RANKING IS NOT ANSWERING: WHEN VLM SCORES REVERSE MODEL COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We score each description of a compositional benchmark pair with one yes/no score and compare two criteria, whether the correct description ranks above the incorrect one, and whether each description gets the right answer on its own. On ARO, answering yes whenever "yes" is more probable than "no" reverses two model comparisons. A threshold fitted on held-out images removes both, yet 20.8 to 62.0% of correctly ranked pairs still contain a wrong standalone decision, and a smaller reversal appears. How much of the gap the threshold recovers varies by model, from most of it for InternVL3 on attribution to almost none for Qwen2.5-VL on relation. Across five models and six subsets, ranking exceeds standalone accuracy, and with both descriptions in one prompt the A/B scores of four models track ranking. Benchmarks that report only paired ranking can misorder models; we recommend also reporting standalone accuracy and the decision threshold behind it.

1 INTRODUCTION

Compositional vision-language benchmarks compare a correct image description with a matched incorrect alternative. Changing an object, attribute, relation, or argument order while preserving much of the surrounding content helps to prevent lexical shortcuts and asks whether a model detects the intended difference (Thrush et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023). Paired ranking is useful because relative scores can expose discrimination even when a shared response tendency or decision boundary affects both. However, it is important to note thata successful ranking does itself establish that the model judged each description correctly on its own.

Consider a correct description whose score favours "no" and an incorrect alternative whose score favours "no" even more strongly. For example, scores of -0.2 and -1.0 give the correct description the higher rank, but only the incorrect description receives a correct standalone decision. The paired result captures a useful relative preference but the standalone result reveals that the model would reject the correct description alone. Pairwise selection and independent judgement are both meaningful as they answer different questions.

This distinction matters for instruction-following vision-language models, which usually judge one user-provided statement rather than choose between labelled alternatives. We present each member of a benchmark pair separately under the same yes/no instruction and compare the resulting scores in two ways: whether the correct description ranks above the incorrect one, and whether each receives its standalone answer. A shared shift in both scores can preserve ordering while ultimately changing the standalone decisions. VALSE already treated ranking as the more permissive measure on separately scored captions (Parcalabescu et al., 2022). NaturalBench noted that a threshold fit inside each sample reflects ranking rather than answering one pair, and that one global threshold helps much less (Li et al., 2024). We ask what these results leave open: when the two criteria disagree on the same scores, which disagreements can a single threshold resolve, and what accounts for the rest?

We evaluate five models: Qwen2.5-VL 7B, Qwen3-VL 8B, LLaVA-1.5 7B, InternVL3 9B, and Idefics3 8B (Bai et al., 2025b;a; Liu et al., 2023; 2024; Zhu et al., 2025; Laurençon et al., 2024) on ARO attribution and relation, three SugarCrepe edit or swap subsets, and per-image Winoground (Yuksekgonul et al., 2023; Hsieh et al., 2023; Thrush et al., 2022). Claims about thresholds, model order, and the accounting of the gap use the ARO scoring with single-token answer groups, annotated image crops, and image-disjoint splits. Based on that scoring, when a description counts as true

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.

whenever "yes" is more probable than "no", the two criteria reverse InternVL3 and Idefics3 on attribution tasks, and Idefics3 and LLaVA on relational tasks. A threshold fit on held-out images removes both these apparent reversals but we find a new, smaller one: on relation, paired ranking favours Qwen2.5-VL by 0.7 points, while standalone accuracy at the fitted threshold favours Qwen3-VL by 0.6. Even with that threshold, 20.8 to 62.0% of correctly ranked pairs still contain a wrong standalone decision. We find thatmuch of the gap the threshold recovers depends on the model, 62% of InternVL3's gap on ARO attribution, and almost none of Qwen2.5-VL's on ARO relation. Ranking each correct description against incorrect descriptions from other items already beats the best threshold by 5.2 to 7.2 points in every combination. Ranking against the incorrect description of the same item adds up to 12.9 points in some combinations as well.

A second scoring with a different prompt covers all six subsets and we use it only to confirm that paired ranking exceeds standalone accuracy beyond ARO, which it does for every model and subset. Among the 24 combinations without LLaVA the gaps range from 5.4 to 23.4 points; LLaVA's larger gaps come from answering yes on most descriptions and are reported separately (Appendix E). Across three prompt wordings, standalone accuracy changes more than paired ranking. When both descriptions appear in one prompt, the A/B scores of the four models other than LLaVA-1.5 come within 6.6 points of paired ranking. Replacing an image with uniform grey lowers paired ranking by 20.0 to 43.9 points.

Our contributions are as follows.

- A same-score reversal analysis showing which model-order disagreements a threshold fixed on held-out images removes and which remain (Table 2, Section 5).
- An exact decomposition of the gap into threshold gain, unpaired contrast, and pairing contrast, showing that the same headline gap is a different combination of these three terms in different models (Eq. 5, Figure 2).
- A reporting recommendation: paired ranking, standalone accuracy at the usual cutoff and at a cutoff fitted on held-out images, and the cutoff itself (Sections 4 and 7).

2 Related Work

Compositional benchmarks test whether a model combines entities, attributes, and relations rather than matching them independently. Winoground presents two images and two captions that use the same words or morphemes in different orders. Its text and image scores each require both within-example comparisons in one matching direction, and its group score requires all four image–caption comparisons (Thrush et al., 2022). ARO (Attribution, Relation, and Order) changes attributes, relations, or argument order in controlled captions and asks whether the appropriate caption ranks higher (Yuksekgonul et al., 2023). SugarCrepe keeps a two-candidate image–caption selection protocol but reconstructs and verifies incorrect captions to reduce benchmark shortcuts (Hsieh et al., 2023), and SugarCrepe++ adds a second, lexically different correct caption to each item (Dumpala et al., 2024). VL-CheckList likewise requires a correct description to outscore generated negatives (Zhao et al., 2022). What's Up introduces synthetic, natural-image, and photographic spatial-relation sets and evaluates whether models select the correct caption under controlled relation changes using pairwise and set-wise accuracy rather than Winoground's text/image/group scores (Kamath et al., 2023). We instead score each benchmark alternative independently with fixed yes/no groups, then compute paired ranking and standalone accuracy from the same scores.

Several evaluations compare paired and pointwise measures. VALSE scores caption and foil separately with image–text matching heads and reports pairwise ranking accuracy, binary accuracy, and AUROC; it calls the pairwise measure more permissive (Parcalabescu et al., 2022). For vision-language decoders, Parcalabescu & Frank (2025) compare a prompt that offers both captions as options A and B with a prompt that shows one caption, and find accuracy 12 to 33 points higher in the first. Our A/B readout follows that format, while our paired ranking scores each caption on its own. NaturalBench scores decoder answers by likelihood and fits a threshold within each sample, which it notes reflects ranking and departs from answering a single image–question pair while a single global threshold helps much less (Li et al., 2024). Chan et al. (2026) show that a bias toward answering yes lowers yes/no accuracy, most of all when both answers of a pair must be correct, and reduce it by redistributing attention. POPE and HallusionBench also measure this bias (Li et al., 2023; Guan et al., 2024).

Generative image–text scores use the probability of a caption, which needs correction for the prior probability of the text (Lin et al., 2024a), or the probability of "yes" to a question about the image (Lin et al., 2024b), close to our score. For language models, prompted judgments are less accurate than direct comparisons of probabilities over minimal pairs (Hu & Levy, 2023), and probability-based and generated answers often disagree (Lyu et al., 2024). Relative to this work, we fit one threshold on held-out images, ask which model-order disagreements it removes, and split what it leaves into unpaired discrimination and pairing using the relation between ranking and AUROC (Green & Swets, 1966).

Benchmark audits further show why a score depends on item construction and the abilities it combines. Diwan et al. (2022) analyze why Winoground is difficult, separating failures associated with visual, linguistic, and compositional demands. Their question is essentially which capabilities and dataset properties make the benchmark hard. The 2025 Good CREPE preprint audits biases in 17 compositional benchmarks, including effects of data sources and negative-example construction (Udandarao et al., 2025), and Kamath et al. (2024) show that benchmarks built from hard negatives omit hard positives, descriptions that differ in form but remain true. These studies audit benchmark difficulty and shortcut sensitivity. In contrast, we hold each item and score fixed and compare paired ranking with standalone accuracy. Our claim concerns outcome relationships, not necessarily whether items are unbiased or difficult.

Instead of parsing free-form answers, which depend on decoding, format compliance, verbosity, and parsing, we read the model's probabilities for a fixed set of "yes" and "no" tokens and take their difference as the score (Section 3). Paired ranking and standalone accuracy then come from the same number. The standalone decision has an explicit boundary: yes when the score is above the threshold. On the ARO scoring, the models' own generated answers match the sign of this score on 98.2 to 99.9% of statements.

3 Evaluation Setup

Thresholds, model order, and the accounting of the gap use a scoring of the two ARO subsets with single-token answer groups, annotated image crops, and image-disjoint splits (Appendix D). A second scoring of all six subsets, with a longer prompt and the images as shipped, is used only to check that the sign of the gap is not specific to ARO. In both, standalone accuracy is agreement with the benchmark's labels, which mark one description of each pair as correct. Section 7 discusses items where both are true.

Each description is presented separately with an instruction to answer yes or no. We compare only fixed groups of next tokens that express those two answers. For image x and description q, let Y and N denote the yes-token and no-token groups, and let ℓ_i be a token's next-token log probability. We call the log probability of the yes group minus that of the no group the *restricted yes/no log margin*:

$$m(x, q) = \log \sum_{i \in Y} e^{\ell_i} - \log \sum_{i \in N} e^{\ell_i}. \tag{1}$$

A positive margin favours yes and a nonpositive margin favours no. For a correct description q_t and its incorrect alternative q_f, write $m_t = m(x, q_t)$ and $m_f = m(x, q_f)$. Standalone correctness and paired ranking are

$$B = \frac{1}{2}\left(\mathbf{1}[m_t > 0] + \mathbf{1}[m_f \leq 0]\right), \qquad C = \mathbf{1}[m_t > m_f], \qquad G = C - B. \tag{2}$$

Thus B is the mean correctness of the two standalone decisions, while C represents whether or notthe correct description's margin exceeds the incorrect one's. The correct description is true when $m_t > 0$; the incorrect alternative is false when $m_f \leq 0$. Paired ties are misses because C uses a strict inequality. A toy pair with $m_t = -0.2$ and $m_f = -1$ has $C = 1$ and $B = 1/2$, and the correct description is ranked higher, while only the incorrect answer receives the correct standalone decision.

We call these *same-answer pairs*: both descriptions are judged yes or both are judged no. Let $S = \mathbf{1}\{\mathbf{1}[m_t > 0] = \mathbf{1}[m_f > 0]\}$ indicate such a pair. If the two decisions differ, the correct and incorrect descriptions contribute the same result to C and B, so $G = 0$. If they agree, exactly one of the two statement decisions is correct and $B = 1/2$. Averaging over pairs,

$$\mathbb{E}[G] = \Pr(S)\left(\mathbb{E}[C \mid S] - \frac{1}{2}\right). \tag{3}$$

A negative mean gap would require same-answer ranking below one half.

Decision thresholds. Standalone accuracy compares each score with zero. For a general threshold τ, let

$$B_\tau = \frac{1}{2}\left(\mathbf{1}[m_t > \tau] + \mathbf{1}[m_f \le \tau]\right), \tag{4}$$

so that $B = B_0$, averaged over pairs, B_τ is the balanced accuracy of the yes/no decisions at τ. For a correctly ranked pair ($m_t > m_f$), both decisions are correct exactly when $m_f \le \tau < m_t$, that is, when τ lies between the pair's two scores. Let τ^* maximize the mean of B_τ over the evaluated pairs. Let AUROC be the probability that a correct description scores higher than an incorrect one drawn independently, with ties counted as one half. We estimate it from all pairs of correct and incorrect descriptions, including the two of the same item. The gap then splits exactly into three parts, which sum as seen below:

$$C - B_0 = \underbrace{(B_{\tau^*} - B_0)}_{\text{threshold gain}} + \underbrace{(\mathrm{AUROC} - B_{\tau^*})}_{\text{unpaired contrast}} + \underbrace{(C - \mathrm{AUROC})}_{\text{pairing contrast}}. \tag{5}$$

The threshold gain is the most one global threshold can add to these scores. The unpaired contrast compares ranking against incorrect descriptions in general with the best balanced accuracy. It depends only on the two score distributions and can take either sign. In the equal-variance normal model of signal detection theory (Green & Swets, 1966), with means d' standard deviations apart, the two accuracies are $\Phi(d'/\sqrt{2})$ and $\Phi(d'/2)$, where Φ is the standard normal distribution function, so the contrast is positive and at most 8.3 points, reached at $d' = 1.67$. Note that we use this model only as an illustration. The pairing contrast compares ranking within an item with ranking across items. It is zero when the two scores of a pair are unrelated, apart from ties, which C counts as misses and AUROC as halves. In practice we fit $\hat{\tau}$ on a calibration split and apply it to a separate test split.

ARO scoring. The threshold analysis uses both ARO subsets. Each description s is placed in the prompt `{s}? Answer exactly yes or no.` The yes and no groups admit only single-tokens, and each image is cropped to the region given in the ARO annotations. ARO items are built from Visual Genome images, and several items can share one image. Items from the same image stay in the same split, a quarter of the images form the calibration split, and $\hat{\tau}$ maximizes standalone accuracy on the calibration pairs. Intervals resample test images with all their items and treat $\hat{\tau}$ as fixed. Greedy decoding with this prompt gives a valid yes or no on 99.7 to 100% of scored statements and agrees with the sign of the score on 98.2 to 99.9%, so standalone accuracy is within 1.4 points of the accuracy of the model's own generated answers (Appendix I). ARO attribution and relation contain 28,748 and 23,937 items.

Six-subset sign check. To test whether a positive gap is specific to this ARO scoring, we also score every description of ARO attribution, ARO relation, SugarCrepe replace-rel, swap-att, and swap-obj, and the two captions of each Winoground image, under the prompt `Look at the image and judge this statement: "{s}" Answer exactly yes or no.`, using the images as shipped. SugarCrepe subsets contain 1,406, 666, and 246 items. Winoground contains 400 examples. Confidence intervals resample source images with all their items for ARO, whole items for SugarCrepe, and whole two-image examples for Winoground. Paired randomization tests use the same units and are adjusted across the 30 total comparisons (Appendix A). Appendix A also compares two other wordings. The two scorings differ in prompt, token groups, and image crops, so their standalone accuracies are not pooled (Appendix D).

4 Behavioural Results

On ARO scoring, the difference is positive in all ten model–subset combinations. Figure 1B shows the six-subset sign check: every 95% confidence interval is above zero. Among the 24 combinations without LLaVA the point gaps range from $+5.4$ to $+23.4$ percentage points. Equation (3) allows for a negative gap whenever same-answer pairs rank below one half. We can see that, when the two standalone answers agree, the two margins usually keep the correct order. Same-answer prevalence and conditional ranking both vary, so we find that the size of $C - B$ is not itself a measure of ranking quality. Appendix G lists ARO items with their margins.

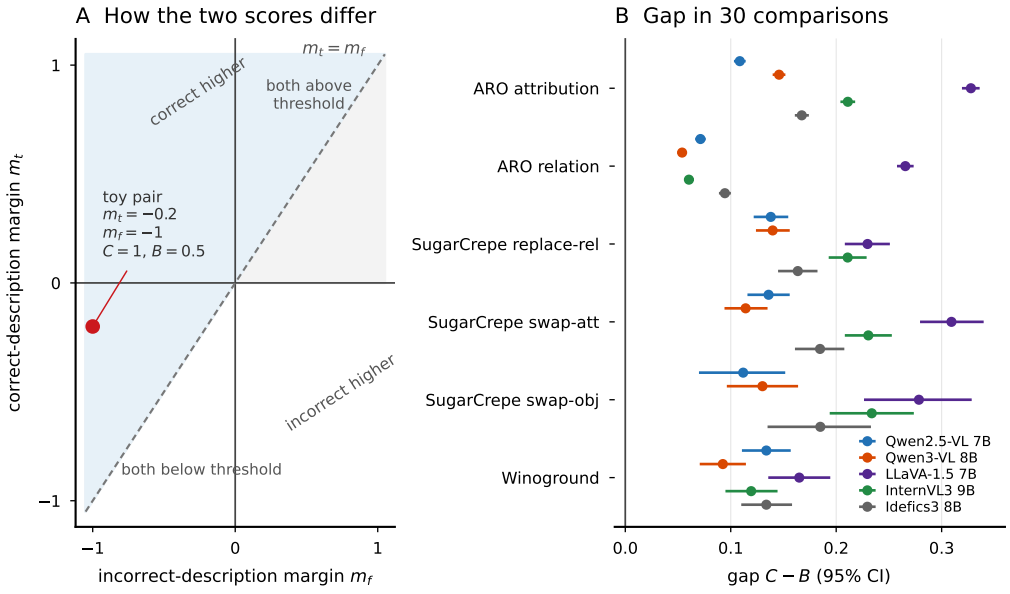


Figure 1: Paired ranking versus standalone decisions. Panel A plots the correct- and incorrect-description margins; the diagonal separates correct from incorrect rankings, while the horizontal and vertical axes determine the two standalone answers. The toy pair $m_t = -0.2$, $m_f = -1$ is ranked correctly although only one standalone answer is correct. Panel B shows $C - B$ for 30 model–benchmark combinations.

LLaVA is a yes-bias case, and does not follow the same pattern as the other models. It answers yes to 89 to 99.8% of descriptions, the bias that POPE and HallusionBench report for earlier models (Li et al., 2023; Guan et al., 2024), and 80.37 to 99.65% of its pairs receive the same answer. Its ranking within those pairs ranges from 68.33% to 82.89%, close to Qwen2.5-VL on ARO attribution (82.15%), while same answers occur on only 33.77% of Qwen2.5-VL pairs. The other four models answer yes for 45 to 80% of descriptions and split same-answer pairs between both yes and both no (Appendix E). On ARO relation their gaps range from 5.4 to 9.4 points. On the ARO scoring LLaVA's gaps are 32.6 and 19.9 points at zero threshold and 18.3 and 9.0 after the held-out threshold.

Standalone accuracy is also more sensitive to the prompt. We scored every description under two further wordings of the yes/no instruction. Across the three wordings, B changes by a median of 2.4 points within a combination and by up to 13.9, while C changes by a median of 1.2 and by at most 5.0. B changes more than C in 23 of the 30 combinations (Appendix A).

We also scored each pair in one prompt that states that exactly one of two statements is true and asks for the letter A or B, once with the correct description as A and once as B. The score is the log probability of the letter A minus that of ''. Averaged over the two orders, A/B accuracy exceeds standalone accuracy in all 30 combinations, by 1.8 to 20.9 points (Table 1). For the four models other than LLaVA it lies within 6.6 points of paired ranking and is closer to paired ranking than to standalone accuracy in 22 of their 24 combinations. The exceptions are Qwen3-VL and Idefics3 on ARO relation. LLaVA favours A on 73 to 87% of prompts in four subsets, like the selection bias that Zheng et al. (2024) report for language models, and trails paired ranking by 7.4 to 21.2 points. The statement that one description is true makes this a different task from judging one statement, and it is stronger than benchmarks built on preferred pairings justify (Section 7). Within that limit, the discrimination measured by paired ranking is available to a restricted A/B readout, and multiple-choice scores of these models mostly track paired ranking.

5 Decision Thresholds

Paired ranking does not depend on where the decision threshold lies while standalone accuracy does. Figure 2 applies the threshold $\hat{\tau}$ to the ARO test pairs (Appendix F). Using $\hat{\tau}$ in place of zero

Table 1: A/B readout with both descriptions in one prompt, averaged over the six subsets of the sign check. The last column gives the range over subsets of the share of prompts scored as A. Per-combination values are in Appendix C.

Model	standalone B	A/B accuracy	paired C	share scored A
Qwen2.5-VL 7B	.789	.902	.905	.376–.496
Qwen3-VL 8B	.797	.905	.910	.373–.496
InternVL3 9B	.726	.869	.904	.495–.598
Idefics3 8B	.703	.840	.858	.396–.517
LLaVA-1.5 7B	.530	.659	.793	.447–.872

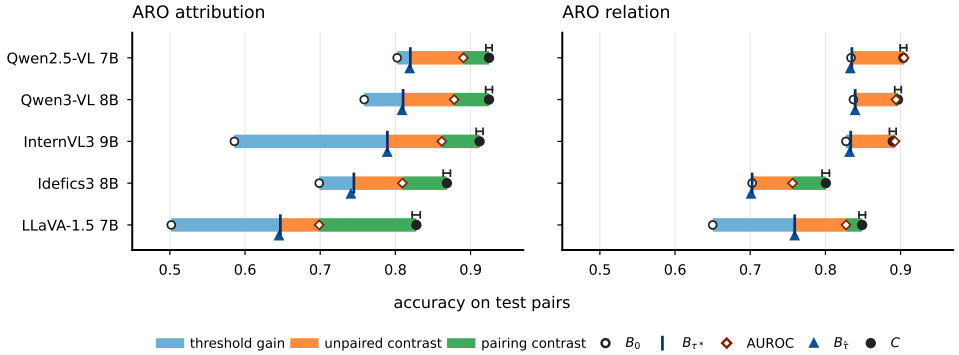


Figure 2: Accounting for the gap on the ARO test pairs (Eq. 5). Each row runs from standalone accuracy at zero (B_0) to paired ranking (C); the vertical bar marks the best threshold (B_{τ^*}) and the diamond marks AUROC. The whisker above each row is the 95% interval of the pairing contrast; triangles mark standalone accuracy at the held-out threshold $\hat{\tau}$.

removes the largest share of the gap for InternVL3 and LLaVA on ARO attribution, 62% and 44%, and between -1% and 8% on ARO relation for every model except LLaVA (55%). After the held-out threshold the gap is 5.7 to 18.3 points, and every 95% interval is above zero. The fitted threshold is close to the best one as its standalone accuracy is within 0.4 points of B_{τ^*} , and τ^* still leaves 5.6 to 18.1 points.

Of what the best threshold leaves, the unpaired contrast is 5.2 to 7.2 points in every combination, below the 8.3-point maximum that the simple normal model of signal detection theory allows for choosing versus judging. It is 3.4 to 5.9 points on ARO attribution for the four models other than LLaVA and 12.9 for LLaVA. On ARO relation it is 4.4 points for Idefics3 and 2.1 for LLaVA, and for Qwen2.5-VL, Qwen3-VL, and InternVL3 it is not distinguishable from zero. After the held-out threshold, 20.8 to 62.0% of correctly ranked test pairs still have one wrong standalone decision. Each gap therefore has different sources. For example, most of InternVL3's attribution gap is recoverable by a threshold, whereas Qwen2.5-VL's relation gap is almost entirely unpaired contrast.

The held-out threshold also changes how models compare (Table 2). At zero, the criteria reverse InternVL3 and Idefics3 on attribution and Idefics3 and LLaVA on relation; with $\hat{\tau}$, both pairs are ordered as by C. Qwen2.5-VL and Qwen3-VL on relation go in the opposite direction as C favours Qwen2.5-VL by 0.7 points and $B_{\hat{\tau}}$ favours Qwen3-VL by 0.6. A threshold fixed on held-out images therefore removes the reversals that come from where each model places its decision boundary, but not necessarily all disagreement between the criteria.

Table 2: Model pairs that the criteria order differently on the ARO test pairs, differences in C, in B_0, and in $B_{\hat{\tau}}$. Every other pair is ordered the same way by all three.

Subset	models	ΔC	ΔB_0	$\Delta B_{\hat{\tau}}$
ARO attr.	InternVL3 9B − Idefics3 8B	+.043 [.037,.050]	-.113 [-.119,-.107]	+.048 [.042,.054]
ARO rel.	Qwen2.5-VL 7B − Qwen3-VL 8B	+.007 [.002,.012]	-.003 [-.007,.001]	-.006 [-.010,-.003]
ARO rel.	Idefics3 8B − LLaVA-1.5 7B	-.048 [-.055,-.041]	+.052 [.047,.058]	-.058 [-.064,-.052]

6 VISUAL CONTROLS

On 506 matched Visual Spatial Reasoning items per model, replacing the original image with uniform grey lowers paired ranking by 20.0 to 43.9 points. Grey replacement changes many visual conditions at once, so the drop is a sensitivity check rather than an additive visual share. Per-model values are in Appendix J. Pairing contrast on ARO still differs across models on the same items, so that term is not a property of the descriptions alone.

A separate set of 37 items holds the two question wordings fixed and defines the direction-score difference $d = m(x, q_{\text{forward}}) - m(x, q_{\text{reverse}})$. After the image is mirrored, the expected reversal is $d_{\text{original}} > 0$ together with $d_{\text{mirror}} < 0$. That reversal occurs in 34 of 37 items for Qwen2.5-VL, Qwen3-VL, and InternVL3.

7 DISCUSSION

Paired ranking is the right score when a system chooses between matched candidates. Standalone accuracy is the right score when a model must judge one claim on its own. Neither must replace the other. Keeping the paired scores supports both analyses and keeps a ranking from being read as a reliable standalone judgment.

A yes/no evaluation should try to state which question it answers. Paired ranking asks whether the model can choose between two matched candidates. Standalone accuracy at the model's own decision rule asks whether it would accept or reject one claim as used. Standalone accuracy at a held-out threshold asks what a rule fixed in advance can provide, the test-optimal threshold bounds what any single threshold could do, while AUROC measures discrimination without pairing. Reporting these, with the decision rule and the tie convention, shows how much of a gap a threshold removes and how much depends on pairing.

Standalone accuracy here is agreement with benchmark labels, and those labels encode preferences. Winoground pairs captions with the images annotators preferred, and its annotators judged at least one caption acceptable for both images (Thrush et al., 2022). In ARO relation, 4.9% of items use a symmetric relation such as *with* or *next to*, so the incorrect description is also true, a hard positive in the sense of Kamath et al. (2024). On these items the models answer yes to both descriptions at $\hat{\tau}$ for 90 to 98% of test pairs, against 13 to 24% on the rest. Excluding them changes the relation gaps by at most 1.0 point (Appendix H).

The main missing control is scoring each description without the image. Without it we describe the pairing contrast and do not assign it a cause. The threshold analysis covers ARO only, with thresholds fitted on the same benchmark and one set of yes/no tokens (Appendix K).

8 CONCLUSION

Paired ranking and standalone accuracy can disagree about the same model, even when both are derived from the same scores. On ARO, fitting the yes/no cutoff on held-out images fixes the model reversals but not necessarily the disagreement, and how much of this difference it recovers depends on the model. Benchmarks that report paired ranking should also report standalone accuracy and the decision threshold supporting it.

AI-use statement. AI assistants were used to refine and reword the text, and to check for clarity and grammar. Further, generative AI was used to help code the underlying experiments detailed in this paper.

REFERENCES

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

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

Tsan Tsai Chan, Varsha Suresh, Anisha Saha, Michael Hahn, and Vera Demberg. System-mediated attention imbalances make vision-language models say yes. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, pp. 31096–31116, 2023.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. The hard positive truth about vision-language compositionality. In *European Conference on Computer Vision (ECCV)*, pp. 37–54, 2024.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

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 *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, 2024b.

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, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. Beyond probabilities: Unveiling the misalignment in evaluating large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pp. 109–131, 2024.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *International Conference on Learning Representations (ICLR)*, 2025.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

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 (CVPR)*, 2022.

Vishaal Udandarao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good CREPE needs more than just sugar: Investigating biases in compositional vision-language benchmarks. *arXiv preprint arXiv:2506.08227*, 2025.

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 *International Conference on Learning Representations (ICLR)*, 2023.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Benchmark Protocol

In the six-subset sign check each benchmark description s is placed in the prompt `Look at the image and judge this statement: "{s}" Answer exactly yes or no.` The yes and no groups contain the tokenizer-derived first-token IDs for the selected surface forms of those answers. The analyzed quantity is their restricted log-margin, not a generated answer sequence. ARO attribution and relation contain 28,748 and 23,937 items, respectively; SugarCrepe replace-rel, swap-att, and swap-obj contain 1,406, 666, and 246 items; and the Winoground rows contain 800 image rows (400 examples). The five models are Qwen2.5-VL 7B, Qwen3-VL 8B, LLaVA-1.5 7B, InternVL3 9B, and Idefics3 8B. Appendix D describes the ARO threshold protocol.

Confidence intervals use 10,000 bootstrap draws that resample complete evaluation units: one source image with all its items for ARO, one item for SugarCrepe, or one full Winoground example containing both images and all four descriptions. ARO item i is item i of the ARO annotation files, which record its image. For Qwen2.5-VL, the six-subset and ARO-protocol margin differences of the same item correlate at 0.80 (relation) and 0.87 (attribution), and below 0.45 when shifted by one item. SugarCrepe exports carry no image identifier, so its intervals resample items. To test whether a

mean gap is positive, paired randomization tests reverse the signs of unit-level differences under a sign-symmetry null. SugarCrepe tests flip item differences and use the exact distribution available for their equal-magnitude differences. ARO tests flip the sums over each source image and Winoground tests flip whole examples, each with 10,000 random sign assignments and a plus-one correction so that the estimated probability is never zero; no assignment reached the observed gap in any ARO or Winoground combination. Holm's step-down procedure adjusts the 30 resulting probabilities for multiple comparisons. The separate four-description Winoground analysis is excluded from that family because its paired-ranking score is mathematically guaranteed to be at least its standalone score.

Table 3: All 30 model–subset comparisons. B is standalone accuracy (the mean correctness of the two separate yes/no decisions), C is paired ranking, and the interval is for $C - B$; it resamples source images for ARO, items for SugarCrepe, and whole examples for Winoground, and Holm p values use the same units. The Winoground comparison ranks the two captions of each image. ARO attr./rel. abbreviate its attribution/relation subsets; SC abbreviates SugarCrepe.

Model	subset	n	B	C	$C - B$	95% interval	Holm p
Qwen2.5-VL 7B	ARO attr.	28748	.816	.925	+.109	[.104,.113]	.0015
Qwen3-VL 8B	ARO attr.	28748	.784	.930	+.146	[.141,.151]	.0015
InternVL3 9B	ARO attr.	28748	.708	.919	+.211	[.205,.217]	.0015
Idefics3 8B	ARO attr.	28748	.732	.899	+.167	[.162,.173]	.0015
LLaVA-1.5 7B	ARO attr.	28748	.501	.828	+.328	[.321,.335]	.0015
Qwen2.5-VL 7B	ARO rel.	23937	.791	.863	+.071	[.068,.075]	.0015
Qwen3-VL 8B	ARO rel.	23937	.798	.852	+.054	[.051,.057]	.0015
InternVL3 9B	ARO rel.	23937	.795	.856	+.060	[.057,.064]	.0015
Idefics3 8B	ARO rel.	23937	.695	.790	+.094	[.090,.099]	.0015
LLaVA-1.5 7B	ARO rel.	23937	.538	.803	+.266	[.259,.272]	.0015
Qwen2.5-VL 7B	SC replace-rel	1406	.782	.920	+.138	[.123,.153]	$< 10^{-4}$
Qwen3-VL 8B	SC replace-rel	1406	.789	.929	+.140	[.125,.155]	$< 10^{-4}$
InternVL3 9B	SC replace-rel	1406	.712	.923	+.211	[.194,.228]	$< 10^{-4}$
Idefics3 8B	SC replace-rel	1406	.712	.876	+.164	[.146,.181]	$< 10^{-4}$
LLaVA-1.5 7B	SC replace-rel	1406	.596	.826	+.230	[.209,.250]	$< 10^{-4}$
Qwen2.5-VL 7B	SC swap-att	666	.833	.968	+.136	[.117,.155]	$< 10^{-4}$
Qwen3-VL 8B	SC swap-att	666	.845	.959	+.114	[.095,.134]	$< 10^{-4}$
InternVL3 9B	SC swap-att	666	.729	.959	+.230	[.209,.252]	$< 10^{-4}$
Idefics3 8B	SC swap-att	666	.748	.932	+.185	[.162,.206]	$< 10^{-4}$
LLaVA-1.5 7B	SC swap-att	666	.511	.820	+.309	[.281,.339]	$< 10^{-4}$
Qwen2.5-VL 7B	SC swap-obj	246	.762	.874	+.112	[.071,.150]	$< 10^{-4}$
Qwen3-VL 8B	SC swap-obj	246	.789	.919	+.130	[.098,.163]	$< 10^{-4}$
InternVL3 9B	SC swap-obj	246	.689	.923	+.234	[.195,.272]	$< 10^{-4}$
Idefics3 8B	SC swap-obj	246	.636	.821	+.185	[.136,.232]	$< 10^{-4}$
LLaVA-1.5 7B	SC swap-obj	246	.506	.785	+.278	[.228,.327]	$< 10^{-4}$
Qwen2.5-VL 7B	Winoground (per image)	800	.749	.882	+.134	[.112,.156]	.0015
Qwen3-VL 8B	Winoground (per image)	800	.779	.871	+.092	[.072,.113]	.0015
InternVL3 9B	Winoground (per image)	800	.723	.843	+.119	[.096,.143]	.0015
Idefics3 8B	Winoground (per image)	800	.695	.829	+.134	[.111,.157]	.0015
LLaVA-1.5 7B	Winoground (per image)	800	.529	.694	+.165	[.137,.193]	.0015

Departures from the original protocols. We score every item of each subset without filtering and report per-item means rather than the benchmarks' aggregate scores. For Winoground we compare the two captions of each image, the direction of its text score, instead of computing its text, image, and group scores. The six-subset scoring uses the images as distributed in the Hugging Face copies; the ARO threshold protocol crops images to the annotated region.

Prompt wording. Every description was also scored under two other prompts, `Based on the image, is the following true? "{s}" Answer yes or no.` and `Statement: "{s}"` followed on a new line by `Is this an accurate description`

of the image? Answer yes or no. Table 4 gives C and B under all three; the main text uses the first. Across the 30 combinations, the range of C over the three prompts has median 1.2 points and maximum 5.0, and the range of B has median 2.4 and maximum 13.9. The gap $C - B$ is between 5.0 and 32.8 points in all 90 prompt–combination cells.

Table 4: Paired ranking C and standalone accuracy B under the three prompts (p1 is the prompt used in the main text), and the range of each over the three.

		C			B			range	
Model	subset	p1	p2	p3	p1	p2	p3	C	B
Qwen2.5-VL 7B	ARO attr.	.925	.925	.922	.816	.821	.817	.003	.005
Qwen3-VL 8B	ARO attr.	.930	.930	.935	.784	.786	.831	.005	.046
InternVL3 9B	ARO attr.	.919	.919	.925	.708	.732	.791	.006	.083
Idefics3 8B	ARO attr.	.899	.908	.889	.732	.755	.736	.019	.023
LLaVA-1.5 7B	ARO attr.	.828	.847	.838	.501	.567	.632	.018	.132
Qwen2.5-VL 7B	ARO rel.	.863	.858	.863	.791	.778	.789	.005	.014
Qwen3-VL 8B	ARO rel.	.852	.852	.857	.798	.794	.807	.005	.014
InternVL3 9B	ARO rel.	.856	.850	.857	.795	.794	.799	.006	.005
Idefics3 8B	ARO rel.	.790	.795	.770	.695	.699	.674	.024	.025
LLaVA-1.5 7B	ARO rel.	.803	.791	.771	.538	.677	.660	.032	.139
Qwen2.5-VL 7B	SC replace-rel	.920	.922	.923	.782	.777	.782	.004	.005
Qwen3-VL 8B	SC replace-rel	.929	.917	.927	.789	.781	.792	.011	.011
InternVL3 9B	SC replace-rel	.923	.915	.925	.712	.718	.738	.011	.026
Idefics3 8B	SC replace-rel	.876	.895	.880	.712	.721	.693	.019	.028
LLaVA-1.5 7B	SC replace-rel	.826	.844	.794	.596	.687	.689	.050	.093
Qwen2.5-VL 7B	SC swap-att	.968	.965	.977	.833	.834	.835	.012	.002
Qwen3-VL 8B	SC swap-att	.959	.956	.964	.845	.842	.845	.008	.004
InternVL3 9B	SC swap-att	.959	.956	.959	.729	.731	.784	.003	.055
Idefics3 8B	SC swap-att	.932	.946	.938	.748	.738	.716	.014	.032
LLaVA-1.5 7B	SC swap-att	.820	.826	.823	.511	.598	.601	.006	.091
Qwen2.5-VL 7B	SC swap-obj	.874	.878	.882	.762	.768	.770	.008	.008
Qwen3-VL 8B	SC swap-obj	.919	.927	.907	.789	.780	.805	.020	.024
InternVL3 9B	SC swap-obj	.923	.907	.927	.689	.681	.738	.020	.057
Idefics3 8B	SC swap-obj	.821	.841	.837	.636	.640	.616	.020	.024
LLaVA-1.5 7B	SC swap-obj	.785	.760	.809	.506	.579	.608	.049	.102
Qwen2.5-VL 7B	Winoground (per image)	.882	.873	.870	.749	.743	.738	.012	.011
Qwen3-VL 8B	Winoground (per image)	.871	.876	.854	.779	.779	.764	.022	.014
InternVL3 9B	Winoground (per image)	.843	.845	.845	.723	.728	.731	.002	.008
Idefics3 8B	Winoground (per image)	.829	.809	.802	.695	.691	.674	.026	.021
LLaVA-1.5 7B	Winoground (per image)	.694	.694	.691	.529	.579	.576	.002	.051

B Model Rankings

Table 5 gives, for each subset, the model ranked first by paired ranking and by standalone accuracy, and Kendall's τ_b between the two orderings of the five models. Table 6 lists the 13 model pairs within a subset that C and B do not order the same way, one of them tied on C. Differences are paired over the items both models scored. Intervals use 10,000 bootstrap draws over source images for ARO, over items for SugarCrepe, and over whole examples for Winoground. They are nominal: not adjusted for multiple comparisons, and the pairs shown are those the point estimates order differently. With item-level intervals on ARO, six pairs would have both intervals excluding zero; resampling images removes ARO relation Qwen2.5-VL versus InternVL3. The top model changes in five subsets by point estimate, and both intervals exclude zero for the pair concerned only on ARO attribution and ARO relation.

C Choice Between Both Descriptions

Each pair is shown twice in the prompt Look at the image. Exactly one of these two statements is true. followed by the lines A. {a}, B. {b}, and Answer with a single letter, A or B., once with the correct description as A and once as B. The score is the log probability of the A tokens minus that of the B tokens at the answer position, and the model picks A when it is positive. This is a restricted preference between the letter tokens, not a generated answer, and the yes/no generation check of Appendix I does not cover it. Accuracy averages the two

Table 5: Top model under each criterion and Kendall's τ_b between the two orderings of the five models, six-subset scoring.

Subset	top by C	top by B	τ_b
ARO attr.	Qwen3-VL 8B	Qwen2.5-VL 7B	0.60
ARO rel.	Qwen2.5-VL 7B	Qwen3-VL 8B	0.20
SC replace-rel	Qwen3-VL 8B	Qwen3-VL 8B	0.80
SC swap-att	Qwen2.5-VL 7B	Qwen3-VL 8B	0.53
SC swap-obj	InternVL3 9B	Qwen3-VL 8B	0.60
Winoground (per image)	Qwen2.5-VL 7B	Qwen3-VL 8B	0.80

Table 6: Model pairs that C and B do not order the same way. ΔC and ΔB are the first model minus the second on the same items, with 95% intervals (over source images for ARO). $*$ marks pairs where both intervals exclude zero.

Subset	first model	second model	ΔC [95% CI]	ΔB [95% CI]	
ARO attr.	Qwen2.5-VL 7B	Qwen3-VL 8B	-.005 [-.010,-.001]	+.032 [.027,.037]	$*$
ARO attr.	InternVL3 9B	Idefics3 8B	+.020 [.014,.026]	-.024 [-.030,-.019]	$*$
ARO rel.	Qwen2.5-VL 7B	Qwen3-VL 8B	+.011 [.006,.015]	-.007 [-.010,-.003]	$*$
ARO rel.	Qwen2.5-VL 7B	InternVL3 9B	+.007 [.002,.012]	-.004 [-.008,.000]	
ARO rel.	Qwen3-VL 8B	InternVL3 9B	-.004 [-.008,.001]	+.003 [-.001,.007]	
ARO rel.	Idefics3 8B	LLaVA-1.5 7B	-.013 [-.020,-.007]	+.158 [.152,.163]	$*$
SC replace-rel	Qwen2.5-VL 7B	InternVL3 9B	-.004 [-.017,.010]	+.069 [.057,.082]	
SC swap-att	Qwen2.5-VL 7B	Qwen3-VL 8B	+.009 [-.008,.026]	-.013 [-.030,.005]	
SC swap-att	Qwen3-VL 8B	InternVL3 9B	.000 [-.017,.017]	+.116 [.095,.137]	
SC swap-att	InternVL3 9B	Idefics3 8B	+.027 [.006,.048]	-.019 [-.040,.002]	
SC swap-obj	Qwen2.5-VL 7B	InternVL3 9B	-.049 [-.090,-.008]	+.073 [.041,.106]	$*$
SC swap-obj	Qwen3-VL 8B	InternVL3 9B	-.004 [-.041,.037]	+.100 [.065,.134]	
Winoground (per image)	Qwen2.5-VL 7B	Qwen3-VL 8B	+.011 [-.013,.034]	-.030 [-.049,-.009]	

orders. Table 7 also gives the share of items answered correctly in both orders, the share with both standalone answers correct, and the share of prompts answered A. Both-orders accuracy exceeds both-standalone accuracy in every combination, by 3.2 to 44.2 points.

D ARO Scoring and the Six-Subset Check

The threshold protocol uses the prompt `{s}? Answer exactly yes or no.`, crops each image to the region given in the ARO annotation files, and keeps only yes and no surface forms that are single tokens (Table 9). The six-subset sign check uses a longer prompt and the images as shipped. In that check, the two LLaVA-1.5 groups share token 29871, a leading space. Adding the same probability to both groups moves a margin toward zero without changing its sign, so standalone decisions at zero are unaffected, but the order of two margins can change. The ARO threshold protocol has no shared tokens.

The six-subset groups take the first token of `yes`, `Yes`, `YES`, and each of these with a leading space, and likewise for no. The ARO threshold protocol keeps only those of `yes`, `Yes`, `yes`, `Yes`, and `YES` (and the matching no forms) that are single tokens. Table 8 summarizes both protocols. Table 10 compares them on the same ARO items. Paired ranking changes by at most 4.6 points; standalone accuracy moves more. SugarCrepe and Winoground were scored only under the six-subset protocol. Qwen3-VL 32B was scored on a fixed random subset of images containing 5,007 attribution and 5,012 relation items.

E Same-answer Decomposition

Let S indicate that both descriptions receive the same standalone answer (both yes or both no). Under the strict ranking rule and the asymmetric zero conventions,

$$\mathbb{E}[G] = \Pr(S)\big(\mathbb{E}[C \mid S] - \tfrac{1}{2}\big). \tag{6}$$

If the decisions differ, either both statement decisions and the ranking are correct ($B = C = 1$), or all three are incorrect ($B = C = 0$); hence $G = 0$. If decisions agree, exactly one of the two statement-level decisions is correct, so $B = 1/2$ and the identity follows. Thus a positive mean gap

Table 7: Choice between both descriptions in one prompt, six-subset items and models. Accuracy averages the two orders; "both orders" requires a correct answer in each; "both standalone" requires both separate yes/no answers to be correct; "picks A" is the share of prompts answered A.

Model	subset	accuracy	both orders	both standalone	picks A	B	C
Qwen2.5-VL 7B	ARO attr.	.929	.880	.647	.476	.816	.925
Qwen3-VL 8B	ARO attr.	.949	.913	.581	.496	.784	.930
InternVL3 9B	ARO attr.	.893	.818	.430	.520	.708	.919
Idefics3 8B	ARO attr.	.897	.817	.482	.517	.732	.899
LLaVA-1.5 7B	ARO attr.	.698	.422	.002	.734	.501	.828
Qwen2.5-VL 7B	ARO rel.	.827	.702	.627	.376	.791	.863
Qwen3-VL 8B	ARO rel.	.816	.686	.654	.373	.798	.852
InternVL3 9B	ARO rel.	.835	.755	.649	.495	.795	.856
Idefics3 8B	ARO rel.	.737	.602	.442	.457	.695	.790
LLaVA-1.5 7B	ARO rel.	.695	.525	.083	.447	.538	.803
Qwen2.5-VL 7B	SC replace-rel	.926	.886	.574	.483	.782	.920
Qwen3-VL 8B	SC replace-rel	.934	.902	.587	.492	.789	.929
InternVL3 9B	SC replace-rel	.857	.745	.426	.598	.712	.923
Idefics3 8B	SC replace-rel	.896	.831	.432	.479	.712	.876
LLaVA-1.5 7B	SC replace-rel	.752	.519	.194	.727	.596	.826
Qwen2.5-VL 7B	SC swap-att	.984	.970	.668	.496	.833	.968
Qwen3-VL 8B	SC swap-att	.975	.952	.695	.492	.845	.959
InternVL3 9B	SC swap-att	.938	.880	.464	.547	.729	.959
Idefics3 8B	SC swap-att	.926	.863	.503	.480	.748	.932
LLaVA-1.5 7B	SC swap-att	.607	.230	.023	.870	.511	.820
Qwen2.5-VL 7B	SC swap-obj	.900	.837	.537	.478	.762	.874
Qwen3-VL 8B	SC swap-obj	.911	.866	.614	.472	.789	.919
InternVL3 9B	SC swap-obj	.868	.780	.382	.559	.689	.923
Idefics3 8B	SC swap-obj	.803	.654	.285	.396	.636	.821
LLaVA-1.5 7B	SC swap-obj	.596	.220	.020	.872	.506	.785
Qwen2.5-VL 7B	Winoground (per image)	.843	.735	.525	.417	.749	.882
Qwen3-VL 8B	Winoground (per image)	.844	.745	.588	.425	.779	.871
InternVL3 9B	Winoground (per image)	.822	.719	.482	.532	.723	.843
Idefics3 8B	Winoground (per image)	.781	.635	.434	.427	.695	.829
LLaVA-1.5 7B	Winoground (per image)	.606	.336	.079	.652	.529	.694

Table 8: The two scorings.

	six-subset	ARO protocol
Prompt	`Look at the image and judge this statement: "{s}" Answer exactly yes or no.`	`{s}? Answer exactly yes or no.`
Yes/no tokens	first token of each surface form; the LLaVA groups share token 29871	single-token surface forms only; no shared tokens
Images	as shipped in the Hugging Face copies	ARO images cropped to the annotated region
Subsets	ARO, SugarCrepe (3), Winoground	ARO
Models	five	five and Qwen3-VL 32B
Interval unit	source image (ARO), item (SugarCrepe), example (Winoground)	source image
Thresholds	zero	zero, held-out $\hat{\tau}$, test-optimal τ^*
Generated answers	not checked	checked (Appendix I)

is empirical, and equals same-answer prevalence times the excess of conditional ranking over one half. This decomposition is descriptive: it does not provide a new confidence interval or identify a calibration mechanism.

The following counts and point estimates were recomputed from the existing saved summary: gains are $m_t > m_f$ among same-answer pairs, losses are $m_t \leq m_f$ among those pairs, prevalence is $(\text{gains} + \text{losses})/n$, and conditional ranking is $\text{gains}/(\text{gains} + \text{losses})$. No conditional confidence intervals are asserted. The different prevalence and conditional-ranking combinations explain why the size of $C - B$ is not itself a measure of relational representation.

Table 9: Yes and no token IDs in the two scorings. Surface forms are listed in the text.

Model	scoring	yes	no
Qwen2.5-VL 7B	six-subset	7414 9454 9693 9834 14004 14080	902 2152 2308 2753 5664 8996
	ARO protocol	7414 9454 9693 9834 14080	902 2152 2308 2753 5664
Qwen3-VL 8B	six-subset	7414 9454 9693 9834 14004 14080	902 2152 2308 2753 5664 8996
	ARO protocol	7414 9454 9693 9834 14080	902 2152 2308 2753 5664
InternVL3 9B	six-subset	1801 4910 5712 6072 6466 10933	1042 1812 2330 3496 3958 7921
	ARO protocol	4910 6072 6466 10933	1042 1812 2330 3958 7921
Idefics3 8B	six-subset	7566 9642 9891 10035 14331 14410	912 2201 2360 2822 5782 9173
	ARO protocol	7566 9642 9891 10035 14410	912 2201 2360 2822 5782
LLaVA-1.5 7B	six-subset	3869 4874 22483 29871	694 1939 11698 29871
	ARO protocol	3582 3869 4874 8241 22483	694 1217 1939 3782 11698

Table 10: Standalone accuracy B and paired ranking C on the same ARO items under the two scorings.

		B			C		
Model	subset	six-subset	ARO prot.	change	six-subset	ARO prot.	change
Qwen2.5-VL 7B	ARO attr.	.816	.802	-.015	.925	.922	-.003
Qwen3-VL 8B	ARO attr.	.784	.757	-.028	.930	.923	-.007
InternVL3 9B	ARO attr.	.708	.585	-.123	.919	.910	-.009
Idefics3 8B	ARO attr.	.732	.700	-.033	.899	.868	-.031
LLaVA-1.5 7B	ARO attr.	.501	.502	+.001	.828	.824	-.004
Qwen2.5-VL 7B	ARO rel.	.791	.836	+.044	.863	.905	+.042
Qwen3-VL 8B	ARO rel.	.798	.837	+.039	.852	.897	+.045
InternVL3 9B	ARO rel.	.795	.827	+.032	.856	.890	+.034
Idefics3 8B	ARO rel.	.695	.703	+.008	.790	.801	+.011
LLaVA-1.5 7B	ARO rel.	.538	.651	+.113	.803	.849	+.046

Table 12 gives the share of yes decisions in each combination. LLaVA-1.5 answers yes to 89.4 to 99.8% of descriptions, and nearly all its same-answer pairs are both yes. The other four models answer yes to 45.4 to 80.3%, and their same-answer pairs include both-no pairs in every combination.

F Threshold Analysis

Table 14 gives the thresholds behind Table 13, with Qwen3-VL 32B added. The share of the gap removed is $1 - (C - B_{\hat{\tau}})/(C - B_0)$; it is negative when calibration widens the gap slightly on the test split. Holm-adjusted bootstrap p values for the calibrated gap range from 0.0006 to 0.003. For every correctly ranked test pair, both standalone decisions are correct exactly when $m_f \leq \tau < m_t$, as stated in Section 3; we checked this pair by pair at $\tau = 0$ and $\tau = \hat{\tau}$.

For the scaling comparison, both Qwen3-VL models are scored on the 32B subset, with thresholds fitted on the same calibration images (Table 15); the last column is the paired change in the calibrated gap.

AUROC is computed from the ranks of all correct- and incorrect-description scores on the test split, with ties counted as one half. The interval for the pairing contrast resamples test images with 1,000 draws. For Qwen3-VL 32B on its subset, the threshold gain, unpaired contrast, and pairing contrast are +.032, +.070, and +.041 [.032,.052] on attribution and +.005, +.047, and +.001 [−.007,.009] on relation. Table 13 gives the values behind Figure 2. Paired ranking counts ties as misses and AUROC counts them as halves. Ties occur in 1.9% of LLaVA attribution test pairs and in at most 0.6% elsewhere; counting them as halves in C would raise the pairing contrast by 0.9 points for LLaVA attribution, 0.3 for LLaVA relation, and at most 0.15 in every other cell.

G Examples

Table 16 lists, for Qwen3-VL 8B on the ARO scoring, the first test item by ARO index of each kind. Symmetric-relation items are excluded from every kind except their own. The held-out thresholds are

Table 11: Descriptive same-answer decomposition of all 30 model–subset comparisons. Counts are correctly/incorrectly ranked pairs within the same-answer population; percentages are point estimates from the saved records, without new intervals.

Model	subset	n	correct/wrong	same-answer	cond. ranking
Qwen2.5-VL 7B	ARO attr.	28748	7976/1733	33.77%	82.15%
Qwen2.5-VL 7B	ARO rel.	23937	5651/2238	32.96%	71.63%
Qwen2.5-VL 7B	SC replace-rel	1406	486/98	41.54%	83.22%
Qwen2.5-VL 7B	SC swap-att	666	200/19	32.88%	91.32%
Qwen2.5-VL 7B	SC swap-obj	246	83/28	45.12%	74.77%
Qwen2.5-VL 7B	Winoground (per image)	800	286/72	44.75%	79.89%
Qwen3-VL 8B	ARO attr.	28748	10028/1642	40.59%	85.93%
Qwen3-VL 8B	ARO rel.	23937	4738/2158	28.81%	68.71%
Qwen3-VL 8B	SC replace-rel	1406	480/87	40.33%	84.66%
Qwen3-VL 8B	SC swap-att	666	176/24	30.03%	88.00%
Qwen3-VL 8B	SC swap-obj	246	75/11	34.96%	87.21%
Qwen3-VL 8B	Winoground (per image)	800	227/79	38.25%	74.18%
LLaVA-1.5 7B	ARO attr.	28748	23746/4901	99.65%	82.89%
LLaVA-1.5 7B	ARO rel.	23937	17239/4525	90.92%	79.21%
LLaVA-1.5 7B	SC replace-rel	1406	888/242	80.37%	78.58%
LLaVA-1.5 7B	SC swap-att	666	531/119	97.60%	81.69%
LLaVA-1.5 7B	SC swap-obj	246	188/51	97.15%	78.66%
LLaVA-1.5 7B	Winoground (per image)	800	492/228	90.00%	68.33%
InternVL3 9B	ARO attr.	28748	14071/1938	55.69%	87.89%
InternVL3 9B	ARO rel.	23937	4944/2055	29.24%	70.64%
InternVL3 9B	SC replace-rel	1406	699/106	57.25%	86.83%
InternVL3 9B	SC swap-att	666	330/23	53.00%	93.48%
InternVL3 9B	SC swap-obj	246	133/18	61.38%	88.08%
InternVL3 9B	Winoground (per image)	800	288/97	48.13%	74.81%
Idefics3 8B	ARO attr.	28748	12004/2382	50.04%	83.44%
Idefics3 8B	ARO rel.	23937	8330/3808	50.71%	68.63%
Idefics3 8B	SC replace-rel	1406	623/163	55.90%	79.26%
Idefics3 8B	SC swap-att	666	286/40	48.95%	87.73%
Idefics3 8B	SC swap-obj	246	132/41	70.33%	76.30%
Idefics3 8B	Winoground (per image)	800	316/102	52.25%	75.60%

Table 12: Share of standalone yes decisions (over correct and incorrect descriptions), six-subset scoring.

Model	ARO attr.	ARO rel.	SC replace-rel	SC swap-att	SC swap-obj	Winoground
Qwen2.5-VL 7B	.521	.458	.617	.589	.661	.487
Qwen3-VL 8B	.624	.520	.614	.548	.585	.530
InternVL3 9B	.724	.530	.761	.744	.774	.649
Idefics3 8B	.635	.454	.728	.680	.803	.624
LLaVA-1.5 7B	.998	.952	.894	.979	.986	.941

Table 13: Decision thresholds and the accounting of Eq. (5) on the ARO scoring. $\hat{\tau}$ is fitted on a quarter of the images and applied to the rest; τ^* is the best single threshold for the test pairs. Test pairs number 21,643 (attribution) and 17,982 (relation). Intervals resample test images and treat $\hat{\tau}$ as fixed.

Model	C	B_0	$B_{\hat{\tau}}$	$C - B_{\hat{\tau}}$ [95% CI]	threshold	unpaired	pairing [95% CI]
ARO attribution							
Qwen2.5-VL 7B	.925	.802	.819	+.106 [.101,.111]	+.017	+.071	+.034 [.030,.038]
Qwen3-VL 8B	.925	.759	.809	+.116 [.110,.121]	+.052	+.068	+.046 [.042,.051]
InternVL3 9B	.912	.586	.789	+.123 [.118,.128]	+.204	+.072	+.050 [.046,.055]
Idefics3 8B	.869	.699	.741	+.128 [.122,.133]	+.046	+.065	+.059 [.054,.064]
LLaVA-1.5 7B	.828	.502	.645	+.183 [.176,.189]	+.145	+.052	+.129 [.124,.134]
ARO relation							
Qwen2.5-VL 7B	.904	.834	.833	+.071 [.067,.074]	+.001	+.069	-.001 [-.005,.003]
Qwen3-VL 8B	.897	.837	.840	+.057 [.053,.061]	+.002	+.054	+.003 [-.002,.007]
InternVL3 9B	.890	.827	.832	+.057 [.054,.061]	+.006	+.059	-.003 [-.007,.001]
Idefics3 8B	.801	.703	.701	+.099 [.094,.105]	.000	+.054	+.044 [.039,.049]
LLaVA-1.5 7B	.849	.650	.759	+.090 [.085,.094]	+.109	+.068	+.021 [.017,.026]

$\hat{\tau} = +2.50$ (attribution) and -0.75 (relation); a description is judged yes when its margin exceeds the threshold in use. Images are the Visual Genome images named in the ARO annotations.

H Symmetric Relations in ARO

We mark an ARO-relation item as symmetric when its correct description has the form "the X is R the Y" with R one of *near*, *next to*, *with*, *by*, *beside*, *close to*, *adjacent to*, *touching*, or *alongside*. This marks 1,173 of 23,937 items (4.9%), 882 of them in the test split. For these relations the incorrect description, which swaps the two objects, is also true. Table 17 gives the share of test items on which both descriptions are judged yes at $\hat{\tau}$, and the relation results with and without the symmetric items. Without them, the pairing contrast of Qwen2.5-VL, Qwen3-VL, and InternVL3 is slightly negative (-0.5 to -0.8 points), and the other results change by at most 1.0 point.

Table 14: Threshold details on the ARO scoring. Held-out and test columns count pairs; images counts test images. Removed is the share of the gap removed by $\hat{\tau}$. Same-answer rate and conditional ranking are computed at $\hat{\tau}$.

Model	subset	cal.	test	images	$\hat{\tau}$	τ^*	removed	same ans.	cond. rank
Qwen2.5-VL 7B	ARO attr.	7105	21643	3405	+1.26	+1.07	14%	.337	.813
Qwen3-VL 8B	ARO attr.	7105	21643	3405	+2.50	+3.00	30%	.358	.823
InternVL3 9B	ARO attr.	7105	21643	3405	+1.67	+1.75	62%	.391	.814
Idefics3 8B	ARO attr.	7105	21643	3405	+1.63	+2.00	25%	.477	.768
LLaVA-1.5 7B	ARO attr.	7105	21643	3405	+1.87	+1.75	44%	.661	.777
Qwen3-VL 32B	ARO attr.	1231	3776	562	+2.25	+2.25	21%	.324	.850
Qwen2.5-VL 7B	ARO rel.	5955	17982	3999	-0.50	-0.14	-1%	.266	.765
Qwen3-VL 8B	ARO rel.	5955	17982	3999	-0.75	-0.75	4%	.259	.720
InternVL3 9B	ARO rel.	5955	17982	3999	+0.25	+0.50	8%	.266	.716
Idefics3 8B	ARO rel.	5955	17982	3999	-0.12	+0.00	-1%	.516	.693
LLaVA-1.5 7B	ARO rel.	5955	17982	3999	+0.87	+0.87	55%	.401	.724
Qwen3-VL 32B	ARO rel.	1198	3814	833	-1.12	-2.25	5%	.219	.728

Table 15: Qwen3-VL 8B and 32B on identical ARO items, with thresholds fitted on the same calibration images.

Model	subset	test pairs	C	$B_{\hat{\tau}}$	$C - B_{\hat{\tau}}$ [95% CI]	change, 32B − 8B
Qwen3-VL 8B	ARO attr.	3776	.923	.814	+.110 [.097,.123]	
Qwen3-VL 32B	ARO attr.	3776	.940	.826	+.113 [.102,.125]	+.004 [-.009,.017]
Qwen3-VL 8B	ARO rel.	3814	.904	.847	+.057 [.050,.065]	
Qwen3-VL 32B	ARO rel.	3814	.915	.865	+.050 [.043,.057]	-.007 [-.016,.002]

Table 16: Rule-selected ARO items for Qwen3-VL 8B, with the margins of the correct (m_t) and incorrect (m_f) descriptions.

Subset	kind	correct description	incorrect description	m_t	m_f
attr.	both yes, ranked correctly	the open door and the crouched man	the crouched door and the open man	+9.50	+1.00
attr.	both no, ranked correctly	the empty boat and the orange wall	the orange boat and the empty wall	-0.75	-2.25
attr.	both correct	the gray floor and the long sleeved shirt	the long sleeved floor and the gray shirt	+5.00	-3.25
attr.	ranked incorrectly	the paved street and the white umbrella	the white street and the paved umbrella	-0.25	+1.00
rel.	both yes, ranked correctly	the man is wearing the shirt	the shirt is wearing the man	+17.00	+13.75
rel.	both no, ranked correctly	the truck is in front of the tree	the tree is in front of the truck	-0.25	-6.25
rel.	both correct	the door is to the left of the shirt	the shirt is to the left of the door	+4.75	-4.50
rel.	ranked incorrectly	the plate is of the sandwiches	the sandwiches is of the plate	+7.50	+9.25
rel.	symmetric relation	the plate is with the banana	the banana is with the plate	+12.50	+11.75

I Generated Answers

Greedy decoding uses the ARO protocol prompt of Appendix D with at most eight new tokens, on about 500 test pairs per subset chosen by image. An answer is valid if it begins with yes or no; invalid answers count as incorrect. Generated answers disagree with the sign of the margin only when the margin is within 0.5 of zero. Table 18 compares the accuracy of the generated answers with that of the margin decisions on the same statements.

J Visual Controls

Real–grey comparisons use 506 matched VSR items per model. The mirror test uses a separate direction-score difference. Holding the forward/reverse question strings and indices fixed, let $d = m(x, q_{\text{forward}}) - m(x, q_{\text{reverse}})$. Mirroring changes which statement is true, so the expected sign reversal is $d_{\text{original}} > 0$ and $d_{\text{mirror}} < 0$, not merely $d_{\text{mirror}} \leq 0$. The mirror set has 37 items per model.

On a separate, earlier set of 194 VSR items, comparing paired ranking with standalone decision accuracy gives the following gain/loss/tie counts: LLaVA-1.5 7B, 154/32/8; Qwen2.5-VL 7B, 37/4/153; and Qwen3-VL 8B, 23/4/167. Here gains and losses indicate whether paired ranking is better or worse than standalone accuracy under the default ranking rule. These values are not pooled with the 506-item grey control or the 37-item mirror control. Grey replacement is an input-sensitivity control, not an additive estimate of visual contribution. The mirror result applies only to the matched 37-item set.

K Limitations

The behavioural score is a restricted yes/no margin under a fixed prompt, token groups, decision boundary, and tie convention. Greedy answers agree with its sign on 98.2 to 99.9% of ARO statements under the threshold protocol, but this was not checked on SugarCrepe or Winoground. SugarCrepe records carry no image identifiers, so their intervals resample items; ARO intervals resample source images. The two scorings differ in prompt, token groups, and image crops, so their standalone accuracies are not pooled.

The threshold analysis covers ARO only. It tests thresholds fitted on the same benchmark and not thresholds transferred from another one; no single threshold can beat τ^*, which still leaves 5.6 to 18.1

Table 17: ARO relation with and without symmetric-relation items (ARO scoring, test split). "Both yes" is the share of items with both descriptions judged yes at $\hat{\tau}$. Each pair of values is all items / without symmetric items; the interval is for the latter and resamples test images with 1,000 draws.

Model	both yes, symmetric	both yes, other	$C - B_0$	$C - B_{\hat{\tau}}$	pairing contrast [95% CI]
Qwen2.5-VL 7B	.948	.149	+.070 / +.072	+.071 / +.073	-.001 / -.005 [-.009,-.001]
Qwen3-VL 8B	.978	.132	+.059 / +.061	+.057 / +.059	+.003 / -.005 [-.010,.000]
InternVL3 9B	.954	.168	+.062 / +.063	+.057 / +.058	-.003 / -.008 [-.012,-.003]
Idefics3 8B	.940	.237	+.098 / +.103	+.099 / +.105	+.044 / +.041 [.036,.046]
LLaVA-1.5 7B	.896	.208	+.199 / +.209	+.090 / +.094	+.021 / +.020 [.016,.025]

Table 18: Greedy answers on the ARO scoring. Agreement is with the sign of the margin. Intervals resample images.

Model	subset	statements	valid	agreement	generated accuracy	margin accuracy
Qwen2.5-VL 7B	ARO attr.	1020	1.000	.995	.811 [.778,.847]	.812
Qwen3-VL 8B	ARO attr.	1020	1.000	.982	.809 [.768,.845]	.799
InternVL3 9B	ARO attr.	1020	1.000	.984	.597 [.569,.625]	.583
Idefics3 8B	ARO attr.	1020	1.000	.994	.716 [.676,.753]	.714
LLaVA-1.5 7B	ARO attr.	1020	1.000	.999	.505 [.501,.512]	.504
Qwen2.5-VL 7B	ARO rel.	1002	1.000	.995	.843 [.809,.874]	.846
Qwen3-VL 8B	ARO rel.	1002	1.000	.992	.825 [.789,.858]	.827
InternVL3 9B	ARO rel.	1002	1.000	.989	.816 [.780,.851]	.819
Idefics3 8B	ARO rel.	1002	.997	.990	.705 [.669,.742]	.703
LLaVA-1.5 7B	ARO rel.	1002	1.000	.991	.639 [.604,.673]	.636

points. Its intervals treat the fitted threshold as fixed. The scaling comparison is one step within one model family. The main missing control is scoring each description without the image, which would show whether the pairing contrast tracks how plausible each description is on its own. Thresholds were not transferred across benchmarks, and only one set of yes/no tokens was used for each scoring. A uniform-grey image is out of distribution relative to natural images and may introduce artifacts.

Table 19: VSR visual controls. Intervals resample whole items. The first mirror column counts every nonpositive mirrored score difference; the expected-reversal column also requires a positive difference before mirroring.

Model	real C	grey C	real–grey	paired 95% CI	mirrored $d \leq 0$	expected reversal
Qwen2.5-VL 7B	0.913	0.713	+0.1996	[0.1581,0.2411]	34/37=0.9189	34/37=0.9189
Qwen3-VL 8B	0.917	0.658	+0.2589	[0.2134,0.3063]	34/37=0.9189	34/37=0.9189
InternVL3 9B	0.913	0.597	+0.3162	[0.2668,0.3656]	34/37=0.9189	34/37=0.9189
Idefics3 8B	0.877	0.650	+0.2273	[0.1818,0.2708]	33/37=0.8919	28/37=0.7568
LLaVA-1.5 7B	0.832	0.393	+0.4387	[0.3794,0.4980]	28/37=0.7568	22/37=0.5946