THE ANSWER FORMAT IS PART OF THE MEASUREMENT: OPTION LABELS CHANGE VISUAL CONTEXT EFFECTS IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly evaluated for how much they rely on cultural and regional context. A common protocol holds a target fixed, changes its visual surroundings, and compares the answers the model chooses among labeled options. However, we argue that such measurements depend on how the options are labeled. We introduce Scene-Orthography, a controlled benchmark built on a cultural convention with an unambiguous answer: a model must identify which spelling of a word, American or British (e.g., COLOR vs. COLOUR), appears on a sign whose pixels are identical in a US and a UK street scene, so the correct answer never depends on the scene. Holding the images fixed, we vary only the answer format: whether the options are labeled with country names or with letters that carry no regional meaning, which spelling each label names, whether the spellings are shown directly, and where the labels are placed. Across eight open models, country-name labels enlarge the measured effect of the scene, and in the two primary models, swapping which spelling each label names reverses its direction. A pre-specified test on held-out words and photographs shows that labels act through the option they sit on: a label beside an option steers the choice even when the prompt declares it unassigned, whereas the same words on a separate line have little effect. This dependence varies widely across models and shrinks with scale within one model family, and on the natural-image benchmark MMVP, the answer interface also changes measured accuracy. Therefore, a visual context effect is a property of both an image *and* an answer format, suggesting that evaluations of cultural context should report both.

1 INTRODUCTION

Cultural and regional differences leave visible traces in images and in writing: street scenes look different from country to country, and so do the conventions for writing dates, units, and spelling. Vision-language models (VLMs) can be swayed by learned priors (Agrawal et al., 2018) and by visual context (Vo et al., 2026), and benchmarks test how they reason about text in relation to its visual surroundings (Wadhawan et al., 2024). The usual way to measure such context effects is to hold a target fixed, change its surroundings, and compare the answers the model chooses from a set of options, as in most multiple-choice VLM benchmarks (Liu et al., 2024; Chen et al., 2024; Tong et al., 2024).

This protocol controls the image but not the answer format. The options must be labeled somehow, and in evaluations of cultural context, the labels can name the very dimension along which the context varies, such as a country or a region. We therefore ask: *does the answer format change the measured effect of cultural context, and if so, which part of the format is responsible?* If it does, the same model evaluated on the same images can appear more context-sensitive, less, or sensitive in the opposite direction, depending only on how the options are written.

We study this question with regional spelling, a cultural convention whose correct answer can be checked exactly. Figure 1a shows the central result. We place the same sign, with identical pixels, in a UK and a US street scene and ask which spelling it shows. When the options are labeled "United States" and "United Kingdom" and each label names its own country's spelling, the US scene pulls

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.

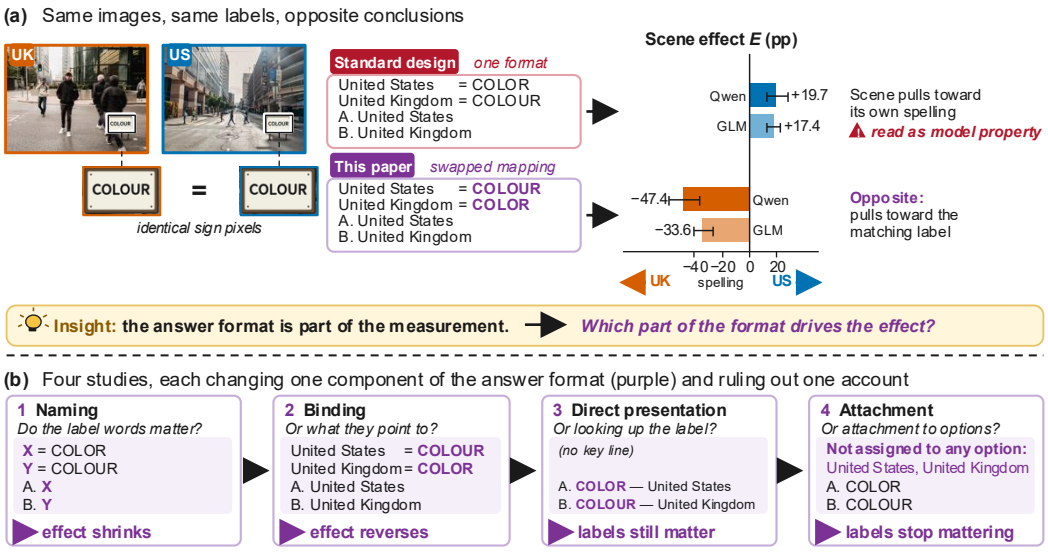

Figure 1: **The answer format is part of the measurement.** (a) The same sign, with identical pixels, appears in a UK and a US street scene. Under a single answer format (top), the scene effect suggests that the scene pulls answers toward its own spelling. Swapping which spelling each label points to, with the same images and label words (bottom), reverses the effect. Bars show the signed effect E in the Binding study (pp; 95% intervals). (b) Four studies each change one component of the answer format and rule out one account of this dependence (§4–§5). Prompts are schematic and omit options C (Neither) and D (Cannot read the sign). The exact prompts are provided in Appendix A.

Qwen3-VL-8B toward the American spelling. When we swap which spelling each label names, the same scene swap pulls it toward the British spelling instead. The images and the label words are identical; only what the labels point to has changed.

Why would the labels matter at all? The model answers by choosing among option lines, and a country name on an option line is itself something a scene can match: a US street makes the option that says "United States" look right, whatever spelling it carries. Models can indeed pick answers by matching the image against the options (Zhang et al., 2025). If this is what happens, the measured context effect is produced jointly by the image and by how the options are written.

Prior work shows that option order, option identifiers, and answer symbols shift multiple-choice decisions (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Robinson & Wingate, 2023; Yang et al., 2025), that prompt formatting changes measured performance (Sclar et al., 2024), and that similar sensitivities appear in vision-language multiple-choice QA (Zong et al., 2024; Atabuzzaman et al., 2025; Loginova et al., 2025). Scoring and answer-extraction choices can even change model comparisons (Alzahrani et al., 2024; Molfese et al., 2025). These studies concern accuracy and rankings. Our question is whether the format changes the estimated effect of a controlled change of context, including its sign.

To answer this question, we propose Scene-Orthography (§3), a benchmark in which the correct answer never depends on the scene. Thus, any scene effect reflects the context rather than a change in the correct answer, allowing accuracy and context sensitivity to be measured separately. We summarize our key findings as follows:

- **The format changes the effect, not only accuracy.** Three studies change one component of the format at a time. Country-name labels enlarge the scene effect in all eight models, swapping what they point to reverses it, and listing the spellings directly does not remove it (§4).
- **The effect follows the label's position.** A pre-specified test on held-out materials confirms that labels matter when attached to the options, and a factorial follow-up shows that a label beside an option steers the choice even when the prompt declares it unassigned (§5).

- **Where it matters for evaluation.** The effect varies widely across models and shrinks with scale within one model family. Formats change effect sizes more reliably than model rankings, and an indirect answer key lowers accuracy on MMVP (Tong et al., 2024) (§6, §7).

2 Related Work

Visual context and learned priors. VQA research exposed reliance on answer priors by constructing test splits with shifted answer distributions (Agrawal et al., 2018). VLMs also make systematic errors driven by contextual cues (Vo et al., 2026), while ConTextual and scene-text benchmarks such as TextVQA and ST-VQA require models to read text in its visual surroundings (Wadhawan et al., 2024; Singh et al., 2019; Biten et al., 2019). These works ask *when* context influences VLM behavior; we ask whether the *measurement* of that influence depends on the answer format.

Prompt and answer-format sensitivity. Beyond the effects of option order, identifiers, and symbols discussed above (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Yang et al., 2025; Zong et al., 2024), the binding between answer content and reporting symbols has been studied directly (Robinson & Wingate, 2023; Xue et al., 2024), and contextual calibration estimates prompt-induced preferences from content-free inputs (Zhao et al., 2021), but cannot capture a preference that depends on the image. These studies measure how formats shift accuracy or rankings; we measure how formats change the estimated effect of a controlled visual intervention.

What image–text evaluations measure. Compositionality diagnostics reveal failures hidden by aggregate scores (Yuksekgonul et al., 2023). SugarCrepe shows that candidate construction can be exploited (Hsieh et al., 2023), Lin et al. (2024) show how language priors and scoring choices shape image–text retrieval, and NaturalBench and MMVP use paired natural images to test visual understanding (Li et al., 2024; Tong et al., 2024). MMStar curates questions that require visual evidence (Chen et al., 2024), and visually easy distractors can let models answer from the image and options without consulting the question (Zhang et al., 2025). While these works ask whether a benchmark's *scores* reflect visual understanding, we ask whether a controlled *effect* does.

3 Scene-Orthography

3.1 Task Setting

Scene-Orthography tests whether a model reads the spelling of a word correctly when the surrounding scene suggests a different regional convention. As shown in Figure 1a, a model sees a street photograph that contains a small sign with one word, and it must identify the exact written form of that word: the American spelling (e.g., COLOR) or the British spelling (e.g., COLOUR). It answers by choosing one of four options, namely the two spellings, *Neither*, and *Cannot read the sign*, and reports only the option letter. We call one such question an *item*.

Each item combines a word that has an American and a British spelling, a rendered sign that shows one of the two, and a real photograph of a US or a UK street into which the sign is pasted. For every word, we paste the same sign, with identical pixels, into one US and one UK photograph, so the two items differ only in the scene (see Figure 1a). The scene changes everything around the sign, including layout, color, and other visible text, and we treat it as a single, bundled manipulation. A scene is *congruent* when its country matches the spelling on the sign (e.g., a US scene with COLOR) and *conflicting* otherwise. The correct answer is the same in both.

We contrast US and UK scenes because British and American English are two major national varieties of English (Algeo, 2006), and many common words are spelled differently in each (Tottie, 2001; Baker, 2017). Australian spelling, for example, generally follows British conventions with some exceptions (Australian Government, 2023); to keep the comparison focused, we use one country per convention. For the selected word pairs, this yields regionally congruent and conflicting scenes, and the two spellings differ in only a few letters, so the task hinges on discriminating the exact written form. Because the task requires exact matching and the sign pixels are identical across paired scenes, changing the scene leaves the correct answer unchanged.

3.2 Answer Formats

The two spellings are offered as answer options, and each option can carry a *label*, a short name attached to it. We vary the labels while keeping the image, the instruction, and the correct answer fixed. *Regional labels* are the country names `United States` and `United Kingdom`. *Neutral labels* are the letters X and Y, which carry no regional meaning and only give the two spellings arbitrary names. Which spelling a label names is its *assignment*. With neutral labels, X names the American spelling and Y the British one in one configuration, and the reverse in the other; we average the two, so neither letter is tied to either spelling. With regional labels, the usual assignment lets each country name its own spelling.

Labels can be presented in two ways. In the *indirect* format, the options show only the labels, and a key line above them states what each label names (e.g., X = COLOR, Y = COLOUR). In the *direct* format, each option shows a spelling followed by its label:

```
Which reference entry exactly matches the visible sign?
Match the written form exactly.
A. DENUCLEARIZE -- X  B. DENUCLEARISE -- Y
C. Neither reference entry  D. Cannot read the sign
Answer with exactly one letter:  A, B, C, or D.
```

Labels can also be moved off the option lines onto a separate line, or removed altogether (*Plain*). Options rotate through all four positions. Figure 1b summarizes the variants, and the exact prompts are provided in Appendix A.

3.3 Evaluation

Words and photographs recur across items, so we first average within each observed combination of a word and a US–UK photograph pair (an *edge*) and then across edges. For a given format, let a^+ and a^- denote the accuracies in congruent and conflicting scenes, respectively. The *scene effect* and the *naming effect* are

$$\tau = 100\,(a^+ - a^-), \qquad J = \tau_{\text{regional}} - \tau_{\text{neutral}}, \tag{1}$$

both in pp, where τ_{neutral} averages the two assignments of X and Y. The scene effect τ measures how much more accurate the model is when the scene agrees with the sign, and J how much regional labels add to that. Balanced accuracy, $\text{BA} = 50\,(a^+ + a^-)$, measures overall performance. The two answer different questions: if regional labels raised accuracy by the same amount across all scenes, BA would rise while J stayed at zero. J is positive only when labels change *how much the scene matters*. To see what J captures, consider a model that, when unsure, picks the option whose label matches the scene. With regional labels, this adds errors only in conflicting scenes, where the matching label names the wrong spelling, so τ rises. With neutral labels, no option matches the scene, and the same uncertainty produces no scene effect. J is the difference between the two.

When labels change what they denote, accuracy no longer tracks a fixed spelling. For these comparisons, we code each response as $+1$ (American spelling), -1 (British spelling), or 0 (anything else), and define the *signed effect* E as the mean change in this code, in pp, when a US scene replaces a UK scene.

3.4 Experimental Settings

We primarily use two models: Qwen3-VL-8B-Instruct (Bai et al., 2025) and GLM-4.6V-Flash (Z.ai, 2025). We further extend the main comparisons to six additional models: Ministral-3-8B (Liu et al., 2026), Qwen3-VL-32B, Gemma-3-12B (Team et al., 2025), Llama-3.2-11B-Vision (Meta Llama, 2024), Idefics3-8B (Laurençon et al., 2024), and InternVL3.5-8B (Wang et al., 2025). We read each choice from the log probabilities of the four single-token answer keys in one forward pass and rotate the options through all positions, as in CircularEval (Liu et al., 2024). A top-two margin below 10^{-6} nats counts as a tie and is scored as incorrect, and every response, including Neither, Cannot read, and ties, remains in every denominator. Each sign is rendered at two levels of evidence: HIGH and a smaller, lighter, slightly blurred MEDIUM.

Words and photographs both recur across edges, so intervals come from a crossed bootstrap over words and photographs (Owen, 2007; Owen & Eckles, 2012) with 100,000 draws, keeping all conditions

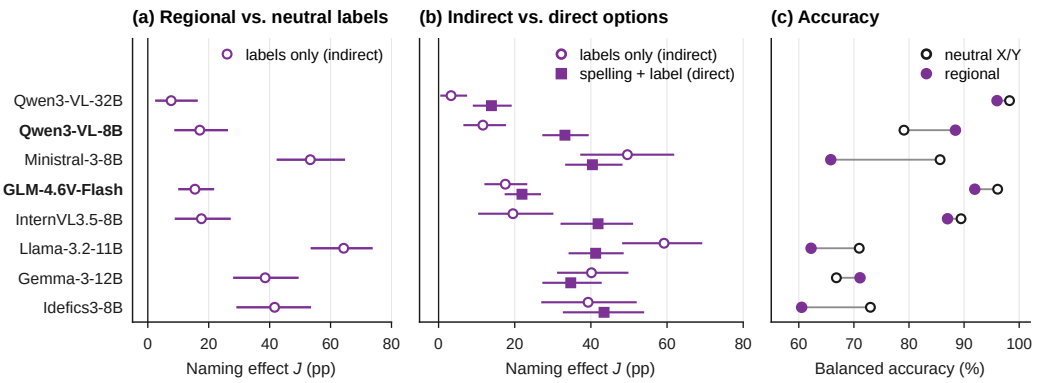

Figure 2: **Regional labels enlarge the scene effect across models, and showing the spellings does not remove it.** (a) Naming effect J (regional minus neutral labels) with indirect options on the Naming panel. (b) Naming effect with indirect options (labels only) and direct options (spelling plus label) on the Direct-presentation panel. (c) Balanced accuracy with neutral and regional labels on the Naming panel. Models are ordered by accuracy without labels (Table 1); primary models (bold) have 95% intervals, the others Bonferroni-adjusted intervals for six comparisons.

paired within each draw. We report 95% intervals for the primary models and Bonferroni-adjusted intervals across the six further models. Each study uses its own set of words and photographs, which we call its *panel*; more details about the panels and evidence levels are provided in Appendix A.

4 Which Part of the Answer Format Drives the Effect?

We begin with the simplest question: which part of the format carries the effect? A label differs from a bare option in three ways. It has a meaning ("United States"), it points to a particular spelling, and it forces the model to look up that spelling. The following three investigations vary these factors one at a time while keeping everything else, including the images, fixed.

4.1 Naming: The Label Words

Naming replaces the regional names with the neutral labels X and Y. Regional labels make the scene matter more. For Qwen3-VL-8B, the scene effect roughly quadruples when neutral labels are replaced by regional ones, and GLM-4.6V-Flash, which shows no scene effect with neutral labels, shows a clear one with regional labels. The naming effect is positive in all eight models (Figure 2a).

Accuracy tells a different story. Regional labels lower accuracy in six of the eight models but raise it in the other two (Figure 2c), so a model can become more accurate and more context-sensitive at the same time. Accuracy alone would hide the shared effect, and the effect alone would hide the divergent accuracy. Regional labels, however, differ from neutral ones both in what they say and in what they point to. Binding separates the two.

4.2 Binding: What the Labels Point To

Binding keeps the regional labels and swaps only the spellings they name. The scene effect reverses in both primary models (Figure 1a): under the usual assignment, a US scene pulls answers toward the American spelling, whereas under the swapped assignment, it pulls them toward the British spelling, which is now labeled "United States". Because the label words do not change, most of the scene's influence follows what the labels denote rather than the spellings themselves; the part of the effect that flips with the assignment is larger than the part that does not. More details about this decomposition are provided in Appendix B.2. The swapped assignment is harder to use, and among the six further models only Qwen3-VL-32B passed the control checks required to run it, so we report Binding for the primary models.

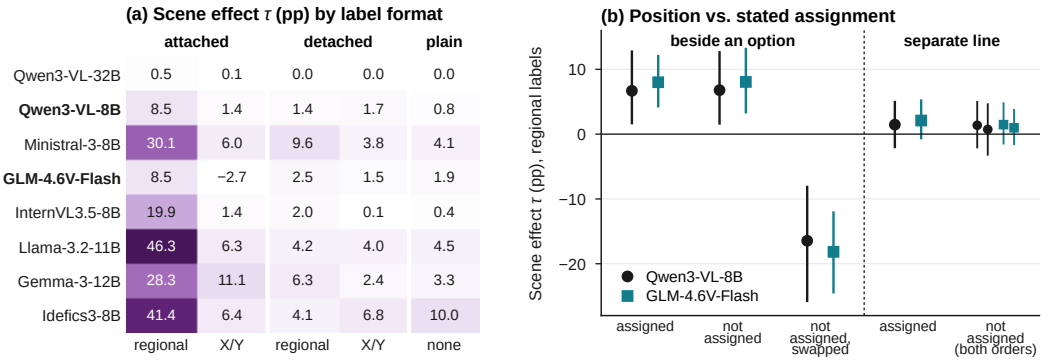

Figure 3: **The scene effect lives in labels that sit on the option lines.** (a) Scene effect τ (pp) on the confirmation set under the five label formats of the attachment test; neutral and detached entries average their two configurations. (b) Position $\times$ stated assignment for the primary models: scene effect under regional labels, with 95% intervals. "Swapped" places each regional name beside the other country's spelling. With neutral labels, the scene effect stays close to zero in every cell; the full results are provided in Appendix C.3.

In both studies so far, the model has to map a label back to a spelling. Perhaps the effect arises in that lookup step. Direct presentation removes it.

4.3 Direct Presentation: The Lookup Step

One might expect the naming effect to vanish once the spellings themselves are shown as options. It does not. With each spelling shown next to its label, regional labels still enlarge the scene effect in all eight models (Figure 2b). Whether direct options strengthen or weaken the effect depends on the model; for Qwen3-VL-8B, they nearly triple it. Direct options also cost accuracy under regional labels in six of the eight models. Even with the correct spelling in plain view, the label beside it changes how much the scene matters.

What the three studies establish. Together, the studies narrow down the source of the effect. It changes the scene effect rather than accuracy alone (§4.1); it follows what the labels denote rather than which words appear (§4.2); and it persists when the spellings are in plain view (§4.3). What remains is the relation between a label and the option it names, which we investigate next.

5 Labels Act Through the Option They Sit On

5.1 A Pre-Specified Test on Held-Out Materials

Two accounts. Section 4 leaves two ways a label could act. It could act as part of its option: the model matches the scene against the option's text, and the label is part of that text. Or it could act on the prompt as a whole, raising the salience of the region wherever the words appear. The two accounts agree whenever labels are attached to options. They come apart when the same words appear in the prompt but are assigned to no option: the first predicts that such labels barely change the scene effect, and the second that they change it as much as attached labels do.

Design. All formats in this test show the spellings directly. We measure the naming effect with labels beside the spellings (J_{att}) and with the same labels on a separate line marked as not assigned to any option (J_{det}), and define the *attachment effect* $K = J_{\text{att}} - J_{\text{det}}$. Because both families show the same spellings and use the same neutral labels, K isolates what attaching the regional names adds. The first account predicts $K > 0$ with J_{det} near zero. We fixed this prediction, the analysis,

Table 1: **Attachment effect and accuracy on the held-out confirmation set** (16 words, 16 photograph pairs, 128 edges). J_{att} and J_{det} are naming effects with labels attached to or detached from the options, and $K = J_{\text{att}} - J_{\text{det}}$, all in pp. BA is balanced accuracy (%) under attached regional labels, detached regional labels, or no labels. Models are ordered by accuracy without labels. Bold K: interval excludes zero; bold BA: the most accurate format for that model. †Primary models, with 95% intervals from the pre-specified test; the other six have Bonferroni-adjusted intervals for six comparisons.

	Effects (pp)				BA (%)		
Model	J_{att}	J_{det}	K	Interval	Attached	Detached	Plain
Qwen3-VL-32B	0.4	0.0	0.4	$[-1.8, 3.4]$	98.9	**100.0**	**100.0**
Qwen3-VL-8B†	7.1	-0.3	**7.3**	$[1.9, 13.1]$	91.5	98.5	**99.2**
Ministral-3-8B	24.1	5.8	**18.3**	$[5.8, 30.2]$	77.3	94.6	**97.4**
GLM-4.6V-Flash†	11.3	1.0	**10.3**	$[4.3, 16.3]$	78.8	91.9	**96.4**
InternVL3.5-8B	18.5	1.9	**16.6**	$[3.6, 28.1]$	84.5	93.7	**94.0**
Llama-3.2-11B	39.9	0.2	**39.8**	$[23.8, 55.9]$	68.3	86.0	**86.9**
Gemma-3-12B	17.2	3.9	**13.3**	$[1.6, 24.2]$	75.1	**84.5**	78.1
Idefics3-8B	35.0	-2.6	**37.6**	$[23.9, 53.3]$	49.3	70.3	**74.9**

and the materials on a small development set and then collected a confirmation set of new words and new photographs.

Results. The prediction held in both primary models. Attached labels produced a naming effect and detached labels almost none, giving attachment effects of 7.3 pp [1.9, 13.1] for Qwen3-VL-8B and 10.3 pp [4.3, 16.3] for GLM-4.6V-Flash. Both models also kept tracking the target on control items without the scene, the second pre-specified condition. More details about the success criteria and controls are provided in Appendix C.1. Figure 3a shows the same picture across all eight models: the scene effect is concentrated in one cell, regional labels beside the options, and small everywhere else. The attachment effect is positive in five of the six further models (Table 1).

5.2 Position, Not Stated Assignment

Detaching a label changes two things at once: where it sits and what the prompt says about it. A factorial follow-up on the confirmation panel crosses the two for the primary models (Figure 3b). Labels stand either beside the options or on a separate line, and the prompt either assigns them to the options or states that they are not assigned.

The stated assignment turns out to matter little. With labels beside the options, declaring them unassigned leaves the scene effect unchanged, and placing each regional name beside the other country's spelling reverses it. With labels on a separate line, stating the assignment does not bring the effect back. A label thus acts on the choice through the option line it occupies, not through what the prompt declares it to mean. For practice, this means that a disclaimer in the instruction cannot neutralize a label that stays beside an option; to remove its influence, the label itself has to move. The attachment effect is also stable across the two levels of sign evidence, as shown in Appendix C.4.

These results fit the reading sketched in the introduction: a model scores each option line as a unit against the image, so a regional name on that line is part of what the scene can match. Our contrasts are consistent with this reading but identify what in the prompt changes behavior, not how the model computes it.

6 How the Effect Varies Across Models

6.1 Eight Models

Table 1 compares the attachment effect across all eight models on the same confirmation set. The effect varies widely, from none in Qwen3-VL-32B to about 40 pp in Llama-3.2-11B. Detached labels leave little naming effect in any model, so in every model the effect comes from the attached labels.

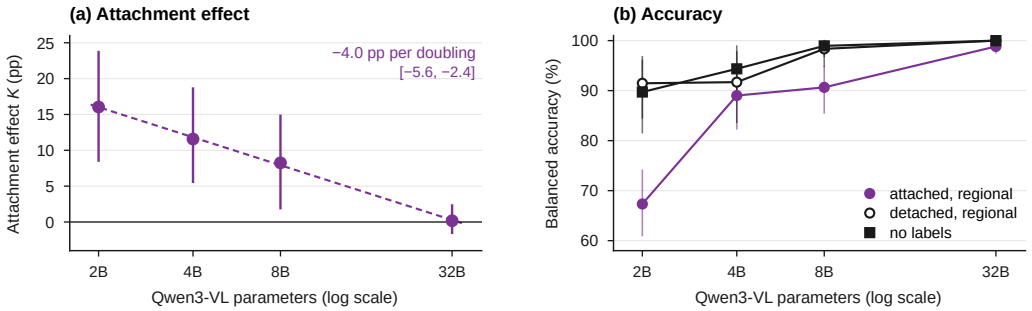

Figure 4: **Within one model family, the attachment effect shrinks with scale.** Qwen3-VL checkpoints run through one matched pipeline on the confirmation panel. (a) Attachment effect K with 95% intervals; the dashed line is the pre-specified least-squares trend on log parameters. (b) Balanced accuracy under attached regional labels, detached regional labels, and no labels.

Table 2: **The interface changes accuracy on MMVP.** Paired accuracy (%) with the answers shown directly and with neutral names plus an answer key, and their difference Δ (pp). Bold Δ: Bonferroni-adjusted interval excludes zero.

	Qwen-32B	Qwen-8B	Ministral	GLM	InternVL	Llama	Gemma	Idefics
Direct	67.0	58.3	29.0	43.3	49.3	38.7	46.7	49.3
Indirect	63.7	55.2	26.5	38.7	37.8	22.5	38.7	40.2
Δ	−3.3	−3.2	−2.5	−4.7	**−11.5**	**−16.2**	**−8.0**	**−9.2**

Attached regional labels also cost accuracy in every model, and removing labels altogether is the most accurate format for all but one. The models with the largest attachment effects are also the least accurate with attached labels, raising the question of how the effect changes as a model improves.

6.2 Scale Within One Model Family

To answer this without mixing model families, we ran four Qwen3-VL checkpoints (2B, 4B, 8B, and 32B) using a single matched pipeline. The attachment effect falls steadily with size, from 16.0 pp at 2B to 0.2 pp at 32B, a trend of -4.0 pp per doubling of parameters $[-5.6, -2.4]$ (Figure 4a). Accuracy under attached labels rises toward the ceiling that the models reach without labels (Figure 4b): the effect shrinks as the model stops paying for the labels. Rerunning the 8B and 32B checkpoints through this pipeline reproduces their original estimates, so the difference between them does not come from how they were run. More details about the matched pipeline and the reruns are provided in Appendix D.1. We read the series as a trend within one family rather than as a scaling law.

7 Consequences for Evaluation

Formats change effect sizes more reliably than rankings. The ordering of the eight models by scene effect shifts from one label format to another, but no pair of models reverses reliably, in the sense that both paired intervals exclude zero with opposite signs. More details about this ranking analysis are provided in Appendix E.1. What the format changes reliably is the size of effects and of the gaps between models. In the Direct-presentation study, the two primary models do not differ reliably under indirect options but differ clearly under direct ones. An evaluator comparing the two models would find no difference under one format and a clear gap under the other.

The interface matters even without a context manipulation. On MMVP (Tong et al., 2024), 150 pairs of natural images that share a question, replacing directly displayed answers with neutral names and an answer key lowers paired accuracy in every model, significantly in four of the eight

(Table 2). The drop ranges from a few points to more than 15 pp, while the ranking of the models stays largely intact. More details about the MMVP design are provided in Appendix E.2. MMVP does not manipulate visual context, so this shows the interface changing measured accuracy on a natural benchmark, not the scene effect transferring to it.

Practices. For evaluations that manipulate visual context and read out a choice among labeled options, our results suggest three practices. First, state the format: which labels are used, what they point to, and where they sit. Second, estimate the interaction with one extra format on the same images, such as neutral labels or labels moved off the option lines; if the options must carry meaningful names, the extra format can keep the names but place them on a separate line. This separates a format that changes how much context matters from one that shifts accuracy everywhere. In our data, an evaluation of Llama-3.2-11B with attached regional labels alone would report a scene effect of 46.3 pp; adding the neutral-label format on the same images shows that only 6.3 pp of it remains without the regional names. Third, report accuracy besides the effect, because a smaller context effect can come with lower accuracy (§4.1).

8 Discussion

What the dependence is. The dependence is neither a preference for one spelling nor a uniform change in accuracy but an interaction: a scene and a label that name the same region reinforce each other, so the model errs mainly when the scene and the sign disagree. In the language of measurement, the answer format is part of the instrument: changing it changes what a context effect measures.

Where to expect it. Two conditions make the dependence likely. The options must carry words that the context can match, as regional names do for street scenes; neutral labels carry no such meaning, and with them the effect is small in every model we test. And the words must sit on the option lines, where they become part of what is chosen (§5.2). Many evaluations meet both conditions, for example questions about places answered with country or city names whenever the context varies the dimension that the names describe.

Why larger models show less. In the scale series, the effect shrinks as accuracy under attached labels approaches the ceiling that the models reach without labels: a model that reads the sign reliably has little room to be pulled by a label. Within a model, however, making the sign harder to read did not change the effect, as shown in Appendix C.4. The dependence therefore reflects how strongly a model weighs labels against visual evidence overall, rather than how difficult a particular item is. Evaluations of strong models may see little of it, while evaluations that compare models of different strength will see it change the size of the gaps between them.

9 Conclusion

Holding a visual target fixed controls the image but not the measurement. In Scene-Orthography, the same scene swap produces a context effect whose size depends on how the options are labeled and whose direction depends on what the labels point to, and the effect travels with the label's position on the option lines rather than with what the prompt says the label means. It is largest when the option labels name the dimension that the context varies and the model reads the target less reliably, and both conditions can be checked before an evaluation is run. Because the answer interface also changes measured accuracy on a natural benchmark, a visual context effect should be treated as a property of the image and the answer format together. Evaluations should therefore document the labels, their mappings, and their placement, compare at least two formats on the same images, and report accuracy alongside the context effect.

Limitations and future work. We study one controlled task family, chosen so that the correct answer never depends on the context; testing the same interaction in other tasks and label vocabularies, such as culturally grounded questions answered with country names, is a natural next step, and our preliminary attempts are described in Appendix F. Our analysis characterizes behavior rather than internal mechanisms, and it reads choices from answer-key probabilities of open-weight models; extending it to generated answers and to other model families would broaden the picture.

AI USE STATEMENT

We used generative AI tools to polish the writing, identify relevant literature, and assist with writing and running code for the experiments and analyses. The authors take full responsibility for the methods, results, references, and final manuscript.

ETHICS STATEMENT

This study examines the reliability of vision-language model evaluation through controlled changes to scene context and answer formats. Regional labels are used as experimental stimuli to test associations between visual context and spelling conventions. The reported effects characterize model behavior under the tested conditions and should not be interpreted as characteristics of national or cultural groups. MMVP images are used from their official source and are not redistributed.

REPRODUCIBILITY STATEMENT

Section 3 describes the task, answer formats, and measures. Appendix A documents the materials, prompts, evidence levels, and response coding; Appendices B–E give full results for every analysis in the main text; and Appendix G describes the resampling design. An anonymized code package is provided in the supplementary materials.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

John Algeo. British or american english?: A handbook of word and grammar patterns. pp. 1–348, 01 2006. doi: 10.1017/CBO9780511607240.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating MCQA selection bias of large vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33548–33562, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1703. URL https://aclanthology.org/2025.emnlp-main.1703/.

Australian Government. Spelling. https://www.stylemanual.gov.au/grammar-punctuation-and-conventions/spelling, 2023. Style Manual. Updated 7 July 2023. Accessed 26 September 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Paul Baker. *American and British English: Divided by a Common Language?* 09 2017. ISBN 9781107088863. doi: 10.1017/9781316105313.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluis Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4290–4300, 2019. doi: 10.1109/ICCV.2019.00439.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: fixing hackable benchmarks for vision-language compositionality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. URL https://arxiv.org/abs/2408.12637.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 17044–17068. Curran Associates, Inc., 2024. doi: 10.52202/079017-0542. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1e69ff56d0ebff0752ff29caaddc25dd-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29914–29934. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lin24c.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang,

Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, pp. 216–233, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72657-6. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. Addressing blind guessing: Calibration of selection bias in multiple-choice question answering by video language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3216–3246, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.162. URL https://aclanthology.org/2025.acl-long.162/.

Meta Llama. Llama-3.2-11B-Vision-Instruct. Model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18477–18494, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.950. URL https://aclanthology.org/2025.findings-acl.950/.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. ISSN 19326157. URL http://www.jstor.org/stable/4537442.

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. ISSN 19326157, 19417330. URL http://www.jstor.org/stable/41713508.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019. doi: 10.1109/CVPR.2019.00851.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914.

Gunnel Tottie. An introduction to american english. 2001. URL https://api.semanticscholar.org/CorpusID:152500583.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DG4S2OlGQA.

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. ConTextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49733–49787. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wadhawan24a.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao

Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4331–4344, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.237. URL https://aclanthology.org/2024.acl-long.237/.

Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1902–1917, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.95. URL https://aclanthology.org/2025.naacl-long.95/.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Z.ai. GLM-4.6V-Flash. Model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V-Flash.

Hao Zhang, Chen Li, and Basura Fernando. Mitigating easy option bias in multiple-choice question answering, 2025. URL https://arxiv.org/abs/2508.13428.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (Vision and) language model with embarrassingly simple permutations. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62892–62913. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zong24b.html.

APPENDIX OVERVIEW

Each appendix section supports one part of the main text. Appendix A documents the task materials and prompts of Section 3. Appendix B gives full results for the three format studies of Section 4, Appendix C for the attachment test and its follow-ups in Section 5, Appendix D for the models and scale series in Section 6, and Appendix E for the ranking and MMVP analyses in Section 7. Appendix F reports the transfer attempts behind our scope limitation, and Appendix G the resampling design.

A MATERIALS AND PROMPTS

Items and panels. Each study draws its own words and photographs (Table A1), so no word or photograph is shared between the Naming, Binding, Direct-presentation, and attachment panels. The Naming panel uses photographs from six US cities (Madison, Pittsburgh, Portland, Providence, Sacramento, Savannah) and six UK cities (Canterbury, Exeter, Norwich, Nottingham, Oxford, York). Within an edge, the sign is pasted with identical pixels into the US and the UK photograph. The factorial follow-up and the scale series reuse the attachment confirmation panel, so every analysis in Sections 5 and 6 uses the same 128 edges.

Table A1: Material panels. Each study uses its own words and photographs; an edge is one word combined with one US–UK photograph pair. The factorial follow-up and the scale series reuse the attachment confirmation panel.

Study	Words	Photograph pairs	Edges	Models
Naming (§4.1)	64	24	384	all eight
Binding (§4.2)	48	24	288	two primary
Direct presentation (§4.3)	48	24	288	all eight
Attachment, development (§5.1)	12	6	36	two primary
Attachment, confirmation (§5.1)	16	16	128	all eight
Position × assignment (§5.2)	16	16	128	two primary
Scale series (§6.2)	16	16	128	Qwen3-VL 2B–32B

Prompts. All formats share one template (Table A2). Regional labels are `United States` and `United Kingdom`; the two neutral assignments use X/Y and Y/X. Indirect prompts add a key line that maps each label to a spelling and list the options by label only. Direct prompts list each spelling beside its label. Detached prompts list the spellings without labels and add the line *Auxiliary task annotations (not assigned to any answer candidate):* followed by the labels in one of two orders. Plain prompts have no labels. Neither and Cannot read the sign are options C and D before rotation.

Table A2: Matched HIGH and GOLD control prompts in the direct format with neutral labels, with original wording and line breaks. HIGH shows the sign alone on a gray background; GOLD gives the target as text with a blank gray image. Scene items use the HIGH wording with a street photograph.

Control	Prompt
HIGH	Which reference entry exactly matches the visible sign? Match the written form exactly. A. DENUCLEARIZE — X B. DENUCLEARISE — Y C. Neither reference entry D. Cannot read the sign Answer with exactly one letter: A, B, C, or D.
GOLD	Target transcription: DENUCLEARISE. Which reference entry exactly matches the target transcription? Match the written form exactly. A. DENUCLEARIZE — X B. DENUCLEARISE — Y C. Neither reference entry D. Cannot read the sign Answer with exactly one letter: A, B, C, or D.

Evidence levels and control items. MEDIUM lowers the maximum font-size factor from 1.00 to 0.82, lightens the text from gray level 24 to 64 (on a 0–255 scale), and applies a Gaussian blur of radius 0.45 to the text layer; the sign geometry and background are unchanged. Because the font is fitted adaptively, the final size ratio can differ slightly by word. Two control items remove the scene: HIGH shows the HIGH-evidence sign alone on a gray background, and GOLD gives the target spelling as text with a blank gray image. On these items we swap the target spelling and measure how far responses move, which we call target responsiveness ρ ($\rho = 1$ means responses always follow the target).

Readout and response coding. The four answer keys are single tokens at the model's native assistant boundary, so one forward pass yields all four scores. Choices are decoded into the American spelling, the British spelling, Neither, Cannot read, or a tie; regional-label choices are mapped through the current assignment to a spelling. No analysis contains missing or non-finite scores. The stimuli are programmatic composites that have not been rated for naturalness by humans, and Figure 1 shows original experimental pixels with vector annotations only.

B The Three Answer-Format Studies

B.1 Naming

Table B1 gives, for every model, the scene effect under regional and neutral labels, the naming effect, and accuracy. The naming effect is positive in all eight models. It does not hinge on a single location or photographer: for the six further models, leaving out any one city, or any photographer who contributed several photographs, changes J by at most 2.6 pp.

Table B1: Naming study (indirect options). Scene effects τ under regional and neutral labels, their difference J, and balanced accuracy. Neutral averages the two X/Y assignments. †95% intervals; other models Bonferroni-adjusted for six comparisons.

	τ (pp)		Naming effect		BA (%)	
Model	regional	neutral	J (pp)	Interval	regional	neutral
Qwen3-VL-32B	7.8	0.2	7.7	[2.4, 16.4]	96.0	98.3
Qwen3-VL-8B†	22.4	5.3	17.1	[8.7, 26.3]	88.4	79.1
Ministral-3-8B	65.9	12.6	53.4	[42.3, 64.8]	65.8	85.6
GLM-4.6V-Flash†	15.3	−0.1	15.5	[10.0, 21.8]	91.9	96.1
InternVL3.5-8B	23.1	5.6	17.6	[8.8, 27.2]	87.0	89.4
Llama-3.2-11B	65.5	1.2	64.3	[53.5, 73.8]	62.2	71.0
Gemma-3-12B	45.3	6.8	38.5	[28.0, 49.5]	71.1	66.8
Idefics3-8B	58.2	16.6	41.6	[29.1, 53.6]	60.5	73.0

B.2 BINDING

How much of the effect follows the labels? Table B2 splits the signed effect into a part that flips with the assignment (B) and a part that does not (R). B dominates in both models, which is the sense in which the scene's influence follows the labels. R is negative because the effect is larger under the swapped assignment than under the usual one, so the two assignments do not combine additively.

Table B2: Binding study. Signed scene effect E (pp toward the American spelling when a US scene replaces a UK scene) under the usual and swapped assignments, with 95% intervals. $B = (E_{\text{usual}} - E_{\text{swap}})/2$ follows the labels; $R = (E_{\text{usual}} + E_{\text{swap}})/2$ does not. BA is balanced accuracy (%).

	Signed effect E (pp)		Components (pp)		BA (%)	
Model	usual	swapped	B	R	usual	swapped
Qwen3-VL-8B	19.7 [12.5, 27.8]	−47.4 [−57.9, −36.8]	33.5 [26.5, 40.8]	−13.8 [−19.7, −7.9]	89.5	72.8
GLM-4.6V-Flash	17.4 [12.6, 22.8]	−33.6 [−40.4, −26.9]	25.5 [20.7, 30.4]	−8.1 [−11.6, −4.7]	91.2	75.5

Why only two models? A model can take the Binding test only if it reads a label that contradicts the label's usual meaning. The study therefore admits models only if they meet fixed control thresholds under both assignments before any scene item is run, and stops the whole roster otherwise. Among the six further models, only Qwen3-VL-32B met the thresholds (Table B3), so no Binding scene items were run for them. This is a limit of the test, not evidence that the effect is absent in these models.

Table B3: Binding controls for the six further models. The study requires, under both assignments, at least 346/384 correct on HIGH and 365/384 on GOLD before any scene item is run, and stops the whole roster if any model fails. *Below threshold. Only Qwen3-VL-32B qualified, so no Binding scene items were run for these models.

	Usual assignment		Swapped assignment		
Model	HIGH	GOLD	HIGH	GOLD	Qualified
Ministral-3-8B	320/384*	306/384*	283/384*	270/384*	no
Qwen3-VL-32B	380/384	378/384	364/384	380/384	yes
Gemma-3-12B	354/384	361/384*	310/384*	333/384*	no
Llama-3.2-11B	224/384*	130/384*	160/384*	153/384*	no
Idefics3-8B	308/384*	321/384*	235/384*	283/384*	no
InternVL3.5-8B	318/384*	375/384	271/384*	369/384	no

B.3 DIRECT PRESENTATION

Table B4 gives the naming effect under indirect and direct options on the same panel. The direct-option effect is positive in all eight models, and leaving out any one city or repeated photographer changes it by at most 2.4 pp. Its difference from the indirect-option effect, $\Lambda = J_I - J_D$, takes both signs across models. The comparison between the two primary models also changes with the format, and the change is reliable (-17.2 pp $[-25.9, -8.3]$ in the Qwen3-VL-8B minus GLM-4.6V-Flash difference of Λ). Both primary models met the pre-specified responsiveness rule on controls (Table B5); GLM-4.6V-Flash is less accurate on GOLD mainly because it often chooses Neither (383 of 1,152 responses) even when the target is given as text.

Table B4: Direct-presentation study. Naming effects with indirect (J_I) and direct (J_D) options on the same panel, their difference $\Lambda = J_I - J_D$ (bold: interval excludes zero), and balanced accuracy under regional labels. †95% intervals; other models Bonferroni-adjusted for six comparisons.

	Naming effect (pp)			Difference (pp)		BA, regional (%)	
Model	J_I	J_D	Interval	Λ	Interval	indirect	direct
Qwen3-VL-32B	3.3	13.9	[9.0, 19.2]	**−10.6**	[−16.1, −5.2]	98.2	91.8
Qwen3-VL-8B†	11.6	33.2	[27.2, 39.4]	**−21.5**	[−29.5, −13.5]	93.1	75.2
Ministral-3-8B	49.6	40.4	[33.2, 48.3]	9.2	[−0.7, 19.1]	68.5	72.9
GLM-4.6V-Flash†	17.5	21.9	[17.4, 26.9]	−4.4	[−8.9, 0.3]	90.3	77.9
InternVL3.5-8B	19.5	41.9	[32.0, 51.1]	**−22.4**	[−33.2, −11.6]	87.2	74.7
Llama-3.2-11B	59.2	41.3	[34.1, 48.6]	**18.0**	[7.4, 27.5]	62.0	69.1
Gemma-3-12B	40.1	34.7	[27.3, 42.8]	5.4	[−4.2, 14.5]	73.7	70.8
Idefics3-8B	39.3	43.5	[32.6, 54.0]	−4.2	[−17.1, 9.3]	56.5	46.5

Table B5: Direct-presentation controls for the primary models: target responsiveness ρ with 95% word-bootstrap intervals, and GOLD exact-match accuracy over 1,152 responses (48 words × 3 label configurations × 2 target spellings × 4 rotations).

Model	Control	ρ	Interval	GOLD accuracy
Qwen3-VL-8B	HIGH	0.923	[0.887, 0.953]	
	GOLD	0.874	[0.833, 0.911]	1,020/1,152 (88.5%)
GLM-4.6V-Flash	HIGH	0.810	[0.781, 0.835]	
	GOLD	0.420	[0.363, 0.477]	619/1,152 (53.7%)

C THE ATTACHMENT TEST AND ITS FOLLOW-UPS

C.1 FORMATS AND PRE-SPECIFIED CRITERIA

The attachment test shows the spellings directly in every format and varies only the labels (Table C1). For each family (Attached or Detached), the naming effect is regional minus neutral, with neutral averaging its two configurations; in the Detached family the regional effect also averages the two label orders. The development set (12 words, 6 photograph pairs, 36 edges) was used only to fix the prediction and the analysis. The pre-specified success criterion had two parts: the lower 95% bound on K had to be positive for each primary model, and target responsiveness had to have a positive lower bound on HIGH and GOLD controls under both Attached and Detached labels, eight checks in all, so that a positive K could not come from a model that had stopped reading the target. Both parts were met (Table C2). The two primary models do not differ reliably in K (-3.0 pp $[-9.5, 3.6]$).

Table C1: The eight label formats of the attachment study. All show the two spellings directly as options; they differ in which labels appear and where. Neutral and detached formats have two counterbalanced configurations each.

Format	Labels	Placement	Relation to the options
Attached, regional	regional names	beside each spelling	assigned
Attached, X/Y (two assignments)	X / Y	beside each spelling	assigned
Detached, regional (two orders)	regional names	separate line	stated as not assigned
Detached, X/Y (two orders)	X / Y	separate line	stated as not assigned
Plain	none	—	—

Table C2: Pre-specified control checks for the attachment test (confirmation set). Target responsiveness ρ (1 means responses always follow the target) with 95% word-bootstrap intervals, exact-match accuracy, and number of control responses. All eight lower bounds are positive.

Model	Labels	Control	ρ	Interval	Accuracy (%)	n
Qwen3-VL-8B	Attached	HIGH	0.979	[0.956, 0.995]	98.4	384
	Attached	GOLD	0.953	[0.927, 0.977]	95.6	384
	Detached	HIGH	1.000	[1.000, 1.000]	100.0	512
	Detached	GOLD	1.000	[1.000, 1.000]	100.0	512
GLM-4.6V-Flash	Attached	HIGH	0.716	[0.648, 0.779]	77.9	384
	Attached	GOLD	0.539	[0.477, 0.602]	63.0	384
	Detached	HIGH	0.834	[0.736, 0.924]	91.2	512
	Detached	GOLD	0.814	[0.744, 0.883]	86.3	512

C.2 Where Does the Scene Effect Occur?

Table C3 gives the scene effect under each format with intervals; Figure 3a shows the point estimates. In every model, the largest scene effect occurs with attached regional labels.

Table C3: Scene effect τ (pp) under each format of the attachment study on the confirmation set, with pointwise 95% intervals; the largest effect in each row is in bold. Neutral and detached-regional entries average their two configurations.

	Attached		Detached		
Model	regional	X/Y	regional	X/Y	Plain
Qwen3-VL-32B	**0.5** [−1.0, 2.6]	0.1 [0.0, 0.3]	0.0 [0.0, 0.1]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Qwen3-VL-8B†	**8.5** [2.5, 14.9]	1.4 [−1.0, 5.2]	1.4 [−0.0, 4.3]	1.7 [0.1, 5.0]	0.8 [−0.6, 3.6]
Ministral-3-8B	**30.1** [19.6, 39.5]	6.0 [2.7, 9.7]	9.6 [4.2, 16.0]	3.8 [1.3, 7.0]	4.1 [1.4, 7.3]
GLM-4.6V-Flash†	**8.5** [3.3, 13.6]	−2.7 [−6.3, 0.3]	2.5 [−1.4, 6.7]	1.5 [−1.9, 4.6]	1.9 [−0.5, 5.2]
InternVL3.5-8B	**19.9** [8.6, 30.2]	1.4 [−1.1, 3.9]	2.0 [−0.4, 4.9]	0.1 [−2.0, 2.3]	0.4 [−2.1, 2.8]
Llama-3.2-11B	**46.3** [35.2, 56.5]	6.3 [1.9, 12.1]	4.2 [−1.0, 9.6]	4.0 [−1.6, 11.1]	4.5 [−1.6, 11.8]
Gemma-3-12B	**28.3** [17.5, 39.2]	11.1 [3.7, 18.9]	6.3 [−1.2, 14.3]	2.4 [−4.1, 8.9]	3.3 [−2.9, 9.5]
Idefics3-8B	**41.4** [29.6, 53.1]	6.4 [−1.9, 15.3]	4.1 [−6.1, 14.6]	6.8 [−3.5, 18.1]	10.0 [1.1, 19.7]

C.3 Does the Stated Assignment Matter?

The factorial follow-up crosses two factors on the confirmation panel for the two primary models: whether the labels stand beside the spellings or on a separate line, and whether the prompt assigns them to the options or states that they are not assigned. When the labels are unassigned but beside the options, their placement is counterbalanced, so each regional name stands either beside its own country's spelling or beside the other spelling. Table C4 gives the scene effect under regional labels

for every placement and the naming effect in each cell. The naming effect of the "beside, unassigned" cell averages the two placements and is negative, because the swapped placement produces the larger effect.

Table C4: Position × stated assignment on the confirmation panel (two primary models, 114,688 responses). Upper block: scene effect τ (pp) under regional labels for each placement, with 95% intervals. Lower block: naming effects J (regional minus neutral within each cell) and their interaction. With unassigned labels beside the options, the two placements put each regional name beside its own country's spelling or beside the other spelling. Bold: interval excludes zero.

	Qwen3-VL-8B		GLM-4.6V-Flash	
Cell	Est.	Interval	Est.	Interval
Beside options, assigned (= Attached)	$\mathbf{6.7}$	$[1.5, 12.9]$	$\mathbf{8.0}$	$[4.1, 12.2]$
Beside options, declared unassigned, own spelling	$\mathbf{6.8}$	$[1.5, 12.8]$	$\mathbf{8.1}$	$[3.2, 13.3]$
Beside options, declared unassigned, other spelling	$\mathbf{-16.5}$	$[-25.9, -8.0]$	$\mathbf{-18.2}$	$[-24.6, -11.9]$
Separate line, assigned	1.5	$[-2.2, 5.1]$	2.1	$[-0.8, 5.4]$
Separate line, unassigned, order 1 (= Detached)	1.4	$[-2.2, 5.1]$	1.5	$[-1.6, 4.9]$
Separate line, unassigned, order 2 (= Detached)	0.7	$[-3.3, 4.8]$	1.0	$[-1.7, 3.9]$
J, beside and assigned	5.5	$[0.8, 11.2]$	6.6	$[2.8, 10.8]$
J, beside and unassigned (both placements)	-5.7	$[-9.7, -2.5]$	-6.4	$[-10.2, -2.6]$
J, separate and assigned	-1.0	$[-4.3, 1.8]$	1.4	$[-0.4, 3.7]$
J, separate and unassigned	-1.2	$[-4.6, 1.4]$	0.6	$[-1.4, 2.9]$
Interaction $(J_{BA} - J_{BU}) - (J_{SA} - J_{SU})$	11.0	$[5.1, 17.3]$	12.2	$[7.2, 17.1]$

C.4 Does the Effect Depend on How Legible the Sign Is?

Table C5 splits the attachment effect by sign evidence. In every model the difference between MEDIUM and HIGH is small and its interval includes zero. This split was not pre-specified and is not an additional confirmation.

Table C5: Attachment effect K (pp) at the two levels of sign evidence, and their paired difference. MEDIUM renders the word smaller, lighter, and slightly blurred (Appendix A). This split was not pre-specified. †95% intervals; other models 99.2% intervals (Bonferroni for six comparisons).

	HIGH		MEDIUM		MEDIUM − HIGH	
Model	K	Interval	K	Interval	Diff.	Interval
Qwen3-VL-32B	-0.2	$[-2.9, 2.8]$	1.0	$[-0.9, 4.4]$	1.2	$[-0.4, 3.7]$
Qwen3-VL-8B†	7.3	$[1.7, 13.4]$	7.4	$[1.9, 13.1]$	0.1	$[-2.6, 2.8]$
Ministral-3-8B	18.3	$[5.5, 30.6]$	18.3	$[6.3, 30.0]$	0.0	$[-3.7, 4.2]$
GLM-4.6V-Flash†	10.2	$[4.0, 16.4]$	10.4	$[4.0, 16.8]$	0.2	$[-2.8, 3.4]$
InternVL3.5-8B	16.2	$[3.3, 27.8]$	17.0	$[3.4, 28.7]$	0.9	$[-3.5, 5.8]$
Llama-3.2-11B	39.6	$[23.5, 56.1]$	40.0	$[23.5, 56.5]$	0.4	$[-5.3, 5.8]$
Gemma-3-12B	14.1	$[2.5, 25.9]$	12.5	$[0.2, 24.0]$	-1.6	$[-8.2, 6.6]$
Idefics3-8B	36.8	$[23.5, 52.7]$	38.5	$[23.3, 54.8]$	1.7	$[-6.2, 10.2]$

D MODELS AND THE SCALE SERIES

Table D1: Evaluated models and where each is used. Several share language backbones (Qwen for the Qwen models and InternVL; Llama for Llama and Idefics), so they are not independent model families.

Name in the paper	Checkpoint	Used in
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	all studies (primary)
GLM-4.6V-Flash	`zai-org/GLM-4.6V-Flash`	all studies (primary)
Ministral-3-8B	`mistralai/Ministral-3-8B-Instruct-2512-BF16`	§§4.1–5.1; MMVP
Qwen3-VL-32B	`Qwen/Qwen3-VL-32B-Instruct`	§§4.1–5.1, 6.2; MMVP
Gemma-3-12B	`google/gemma-3-12b-it`	§§4.1–5.1; MMVP
Llama-3.2-11B	`meta-llama/Llama-3.2-11B-Vision-Instruct`	§§4.1–5.1; MMVP
Idefics3-8B	`HuggingFaceM4/Idefics3-8B-Llama3`	§§4.1–5.1; MMVP
InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B-Instruct`	§§4.1–5.1; MMVP
Qwen3-VL-2B, 4B	`Qwen/Qwen3-VL-{2B,4B}-Instruct`	§6.2

D.1 IS THE 8B–32B GAP AN ARTIFACT OF THE PIPELINE?

The four Qwen3-VL checkpoints were run through one pipeline: BF16 weights, eager attention in the text and vision towers, one NVIDIA A100 80GB per process, batch size 1, the native processor and chat template, and the literal prefix `Answer:` followed by a newline after the assistant header. Each checkpoint received the same 32,768 presentations of the confirmation panel. Parameter counts include the vision encoder. The 2B and 4B checkpoints use a smaller vision encoder than 8B and 32B, and training differs across checkpoints, so the series is observational. Table D2 gives all estimates. Table D3 compares the reruns of 8B and 32B with their original runs on exactly the same presentations: the reruns reproduce the originals, while the 8B–32B gap remains.

Table D2: Scale series: Qwen3-VL checkpoints run through one matched pipeline on the confirmation panel. Effects in pp with 95% intervals for K; BA in %. The pre-specified trend (OLS of K on log total parameters) is -4.0 pp per doubling $[-5.6, -2.4]$.

	Params	Effects (pp)				BA (%)		
Model	(B)	K	Interval	J_{att}	J_{det}	Attached	Detached	Plain
Qwen3-VL-2B	2.13	**16.0**	[8.6, 23.7]	17.4	1.3	67.3	91.5	89.7
Qwen3-VL-4B	4.44	**11.6**	[5.6, 18.6]	11.9	0.3	89.0	91.7	94.4
Qwen3-VL-8B	8.77	**8.3**	[1.9, 14.8]	8.1	−0.2	90.6	98.4	99.0
Qwen3-VL-32B	33.36	0.2	[−1.5, 2.3]	0.2	0.0	98.8	100.0	100.0

Table D3: Matched-pipeline reruns of Qwen3-VL-8B and 32B on exactly the same presentations as the original runs (new minus original), and the 8B minus 32B difference under the matched pipeline, in pp with 95% intervals.

	8B: rerun − original		32B: rerun − original		Rerun: 8B − 32B	
Quantity	Est.	Interval	Est.	Interval	Est.	Interval
K	0.9	[−1.2, 3.4]	−0.2	[−0.8, 0.4]	8.1	[2.4, 14.1]
J_{att}	1.0	[−1.1, 3.4]	−0.2	[−0.9, 0.3]	7.9	[2.3, 13.8]
BA, attached	−0.8	[−2.1, 0.6]	−0.0	[−0.3, 0.2]	−8.2	[−12.9, −4.5]
BA, plain	−0.2	[−0.8, 0.1]	0.0	[0.0, 0.0]	−1.0	[−3.1, 0.0]

E RANKINGS AND THE MMVP BENCHMARK

E.1 DO FORMATS REORDER MODELS?

We rank the eight models by scene effect under each of the five formats of the attachment test and compare every pair of formats (Table E1). A reversal is reliable only if the two models' paired difference has intervals excluding zero with opposite signs in the two formats. Point orderings flip for up to eight of the 28 model pairs, but no flip is reliable in any of the ten format comparisons. This analysis was not pre-specified.

Table E1: Model rankings by scene effect τ across the five formats of the attachment study (eight models). Kendall τ_b with pointwise 95% intervals, the number of model pairs whose point ordering flips (of 28), and the number of reliable reversals, which require both paired intervals to exclude zero with opposite signs. This analysis was not pre-specified.

Format pair	τ_b	Interval	Point flips	Reliable
Attached regional vs. Attached X/Y	0.50	$[0.21, 0.79]$	7	0
Attached regional vs. Detached regional	0.64	$[-0.14, 0.79]$	5	0
Attached regional vs. Detached X/Y	0.71	$[-0.21, 0.79]$	4	0
Attached regional vs. Plain	0.79	$[0.00, 0.86]$	3	0
Attached X/Y vs. Detached regional	0.43	$[0.00, 0.86]$	8	0
Attached X/Y vs. Detached X/Y	0.64	$[-0.14, 0.79]$	5	0
Attached X/Y vs. Plain	0.57	$[-0.07, 0.79]$	6	0
Detached regional vs. Detached X/Y	0.50	$[0.14, 0.86]$	7	0
Detached regional vs. Plain	0.57	$[0.00, 0.86]$	6	0
Detached X/Y vs. Plain	0.93	$[0.18, 0.98]$	1	0

E.2 DOES THE INTERFACE MATTER ON A NATURAL BENCHMARK?

MMVP (Tong et al., 2024) consists of 150 pairs of natural images that share a question and have opposite answers. We keep the official question and the two answer contents and compare two presentations under the same A/B readout. The direct format shows the two contents as options (two orders). The indirect format names the contents X and Y in a key line and lists X and Y as options (two key assignments $\times$ two orders). A pair counts as correct only if both images are answered correctly under the same presentation. The pre-specified equivalence margin was ± 3 pp; no model met it, and four of the eight Bonferroni-adjusted intervals exclude zero (Table E2). MMVP pairs are different original images rather than target-preserving counterfactuals, so this study concerns the interface's effect on accuracy, not the scene effect.

Table E2: MMVP with direct options versus neutral names plus an answer key (150 image pairs, 14,400 requests). Pair accuracy counts a pair as correct only if both images are answered correctly under the same presentation. Δ is indirect minus direct pair accuracy with Bonferroni-adjusted intervals for eight models (bold: interval excludes zero). Kendall τ_b between the two rankings is 0.74 $[0.57, 0.93]$; no model pair reverses reliably.

	Pair accuracy (%)		Difference (pp)		Item accuracy (%)	
Model	direct	indirect	Δ	Interval	direct	indirect
Qwen3-VL-32B	67.0	63.7	-3.3	$[-7.3, 0.3]$	82.5	80.9
Qwen3-VL-8B	58.3	55.2	-3.2	$[-8.5, 2.3]$	78.0	75.6
Ministral-3-8B	29.0	26.5	-2.5	$[-7.7, 2.2]$	62.2	61.2
GLM-4.6V-Flash	43.3	38.7	-4.7	$[-10.8, 1.3]$	70.8	68.5
InternVL3.5-8B	49.3	37.8	$\mathbf{-11.5}$	$[-17.7, -5.3]$	73.3	67.0
Llama-3.2-11B	38.7	22.5	$\mathbf{-16.2}$	$[-23.2, -9.5]$	68.2	60.2
Gemma-3-12B	46.7	38.7	$\mathbf{-8.0}$	$[-14.2, -2.0]$	72.2	68.4
Idefics3-8B	49.3	40.2	$\mathbf{-9.2}$	$[-15.8, -2.8]$	71.7	67.2

F Preliminary Transfer Attempts

We ran four small development batches to see whether a comparable answer-format contrast could be measured in other tasks, using questions from CLEVR (Johnson et al., 2017) and VQAv2 (Goyal et al., 2017). Each batch had a pre-specified criterion for continuing to a held-out confirmation; none met it, so no confirmation was run (Table F1). The criteria required agreement between a choice readout and a free-generation readout, minimum effect sizes, and adequate performance on cue checks. Batches fell short for different reasons: effects below the required size, worst-case bounds that included zero once invalid free-generation outputs were counted, parser-invalid outputs, or cue accuracy below threshold. These outcomes do not show that the effect is absent in those tasks, but they provide no evidence that it transfers. In these tasks the answer labels do not share a semantic dimension with a context manipulation, the condition under which Scene-Orthography produces its effect.

Table F1: Preliminary transfer batches. None met its pre-specified criterion for continuing to a held-out confirmation, so no confirmation was run. Call counts include repeated conditions.

Batch	Tasks and materials	Models	Calls
CLEVR counting and size	Count and `equal_size`; SHORT/FULL conditions; 32 roots	8	28,064
VQAv2 binary questions	Yes/No; 32 components	8	4,672
VQAv2 small counts	32 development units; 128 reserved units	8	23,392
Cross-task attachment	CLEVR `exist` and `equal_size`; 32 development and 128 reserved roots per task	8	9,472

G Resampling Design

Every scene analysis treats the observed word–photograph edges as its units. Words and photograph pairs each appear in several edges, so we resample both jointly with a crossed bootstrap and recompute every estimate from edge means in each draw. Option rotations, evidence levels, and repeated label configurations add precision within an edge but are not treated as independent samples. All conditions share the same draws, so every contrast has a paired bootstrap distribution, and intervals are percentile intervals. Intervals for the six further models are Bonferroni-adjusted for six comparisons within each study, and MMVP intervals for eight models (Table G1).

Table G1: Resampling units and interval families. All scene analyses resample words and photograph pairs jointly over the observed edges with 100,000 draws, keeping every condition paired within a draw; MMVP resamples image pairs.

Analysis	Units	Intervals
Naming, Binding, Direct (primary models)	words $\times$ photograph pairs of each panel	95%
Naming, Direct (six further models)	same panels	Bonferroni, six models
Attachment test	16 words $\times$ 16 photograph pairs	95%; Bonferroni, six models
Scene effect by format, rankings	same 128 edges	pointwise 95%
Evidence levels	same 128 edges	95%; 99.2% for six models
Position $\times$ assignment, scale series	same 128 edges	95%
MMVP	150 image pairs	95%; Bonferroni, eight models