converted into fixation with an Identification by Dispersion Threshold (I-DT) procedure (Salvucci & Goldberg, 2000) using a minimum duration

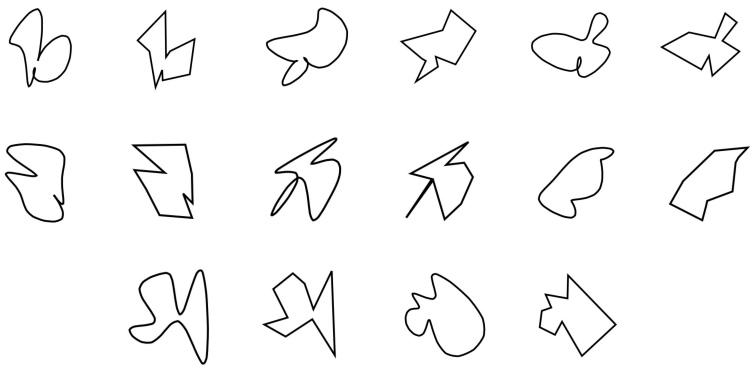


Figure 8: Image pairs from Kouwenhoven et al. (2025)

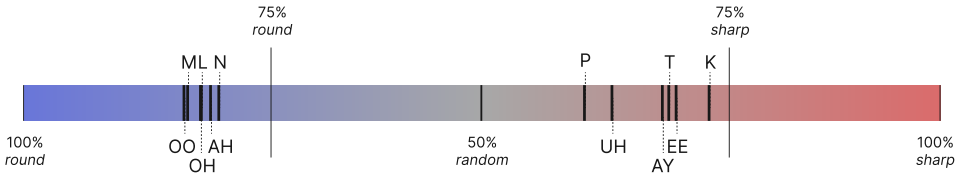


Figure 9: Human intended-category agreement by phoneme.

of 100 ms and a dispersion of at most 50 screen pixels, with recording gaps over 100 ms breaking continuous runs. Table 4 defines "the image attended to more" in three ways: the image containing more exported gaze samples, the image with the longer summed fixation duration (assigned by fixation centroid), and the image containing the first detected in-image fixation. We report how often each agrees with the participant's choice. Accumulated gaze aligns strongly with their choices, whereas early gaze does not. Under either accumulation measure, the image a participant looked at more is the image they chose in 77–80% of trials, while the first in-image fixation agrees at chance level: the coupling reflects where gaze settles, not an initial side bias. This first fixation is defined within the recorded image interval and may continue a fixation that began before image onset; it need not be the first newly initiated fixation after onset.

The image-pair slide was advanced immediately after each verbal response. These statistics describe gaze–choice association and do not establish its causal direction. The coupling replicates the well-documented tendency for gaze to accumulate on the option that is eventually chosen (Shimojo et al., 2003; Krajbich et al., 2010), and it matters here for one reason: a participant's gaze carries information about that participant's choice on that trial, which is the premise for using it as a supervision signal in Section 5. This association motivates gaze supervision in Section 5, as participants tend to look more at the image they choose. Gaze–choice agreement alone does not establish that round- and sharp-associated words produce distinct patterns. For the model comparisons, we later smooth gaze samples into density maps (Section 5.2); the statistics in Table 4 use raw samples and fixations only.

E.5 Participants and cleaning

We recruited 55 participants (undergraduate, graduate students and staff at our institution), yielding 9,688 recorded trials. We retain trials with a verified complete image presentation, a valid A/B choice, and at least two fixations within the two images combined. Gaze recording must also cover at least 80% of the time the images were displayed. The cleaned dataset contains 8,208 trials from 53 participants: 7,141 constructed pseudo-words, 910 adjectives, and 157 canonical trials. One participant had a calibration failure on gaze recording, and another participant's data was lost. Among pseudo-word trials with valid responses, intended-category agreement was 70.14% (881/1,256) in the excluded trials and 73.98% (5,283/7,141) in the retained trials; it was 73.41% (6,164/8,397) before cleaning. Unless stated otherwise, all human statistics use the cleaned dataset.

Table 4: Agreement (%) between the image a participant attended to more and the image they chose, on the cleaned dataset (53 participants; 8,208 trials: 7,141 constructed pseudo-words, 910 adjectives, 157 canonical trials). Parentheses show matches/valid trials; trials tied in a measure are excluded only from that measure's denominator. Gaze samples and fixation durations are summed inside each image over the full image-display interval; the first-fixation row uses the first detected in-image fixation. These direct image-region statistics use no Gaussian smoothing or model token pooling. 95% participant-cluster bootstrap intervals (20,000 resamples): 80.21 [77.02, 83.00], 76.97 [73.95, 79.75], 51.07 [49.70, 52.41].

"Attended to more" defined by	Total	Pseudo-words	Adjectives	Canonical
More gaze samples inside the image	80.21 (6,568/8,189)	80.58 (5,742/7,126)	77.59 (703/906)	78.34 (123/157)
Longer fixation duration inside the image	76.97 (6,318/8,208)	77.37 (5,525/7,141)	73.85 (672/910)	77.07 (121/157)
Image containing the first fixation	51.07 (4,192/8,208)	51.09 (3,648/7,141)	51.76 (471/910)	46.50 (73/157)

Table 5: CLIP (RN50) caption comparison for the overall subset. Each cell reports the point estimate with the 95% confidence interval below it. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Maj. and ICA use 95% Clopper–Pearson confidence intervals; correlation intervals use 95% bootstrap confidence intervals over labels. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

Caption	Maj.	r	ρ	ICA
{label}	0.57 [0.50, 0.65]	0.42* [0.29, 0.53]	0.42* [0.29, 0.53]	0.55 [0.54, 0.55]
The label for this image is {label}	0.54 [0.46, 0.61]	0.27* [0.11, 0.42]	0.27* [0.12, 0.42]	0.54 [0.53, 0.54]
This is a {label}	0.58 [0.51, 0.65]	0.36* [0.21, 0.50]	0.32* [0.17, 0.46]	0.55 [0.54, 0.55]
This looks like a {label}	0.63 [0.56, 0.70]	0.40* [0.25, 0.54]	0.37* [0.23, 0.51]	0.55 [0.55, 0.56]
An image of {label}	0.57 [0.50, 0.65]	0.49* [0.39, 0.59]	0.49* [0.38, 0.60]	0.56 [0.55, 0.56]
This is a {label} thing	0.60 [0.53, 0.67]	0.45* [0.32, 0.56]	0.44* [0.32, 0.56]	0.55 [0.55, 0.56]
A {label} object	0.56 [0.48, 0.63]	0.45* [0.30, 0.56]	0.48* [0.34, 0.59]	0.56 [0.55, 0.56]
A {label} shape	0.61 [0.54, 0.68]	0.51* [0.38, 0.61]	0.46* [0.33, 0.57]	0.57 [0.57, 0.58]
A {label} visual	0.53 [0.46, 0.61]	0.32* [0.18, 0.46]	0.32* [0.18, 0.46]	0.53 [0.53, 0.54]
A picture of a {label} object	0.56 [0.49, 0.64]	0.39* [0.23, 0.53]	0.43* [0.28, 0.55]	0.55 [0.54, 0.55]
Combined	0.58 [0.50, 0.65]	0.42* [0.30, 0.53]	0.42* [0.29, 0.53]	0.55 [0.55, 0.55]

F Image Type and Caption Type Analysis

Tables 5–7 and 8–10 show the results split by caption, for CLIP (RN50) and SigLIP-so400m.

Tables 11 and 12 show the results split by image types: line-based (original and generated minimal pairs) and model-based (Generated with Gemini), for CLIP (RN50) and SigLIP-so400m.

G Full Statistics for Table 1

Tables 13–15 report the same cleaned-cohort point estimates as Table 1, with 95% confidence intervals. The six embedding models, six completed original-prompt generative models, and the two neutral-prompt runs each cover 8,208 trials from 53 participants: 7,141 pseudo-word trials (162 labels), 910 adjective trials (20 labels), and 157 canonical trials (4 labels). Overall includes all 186 labels; the pseudo-word subset excludes canonical labels. For embedding models, each human trial is matched to all ten captions, and ICA and label-level round-choice rates use the same trial–caption

Table 6: CLIP (RN50) caption comparison for the pseudo-word subset. Each cell reports the point estimate with the 95% confidence interval below it. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Maj. and ICA use 95% Clopper–Pearson confidence intervals; correlation intervals use 95% bootstrap confidence intervals over labels. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

Caption	Maj.	r	ρ	ICA
{label}	0.53 [0.45, 0.60]	0.20 [0.06, 0.33]	0.23* [0.08, 0.36]	0.52 [0.51, 0.52]
The label for this image is {label}	0.51 [0.43, 0.59]	0.21* [0.06, 0.34]	0.17* [0.02, 0.32]	0.52 [0.51, 0.53]
This is a {label}	0.54 [0.46, 0.62]	0.22* [0.08, 0.35]	0.18* [0.03, 0.32]	0.52 [0.52, 0.53]
This looks like a {label}	0.60 [0.53, 0.68]	0.23* [0.08, 0.37]	0.20* [0.06, 0.34]	0.53 [0.52, 0.53]
An image of {label}	0.53 [0.45, 0.60]	0.31* [0.18, 0.44]	0.33* [0.18, 0.46]	0.53 [0.52, 0.53]
This is a {label} thing	0.57 [0.49, 0.65]	0.31* [0.16, 0.45]	0.30* [0.16, 0.44]	0.53 [0.53, 0.54]
A {label} object	0.51 [0.43, 0.59]	0.29* [0.15, 0.44]	0.33* [0.19, 0.47]	0.53 [0.52, 0.53]
A {label} shape	0.58 [0.50, 0.66]	0.36* [0.23, 0.49]	0.36* [0.23, 0.49]	0.55 [0.54, 0.55]
A {label} visual	0.49 [0.41, 0.57]	0.15 [-0.00, 0.29]	0.14 [-0.02, 0.28]	0.51 [0.50, 0.51]
A picture of a {label} object	0.51 [0.43, 0.59]	0.25* [0.09, 0.39]	0.28* [0.12, 0.42]	0.52 [0.51, 0.52]
Combined	0.53 [0.45, 0.60]	0.20* [0.06, 0.33]	0.23* [0.08, 0.36]	0.52 [0.52, 0.53]

weighting. Maj. compares human and model majority choices per label, treating a rate of 0.5 as a tie; ties match only when both rates are tied. Generative model round-choice rates average hard A/B decisions derived from FP32 logits over the corresponding human trials; an exact A/B logit tie contributes 0.5. These rates are not averages of $P(\text{round})$. Confidence intervals use 10,000 participant-cluster bootstrap resamples with seed 20260922, applying the same participant weights to all models and retaining the observed labels, images, and captions. Replicates missing any observed label are omitted for Maj., r, and ρ; undefined correlations are also omitted. These intervals are conditional on the observed stimulus inventory and checkpoint and do not measure generalization to new labels or variation across training seeds. Because a resample repeats some participants and omits others, label-level human rates are noisier within resamples than in the full sample, which attenuates label-level correlations; percentile bootstrap intervals need not contain the observed point estimate, as illustrated by Qwen3-VL-30B's pseudo-word correlation (r = .571, 95% CI [.414, .566]).

G.1 Neutral-prompt robustness and paired changes

Both large models were evaluated on the same 7,583 unique inputs, mapped to all 8,208 human trials, under original and neutral prompts. The checkpoint revision, processor, 512×512 input, eager backend, and FP32 A/B projection were identical within each model; Gemma's output softcap was retained.

This replaces the original task/rule block and omits the label type, including the instruction to judge pseudowords by "visual/sound-symbolic fit." It is a comparison of two prompt formulations, not an isolated intervention on one phrase. Neutral runs record choices only; they do not provide a neutral-prompt spatial-alignment comparison.

Table 1 and Tables 13–15 report the completed neutral point estimates and participant-bootstrap intervals. On pseudo-words, neutral Pearson correlations are .534 for Gemma4-26B (95% CI [.426, .549]) and .535 for Qwen3-VL-30B ([.410, .541]). Neutral adjective ICA remains high at .945 for both models (original: .970 and .937, respectively).

Table 7: CLIP (RN50) caption comparison for the adjective subset. Each cell reports the point estimate with the 95% confidence interval below it. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Maj. and ICA use 95% Clopper–Pearson confidence intervals; correlation intervals use 95% bootstrap confidence intervals over labels. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

Caption	Maj.	r	ρ	ICA
{label}	0.90 [0.68, 0.99]	0.87* [0.76, 0.95]	0.73* [0.43, 0.85]	0.78 [0.77, 0.80]
The label for this image is {label}	0.75 [0.51, 0.91]	0.67* [0.44, 0.87]	0.53* [0.17, 0.76]	0.69 [0.67, 0.70]
This is a {label}	0.90 [0.68, 0.99]	0.84* [0.73, 0.93]	0.70* [0.35, 0.86]	0.75 [0.74, 0.76]
This looks like a {label}	0.90 [0.68, 0.99]	0.92* [0.86, 0.96]	0.77* [0.45, 0.91]	0.79 [0.77, 0.80]
An image of {label}	0.90 [0.68, 0.99]	0.87* [0.75, 0.97]	0.78* [0.48, 0.89]	0.79 [0.78, 0.81]
This is a {label} thing	0.85 [0.62, 0.97]	0.83* [0.72, 0.92]	0.68* [0.35, 0.83]	0.74 [0.73, 0.76]
A {label} object	0.90 [0.68, 0.99]	0.89* [0.78, 0.96]	0.74* [0.40, 0.87]	0.78 [0.77, 0.80]
A {label} shape	0.90 [0.68, 0.99]	0.90* [0.79, 0.97]	0.81* [0.59, 0.90]	0.79 [0.78, 0.80]
A {label} visual	0.85 [0.62, 0.97]	0.81* [0.67, 0.93]	0.67* [0.34, 0.81]	0.72 [0.71, 0.74]
A picture of a {label} object	1.00 [0.83, 1.00]	0.94* [0.89, 0.98]	0.78* [0.54, 0.89]	0.80 [0.78, 0.81]
Combined	0.90 [0.69, 0.99]	0.87* [0.76, 0.95]	0.73* [0.43, 0.85]	0.76 [0.76, 0.77]

You will see one image containing two options:
- Image A is on the LEFT.
- Image B is on the RIGHT.

Label: "{word}"

Choose the image that better matches this word.

Choose exactly one option, A or B. Return only valid JSON:
{"choice":"A"}
or
{"choice":"B"}

Figure 10: Neutral prompt used for forced choice analysis.

For the exploratory paired contrasts in Table 16, we resample the 53 participants 10,000 times with replacement (seed 20260922), applying identical weights to original and neutral trials and to both models. We recompute each prompt's human/model label rates and trial-average metrics before taking the neutral-minus-original difference. These are nominal 95% percentile intervals, conditional on the observed labels, images, and checkpoints; they are not adjusted for multiple comparisons. They differ from the word–pair–participant bootstrap used for the spatial Center contrasts. Marginal confidence-interval overlap is not used as a test of prompt differences.

Gemma4-26B's pseudo-word ICA changes by −3.99 pp (paired 95% CI $[-5.29, -2.74]$), while its agreement with the actual participant response changes from 57.30% to 57.74% (+0.43 pp, $[-0.96, 1.84]$). Qwen3-VL-30B's ICA has the same point estimate under both prompts, but its actual-response agreement changes from 55.73% to 57.72% (+1.99 pp, $[0.76, 3.22]$). Neither

Table 8: SigLIP-so400m caption comparison for the overall subset. Each cell reports the point estimate with the 95% confidence interval below it. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Asterisks indicate two-sided approximate tests of zero correlation at $p < .05$.

Caption	Maj.	r	ρ	ICA
{label}	0.67 [0.59, 0.73]	0.45* [0.31, 0.56]	0.47* [0.34, 0.58]	0.59 [0.58, 0.59]
The label for this image is {label}	0.68 [0.61, 0.74]	0.55* [0.42, 0.67]	0.59* [0.47, 0.68]	0.58 [0.57, 0.58]
This is a {label}	0.69 [0.62, 0.75]	0.48* [0.34, 0.61]	0.54* [0.41, 0.64]	0.57 [0.56, 0.58]
This looks like a {label}	0.59 [0.52, 0.66]	0.38* [0.23, 0.52]	0.43* [0.30, 0.55]	0.56 [0.55, 0.56]
An image of {label}	0.73 [0.66, 0.79]	0.53* [0.40, 0.65]	0.55* [0.42, 0.66]	0.60 [0.59, 0.60]
This is a {label} thing	0.62 [0.54, 0.69]	0.42* [0.28, 0.56]	0.48* [0.34, 0.59]	0.56 [0.55, 0.57]
A {label} object	0.54 [0.47, 0.62]	0.29* [0.14, 0.43]	0.37* [0.22, 0.50]	0.55 [0.54, 0.55]
A {label} shape	0.66 [0.58, 0.72]	0.39* [0.24, 0.54]	0.42* [0.29, 0.56]	0.57 [0.57, 0.58]
A {label} visual	0.72 [0.65, 0.78]	0.54* [0.42, 0.65]	0.57* [0.47, 0.67]	0.59 [0.58, 0.59]
A picture of a {label} object	0.61 [0.53, 0.68]	0.41* [0.27, 0.54]	0.46* [0.32, 0.57]	0.56 [0.55, 0.57]
Combined	0.67 [0.59, 0.73]	0.45* [0.31, 0.56]	0.47* [0.34, 0.58]	0.57 [0.57, 0.57]

Pearson-r change has an interval excluding zero; these intervals do not establish equivalence. The original/neutral choice differs on 33.23% of the 6,720 unique pseudo-word inputs for Gemma4-26B and 20.55% for Qwen3-VL-30B (trial-weighted: 33.26% and 20.70%). In particular, unchanged aggregate Qwen ICA does not mean unchanged individual decisions. Together, these results support persistence of positive choice alignment under the neutral formulation, while qualifying any claim of prompt invariance.

H Attention Alignment

H.1 Choice evaluation

Generative models use the original forced-choice prompt (Appendix C), as in fine-tuning, with the composite panel resized to 512×512 before native processing. All six completed generative runs use 512×512 inputs (resolution sensitivity: *Input resolution* below). Generative choices and attention come from one eager forward pass per input. A/B probabilities use FP32 projections of the final hidden state, with Gemma's output softcap retained; remaining exact ties count as half in trial agreement measures. ICA and same-trial human agreement describe individual responses; label-level majority match and correlations compare response distributions. Overall, pseudo-word, and adjective summaries contain 186, 162, and 20 labels, respectively; the four canonical labels enter Overall only.

H.2 Trial-wise native-grid spatial evaluation

The full-cohort evaluation contains 8,208 human trials and 7,583 unique model inputs; the main spatial table uses 7,141 constructed pseudo-word trials (6,720 unique inputs). Each model map is matched back to every corresponding human trial. We do not average maps across participants or image–category cells before computing the metrics.

Generative readout and layers. Each generative model receives the original forced-choice prompt, and we read attention from the last prompt position, the position whose output predicts the A/B answer; the answer token itself is not part of the input. This is the decision-token attention readout, and it is

Table 9: SigLIP-so400m caption comparison for the pseudo-word subset. Each cell reports the point estimate with the 95% confidence interval below it. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Maj. and ICA use 95% Clopper–Pearson confidence intervals; correlation intervals use 95% bootstrap confidence intervals over labels. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

Caption	Maj.	r	ρ	ICA
{label}	0.66 [0.58, 0.73]	0.47* [0.35, 0.59]	0.47* [0.34, 0.58]	0.58 [0.58, 0.59]
The label for this image is {label}	0.67 [0.60, 0.74]	0.57* [0.46, 0.66]	0.54* [0.40, 0.64]	0.56 [0.56, 0.57]
This is a {label}	0.70 [0.62, 0.77]	0.53* [0.42, 0.63]	0.52* [0.39, 0.62]	0.56 [0.56, 0.57]
This looks like a {label}	0.59 [0.51, 0.66]	0.37* [0.23, 0.48]	0.35* [0.20, 0.49]	0.55 [0.54, 0.55]
An image of {label}	0.72 [0.64, 0.78]	0.52* [0.40, 0.63]	0.52* [0.39, 0.62]	0.59 [0.58, 0.59]
This is a {label} thing	0.62 [0.54, 0.69]	0.39* [0.27, 0.50]	0.38* [0.25, 0.50]	0.55 [0.54, 0.55]
A {label} object	0.53 [0.45, 0.61]	0.24* [0.09, 0.37]	0.30* [0.16, 0.44]	0.53 [0.53, 0.54]
A {label} shape	0.64 [0.56, 0.72]	0.36* [0.21, 0.50]	0.36* [0.21, 0.50]	0.56 [0.55, 0.56]
A {label} visual	0.72 [0.65, 0.79]	0.60* [0.50, 0.69]	0.59* [0.48, 0.69]	0.58 [0.58, 0.59]
A picture of a {label} object	0.59 [0.51, 0.66]	0.35* [0.21, 0.48]	0.36* [0.22, 0.49]	0.54 [0.54, 0.55]
Combined	0.66 [0.58, 0.73]	0.47* [0.35, 0.59]	0.47* [0.34, 0.58]	0.56 [0.56, 0.56]

the same attention query used in fine-tuning. Maps are averaged over heads and over full-attention layers with zero-based index at least $\lfloor L/2 \rfloor$; for Gemma, only the global-attention layers in that half contribute. The resulting layers are L18–35 for Qwen3-VL-8B, Molmo2-4B, and Molmo2-8B; L24–47 for Qwen3-VL-30B; L23/29/35/41 for Gemma4-E4B; and L17/23/29 for Gemma4-26B.

Inputs are resized to 512×512 before each model's standard image preprocessing. At this size, Qwen and Gemma produce a 16×16 image-token grid; Molmo uses its global and high-resolution tokens together (725 image tokens), as in fine-tuning, with each token mapped to its actual image region. Model maps are compared on these native cells without upsampling, and tokens that straddle the boundary between the two images are excluded.

A/B logits are computed in FP32 from the final hidden state using the two corresponding rows of the output projection, with Gemma's final-logit softcap retained; exact ties count as half in agreement measures. Because choices on near-indifferent items can depend on the numerical backend, choices and attention maps for each input come from a single forward pass with a fixed eager backend, and all generative entries in Table 1, the Choice r entries in Table 2, and the participant-bootstrap intervals in Appendix G are computed from these choices. The neutral-prompt runs (Gemma4-26B and Qwen3-VL-30B; Section 4.2) use the same input size and backend but record choices only, so all spatial results use the original prompt.

Embedding readouts and coordinates. All six embedding models use the fixed caption "The label for this image is {word}" and the original composite panel resized to 512×512 before native image processing. CLIP RN50 uses layer4 feature Grad-CAM; CLIP ViT-B/32 uses positive final-layer attention–gradient products, head-averaged at CLS-to-patch positions. SigLIP and SigLIP2 use feature Grad-CAM at vision post-layernorm with gradients of their native pooled image–text cosine score. Native positive maps and coordinates are saved directly. CLIP uses normalized feature-cell bins, not receptive-field masks; SigLIP coordinates follow the actual patch convolution kernel, stride, padding, and processor size, including uncovered edge strips where applicable. Trials whose Grad-CAM map is identically zero after retaining positive evidence are excluded and counted explicitly; such maps

Table 10: SigLIP-so400m caption comparison for the adjective subset. Each cell reports the point estimate with the 95% confidence interval below it. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Maj. and ICA use 95% Clopper–Pearson confidence intervals; correlation intervals use 95% bootstrap confidence intervals over labels. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

Caption	Maj.	r	ρ	ICA
{label}	0.75 [0.51, 0.91]	0.64* [0.33, 0.84]	0.57* [0.12, 0.82]	0.63 [0.61, 0.64]
The label for this image is {label}	0.75 [0.51, 0.91]	0.82* [0.68, 0.93]	0.75* [0.48, 0.88]	0.70 [0.69, 0.72]
This is a {label}	0.65 [0.41, 0.85]	0.70* [0.53, 0.83]	0.63* [0.31, 0.81]	0.63 [0.61, 0.64]
This looks like a {label}	0.65 [0.41, 0.85]	0.76* [0.65, 0.88]	0.73* [0.46, 0.85]	0.63 [0.62, 0.65]
An image of {label}	0.90 [0.68, 0.99]	0.80* [0.64, 0.91]	0.69* [0.37, 0.85]	0.68 [0.66, 0.69]
This is a {label} thing	0.65 [0.41, 0.85]	0.83* [0.72, 0.93]	0.69* [0.42, 0.83]	0.67 [0.65, 0.68]
A {label} object	0.65 [0.41, 0.85]	0.68* [0.51, 0.84]	0.66* [0.34, 0.83]	0.66 [0.65, 0.68]
A {label} shape	0.85 [0.62, 0.97]	0.80* [0.66, 0.92]	0.73* [0.46, 0.89]	0.70 [0.69, 0.72]
A {label} visual	0.75 [0.51, 0.91]	0.64* [0.48, 0.80]	0.56* [0.18, 0.81]	0.63 [0.62, 0.65]
A picture of a {label} object	0.80 [0.56, 0.94]	0.86* [0.76, 0.94]	0.71* [0.46, 0.84]	0.70 [0.68, 0.71]
Combined	0.75 [0.51, 0.91]	0.64* [0.33, 0.84]	0.57* [0.12, 0.82]	0.66 [0.66, 0.67]

Table 11: Model–human alignment by subset for CLIP (RN50), comparing image types. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

	Overall				Pseudo-word				Adjective			
Image Type	Maj.	r	ρ	ICA	Maj.	r	ρ	ICA	Maj.	r	ρ	ICA
Lines	0.551	0.425*	0.423*	0.558	0.512	0.277*	0.269*	0.530	0.857	0.868*	0.701*	0.787
Realistic	0.577	0.408*	0.333*	0.542	0.543	0.222*	0.170*	0.518	0.850	0.820*	0.731*	0.741
Combined	0.575	0.419*	0.418*	0.550	0.525	0.202*	0.227*	0.524	0.900	0.868*	0.726*	0.764

cannot support a spatial correlation. These spatial readouts use one caption, whereas the embedding choice survey retains ten-caption averaging.

Common evaluation and references. For each trial, sample-density gaze maps are smoothed with $\sigma = 40$ screen pixels, integrated over each model's native token regions, and normalized within each valid image with the same target construction as in fine-tuning. Spearman correlation is computed separately within images A and B, averaged over valid image correlations within trial, and then over trials with a defined correlation. Constant maps yield undefined correlations. The center reference (the center baseline of Section 3.3) is an isotropic Gaussian with $\sigma = 0.125$ in normalized composite-panel coordinates, centered at $(0.25, 0.5)$ for A and $(0.75, 0.5)$ for B. Its mass is integrated over the same cells and evaluated with the same human masks; valid image and correlation denominators are reported alongside every summary. Relative mass on A/B and the gap cannot determine this within-image statistic.

As a consistency check, the 1,030 Test trials of each small-model run reproduce the zero-shot rows of Table 3: the same trial identities, human masks, 2,028 valid images, valid-correlation counts, and metrics within a prespecified tolerance of 0.0005 (Gemma4-E4B $\rho = .349$, $\Delta\mathrm{KL} = -2.398$ bits; Qwen3-VL-8B .032, −3.776; Molmo2-8B .010, −5.548). These Test values differ from the full-cohort pseudo-word means in Table 2 because the latter use all splits.

Table 12: Model–human alignment by subset for SigLIP-so400m, comparing image types. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. Asterisks indicate $p < .05$ under two-sided tests of zero Pearson/Spearman correlation.

	Overall				Pseudo-word				Adjective			
Image Type	Maj.	r	ρ	ICA	Maj.	r	ρ	ICA	Maj.	r	ρ	ICA
Lines	0.647	0.393*	0.418*	0.561	0.648	0.362*	0.370*	0.550	0.667	0.661*	0.644*	0.654
Realistic	0.769	0.663*	0.657*	0.583	0.753	0.604*	0.603*	0.572	0.900	0.846*	0.763*	0.672
Combined	0.667	0.452*	0.473*	0.572	0.660	0.469*	0.471*	0.561	0.750	0.639*	0.569*	0.663

Table 13: Uncertainty statistics for the overall subset: 8,208 trials and 186 labels. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. All intervals are 95% percentile participant-cluster bootstrap intervals (10,000 resamples of the 53 participants). Human and model label rates and ICA are recomputed in each resample. Asterisks indicate nominal $p < .05$ in two-sided label-level correlation tests, unadjusted for multiple comparisons; these tests differ from the bootstrap intervals. Dashes indicate unavailable runs.

Model	Maj.	95% CI	r	95% CI	ρ	95% CI	ICA	95% CI
CLIP RN50	0.618	[0.559, 0.640]	0.452*	[0.374, 0.474]	0.424*	[0.334, 0.446]	0.548	[0.541, 0.555]
CLIP ViT-B/32	0.597	[0.548, 0.629]	0.375*	[0.296, 0.402]	0.344*	[0.252, 0.377]	0.534	[0.527, 0.540]
SigLIP-base	0.613	[0.565, 0.640]	0.381*	[0.314, 0.411]	0.372*	[0.296, 0.408]	0.548	[0.540, 0.556]
SigLIP-so400m	0.731	[0.667, 0.753]	0.534*	[0.438, 0.547]	0.603*	[0.474, 0.601]	0.574	[0.566, 0.582]
SigLIP2-base	0.495	[0.446, 0.516]	0.324*	[0.240, 0.354]	0.162*	[0.110, 0.262]	0.531	[0.524, 0.538]
SigLIP2-so400m	0.478	[0.446, 0.559]	0.018	[-0.079, 0.110]	-0.043	[-0.109, 0.074]	0.502	[0.493, 0.510]
Gemma4-E4B	0.565	[0.527, 0.624]	0.447*	[0.350, 0.464]	0.326*	[0.259, 0.395]	0.554	[0.545, 0.564]
Gemma4-26B	0.758	[0.694, 0.763]	0.696*	[0.626, 0.704]	0.695*	[0.612, 0.701]	0.663	[0.654, 0.672]
Qwen3-VL-8B	0.613	[0.543, 0.651]	0.527*	[0.421, 0.531]	0.456*	[0.314, 0.471]	0.566	[0.556, 0.575]
Qwen3-VL-30B	0.715	[0.661, 0.747]	0.707*	[0.612, 0.702]	0.693*	[0.563, 0.689]	0.613	[0.603, 0.623]
Molmo2-4B	0.570	[0.500, 0.613]	0.417*	[0.251, 0.412]	0.349*	[0.173, 0.367]	0.530	[0.519, 0.541]
Molmo2-8B	0.656	[0.597, 0.694]	0.622*	[0.514, 0.618]	0.580*	[0.429, 0.577]	0.584	[0.573, 0.594]
Neutral prompt								
Gemma4-26B	0.726	[0.677, 0.753]	0.678*	[0.599, 0.683]	0.656*	[0.566, 0.663]	0.625	[0.614, 0.636]
Qwen3-VL-30B	0.624	[0.608, 0.677]	0.665*	[0.587, 0.667]	0.644*	[0.550, 0.657]	0.609	[0.600, 0.618]

Completed embedding grids and denominators. Table 17 lists the denominators for the pseudo-word entries in Table 2. CLIP RN50 excludes 42 pseudo-word trials with zero positive-evidence maps. Of the remaining 7,099 trials, 22 have no defined model correlation in either valid image, leaving 7,077 trials for model ρ; Center remains defined on 7,099 trials. Undefined image-level correlations are omitted before the within-trial average, so model and Center image counts can also differ (including seven images for SigLIP2-so400m). These are descriptive means with the displayed denominators, not a paired significance test against Center.

Input resolution. To check sensitivity to input size, we compared the two large models at 512×512 with a 1000×1000 input on 500 stratified pseudo-word inputs (539 human trials). For Qwen3-VL-30B, ICA changed by -0.93 pp and human-choice agreement by $+0.56$ pp, and input-level $P(\text{round})$ correlated at Spearman .828 across resolutions; for Gemma4-26B, the changes were $+2.04$ and $+2.23$ pp, with correlation .892. Both changes are within a tolerance of 3 pp fixed before the comparison. The 1000×1000 outputs use native BF16 logits rather than the FP32 A/B projection, so these differences bound the combined effect of resolution and numerical backend rather than isolating resolution.

Generative results and denominators. For all six generative models, each pseudo-word spatial mean uses 7,141 trials (6,720 unique model inputs, 162 labels) and 14,070 valid image-level correlations for both the model and Center, so the two denominators match. Qwen and Gemma use 16×16 grids; both Molmo models use the combined 725-token global and high-resolution geometry. The model/Center means are .340/.637 (Gemma4-E4B), .265/.637 (Gemma4-26B), .022/.637 (Qwen3-VL-8B), $-.003$/.637 (Qwen3-VL-30B), $-.074$/.601 (Molmo2-4B), and .004/.601 (Molmo2-8B). For the three fine-tuned models, the Test subset of these runs reproduces the zero-shot rows of Table 3 (consistency check above); Molmo2-4B is not fine-tuned, so this check does not apply to it. Paired comparisons against Center for all six models are reported below.

Table 14: Uncertainty statistics for the pseudo-word subset: 7,141 trials and 162 labels. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. All intervals are 95% percentile participant-cluster bootstrap intervals (10,000 resamples of the 53 participants). Human and model label rates and ICA are recomputed in each resample. Asterisks indicate nominal $p < .05$ in two-sided label-level correlation tests, unadjusted for multiple comparisons; these tests differ from the bootstrap intervals. Dashes indicate unavailable runs.

Model	Maj.	95% CI	r	95% CI	ρ	95% CI	ICA	95% CI
CLIP RN50	0.580	[0.518, 0.611]	0.285*	[0.192, 0.318]	0.274*	[0.168, 0.310]	0.520	[0.513, 0.528]
CLIP ViT-B/32	0.549	[0.494, 0.586]	0.116	[0.036, 0.170]	0.132	[0.031, 0.192]	0.505	[0.497, 0.512]
SigLIP-base	0.574	[0.519, 0.611]	0.209*	[0.136, 0.256]	0.197*	[0.114, 0.259]	0.521	[0.512, 0.529]
SigLIP-so400m	0.716	[0.648, 0.747]	0.560*	[0.427, 0.566]	0.561*	[0.419, 0.568]	0.565	[0.557, 0.573]
SigLIP2-base	0.438	[0.383, 0.463]	-0.095	[-0.165, 0.038]	-0.121	[-0.174, 0.042]	0.497	[0.489, 0.506]
SigLIP2-so400m	0.500	[0.451, 0.580]	0.011	[-0.101, 0.120]	-0.019	[-0.115, 0.103]	0.503	[0.494, 0.512]
Gemma4-E4B	0.512	[0.469, 0.580]	0.153	[0.031, 0.206]	0.085	[0.002, 0.193]	0.511	[0.501, 0.522]
Gemma4-26B	0.728	[0.654, 0.741]	0.580*	[0.493, 0.594]	0.579*	[0.470, 0.592]	0.622	[0.612, 0.632]
Qwen3-VL-8B	0.562	[0.488, 0.605]	0.262*	[0.078, 0.279]	0.242*	[0.054, 0.275]	0.516	[0.505, 0.526]
Qwen3-VL-30B	0.679	[0.617, 0.716]	0.571*	[0.414, 0.566]	0.562*	[0.387, 0.564]	0.566	[0.554, 0.577]
Molmo2-4B	0.525	[0.451, 0.574]	0.131	[-0.021, 0.186]	0.143	[-0.025, 0.189]	0.507	[0.496, 0.518]
Molmo2-8B	0.605	[0.543, 0.654]	0.413*	[0.218, 0.410]	0.390*	[0.190, 0.400]	0.532	[0.521, 0.543]
Neutral prompt								
Gemma4-26B	0.704	[0.642, 0.728]	0.534*	[0.426, 0.549]	0.516*	[0.400, 0.535]	0.582	[0.569, 0.594]
Qwen3-VL-30B	0.574	[0.562, 0.642]	0.535*	[0.410, 0.541]	0.524*	[0.400, 0.546]	0.566	[0.556, 0.575]

Table 15: Uncertainty statistics for the adjective subset: 910 trials and 20 labels. Maj. = majority-choice match per label; r = Pearson correlation; ρ = Spearman correlation; ICA = intended-category agreement. All intervals are 95% percentile participant-cluster bootstrap intervals (10,000 resamples of the 53 participants). Human and model label rates and ICA are recomputed in each resample. Asterisks indicate nominal $p < .05$ in two-sided label-level correlation tests, unadjusted for multiple comparisons; these tests differ from the bootstrap intervals. Dashes indicate unavailable runs.

Model	Maj.	95% CI	r	95% CI	ρ	95% CI	ICA	95% CI
CLIP RN50	0.950	[0.850, 1.000]	0.879*	[0.829, 0.901]	0.715*	[0.667, 0.810]	0.750	[0.733, 0.768]
CLIP ViT-B/32	1.000	[0.900, 1.000]	0.918*	[0.863, 0.942]	0.752*	[0.702, 0.853]	0.760	[0.739, 0.781]
SigLIP-base	0.950	[0.850, 1.000]	0.917*	[0.872, 0.935]	0.847*	[0.750, 0.903]	0.762	[0.744, 0.780]
SigLIP-so400m	0.900	[0.700, 0.950]	0.819*	[0.704, 0.861]	0.685*	[0.568, 0.815]	0.657	[0.634, 0.679]
SigLIP2-base	0.950	[0.850, 1.000]	0.923*	[0.884, 0.940]	0.756*	[0.713, 0.847]	0.786	[0.766, 0.807]
SigLIP2-so400m	0.350	[0.300, 0.600]	-0.022	[-0.259, 0.211]	-0.032	[-0.225, 0.256]	0.497	[0.477, 0.517]
Gemma4-E4B	1.000	[0.950, 1.000]	0.992*	[0.973, 0.994]	0.854*	[0.780, 0.915]	0.909	[0.891, 0.927]
Gemma4-26B	1.000	[1.000, 1.000]	0.994*	[0.981, 0.996]	0.775p*	[0.734, 0.875]	0.970	[0.960, 0.980]
Qwen3-VL-8B	1.000	[0.950, 1.000]	0.990*	[0.973, 0.993]	0.829*	[0.750, 0.891]	0.942	[0.928, 0.955]
Qwen3-VL-30B	0.950	[0.950, 1.000]	0.985*	[0.964, 0.992]	0.905*	[0.818, 0.940]	0.937	[0.925, 0.950]
Molmo2-4B	0.950	[0.800, 1.000]	0.866*	[0.760, 0.901]	0.817*	[0.674, 0.876]	0.716	[0.688, 0.745]
Molmo2-8B	1.000	[1.000, 1.000]	0.987*	[0.970, 0.992]	0.825*	[0.761, 0.888]	0.936	[0.920, 0.951]
Neutral prompt								
Gemma4-26B	0.950	[0.950, 1.000]	0.982*	[0.960, 0.991]	0.771*	[0.735, 0.852]	0.945	[0.930, 0.958]
Qwen3-VL-30B	0.950	[0.950, 1.000]	0.983*	[0.962, 0.991]	0.869*	[0.809, 0.911]	0.945	[0.933, 0.957]

I Probability-based Analysis

Tables 18, 19, and 20 report the probability-based analysis following Kouwenhoven et al. (2025). This method produces a model choice for each image by comparing the probabilities assigned to candidate labels. However, it differs substantially from the human forced-choice task: human participants choose between two images for a single label, whereas the probability-based method chooses among all labels for a given image. As a result, this analysis is not directly comparable to the human study. We therefore use it only to report ICA, rather than model–human agreement.

This distinction is important because the probability-based method can introduce label-side biases. In particular, we observe a strong tendency toward sharp labels. Even when overall accuracy is close to chance, round and sharp accuracy can be highly asymmetric. If a model is slightly more likely to choose sharp labels regardless of the image, then sharp-label accuracy will be inflated while round-label accuracy will be correspondingly reduced. Such an asymmetry does not necessarily

Table 16: Exploratory neutral-minus-original changes on the same 7,141 pseudo-word trials. Intervals use paired participant bootstrap and are nominal 95% intervals without multiplicity adjustment. Maj., ICA, and human agreement differences are percentage points (pp); correlation differences are on their original scale. Human agreement scores the actual participant response on each trial.

Model	Metric	Difference	Paired 95% CI
Gemma4-26B	Maj. (pp)	−2.47	[−6.79, 4.94]
Gemma4-26B	r	−0.047	[−0.113, 0.001]
Gemma4-26B	ρ	−0.063	[−0.123, −0.002]
Gemma4-26B	ICA (pp)	−3.99	[−5.29, −2.74]
Gemma4-26B	Human agreement (pp)	+0.43	[−0.96, 1.84]
Qwen3-VL-30B	Maj. (pp)	−10.49	[−12.35, −1.85]
Qwen3-VL-30B	r	−0.037	[−0.068, 0.039]
Qwen3-VL-30B	ρ	−0.038	[−0.066, 0.058]
Qwen3-VL-30B	ICA (pp)	+0.00	[−0.93, 0.93]
Qwen3-VL-30B	Human agreement (pp)	+1.99	[0.76, 3.22]

Table 17: Native grids and defined-correlation denominators for the embedding pseudo-word spatial results. Counts follow zero-map exclusions; image-level values are averaged within each trial before averaging trials.

		Trials		Images	
Model	Grid	ρ	Center	ρ	Center
CLIP RN50	7×7	7,077	7,099	13,233	13,987
CLIP ViT-B/32	7×7	7,141	7,141	14,070	14,070
SigLIP-base	14×14	7,141	7,141	14,070	14,070
SigLIP-so400m	27×27	7,141	7,141	14,070	14,070
SigLIP2-base	14×14	7,141	7,141	14,070	14,070
SigLIP2-so400m	32×32	7,141	7,141	14,063	14,070

indicate human-aligned sensitivity to the round–sharp contrast; instead, it may reflect a bias induced by the label-selection procedure. For this reason, we treat the probability-based results as a diagnostic for the ICA rather than as an approximation of human forced-choice behavior.

J Choice agreement and majority-vote reference

Table 21 reports individual 95% confidence intervals for the two percentage-valued scores in Table 3: ICA and human-choice agreement. The human majority-vote reference predicts each participant's response from the other participants' responses to the same word on Test; ties receive an expected agreement score of .5. All rows use the same 1,030 test trials.

Model score intervals are percentile intervals from 20,000 word×pair×participant product-cluster bootstrap draws, with the selected checkpoints held fixed. The reference score and paired model–reference differences use 20,000 draws with the original majority-vote predictions held fixed; their intervals are centered on the observed estimate using the 95th percentile of absolute bootstrap errors. Within each analysis, resampling weights are shared across conditions. All intervals are individual and unadjusted; they exclude uncertainty from training runs, checkpoint selection, and, for the reference, estimation of the majority-vote predictions.

Table 22 separately reports the paired differences between Choice-only and the human majority-vote reference. These intervals are computed directly from differences on matched trials, rather than by subtracting the bounds of the individual score intervals. They describe uncertainty in the comparison and do not establish equivalence or a ceiling on achievable agreement.

Table 18: Probability-based top-choice shape-match accuracy, overall, following Kouwenhoven et al. (2025). Values in brackets are Wilson 95% confidence intervals.

Caption	Total	Round	Sharp
CLIP (RN50)	0.712 [0.702, 0.722]	0.644 [0.629, 0.660]	0.780 [0.766, 0.793]
CLIP (ViT)	**0.720** [0.709, 0.730]	**0.701** [0.686, 0.715]	0.739 [0.724, 0.753]
SigLIP-base	0.521 [0.510, 0.533]	0.464 [0.448, 0.480]	0.579 [0.563, 0.595]
SigLIP-so400m	0.635 [0.624, 0.646]	0.467 [0.451, 0.483]	0.804 [0.790, 0.816]
SigLIP2-base	0.516 [0.504, 0.527]	0.142 [0.131, 0.153]	**0.890** [0.879, 0.900]
SigLIP2-so400m	0.511 [0.500, 0.523]	0.345 [0.329, 0.360]	0.678 [0.662, 0.693]

Table 19: Probability-based top-choice shape-match accuracy, pseudo-word captions only. Values in brackets are Wilson 95% confidence intervals

Caption	Total	Round	Sharp
CLIP (RN50)	0.552 [0.536, 0.568]	0.537 [0.515, 0.560]	0.566 [0.543, 0.588]
CLIP (ViT)	0.583 [0.567, 0.599]	**0.666** [0.644, 0.687]	0.499 [0.477, 0.522]
SigLIP-base	0.523 [0.507, 0.539]	0.291 [0.270, 0.312]	0.756 [0.736, 0.775]
SigLIP-so400m	**0.604** [0.588, 0.619]	0.573 [0.550, 0.595]	0.634 [0.612, 0.656]
SigLIP2-base	0.532 [0.515, 0.548]	0.283 [0.263, 0.304]	**0.780** [0.760, 0.798]
SigLIP2-so400m	0.513 [0.497, 0.529]	0.366 [0.345, 0.389]	0.660 [0.638, 0.682]

K Gaze controls and allocation comparisons

K.1 Training details

Optimization and checkpoint selection. All conditions train LoRA adapters (rank 16, scaling 32, dropout .05) on the language decoder's attention and MLP projections, with the vision encoder and projector frozen, using AdamW (2×10^{-5}, 5% warmup, cosine decay, effective batch 8, bf16) for at most six epochs of 242 updates with patience two, at a single optimization seed and a single gaze-loss weight. Every condition selects its checkpoint by Val agreement with the actual human choice, never by gaze fit or test performance.

Gaze targets. We construct sample-density gaze maps from the verified image-display interval using Gaussian smoothing ($\sigma = 40$ screen pixels), integrate them over each model's image-token regions, and normalize separately within each valid image. For +Gaze, $H_{t,i}$ is the map of the participant who supplied trial t's choice; for +AvgGaze, it is replaced by the fixed Train average for the same image position $i \in \{A, B\}$, with the original trial masks retained. +Gaze+Side instead jointly normalizes the human token masses and the readout over the union of both images, excluding boundary-crossing tokens. Its joint KL decomposes into a KL between the two-image mass distributions and within-image KL terms weighted by the corresponding human mass. Unlike +Gaze, +Gaze+Side does not mask an image solely because it contains no raw gaze samples; any mass contributed by Gaussian smoothing is retained.

Loss details. The gaze term is zero if no image is valid; the choice loss is retained, and the training pool has at least one valid image per trial. KL uses natural logarithms during training, with zero human bins contributing zero. Because gaze mass and choice coincide on most trials, +Gaze+Side partly re-supplies the choice label through the gaze channel; a choice gain under +Gaze+Side would

Table 20: Probability-based top-choice shape-match accuracy, adjective captions only. Values in brackets are Wilson 95% confidence intervals

Caption	Total	Round	Sharp
CLIP (RN50)	**0.872** [0.861, 0.883]	**0.751** [0.731, 0.770]	**0.993** [0.989, 0.996]
CLIP (ViT)	0.857 [0.845, 0.868]	0.735 [0.715, 0.755]	0.978 [0.971, 0.984]
SigLIP-base	0.519 [0.503, 0.535]	0.636 [0.614, 0.658]	0.402 [0.380, 0.425]
SigLIP-so400m	0.667 [0.651, 0.682]	0.361 [0.339, 0.383]	0.973 [0.964, 0.979]
SigLIP2-base	0.856 [0.844, 0.867]	0.744 [0.724, 0.763]	0.968 [0.959, 0.975]
SigLIP2-so400m	0.509 [0.493, 0.525]	0.323 [0.302, 0.345]	0.695 [0.674, 0.716]

Table 21: Individual ICA and human-choice agreement scores (%) with 95% confidence intervals on Test. These are intervals for each score, not for differences between scores. ICA measures agreement with the designed category; human-choice agreement measures agreement with the observed participant response. The reference row reports the human majority-vote predictor's agreement with participant responses. Model and reference intervals use the respective bootstrap constructions described above.

Model / reference	Condition	ICA (%) [95% CI]	Human agr. (%) [95% CI]
Gemma4-E4B	Zero-shot	53.69 [44.95, 62.21]	52.91 [44.90, 61.16]
	Choice-only	92.72 [86.36, 97.33]	72.72 [64.13, 80.61]
	+Gaze	92.04 [85.45, 96.92]	72.82 [64.78, 80.45]
	+AvgGaze	94.37 [88.45, 98.34]	73.01 [64.90, 80.60]
	+Gaze+Side	94.56 [88.86, 98.42]	72.62 [64.39, 80.39]
Qwen3-VL-8B	Zero-shot	49.90 [41.56, 58.16]	48.54 [40.43, 56.50]
	Choice-only	94.08 [88.16, 98.18]	72.91 [64.49, 80.62]
	+Gaze	92.43 [86.25, 96.99]	72.82 [64.41, 80.46]
	+AvgGaze	89.61 [81.87, 95.50]	71.75 [63.32, 79.50]
	+Gaze+Side	90.97 [84.29, 96.00]	71.94 [63.35, 79.74]
Molmo2-8B	Zero-shot	53.50 [44.86, 61.99]	50.00 [41.90, 57.94]
	Choice-only	92.52 [86.66, 96.89]	71.94 [63.45, 79.76]
	+Gaze	89.03 [81.36, 95.00]	73.11 [65.00, 80.67]
	+AvgGaze	89.71 [82.42, 95.52]	72.43 [64.21, 80.15]
	+Gaze+Side	89.22 [81.52, 95.30]	72.72 [64.52, 80.43]
Human majority-vote	—	—	72.86 [64.76, 80.96]

therefore not isolate gaze information beyond the label, whereas its absence is the stronger result (Section 5.4).

Alignment measures. We evaluate every condition on Test along three groups of measures. *Choice:* human-choice agreement against the actual choice in the same trial, and mean human-response negative log-likelihood (NLL) from the saved A/B probabilities (Section 4.1). NLL averages

Table 22: Paired differences in human-choice agreement between Choice-only and the human majority-vote reference, in percentage points. Positive values favor the model. Individual 95% intervals are conditional on the fixed model and reference predictions.

Model	Model minus reference (pp)	95% CI (pp)
Gemma4-E4B	−0.15	[−5.29, +5.00]
Qwen3-VL-8B	+0.05	[−4.80, +4.90]
Molmo2-8B	−0.92	[−6.20, +4.36]

$-\log_2 p(y_{\text{human}})$ across trials, where $p(y_{\text{human}})$ is the probability assigned to the participant's actual response; lower values indicate better probability predictions, and assigning A and B equal probability gives 1 bit. *Within-image distribution:* within-image gaze correlation and gaze KL. We compute KL using base-2 logarithms and report KL gain over train-mean in bits. Image-level values are averaged within each trial and then across trials; positive KL gain over train-mean indicates lower divergence than the fixed training-mean map (Eq. (2)). Equation 2 defines the KL reduction relative to the train-mean reference. KL matches the direction of the supervised objective. To assess whether the findings extend to other aspects of spatial alignment, we also compare Jensen–Shannon divergence (JS; symmetric distribution mismatch), histogram intersection (overlap in probability mass), and center-of-mass distance (separation between the maps' spatial centers). Lower JS and center-of-mass distance, and higher histogram intersection, indicate closer alignment. *Allocation between images:* We measure how human gaze and model attention are distributed between images A and B. First, the between-image JS compares the proportions of total gaze and attention assigned to the two images; lower values indicate more similar allocation. We also report two agreement rates: choice–attention side agreement, whether the model chooses the image receiving more model attention, and gaze–attention side agreement, whether the image receiving more model attention is also the one receiving more human gaze. These agreement rates can be high simply because both quantities favor the same side across trials. For each rate, we therefore compute the agreement expected if the two quantities were independent, using their respective overall A/B frequencies:

$$E = p_B\, q_B + (1 - p_B)(1 - q_B), \tag{3}$$

where p_B and q_B are the rates at which the two compared quantities (e.g., the model's choice and the image with more model attention) favor B. The observed-minus-expected gap indicates whether they agree more often than these side preferences alone would predict. It does not establish that attention to an image causes the model to choose it.

K.2 Reference maps and uncertainty

Reference maps. Alongside the train-mean map in Eq. (2), we use the two gaze references of Section 3.3: a center baseline (the Gaussian defined in Appendix H.2, centered within each image) and a population-mean map (the average gaze map of the other Test trials, computed separately for A and B, excluding the evaluated trial). These assess alignment with a central spatial prior and a shared gaze pattern, respectively. Reference maps use each model's native image-token geometry, with direct comparisons evaluated on the same valid images; they are empirical references, not performance ceilings.

Uncertainty. For model-condition comparisons, we use paired word×pair×participant cluster bootstraps with shared trial weights across conditions. The 95% confidence intervals are individual and unadjusted, conditional on the fitted seed and selected checkpoints; they do not capture variability from retraining or checkpoint selection. We apply Holm correction to exploratory p-values separately within two families: the 18 +Gaze–+AvgGaze comparisons (three models, six measures) and the nine choice-NLL comparisons among Choice-only, +Gaze, and +AvgGaze (three models, three condition pairs).

K.3 Trial-specific versus average-gaze supervision

The within-image correlations under +Gaze (.733/.729/.673) and +AvgGaze (.733/.731/.679) are close to the population-mean reference (.731/.731/.680) on the corresponding native token geometries. A fixed training-average target thus reproduces high correlation without trial-specific gaze targets, although similar correlations do not establish statistical equivalence. Table 23 reports mean gaze KL under +AvgGaze and +Gaze, providing the denominator for the relative reductions of 7.64/6.97/6.84%. Here, "mean gaze KL" denotes image-level KL averaged within and then across trials; the denominator is the KL of the +AvgGaze model's attention maps, not the KL of the fixed train-mean target or its ΔKL value. Improvements are not uniform across metrics: for Molmo2-8B, within-image correlation is higher under +AvgGaze, while KL point estimates favor +Gaze. Because +AvgGaze uses one map per image position, the remaining +Gaze advantage in KL may reflect image-specific gaze structure shared across participants rather than information specific to individual trials; a per-image average target would separate the two, and these single-seed comparisons do not.

Table 23: Mean gaze KL under +AvgGaze and +Gaze and the reduction relative to +AvgGaze. KL is computed per valid image, averaged within each trial and then across trials, as in Section 5.3. Relative reduction is $100 \times (\mathrm{KL}_{+\mathrm{AvgGaze}} - \mathrm{KL}_{+\mathrm{Gaze}})/\mathrm{KL}_{+\mathrm{AvgGaze}}$, calculated from unrounded values. These are point estimates; paired intervals and significance tests are in Table 24.

Model	+AvgGaze KL (bits)	+Gaze KL (bits)	Reduction (bits)	Reduction (%)
Gemma4-E4B	1.0624	0.9812	0.0812	7.64
Qwen3-VL-8B	1.0608	0.9869	0.0739	6.97
Molmo2-8B	1.1211	1.0444	0.0767	6.84

Table 24 reports all 18 exploratory +Gaze–+AvgGaze comparisons on the full Test set (1,030 trials; 2,028 valid images). The two-sided tests use centered errors from 1,000,000 paired word×pair×participant product-cluster bootstrap resamples, with weights shared across models and metrics. Holm correction is applied across the 18 tests; the centered 95% intervals are individual and unadjusted. The selected checkpoints are held fixed, so these results exclude training-seed and checkpoint-selection uncertainty.

The stars in Table 3 refer specifically to the +Gaze–+AvgGaze contrast in ΔKL. Since both conditions use the same fixed train-mean reference, it cancels in this difference:

$$\Delta\,\mathrm{KL}_{+\mathrm{Gaze}} - \Delta\,\mathrm{KL}_{+\mathrm{AvgGaze}} = \mathrm{KL}(H \,\|\, A_{+\mathrm{AvgGaze}}) - \mathrm{KL}(H \,\|\, A_{+\mathrm{Gaze}}),$$

where $A_{+\mathrm{Gaze}}$ and $A_{+\mathrm{AvgGaze}}$ denote the two models' attention maps, with the same averaging as in Eq. (2). Thus, the starred values indicate lower gaze KL than the +AvgGaze model, not significance against the fixed train-mean map or a test of ΔKL against zero. Markers are assigned only to the three +Gaze rows for this contrast.

Table 24: Exploratory +Gaze–+AvgGaze differences with individual 95% intervals, two-sided bootstrap p-values, and Holm-adjusted p-values across all 18 tests. Positive differences favor +Gaze for human agreement, ρ, histogram intersection, and ΔKL; negative differences favor +Gaze for JS and center-of-mass distance. Human agreement is in percentage points and ΔKL in bits.

Model	Measure	Difference [95% CI]	p	p_{Holm}
Gemma4-E4B	Human agreement (pp)	-0.19 [-3.89, +3.50]	0.9145	1.0000
	Gaze ρ	-0.0004 [-0.0035, +0.0027]	0.8012	1.0000
	Gaze JS	-0.0147 [-0.0215, -0.0079]	0.0003	0.0048
	Histogram intersection	+0.0232 [+0.0128, +0.0335]	0.0001	0.0023
	Center-of-mass distance	-0.0034 [-0.0054, -0.0014]	0.0020	0.0286
	ΔKL (bits)	+0.0812 [+0.0298, +0.1326]	0.0040	0.0478
Qwen3-VL-8B	Human agreement (pp)	+1.07 [-3.41, +5.54]	0.6063	1.0000
	Gaze ρ	-0.0022 [-0.0045, +0.0000]	0.0504	0.2520
	Gaze JS	-0.0105 [-0.0171, -0.0039]	0.0034	0.0447
	Histogram intersection	+0.0176 [+0.0074, +0.0278]	0.0014	0.0211
	Center-of-mass distance	-0.0028 [-0.0049, -0.0008]	0.0083	0.0660
	ΔKL (bits)	+0.0739 [+0.0254, +0.1223]	0.0048	0.0497
Molmo2-8B	Human agreement (pp)	+0.68 [-2.92, +4.28]	0.6778	1.0000
	Gaze ρ	-0.0061 [-0.0089, -0.0034]	0.0003	0.0048
	Gaze JS	-0.0096 [-0.0170, -0.0022]	0.0133	0.0801
	Histogram intersection	+0.0158 [+0.0049, +0.0266]	0.0062	0.0557
	Center-of-mass distance	-0.0031 [-0.0052, -0.0011]	0.0045	0.0497
	ΔKL (bits)	+0.0767 [+0.0202, +0.1331]	0.0109	0.0762

K.4 Allocation between images

Similar choice scores can coexist with different choice–attention relationships (Table 3). Under Choice-only, choice–attention side agreement is 93.11/59.32/47.48%, with gaps over independence of +43.12/+9.59/+0.51 pp. For Molmo2-8B under Zero-shot, the observed 80.68% agreement is below the 82.54% expected from its shared side preferences (Eq. (3)). These associations describe the measured readout and do not identify which image regions caused a choice.

For Gemma4-E4B, +Gaze and +AvgGaze both round to $\rho = .733$, yet choice–attention side agreement is 92.14% versus 53.20% (gaps over independence: +42.11 versus +3.83 pp). Their gaze–attention side agreement is 64.37% versus 46.21%, a difference of +18.16 pp [7.40, 29.03]. Under +AvgGaze, attention favors B on 96.12% of trials in this run. The corresponding between-image JS difference is −.00311 [−.00870, +.00250] nats, so the directional-agreement result does not extend to every measure of allocation. For Qwen3-VL-8B and Molmo2-8B, both +Gaze and +AvgGaze increase choice–attention side agreement relative to Choice-only; this change is therefore not unique to trial-specific gaze supervision.

+Gaze+Side reduces between-image JS relative to Choice-only in all three models, with individual paired intervals excluding zero (Table 25). Relative to +Gaze, only Gemma4-E4B has an interval entirely below zero. These allocation improvements are distinct from the uncertain choice differences reported in Section 5.4. The side comparisons use 20,000 paired word×pair×participant product-cluster bootstrap resamples, with shared weights across conditions and individual, unadjusted 95% percentile intervals conditional on the selected checkpoints. They are not included in the 18-test Holm family above.

Table 25: Paired differences in Side JS (nats), with individual 95% intervals. Negative values indicate closer alignment with the human two-image mass split under +Gaze+Side. These intervals are unadjusted and conditional on a single fitted seed.

Model	+Gaze+Side minus Choice-only [95% CI]	+Gaze+Side minus +Gaze [95% CI]
Gemma4-E4B	-0.00805 [-0.01350, -0.00328]	-0.00368 [-0.00664, -0.00094]
Qwen3-VL-8B	-0.02213 [-0.02997, -0.01477]	-0.00052 [-0.00237, +0.00127]
Molmo2-8B	-0.04174 [-0.05647, -0.02834]	-0.00078 [-0.00201, +0.00035]

L Position-swap evaluation

We evaluate the 15 small-model conditions and two additional larger zero-shot models on the same 1,030 Test trials. For each trial, we exchange the two image panels without mirroring either panel, keep the word and the A/B position labels fixed, and match the original and exchanged predictions by trial. Both layouts use the same 512×512 input size, prompt, and greedy decoding protocol. Responses are accepted as a JSON choice, an exact single A/B, or a whole JSON object enclosed in a code fence; otherwise, we use the A/B probability fallback. The same response rule is applied to all 17 conditions and both layouts.

Swap consistency is the fraction of trials selecting the same shape in both layouts, which requires opposite A/B answers after the exchange. An always-A or always-B strategy scores 0%; independent uniform random answers have expected consistency of 50%. Position-balanced ICA is the mean of original-layout and exchanged-layout ICA, with the intended side reversed in the exchanged layout. Each original trial has equal weight, and its two layouts contribute equally; the two layouts are not treated as independent trials. These controls assess sensitivity to position: even perfect swap consistency does not establish a word-dependent association, and averaging layouts does not remove every possible shortcut. The larger models are additional zero-shot references; comparisons with the smaller models do not isolate the effect of parameter count.

Table 26 reports swap consistency and position-balanced ICA for all conditions. Table 27 additionally reports position-balanced agreement with the original human choice, together with the model's B-choice rate in each layout. For the exchanged layout, the human reference is the originally selected shape moved to its new position; it is not a newly observed human response to the exchanged stimulus.

We do not evaluate human-gaze alignment in the exchanged layout because no human gaze was recorded for that presentation.

All intervals use 20,000 product-cluster bootstrap resamples over words, image pairs, and participants (seed 20260921), with weights shared across conditions and across the two layouts of each trial. The reported intervals are individual, unadjusted 95% percentile intervals, conditional on the fixed checkpoints; they exclude training-seed and checkpoint-selection uncertainty.

Under Choice-only, swap consistency rises to 90.10/96.89/88.45% (from 30.97/8.25/19.51% under Zero-shot), while position-balanced ICA reaches 92.72/94.56/91.50%. These results show that the learned choices largely persist across positions and cannot be explained by always choosing A or always choosing B. The two larger zero-shot models, Gemma4-26B and Qwen3-VL-30B, reach swap consistency of 57.38/54.56% and position-balanced ICA of 61.99/60.10%, respectively. Swap consistency alone does not establish a word-dependent shape association.

Table 26: Position-swap results for all 17 conditions on Test. Values are percentages with individual 95% confidence intervals. Balanced ICA averages intended-category agreement across the original and exchanged layouts.

Model	Condition	Swap consistency [95% CI]	Balanced ICA [95% CI]
Gemma4-E4B	Zero-shot	30.97 [20.97, 41.73]	52.48 [47.34, 57.86]
	Choice-only	90.10 [82.76, 95.66]	92.72 [87.39, 96.72]
	+Gaze	93.01 [87.26, 97.20]	91.84 [85.76, 96.46]
	+AvgGaze	94.95 [89.43, 98.63]	94.76 [89.70, 98.26]
	+Gaze+Side	94.66 [88.92, 98.43]	94.42 [89.25, 98.05]
Qwen3-VL-8B	Zero-shot	8.25 [2.65, 15.88]	51.31 [48.70, 54.61]
	Choice-only	96.89 [93.06, 99.30]	94.56 [88.97, 98.32]
	+Gaze	89.22 [81.30, 95.42]	92.18 [86.79, 96.43]
	+AvgGaze	91.26 [85.09, 95.97]	90.39 [83.46, 95.70]
	+Gaze+Side	87.38 [79.46, 93.70]	91.36 [85.92, 95.68]
Molmo2-8B	Zero-shot	19.51 [10.90, 29.37]	54.22 [49.91, 58.97]
	Choice-only	88.45 [80.57, 94.58]	91.50 [85.92, 95.90]
	+Gaze	82.72 [72.77, 91.07]	87.57 [80.68, 93.35]
	+AvgGaze	82.91 [72.99, 91.34]	88.35 [81.65, 94.06]
	+Gaze+Side	84.37 [74.69, 92.43]	88.30 [81.40, 94.16]
Gemma4-26B-A4B	Zero-shot	57.38 [46.33, 67.85]	61.99 [54.48, 69.33]
Qwen3-VL-30B-A3B	Zero-shot	54.56 [43.74, 65.16]	60.10 [53.04, 67.10]

Table 27: Position-balanced human-reference agreement and B-choice rates for all 17 conditions. Values are percentages with individual 95% confidence intervals. Human-reference agreement uses the participant's originally selected shape in both layouts; the exchanged-layout reference is not a new human observation. Original B and swapped B are the model's B-choice rates before and after exchanging the images.

Model	Condition	Balanced human ref.	Original B	Swapped B
Gemma4-E4B	Zero-shot	51.99 [47.39, 56.80]	18.74 [11.42, 27.42]	13.40 [7.63, 20.42]
	Choice-only	72.62 [64.52, 80.14]	45.15 [36.67, 53.48]	46.31 [38.11, 54.77]
	+Gaze	72.52 [64.64, 79.89]	49.13 [40.92, 57.24]	49.71 [41.73, 57.90]
	+AvgGaze	72.82 [64.75, 80.26]	49.32 [41.14, 57.20]	49.71 [41.63, 57.99]
	+Gaze+Side	72.48 [64.49, 80.02]	47.96 [39.68, 56.00]	48.25 [40.17, 56.63]
Qwen3-VL-8B	Zero-shot	50.92 [48.65, 53.71]	96.12 [91.59, 99.12]	95.63 [90.66, 98.96]
	Choice-only	72.91 [64.59, 80.55]	48.83 [40.77, 56.79]	50.00 [42.06, 57.99]
	+Gaze	72.18 [64.28, 79.54]	45.44 [37.19, 53.52]	45.15 [36.80, 53.73]
	+AvgGaze	71.75 [63.79, 79.21]	51.94 [43.76, 59.88]	52.14 [43.96, 60.38]
	+Gaze+Side	72.62 [64.77, 79.89]	56.21 [47.82, 64.26]	56.02 [47.79, 64.29]
Molmo2-8B	Zero-shot	52.28 [48.40, 56.54]	89.81 [83.25, 95.11]	90.10 [83.66, 95.20]
	Choice-only	71.02 [62.98, 78.63]	46.70 [37.92, 55.48]	47.77 [39.25, 56.45]
	+Gaze	71.55 [64.00, 78.72]	41.84 [33.18, 50.46]	41.84 [33.26, 50.79]
	+AvgGaze	71.46 [63.82, 78.73]	42.14 [33.62, 50.85]	41.55 [32.99, 50.61]
	+Gaze+Side	71.70 [63.92, 79.08]	42.43 [33.69, 51.17]	42.72 [34.20, 51.59]
Gemma4-26B	Zero-shot	56.84 [50.33, 63.36]	29.03 [21.16, 37.44]	30.29 [22.08, 39.27]
Qwen3-VL-30B	Zero-shot	57.86 [51.65, 64.24]	69.32 [60.21, 77.68]	68.54 [59.95, 76.62]