PROBING FINE-GRAINED SEMANTIC ALIGNMENT ACROSS VISION AND LANGUAGE WITH MINIMAL PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

To what extent does alignment between vision and language preserve specific semantic distinctions? We introduce a new benchmark that examines this question across shallow mappings between independently trained vision and language models, CLIP-style contrastive models, and generative vision-language models. Each benchmark item leverages a minimal pair (of pairs) design comprising two minimally contrasting images and two corresponding captions, with contrasts spanning object properties and relations, physical events, and social interactions. By testing whether similarity scores or matching decisions consistently favor the correct image–caption correspondences over the mismatched alternatives, the benchmark probes whether alignment preserves the specific semantic distinction targeted by each item. Across these settings, we find that alignment is highly conditional. Shallow unimodal alignment recovers some shared object and scene structure, but performs poorly on minimal pairs that require preserving specific relational, causal, or social distinctions. Multimodal training improves performance, but gaps remain even for stronger models compared to human baselines. Further analyses show that implicit captions are consistently harder than explicit captions, suggesting that current models are better at matching named visual properties than inferring latent consequences from world knowledge. Together, our results suggest that broad semantic structure may support cross-modal matching, but fine-grained inferential distinctions needed to match closely related images and captions remain fragile even in multimodal systems.

1 INTRODUCTION

Connecting images with language is central to image retrieval, captioning, and visual question answering. Contrastive models such as CLIP learn to assign higher similarity to matching images and captions (Radford et al., 2021). Generative vision-language models (VLMs) learn to use images when producing descriptions or answering questions (Liu et al., 2023). Image-text associations can also be established through shallow mappings between vision and language models trained separately (Merullo et al., 2023; He et al., 2025), suggesting that shared semantic structure can emerge without joint training. This is consistent with the Platonic Representation Hypothesis, which draws on evidence of greater alignment in larger models to propose that models converge toward a shared representation of the world as they scale (Huh et al., 2024). These approaches connect images and language in different ways, but raise a common question. How reliably does this alignment preserve specific differences in meaning?

Broad agreement between visual and linguistic representations may be driven by shared information about object identity and scene category. Such information can support image-caption matching when the alternatives describe very different scenes, without requiring alignment of finer semantic distinctions. For example, recognizing the people and setting may help match an image to a caption among unrelated alternatives, but cannot resolve a choice between captions describing comforting and teasing in the same setting. Even if visual and linguistic representations each distinguish these interactions, the representational changes associated with that distinction may not align across modalities. Testing this requires comparisons that hold the broader scene content fixed, so that the specific semantic distinction determines the correct match.

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.

Minimal-pair benchmarks offer a way to make these comparisons. Winoground tests whether models can match images to captions that use the same words in different orders (Thrush et al., 2022). VisMin tests changes in object identity, attributes, counts, and spatial relations using minimally different images and captions (Awal et al., 2024). We build on this approach to test whether models correctly match minimally contrasting images to their corresponding captions. Our new benchmarks provide a systematic evaluation of perspectival, physical-causal, and social knowledge. Perspectival tests examine whether models connect viewing distance and action context to an object's apparent size. Physical-causal tests cover the states of individual objects and entire scenes, including what produces a state, what it implies about preceding events, and which actions it permits. Social tests span individual emotional states, interactions such as comforting versus teasing and cooperating versus competing, and qualities such as friendliness versus hostility. Even when captions explicitly name a social category, matching requires interpreting people's behavior in context.

A second dimension examines how alignment depends on the inferences needed to connect a caption to an image. We pair the same images with captions that either describe the relevant distinction directly or leave it to be inferred from an event, a prior location, an intended action, or a social context. This allows us to ask both which semantic distinctions models reliably match across modalities and whether they can still make those matches when the relevant meaning is implicit, while holding the visual evidence fixed.

We compare contrastive and generative vision-language models with shallow mappings between separately trained vision and language models. Across these approaches, broad agreement about objects and scenes does not always extend to the finer distinctions needed to interpret physical events or social interactions. Multimodal models generally do better, but still struggle when captions leave the relevant meaning implicit. Together, these findings challenge strong interpretations of the Platonic Representation Hypothesis and call for multimodal evaluation that tests shared world knowledge one semantic distinction at a time, using minimal pairs of pairs across knowledge domains.

2 The Benchmark

2.1 Minimal pairs of pairs

Each item consists of two images and two captions. The images differ in one respect; the captions differ in one respect; caption c_0 describes image v_0 and caption c_1 describes image v_1. Everything else — the scene, the people, the camera, the lighting — is held fixed. This ensures that only the target distinction decided the correct match. We call this a *minimal pair of pairs*.

We evaluate each item by matching images to captions and captions to images. Each directional score requires both matches to be correct, and the group score requires all four (§3.2). This tests whether the targeted distinction supports consistent matching across modalities, preventing a fixed preference for one image or caption from earning group credit. The suite varies both the knowledge needed to distinguish the images (§2.2), and whether that distinction is named in the caption or must be inferred (§2.3).

2.2 Dimension 1: Knowledge Domains

Our new items cover three domains, complementing the object, attribute, counting, and spatial contrasts from VisMin (Awal et al., 2024).

Physical-causal. *Object state* items contrast two states of the same object in the same scene, such as a folded versus unfolded umbrella or a melted versus unmelted chocolate bar. *Scene event* items contrast event outcomes within a shared scene, including which object changes state, how a material responds to the same exposure, and what trace an event leaves. Together, they test whether visible states and outcomes can be matched to descriptions of those states, their causes, or the actions they allow.

Perspectival. These items test whether models relate an object's apparent size to its distance from the viewer. Image pairs show the same object near or far from a fixed camera. Captions describe the viewing distance or an action that implies it, such as reaching for handlebars.

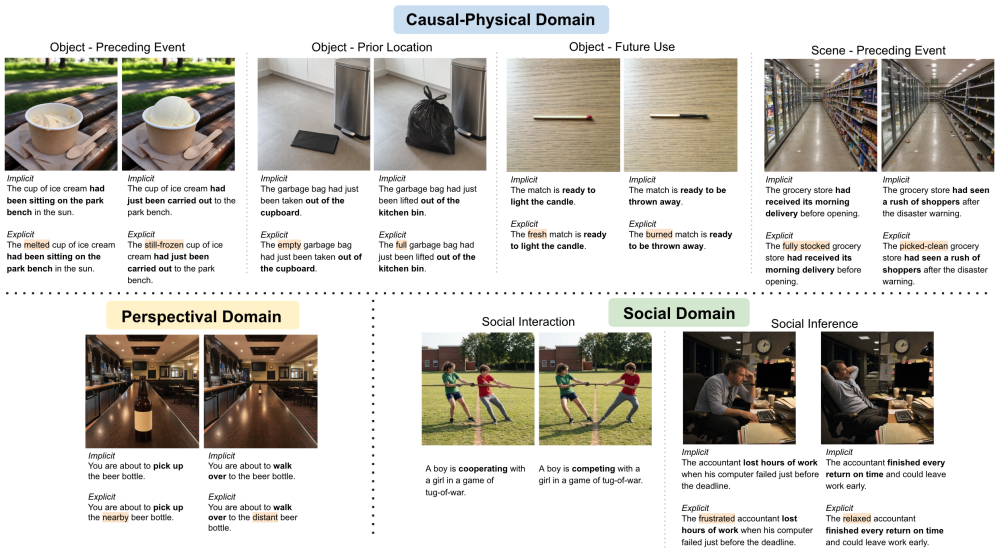

Figure 1: Overview of our benchmark, which consists of image-text pairs of pairs with minimal changes in three domains of knowledge: causal-physical (object and scene), perspective, and social. The causal-physical domain is further divided into three distinct subgroups based on the kind of inference: past event, past location, and future use. The social domain is further divided into social interactions and social inference.

Social. We distinguish two subdomains. *Social interaction* items contrast comforting versus teasing, cooperating versus competing, and friendly versus hostile behavior between the same people in the same setting. Their explicit caption pairs differ only in the social label, with the surrounding context held fixed. *Social inference* items cover both individual emotional states and interpersonal interactions, with captions that relate the depicted emotion or behavior to a preceding event. Their explicit caption pairs differ in both the event described and the named emotion or interaction. Implicit captions retain the event description but omit the social label, testing whether models can connect what happened earlier to the emotion or behavior shown.

2.3 Dimension 2: Explicit and Implicit Captions

The second dimension varies how much inference a caption asks for, with the images held fixed. *Explicit* captions name the relevant state, property, or social category. *Implicit* captions describe a context from which it must be inferred. For images of a melted and an unmelted chocolate bar, an explicit caption reads *The melted chocolate bar had been left on the dashboard during a hot afternoon."* Its implicit counterpart reads *The chocolate bar had been left on the dashboard during a hot afternoon."* Adding only *melted* allows us to measure how naming the state changes matching performance, with the images and remaining wording unchanged.

Implicit captions differ in *which* inference they ask for. Within the physical-causal domain, we distinguish three types of implicit captions. *Past event* captions describe something that happened to an object, such as a match having just been struck, from which its resulting state must be inferred. *Past location* captions describe where an object was previously, such as a match having been taken from an ashtray, leaving its current state unstated. *Future use (affordance)* captions describe an intended action, such as using a match to light a candle, requiring the state needed for that action to be inferred. Within the perspectival domain, implicit captions describe a *viewpoint* or an action that implies one, such as preparing to take hold of handlebars. Matching requires inferring the object's apparent size from the described distance or action. Within the social domain, implicit captions describe a *social context*, including preceding events, from which the depicted emotion or interaction must be inferred.

For physical and perspectival items, explicit captions name a visible property that distinguishes the images. For social items, naming a category such as teasing does not specify the visual cues

that distinguish it from other interactions. Models must still connect the named category to the expressions, gestures, and actions shown in the image.

2.4 CONSTRUCTION

Of the five subsets, two adapt stimuli from sentence-picture verification studies (Pecher et al., 2009; Connell, 2007; Winter & Bergen, 2012), one depicts social concepts from a text-only world-knowledge benchmark (Ivanova et al., 2025), and two were developed for this suite. All images are generated.

Construction follows five constraints. Each image pair shares a setting, identities, camera position, and lighting. Unrelated differences in wording, objects, or props must not provide alternative matching cues. Each implicit caption must favor its intended image on its own, without naming the target, a close synonym, or its diagnostic visual feature. Its explicit counterpart adds one everyday word naming the target. For causal items, the other image must not depict the described cause.

We generate a base image and edit it to produce both members of a pair, constraining the edits to preserve identities, layout, camera position, and lighting apart from changes required by the target distinction. Apparent-size items use a base image containing near and far instances of the same object. Removing one instance at a time produces the pair while preserving the shared background (Appendix A.1).

We select contrasts that can be shown in natural scenes while leaving the context described by an implicit caption outside the image. An object's shape, state, or color can be inferred from its history or intended use, and apparent size from viewing distance. When adapting earlier stimuli, we replace present-location descriptions such as "in the bag" versus "in the salad" with history or intended use, removing lexical shortcuts such as *salad* $\rightarrow$ *sliced*. Every item was audited for asymmetric props, duplicate captions, and template defects, and every image is verified against a content hash recorded at build time.

3 EVALUATION

3.1 MODELS AND HUMAN PARTICIPANTS

We compare three ways of connecting images and language using the same benchmark items and matching criteria. Human judgments provide a reference for each comparison.

Post-hoc shallow alignment of separately trained unimodal models. We align a pretrained vision-only and a pretrained language-only model into a shared representational space using only a lightweight post-hoc alignment. Here, we use orthogonal barycentric alignment, fitted on 5,000 MSCOCO image-caption pairs. This preserves within-model geometry while reconciling the representational axes across modalities. Crucially, the alignment uses no benchmark labels or task-specific supervision. Images are represented by the CLS token from a vision-transformer block, and captions by the mean token representation from a language-transformer block. Features are z-scored dimension-wise before fitting the rotations. After alignment, image-caption compatibility is scored by cosine similarity in the shared space, allowing each benchmark item to be evaluated with the same text, image, and group criteria described above.

Our primary unimodal pair is DINOv2-giant and Llama-3.1-8B-Instruct. Because the optimal representational depth is not known a priori, we evaluate six depth combinations: language depths 70%, 90%, and 100%, crossed with vision depths 90% and 100%. We report both the mean across depth cells and the best cell as an oracle score selected by mean explicit group score. To assess scaling, we vary one model family while holding the other fixed: DINOv2-S/B/L/g with Llama-3.1-8B-Instruct, and Qwen2.5-0.5B/1.5B/3B/7B with DINOv2-giant. Additional alignment methods, full depth grids, and model variants are reported in Appendix A.4.

Contrastive models. We evaluate CLIP ViT-B/32, ViT-L/14 and ViT-L/14-336 (Radford et al., 2021). Captions and images are embedded with the model's projection heads and compared by cosine similarity in the joint space the model was trained to produce (Appendix A.2).

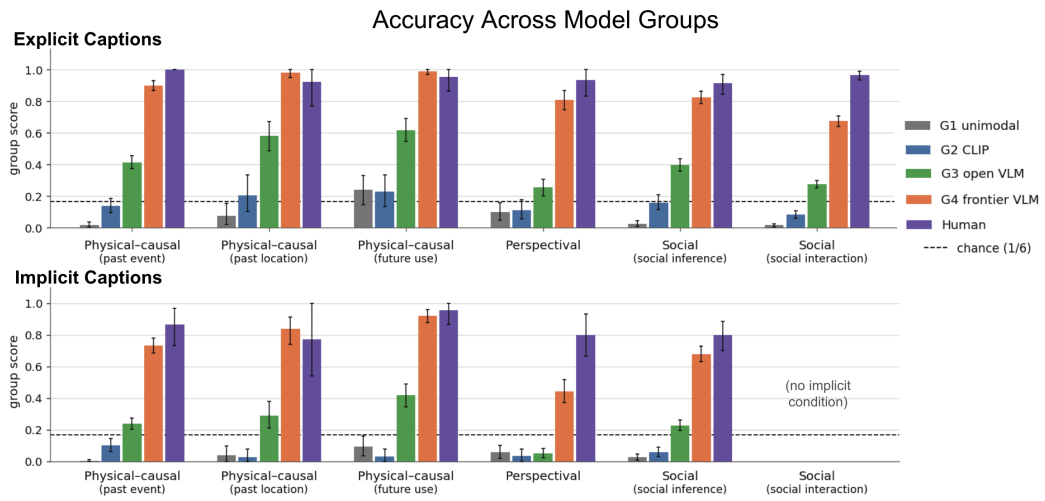

Figure 2: Group accuracy by approach across subsets and caption inference conditions. Top: explicit captions; bottom: implicit captions. Bars are mean over models within each approach - post-hoc alignment between unimodal models, CLIP, open VLM, and frontier VLM–and human raters; error bars are 95% item-boostrap intervals.

Generative models. We evaluate Qwen2.5-VL 3B, 7B and 32B, Idefics2-8B, LLaVA-OneVision-7B and LLaVA-1.6-Mistral-7B alongside four closed-source frontier models accessed through their APIs: Claude Sonnet 4.5, Claude Fable 5.1, GPT-5.6 Terra and Gemini 3.6 Flash. Each model receives either an image with two candidate captions or a caption with two candidate images and selects the matching option (Appendix A.2).

Human judgments. Participants recruited on Prolific answered the same forced-choice questions for every subset, with confidence ratings and response times recorded (Appendix A.3).

3.2 Scoring

Following Winoground and VisMin, each item is scored in both directions. An item is **text**-correct if each image is matched to its own caption rather than the other one, **image**-correct if each caption is matched to its own image, and **group**-correct if both hold. Matching is by cosine similarity for similarity-based models and by the forced-choice answer for generative models and humans. Because a fixed preference for one caption or one image is right for one member of the pair and wrong for the other, only an item whose distinction is preserved across modalities can be group-correct, which makes the group score our primary measure. Prompts follow VisMin with an added answer-format instruction, decoding is greedy, and unparseable answers count as missing rather than incorrect (none occurred).

3.3 Human references and statistical comparisons

We compare model performance with human consensus and a single-rater baseline (§ 4.3). For similarity-based models, conventional chance is $1/4$ for text and image accuracy and $1/6$ for group accuracy, assuming all rankings of the four image-caption scores are equally likely. Because this may not hold for minimal pairs, we use an empirical permutation baseline: caption pairs are randomly reassigned to image pairs 200 times within each condition, preserving within-pair representations while breaking the cross-modal correspondence. We report observed accuracy, the permutation baseline, and a permutation p-value. For forced-choice models, random guessing gives text and image accuracy of $1/4$ and group accuracy of $1/16$, since group correctness requires four correct binary choices. Averaging across option orders leaves these expectations unchanged.

4 RESULTS

4.1 SHALLOW UNIMODAL ALIGNMENT IS INSUFFICIENT FOR FINE-GRAINED SEMANTIC DISTINCTIONS

We first test whether a shallow mapping between independently trained vision and language models supports fine-grained image-caption matching. We compare performance with the permutation baseline for each condition (for text and image accuracy figures, see Appendix A.4).

On VisMin, the map recovers object identity well above its null (text 0.54 vs. 0.14, image 0.37 vs. 0.08, group 0.025 vs. 0.03; all $p < 0.005$) and attributes reliably (text 0.30 vs. 0.12, image 0.13 vs. 0.06, group 0.09 vs. 0.02), while counting and spatial relation sit at their nulls on every score (Figure 3, left). Across our new benchmarks (Figure 3, right), physical–causal items with caption pointing to *future use* of an object state is the only benchmark with explicit-caption performance above baseline in both directions and in the group score (text 0.50 [0.22], $p < 0.005$; image 0.45 [0.18], $p = 0.005$; group 0.35 [0.11], $p < 0.005$).

The remaining subsets show little evidence of consistent matching across modalities. Past-location items exceed the null only on the text score (0.54 vs. 0.15, $p < 0.005$), with image and group scores of 0.08. Past-event items exceed the null only on the text score (0.19 vs. 0.10, $p < 0.05$), with a group score of 0.01. Perspectival and social-interaction items do not exceed the null on any score. Social-inference items show an advantage only on the image score (0.20 vs. 0.12, $p < 0.05$), which does not extend to the group score.

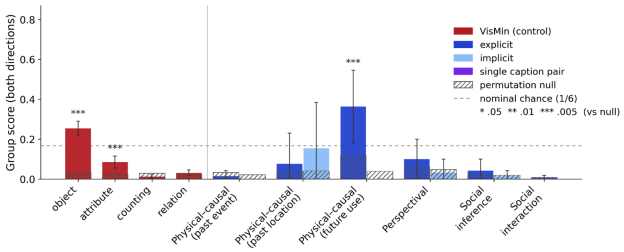

Figure 3: Unimodal model post-hoc alignment: group scores on VisMin and on our benchmarks. Llama-3.1-8B-Instruct aligned with DINOv2-giant, best depth cell. Left of the divider: VisMin by category; right: our subsets with explicit and implicit captions.

Explicit vs. implicit. Leaving the target property implicit substantially reduces performance on future-use items (Figure 3). The group score falls to 0.00 against a null of 0.04, while text and image scores fall to 0.25 and 0.05, respectively. A paired within-item analysis at the same layer combination shows declines of 40 percentage points for image (CI: [20, 60]; exact McNemar $p = .008$), 35 for group (CI: [15, 55]; $p = .016$), and 25 for text (CI: [0, 50]; $p = .13$). For past-location and past-event items, whose explicit scores exceeded the null only on the text score, the declines are not significant (text: 23 points, $p = .25$, and 7 points, $p = .18$). Across the evaluated model pairs and layer combinations, implicit scores do not exceed the null in 87–92% of the conditions in which explicit scores do. Further, the explicit–implicit gap is concentrated in conditions where explicit captions already support above-null performance. Together, these results suggest that naming the relevant property helps matching, but the mapping does not support consistent matches when that property must be inferred from the description.

Scaling improves performance selectivity. Scaling the language model (Qwen2.5 0.5B $\rightarrow$ 7B) improves the explicit physical-causal signal for prior location and future use (future use: text 0.32 $\rightarrow$ 0.47, image 0.18 $\rightarrow$ 0.33, group 0.14 $\rightarrow$0.22; past location: text 0.31 $\rightarrow$ 0.40, image 0.13 $\rightarrow$ 0.28, group 0.08$\rightarrow$ 0.21; Figure 4, top). Implicit captions and the other benchmarks remain at baseline. Increasing the vision encoder from DINOv2-S to giant with Llama fixed raises VisMin object and attribute scores monotonically (object group 0.19 $\rightarrow$ 0.28; attribute text 0.26 $\rightarrow$ 0.34) and the future-use scores (group 0.14 $\rightarrow$ 0.22), but not the past-location scores, which decline on the image and group score (0.23 $\rightarrow$ 0.14 and 0.12 $\rightarrow$ 0.06) while rising on the text score (0.35 $\rightarrow$ 0.46); past-event items and every other subset remain at their nulls (Figure 4, bottom). Scaling the vision encoder (DINOv2-S $\rightarrow$ giant) consistently improves VisMin object and attribute matching (object group 0.19 $\rightarrow$ 0.28) and the future-use scores (group 0.14 $\rightarrow$0.22), but not the past-location and past-event scores. Every other subset remain at their nulls. Scale therefore amplifies contrasts that are already linearly readable and creates none. These findings challenge the Platonic Representation

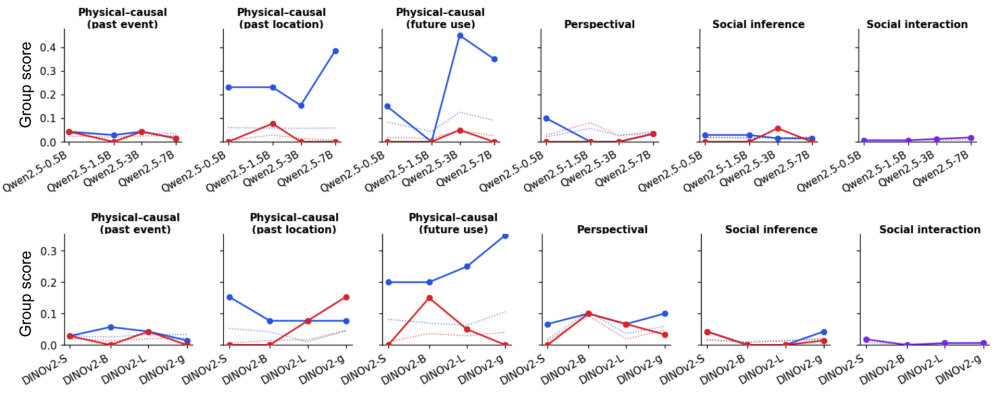

Figure 4: **Shallow alignment does not become more general with model size.** Group score as a function of language-model size (Qwen2.5 0.5B–7B, DINOv2-giant fixed; top) and vision-model size (DINOv2 small–giant, Llama-3.1-8B-Instruct fixed; bottom), per subset, for explicit (blue) and implicit (red) captions; social interaction has a single caption pair. Dotted lines are the permutation null of each condition. Scores are at the best depth cell for each model pair.

Hypothesis (Huh et al., 2024) and suggest that while vision and language models may appear aligned when evaluated on broadly sampled images and text, this alignment does not extend to fine-grained semantic distinctions.

4.2 Multimodal training improves matching, but implicit captions remain difficult

We next evaluate CLIP-style contrastive models and generative VLMs trained on paired images and text. We score CLIP using the same similarity-based protocol and permutation baselines as the shallow mappings. For generative VLMs, we average forced-choice matching scores across option orders, following the balanced-position aggregation of Wang et al. (2024).

CLIP-style contrastive training improves explicit matching but leaves perspectival and inferential gaps. Contrastive training improves over shallow alignment on most explicit benchmark components, including physical-causal, social-pair, and social-inference items. Perspectival items are a clear exception, with no reliable improvement over the shallow mapping (Figure 2).

The same pattern appears in VisMin, which we use as an existing benchmark comparison. CLIP strongly improves over the shallow map for object identity and attributes, but spatial relations remain at the permutation null (text 0.10 [0.07], image 0.05 [0.04], group 0.01 [0.01], Awal et al. (2024)) and close to the shallow alignment (0.09, 0.11, 0.03). These captions explicitly name the relation, showing that the difficulty persists even when the relevant distinction is stated directly. Both VisMin and our perspectival benchmark point to a more general limitation of CLIP-style similarity spaces for spatial or perspectival distinctions.

CLIP's advantage also shrinks when captions leave the target distinction implicit. Much of the improvement observed with explicit captions disappears, particularly in subsets where explicit performance is only modestly above baseline. Paired image-text training therefore supports more of the tested distinctions, but does not ensure reliable matching when their meaning must be inferred.

Larger VLMs improve implicit matching but retain a substantial gap. Within the Qwen2.5-VL family, increasing model size from 3B to 7B to 32B raises group scores on every benchmark with both explicit and implicit captions (Figure 5). Improvements are larger for image selection than caption selection, with image scores increasing by roughly three times as much as text scores. Despite these gains, explicit captions retain an advantage. The absolute explicit-implicit gap increases with model size (0.149, 0.248, and 0.279), while the gap relative to explicit performance decreases (49%, 40%, and 37%). Scaling therefore improves implicit matching, including relative

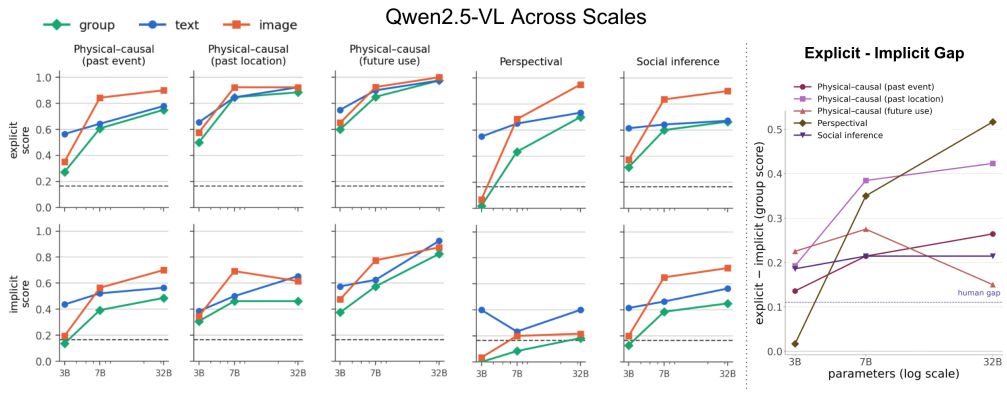

Figure 5: **Scale improves matching in Qwen2.5-VL but does not close the explicit–implicit gap.** Left: text, image and group scores for Qwen2.5-VL 3B, 7B and 32B on each subset, with explicit (top) and implicit (bottom) captions. Right: the explicit–implicit difference in group score at each size; the dashed line is the same difference for human raters.

	Physical–causal (past event)	**Physical–causal (past location)**	**Physical–causal (future use)**	**Perspectival**	**Social inference**	**Social interaction**
Explicit						
Claude Fable 5.1	.971	1.000	1.000	1.000	.900	**.678**
GPT-5.6-terra	**.914**	.962	1.000	.883	.871	**.780**
Gemini 3.6 Flash	**.964**	1.000	1.000	.850	.893	**.901**
Claude Sonnet 4.5	**.743**	.962	.950	**.500**	**.636**	**.337**
Human consensus	*1.000*	*.923*	*.950*	*.933*	*.914*	*.964*
Single rater	*.827*	*.838*	*.913*	*.857*	*.830*	*.908*
Implicit						
Claude Fable 5.1	.871	.885	1.000	**.500**	.850	–
GPT-5.6-terra	.829	.923	.975	.633	.764	–
Gemini 3.6 Flash	**.736**	.885	.950	**.400**	.693	–
Claude Sonnet 4.5	**.493**	.654	**.725**	**.233**	**.407**	–
Human consensus	*.867*	*.769*	*.950*	*.800*	*.800*	–
Single rater	*.724*	*.732*	*.898*	*.611*	*.722*	–

Table 1: Group score by knowledge domain, inference route, and caption condition. Bold values are significantly below the human consensus under the paired within-item test.

to explicit performance, but does not eliminate the difficulty of matching distinctions that captions leave unstated.

4.3 Frontier Models and Human Baselines

We evaluate four closed-API vision-language models, Claude Fable 5.1, Claude Sonnet 4.5, Gemini 3.6 Flash, and GPT-5.6-terra, against human responses collected on the same items using the same matching protocol (§4). Each matching question received three independent human ratings. The *human consensus* uses their majority vote. The *single-rater baseline* samples one rating per question and averages the resulting scores over 400 draws. Both references use the same text, image, and group correctness criteria as the models. Model-human comparisons are paired within item and restricted to the items covered by each human study, using bootstrap confidence intervals and exact sign tests. (§3). Results are shown in Table 1.

Strong models approach human consensus on explicit captions. With the property word present, three of the four frontier models are near the human-consensus ceiling on most domains, except social interaction, where all four fall significantly below it. The strongest models often exceed

the single-rater baseline: GPT-5.6-terra does so in ten of eleven conditions and Fable 5.1 in nine. However, none significantly exceeds the three-rater consensus in any condition; across 44 model $\times$ domain $\times$ level comparisons, 28 are indistinguishable from consensus and 16 are below it. Thus, the best frontier models outperform an average individual respondent, but match rather than exceed human consensus.

Implicit captions reveal a persistent perspectival gap. Implicit captions reduce accuracy for both models and humans, but the remaining model-human differences vary across domains (Figure 7a). GPT-5.6-terra does not differ significantly from consensus in any of the five domains, while Fable falls significantly below it only on perspectival items. Across the five domains in the tables, models fall below consensus in seven of twenty implicit conditions, compared with five of twenty explicit conditions. Perspectival items show the most consistent human advantage. Model group scores range from 0.233 to 0.633, compared with 0.800 for consensus and 0.611 for the single-rater baseline. Three of the four models fall significantly below consensus. In contrast, the three strongest models score at or above consensus on both object-state routes with both caption types. The difficulty of inferring apparent size from viewpoint therefore persists in several frontier models, extending the weakness observed in CLIP.

Human consensus shows the smallest explicit-implicit decline. On the 163 items evaluated by every model and by humans under both caption conditions, the mean decline in group accuracy is 0.110 for human consensus (95% CI [.055, .172]), $+.113$ [.071, .160] for GPT-5.6-terra, $+.135$ [.080, .193] for Fable 5.1, $+.236$ [.178, .298] for Gemini 3.6 Flash, and $+.236$ [.175, .298] for Sonnet 4.5. Human consensus has the smallest estimated decline, though GPT and Fable are statistically indistinguishable from it.

5 Discussion

Our results challenge the interpretation of broad convergence between vision and language models (Huh et al., 2024) as evidence of a shared representation of the world. Prior work has demonstrated correspondences between separately trained vision and language models (He et al., 2025), and increasing agreement with scale has motivated the Platonic Representation Hypothesis (Huh et al., 2024). Yet the shallow mappings tested here rarely support consistent matching when images and captions differ in specific physical, perspectival, or social distinctions. Models may organize objects and scenes similarly without aligning the differences that determine what happened, what could happen next, or how people are interacting.

The benchmark connects this question about representations to controlled evaluations of world knowledge. Winoground and VisMin test image-text contrasts, while EWoK uses controlled text contrasts to assess knowledge across domains (Thrush et al., 2022; Awal et al., 2024; Ivanova et al., 2025). Our minimal pairs of pairs extend this approach to whether physical-causal, perspectival, and social distinctions support consistent matching across modalities, both when captions name the relevant meaning and when they leave it implicit. This contrast matters even for multimodally trained systems: CLIP improves explicit matching but retains spatial and inferential weaknesses, while stronger generative models approach human consensus on some comparisons but still struggle with implicit perspectival judgments. This motivates evaluations of which aspects of world knowledge models connect across vision and language, especially when the connection must be inferred.

We identify several limitations and future directions for this work. First, generated images allow closely controlled comparisons but may introduce artifacts or depict physical and social situations imperfectly and testing the same contrasts in photographs and videos would assess how well the findings generalize. Matching performance also combines visual discrimination, language interpretation, and cross-modal comparison. Hence, failure alone cannot establish whether a distinction is absent from an individual model or present but poorly aligned across models. Future work should pair the benchmark with within-modality analyses of model representations to separate these possibilities. A promising next step is to generate minimal pairs at scale and use them in controlled training studies to identify which kinds of examples help models learn distinctions they currently struggle to match. Ultimately, evidence for shared world knowledge must show that vision and language agree on the differences that change a scene's meaning.

6 Ethics Statement

Human judgments for the different sub-benchmarks were collected through separate studies on Prolific. All participants were adults, provided informed consent before participation, and could withdraw at any time by exiting the study. The task consisted of two-alternative forced-choice judgments over everyday scenes and posed no more than minimal risk; all visual stimuli were synthetic images generated by text-to-image models and did not depict real identifiable individuals. Across studies, participants were compensated at effective hourly rates ranging from $13.11 to $19.30 based on observed completion times. The study protocol was approved by the authors' institutional review board (IRB). We collected only task responses, confidence ratings, response times, and Prolific IDs required for payment; Prolific IDs were removed before data release, and released human data were de-identified.

References

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107795–107829. Curran Associates, Inc., 2024. doi: 10.52202/079017-3423. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c3070c3388552a08a3326f0d28dc2af9-Paper-Conference.pdf.

Louise Connell. Representing object colour in language comprehension. *Cognition*, 102(3):476–485, 2007. doi: 10.1016/j.cognition.2006.02.009. URL https://doi.org/10.1016/j.cognition.2006.02.009.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35657–35672. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1806. URL https://aclanthology.org/2025.emnlp-main.1806/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U. Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian C. Paulun, Maria Ryskina, Ekin Akyürek, Ethan G. Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua B. Tenenbaum, and Jacob Andreas. Elements of world knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, 13:1245–1270, 2025. doi: 10.1162/tacl.a.38. URL https://doi.org/10.1162/tacl.a.38.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8tYRqb05pVn.

Diane Pecher, Saskia van Dantzig, Rolf A. Zwaan, and René Zeelenberg. Language comprehenders retain implied shape and orientation of objects. *The Quarterly Journal of Experimental Psychology*, 62(6):1108–1114, 2009. doi: 10.1080/17470210802633255. URL https://doi.org/10.1080/17470210802633255.

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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. doi: 10.1109/CVPR52688.2022.00517. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Bodo Winter and Benjamin Bergen. Language comprehenders represent object distance both visually and auditorily. *Language and Cognition*, 4(1):1–16, 2012. doi: 10.1515/langcog-2012-0001. URL https://doi.org/10.1515/langcog-2012-0001.

A Appendix

A.1 Benchmark: Image Generation and Validation

A.1.1 Image Sources

Images were generated using a small set of text-to-image and image-editing models, with the generator chosen by benchmark subset. Table 2 summarizes the image sources and generation procedures. Full generation prompts, accepted image identifiers, and revision logs are provided in the supplementary materials.

Subset	Image model	Procedure
Object state	`gpt-image-2`	Base-and-edit
Scene events	`gpt-image-2`	Base-and-edit
Apparent size	`gpt-image-2`	Two-instance removal / fixed-camera edit
Social interactions	`gemini-3.1-flash-image-preview`	Base-and-edit
Social inference	`gpt-image-2`	Base-and-edit

Table 2: Image-generation sources and procedures by benchmark subset.

Because some evaluated models come from the same provider family as one image generator, we treat image provenance as a potential confound. Generation prompts were never provided to evaluated models, and all images were accepted or rejected according to pre-specified visual criteria before model evaluation.

A.1.2 Base-and-Edit Procedure

Most image pairs were created using a base-and-edit procedure. A base image was first generated with fixed camera position, setting, lighting, props, and target object or scene state. The paired image was then produced by editing only the target semantic variable while holding the rest of the scene fixed. This procedure was intended to make the two images differ minimally along the targeted contrast rather than through unrelated scene variation.

For object-state items, the edit was restricted to the target object state. For scene-event and social-inference items, the edit targeted the relevant scene-level consequence. For social-interaction items,

Model	Blocks	70%	90%	100%
Llama-3.1-8B-Instruct	32	22	28	31
Qwen2.5-0.5B	24	16	21	23
Qwen2.5-1.5B	28	19	24	27
Qwen2.5-3B	36	24	32	35
Qwen2.5-7B	28	19	24	27
DINOv2-S / -B	12	–	10	11
DINOv2-L	24	–	21	23
DINOv2-giant	40	–	35	39

edits were restricted to posture, gesture, expression, hand placement, interpersonal distance, relative position, and visible consequences needed to express the social contrast.

A.1.3 Apparent-Size Images

Apparent-size items required additional controls because the same object had to appear near or far while preserving the surrounding scene. For twelve items, we generated a single base image containing two instances of the object, one near and one far, and then created each pair member by removing one instance. Edited images were registered back to the base image before compositing. Outside the object regions, the median mean absolute pixel difference between paired images was 8.

A.2 Evaluation details

A.2.1 Shallow alignment

Representations. Activations are taken from the residual stream at the output of a transformer block. Captions are represented by the mean over all tokens, images by the CLS token. Blocks are indexed from the first block; the six depth cells correspond to the following blocks.

Training set. A fixed random sample of 5,000 MSCOCO train2017 images with their first caption. The same sample is used for every pair and every depth cell.

Alignment. Features of the two models are zero-padded to a common dimensionality (the larger of the two), centred, and divided by the per-dimension standard deviation computed on the training sample. Barycentric alignment then alternates between (i) solving an orthogonal Procrustes problem for each model against the current template (via the SVD of $X_i^\top M$) and (ii) replacing the template by the mean of the aligned representations, until the relative change of the template falls below 10^{-9} or 100 iterations are reached. The two rotations, means, and standard deviations are stored; a benchmark caption or image is standardized with the training statistics and rotated with its model's transformation.

Scoring. Each benchmark caption and image is projected as above and $s(c, v)$ is the cosine similarity of the projected vectors. Nothing is fit on benchmark data.

Permutation null. For a given subset, caption condition, and depth cell, the n image pairs are permuted against the n caption pairs 200 times with a fixed seed, keeping each pair intact; the text, image, and group scores are recomputed for each permutation. The null mean and its 95th percentile are reported, and the permutation p-value is the fraction of permutations whose score is at least the observed score (resolution 0.005). Where a subset is divided by contrast, items are permuted within the division.

Depth statistics. The grid mean averages the six cells; its significance is summarized as the number of cells whose score exceeds their own null at $p < .05$. The best cell is chosen once per model pair as the cell with the highest mean explicit group score over the five subsets, and all its scores are then reported.

A.2.2 CONTRASTIVE MODELS

`openai/clip-vit-base-patch32`, `openai/clip-vit-large-patch14`, and `openai/clip-vit-large-patch14-336`. Captions and images are passed through the text and image towers and projection heads, the embeddings are L2-normalized, and $s(c, v)$ is their cosine similarity. No prompt template is added to the captions. The permutation null is computed as in B.1.

A.2.3 GENERATIVE MODELS

Models. Open models were run locally in bfloat16 with the Hugging Face checkpoints `Qwen/Qwen2.5-VL-3B-Instruct`, `Qwen/Qwen2.5-VL-7B-Instruct`, `Qwen/Qwen2.5-VL-32B-Instruct`, `HuggingFaceM4/idefics2-8b`, `llava-hf/llava-onevision-qwen2-7b-ov-chat-hf`, and `llava-hf/llava-v1.6-mistral-7b-hf`. Frontier models were accessed through their APIs: `claude-sonnet-4-5-20250929`, `claude-fable-5-1`, `gpt-5.6-terra`, and `gemini-3.6-flash` (thinking disabled), between [TK: dates].

Prompts. The text task shows one image and asks:

> *Does this image best match: (A) {caption 0}, or (B) {caption 1}? Choices: A, B. Answer with just the letter.*

The image task shows two images and asks:

> *Which image better aligns with the description: "{caption}"? Choices: First, Second. Answer with just First or Second.*

The wording before the final sentence is VisMin's; the final sentence is the answer-format instruction that standard multiple-choice harnesses append. Each question is asked twice with the options in both orders. Two further modes—VisMin's wording without the instruction, and a bounded chain-of-thought mode ending in "Answer: X"—are run as sensitivity conditions on a subset of models [TK: which, and where reported].

Decoding. Greedy decoding for open models; temperature 0 for API models where the API accepts it (the reasoning-model endpoints reject the parameter and are run at their default). Output budget 1,024 tokens. Images are passed at native resolution through each model's own processor; no downscaling is applied, because the far object in apparent-size items occupies 6–10% of the frame.

Parsing. The reply is matched against the two admissible answers (A/B or First/Second) with case- and punctuation-insensitive matching, taking the first admissible token after any reasoning text. A reply with no admissible token is recorded as missing, not as incorrect. On the frozen suite, no reply from any of the ten models was missing or truncated (0 of 45,440 queries).

Scoring. Each of the four comparisons in §4.2 is decided by the parsed answer. The reported score averages the two option orders; the strict conjunction over both orders is given in Table [TK].

A.2.4 STATISTICS

Confidence intervals are 95% percentile bootstrap intervals over items (1,000 resamples; 2,000 for paired differences). The explicit–implicit difference is tested within item with an exact McNemar test (binomial test on the items whose outcome differs between the two captions) at one depth cell; when the six cells are pooled, the per-item outcome is the mean over cells and the same test becomes a sign test. Stars in figures denote the permutation p-value against the condition's own null: $* < .05$, $** < .01$, $*** < .005$, the last being the resolution of 200 permutations.

A.3 HUMAN BASELINES

Participants and recruitment. Participants were recruited on Prolific and completed a browser task hosted on JATOS. Recruitment was restricted to fluent English speakers aged 18–70, with a Prolific approval rating of at least 95%, on desktop browsers. All participants were adults and gave informed consent on the first screen, which stated the purpose and expected duration of the task, the compensation, what data would be recorded, and the requirement to be over 18; the task could not begin until consent was given, and participants could withdraw at any point by closing the study. The study protocol was reviewed and approved by the authors' institutional review board prior to data collection (identifiers withheld for anonymous review). Studies were advertised at 8-12 minutes for $2.00–$3.00 ($13.11–$19.30 per hour); median completion was 5.4–9.9 minutes, so realized pay met or exceeded the advertised rate. Payment was not contingent on accuracy.

Task. Each trial is one of the four probes that constitute an item: either one caption with two images, or one image with two captions. Option display order was varied between presentations, and both the semantic target and the displayed position of the correct option are recorded, so position bias can be separated from the contrast itself. After each choice, participants rated confidence on a four-point scale ($1 =$ *just guessing*, $4 =$ *completely sure*); response times were recorded. Each study opened with instructions and a worked example of each trial type, in which the correct button is highlighted (Figure 6); the study then ran in two parts, one per trial type, in counterbalanced order.

Design. Trials were pre-assigned to fixed packets of 30–38 trials rather than sampled at run time. **No item appears twice within a packet**, so no participant sees the same item under two probes or two caption levels; probe type and correct-option position are balanced within packets, and across the packet set every (item $\times$ caption level $\times$ probe) cell reaches its target number of ratings. Participants were allocated to the least-used packet through the JATOS batch session.

Sample. Seven studies covering all five subsets yielded 209 sessions; the six audited studies contribute 6,196 ratings over 2,008 cells, with 13–50 participants per study. Every cell has at least three ratings (some four), and none is short.

Exclusions. Exclusions were fixed before analysis and applied in code. Three sessions are excluded: their packets had been flagged for a large imbalance between the two probe types and were deliberately re-run with different participants, so each original is dropped in favour of its replacement. No participant was excluded for low accuracy. A few packets were run twice because of a fault in packet assignment; both runs are kept, so some cells carry a fourth rating.

Data quality. Per-study accuracy ranges .923–.984 and mean confidence 3.59–3.80 on the four-point scale. No participant had a median response time below two seconds. **No item in any study is at or below chance**, so a model's failure on an item cannot be attributed to the item being unanswerable.

Scoring. Models answer all four probes of an item; each participant answers at most one probe per item. Human data are therefore folded to the same conjunctive rule: a **cell** is the majority vote of the three (occasionally four) participants who saw that probe, with ties counted incorrect; the text and image scores each require both of their cells correct; the group score requires both. This yields a **consensus ceiling**—the performance of a three-rater panel, not an individual. Because a panel is stronger than any of its members, we also report a **single-rater baseline**: one rating sampled per cell, folded identically, averaged over 400 draws. The consensus exceeds the single rater by .06–.19, the gap widening on harder items. Comparisons in §5.6 are reported against both, since they answer different questions: whether a model matches a panel, and whether it matches a person.

A.4 POST-HOC ALIGNMENT MAPS

The main text reports one map. This appendix defines all four and shows that the conclusions do not depend on which is used.

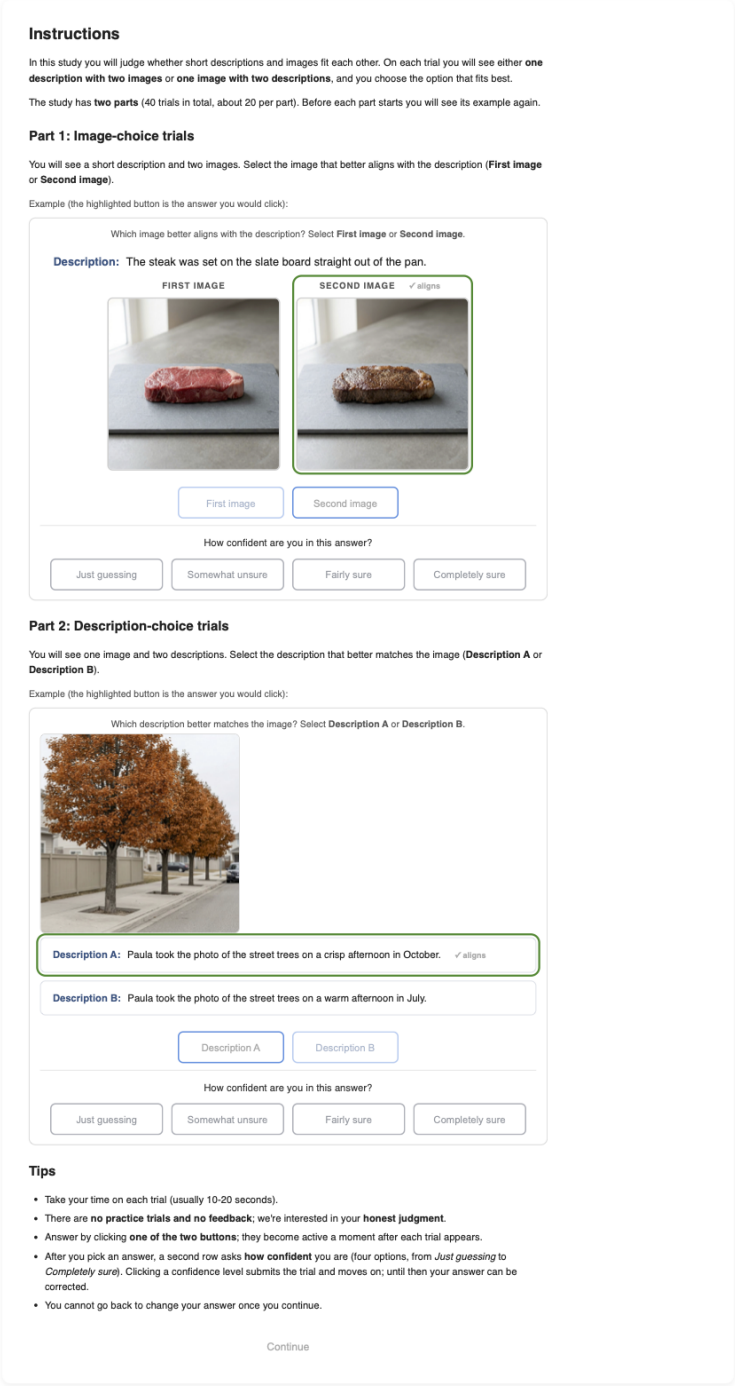

Figure 6: Human baseline task instructions and worked examples.

Map	VisMin	Suite, explicit	Suite, implicit	Suite cells above null
Procrustes	.083	.076	.032	0 / 11
Procrustes (z-scored)	.095	.098	.040	1 / 11
Ridge L→V	.092	.083	.034	2 / 11
Ridge V→L	.050	.055	.044	0 / 11

A.4.1 THE FOUR MAPS

Every map is fitted once on 5,000 MSCOCO image–caption pairs and then applied to benchmark items without refitting. No benchmark item, label, or task-specific supervision enters the fit. Writing $L \in \mathbb{R}^{n \times d}$ for the language features of the training pairs and $V \in \mathbb{R}^{n \times d}$ for the vision features (zero-padded to a common d), the maps are:

Procrustes. Generalized (barycentric) Procrustes: both sides are mean-centred and rotated by orthogonal W_L, W_V into a shared barycentre, fitted by alternating SVD until the template stops moving. Because the transform is orthogonal, it can reconcile the two coordinate systems but cannot rescale or reweight individual dimensions, so within-model geometry is preserved exactly. For two models scored by cosine, this is equivalent to rotating one model onto the other.

Procrustes (z-scored). The same rotation, applied after per-dimension standardisation of each side. Standardising removes the effect of dimensions with unusually large variance dominating the fit; it is the map used in the main text.

Ridge L→V. Ridge regression predicting vision features from language features, with the penalty chosen by generalized cross-validation over $\alpha \in [10^{-8}, 10^{8}]$. Unlike Procrustes, this is unconstrained: it may rescale, reweight, and discard dimensions freely. It therefore fits the training pairs strictly better, at the cost of not preserving geometry.

Ridge V→L. The same regression in the reverse direction.

Both sides are centred with training statistics before scoring. An earlier implementation compared ridge predictions against uncentred vision features; DINOv2's large common mean compressed every cosine into a narrow band near 0.96 and produced exact ties. The results here use the corrected version.

After fitting, each map projects language and vision embeddings independently into the common space, and image–caption compatibility is cosine similarity there—identical scoring, so the text / image / group criteria of §4.2 apply unchanged.

A.4.2 DEPTH SELECTION

Each map is fitted separately at each of the six depth cells (language 70/90/100% × vision 90/100%). The reported cell is the one maximising mean explicit group score across the six suite benchmarks, which is a selection made once per map, not per condition. The selected cell differs by map—Procrustes L0.7/V0.9, Procrustes z-scored L0.9/V1.0, Ridge L→V L1.0/V1.0, and Ridge V→L L0.7/V0.9—which is itself a reason to report the map alongside the depth rather than treating depth as a property of the model pair.

A.4.3 RESULTS

Figure 7 gives the full panel per map: the VisMin control, then the suite by domain and inference route, explicit against implicit, each bar against its own permutation null (200 shuffles). **Figure 8** gives the vision-side and language-side scale ladders with one panel per map. All panels share a y axis so the four are directly comparable.

Mean group score, and the number of conditions significantly above their own null:

Three things hold for every map:

1. **The VisMin control is recoverable and the suite is not.** Object and attribute contrasts clear their null under all four maps; across the suite, at most 2 of 11 conditions do.
2. **Implicit captions score below explicit ones.** The ordering never reverses, under any map.
3. **An unconstrained map does not rescue the suite.** Ridge can reweight dimensions freely and still lands in the same range as the orthogonal maps. The suite contrasts are therefore not merely mis-scaled in the unimodal representations—a linear read-out does not recover them either, which is what the main text's claim rests on.

The maps do differ in level. Ridge V→L is the weakest throughout, including on the VisMin control (.050 against .083–.095), so its flat suite performance is uninformative about the suite. Among the three maps that pass the control, z-scored Procrustes is the strongest on the suite, which is why it is the main-text map; Ridge L→V is close behind and clears the null in two more conditions, both on object state.

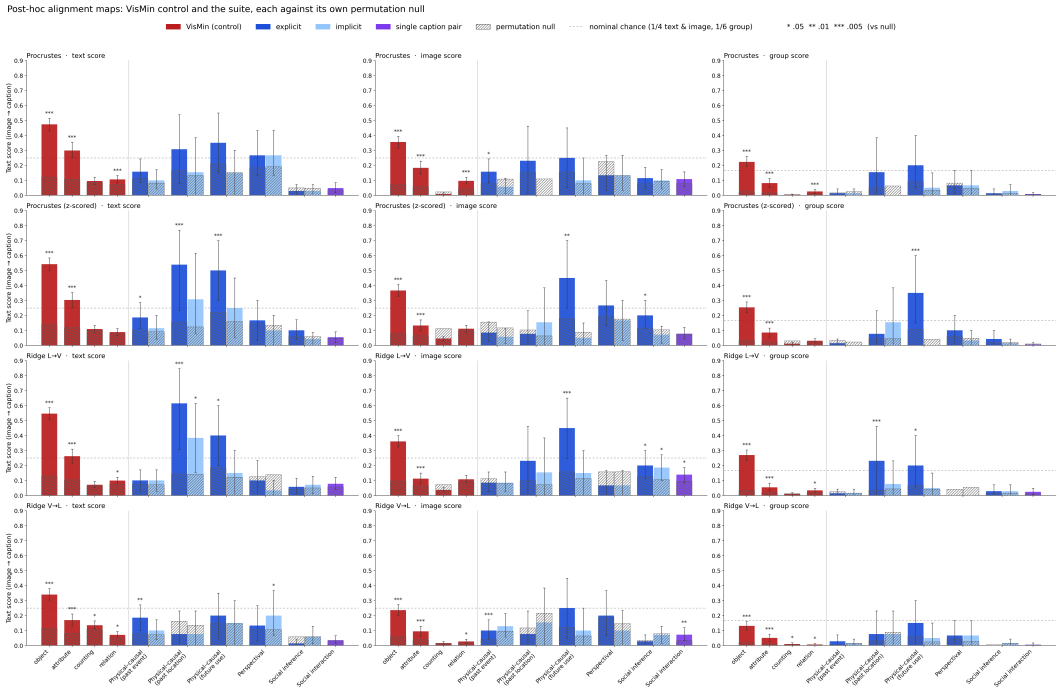

Figure 7: **All four post-hoc alignment maps, on the VisMin control and the full suite.** Rows are alignment maps (generalized Procrustes; the same with per-dimension z-scoring; ridge regression language→vision and vision→language); columns are the three scoring criteria. Within each panel, the four VisMin categories sit left of the rule as a lexical positive control, and the suite follows in canonical order, split by knowledge domain and inference route. Dark and light bars are explicit and implicit captions on the same items; social interaction carries a single literal caption pair and so has one bar. Hatched bars are each condition's own permutation null (200 shuffles of the image side within that condition), error bars are 95% bootstrap confidence intervals over items (1,000 resamples), and stars give the permutation p-value against that condition's null ($^{*} < .05$, $^{**} < .01$, $^{***} < .005$, the last being the resolution of 200 permutations). The dashed line is nominal chance ($1/4$ for text and image, $1/6$ for group). All twelve panels share a y axis. Each map is read at its own best depth cell, selected once per map on mean explicit group score across the six suite benchmarks; the model pair is Llama-3.1-8B-Instruct × DINOv2-giant throughout. The control clears its null under every map while at most 2 of 11 suite conditions do, and implicit captions score below explicit ones under every map.

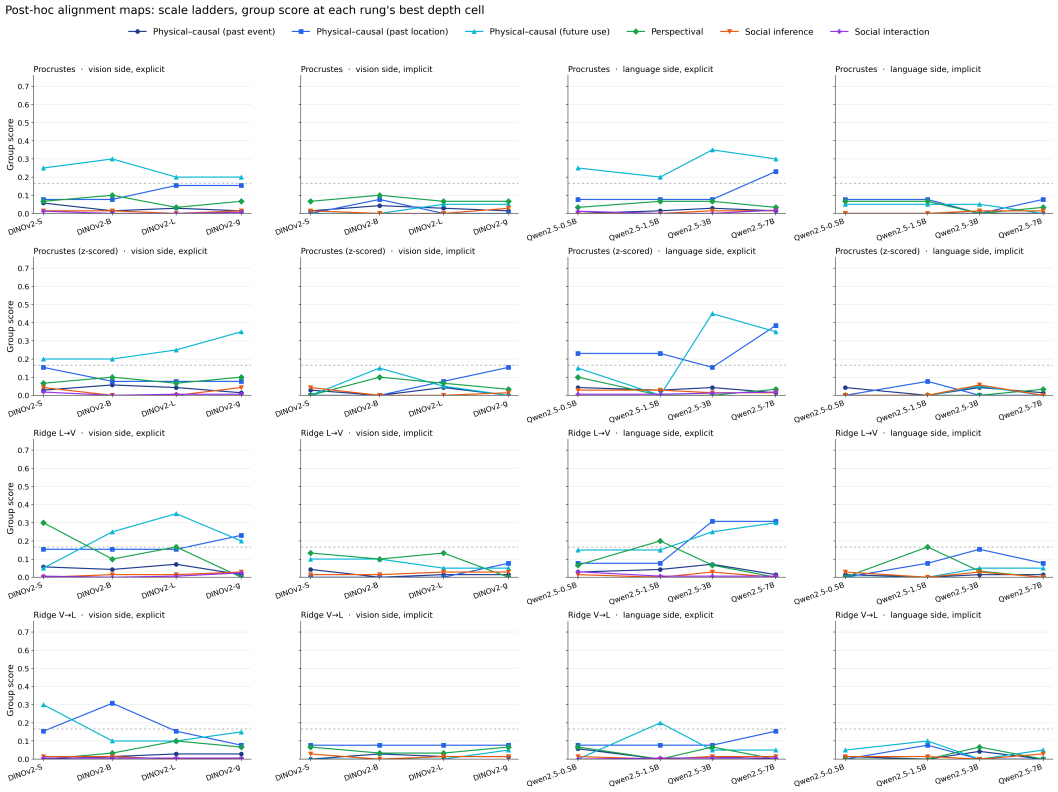

Figure 8: **Encoder scale does not recover the suite, under any alignment map.** Row are the four alignment maps; columns cross the two scale ladders with the two caption levels. The vision-side ladder varies DINOv2-S/B/L/g with Llama-3.1-8B-Instruct held fixed; the language-side ladder varies Qwen2.5-0.5B/1.5B/3B/7B with DINOv2-giant held fixed. The x axis is encoder parameters on a log scale, labelled by model. One line pe domain and inference route; social interaction appears only in the explicit columns, as it has a single literal caption pair. The dashed line is nominal chance ($1/6$). Each rung read at its own best depth cell, selected on mean explicit group score across the six suite benchmarks, so the implicit panels are read at a cell chosen without reference to them. sixteen panels share a y axis. Scores are item means; no confidence intervals are drawn, as each rung is a single condition.