Do VLMs Reason About Faces? Probing the Perception-Reasoning Gap in Identity Judgment

Mahsa Khoshnoodi
Georgetown University

Sarah Adel Bargal
Georgetown University

Abstract

Vision-Language Models (VLMs) achieve high accuracy on face verification tasks, but do they truly understand facial identity or merely exploit response biases? We introduce an evaluation framework that addresses this question along four complementary dimensions: accuracy, cross-image attribute agreement, identity-attribute coherence, and response bias. Given a pair of facial images, a model is first asked to judge whether they depict the same person; the model is then independently asked, for each of the two images, fine-grained binary questions on identity attributes such as nose shape, eye shape, and jawline structure. A model that truly reasons about identity should produce attribute answers that are logically coherent with its identity judgment. We evaluate five open-source VLMs across four controlled setups that vary both identity and visual similarity. Our results reveal a striking perception-reasoning gap: models perceive facial attributes with high cross-image agreement (up to 88.7%), yet their identity judgments are largely decoupled from this evidence, driven instead by severe response biases. Our findings highlight that accuracy alone masks fundamental failures that only emerge when models are evaluated for reasoning capabilities.

1. Introduction

Verifying whether two facial images depict the same person appears simple, but requires sophisticated visual reasoning. Humans perform compositional visual reasoning: (1) analyze fine-grained facial attributes, (2) compare corresponding features between images, and (3) integrate these comparisons for coherent judgment. [6, 7] Vision-Language Models (VLMs) have recently been applied to face verification, achieving impressive accuracy [13], but accuracy alone does not guarantee sound reasoning. VLMs can produce correct answers by relying on biases rather than performing genuine visual reasoning [10, 14]. This raises a fundamental question: *do VLMs reason systematically, or bypasses analytical reasoning?* We investigate this question through a diagnostic evaluation: we ask VLMs to make identity judgments (same *vs.* different person) and to answer fine-grained questions about identity defining facial attributes (nose shape would qualify as an identity defining attribute, where as hair color would not). By examining the consistency between these responses, we assess whether judgments align with attribute-level assessments. If a model judges two faces as the same person, it should also describe their identifying attributes similarly; violations of this expectation expose cases where identity judgments bypass attribute-level analysis. We propose a four-metric evaluation framework and test five open-source VLMs across controlled setups that vary both identity and visual similarity.

Our findings reveal significant reasoning inconsistencies: models often assert that two images depict the same individual while simultaneously identifying differences in key facial features, and conversely. A model might state that two faces have different noses, lips, and eyes, yet conclude they depict the same individual. Such contradictions suggest correct judgments despite short-circuiting compositional reasoning. This leads to potential unpredictable failure in novel contexts, and has direct implications for model reliability and interpretability where reasoning quality is as important as task accuracy.

2. Evaluation Framework

We design an evaluation framework that probes VLMs at two levels: (1) a holistic identity judgment ('are these the same person?') and (2) fine-grained attribute-level questions about each image independently. By comparing these two levels, we assess whether a model's identity judgment is consistent with its perceived identity attributes. Our evaluation focuses on four complementary metrics, each targeting a distinct failure mode.

2.1. Holistic *vs.* Attribute-Level Probing

Given a face pair (f_i, f_j) and a VLM $\mathcal{M}$, our probing proceeds in two steps:

1. **Holistic Identity Judgment.** The model is presented with both images and asked whether they depict the same

person. This produces a binary judgment $I(f_i, f_j) \in \{\textit{same}, \textit{different}\}$.

2. **Attribute-Level Probing.** Each image is presented *individually* to the model with $K = 50$ binary questions about specific facial attributes (*e.g.*, "Does this person have a narrow nose?" or "Is the face shape round?"). This produces attribute response vectors $\{A_k(f_i)\}_{k=1}^{K}$ and $\{A_k(f_j)\}_{k=1}^{K}$.

A model that claims two faces belong to the same person yet answers questions on corresponding facial attributes differently is exhibiting a reasoning inconsistency, its holistic judgment is not supported by its own fine-grained analysis. Conversely, a model that produces nearly identical attribute descriptions for two faces yet claims they are different people is similarly incoherent.

2.2. Experimental Setup

To systematically test model behavior under varying difficulty, we construct four setups that cross two dimensions: identity (same *vs.* different person) and visual similarity (similar *vs.* varying appearance).

Pair Construction via ArcFace Embeddings. We chose all face images from the CelebAMask-HQ dataset [11] and use ArcFace [5] to construct controlled pair setups based on cosine similarity in the embedding space. We extract face embeddings for all images and compute pairwise cosine similarity. All pairs are additionally matched on gender, age, and skin tone to isolate visual similarity from ground-truth attributes. We define four setups of 50 pairs each (200 pairs total, no image is reused across setups):

1. **Same identity, different appearance (SameDiff).** Same identity, *low* similarity (mean cos $= -0.10$). The same person under substantial appearance variation (pose, lighting, expression), testing whether VLMs can recognize identity beyond surface-level cues.
2. **Same identity, similar appearance (SameSim).** Same identity, *high* similarity (mean cos $= 0.93$). Near-identical views of the same person, serving as a ceiling estimate for basic face matching ability.
3. **Different identity, similar appearance (DiffSim).** Different identity, *high* similarity (mean cos $= 0.75$). Visually similar "lookalike" pairs that require fine-grained discrimination.
4. **Different identity, different appearance (DiffDiff).** Different identity, *low* similarity (mean cos $= -0.30$). Visually distinct pairs, serving as a floor estimate for basic discrimination ability.

This 2×2 design disentangles perceptual ability from response bias: a model that truly discriminates faces should perform well across all four setups, while a biased model will excel only on setups aligned with its default answer.

Attribute question generation. We manually select a subset of CelebA attributes that are identity-defining (e.g., nose shape, eye shape) and prompt Claude [1] to convert them into $K = 50$ natural-language yes/no questions (e.g., Narrow_Eyes $\rightarrow$ "Does this person have narrow eyes?").

2.3. Evaluation Metrics

2.3.1. Accuracy

Accuracy measures the correctness of the model's identity judgment against the ground-truth label $\hat{I}(f_i, f_j)$. We report per-setup accuracy and accuracy averaged across setups.

2.3.2. Attribute Agreement

Attribute Agreement quantifies how similarly a model describes the two faces. For each pair (f_i, f_j), we compute:

$$\text{Agr}(f_i, f_j) = \frac{1}{K} \sum_{k=1}^{K} \left[A_k(f_i) == A_k(f_j) \right] \tag{1}$$

where $K = 50$ is the number of attribute questions. For same-identity pairs, we expect high agreement (Agr $\rightarrow 1$), since both images depict the same person and should elicit matching attribute descriptions. For different-identity pairs, agreement may be lower, though pairs matched on demographic attributes can still yield moderate values.

2.3.3. Coherence Score

The Coherence Score measures whether a model's identity judgment is logically compatible with its own attribute-level responses. A pair is *coherent* if it satisfies:

$$\text{Coh}(f_i, f_j) = \begin{cases} \text{Agr}(f_i, f_j) \geq \tau_s & I = same \\ \text{Agr}(f_i, f_j) \leq \tau_d & I = different \end{cases} \tag{2}$$

where τ_s is a threshold for attribute agreement for same identities and τ_d is a threshold for attribute agreement for different identities. If a model claims "same person," its attribute-level responses should agree across both images (Agr $\geq \tau_s$); if it claims "different person," they should not be highly consistent (Agr $\leq \tau_d$). We set $\tau_d > \tau_s$ because individuals who share demographic attributes can exhibit moderate agreement even when they are different people; specifically, τ_s=0.80 and τ_d=0.88, calibrated on the attribute agreement distribution in Setup 3 (DiffSim). The asymmetry $\tau_d > \tau_s$ is intentional: demographically matched different-identity pairs can exhibit moderate attribute agreement even across distinct individuals, so a symmetric threshold would generate false incoherence on visually similar pairs. Coherence must therefore be read jointly with the bias analysis (Section 2.3.4); a model that achieves high coherence solely through a strong response bias is identified as such by d' and c (see Appendix A.1). The overall coherence score is the fraction of pairs satisfying these constraints:

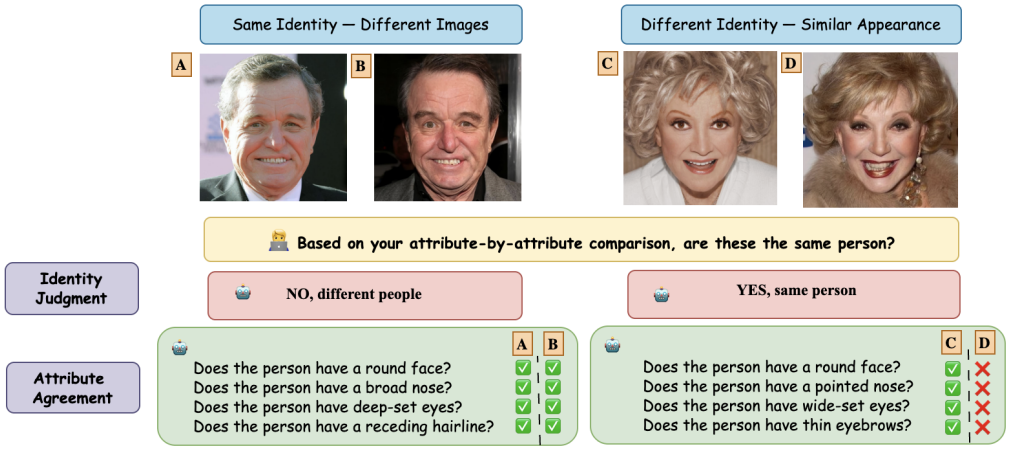

Figure 1. Overview of our evaluation framework. Left: the model incorrectly judges two images of the same person as different, despite producing matching attribute responses. Right: the model incorrectly judges two different people as the same, despite producing conflicting attribute responses. Both cases reveal inconsistencies between judgments and attribute-level reasoning.

$$\text{Coherence} = \frac{1}{N} \sum_{i,j=1}^{N} \text{Coh}(f_i, f_j) \tag{3}$$

where N is the number of image pairs. A score of 1.0 indicates full attribute-judgment consistency. Importantly, this metric requires no ground-truth attribute annotations; it is entirely self-referential.

2.3.4. Response Bias

To separate a model's ability to discriminate faces from its tendency to favor a default answer, we adopt signal detection theory [9]. Treating same-identity pairs as signal-present and different-identity pairs as signal-absent, we compute:

$$d' = z(H) - z(FA), \qquad c = -\tfrac{1}{2}[z(H) + z(FA)] \tag{4}$$

where z denotes the z-score, H is the hit rate (proportion of same-identity pairs correctly judged "same"), and FA is the false alarm rate (proportion of different-identity pairs incorrectly judged "same"). d' measures discriminability independent of bias: $d' = 0$ indicates no ability to distinguish same from different faces, with higher values reflecting stronger discrimination. c captures response tendency: $c < 0$ indicates a liberal bias, where the model favors responding "same"; $c > 0$ indicates a conservative bias, where the model favors responding "different"; and $c = 0$ indicates a neutral criterion.

3. Results

We evaluate five open-source VLMs spanning diverse architectures and parameter scales: InternVL3-8B [3], Llama-3.2-11B [8], LLaVA-OV-1.5-8B [12], Molmo-O-7B [4], and Qwen2.5-VL-7B [2]. All models are evaluated under zero-shot prompting without fine-tuning.

Table 1 reports per-setup accuracy. While overall scores range from 50.0% to 80.0%, decomposing by setup reveals that accuracy conflates perceptual ability with response bias. Llama scores 100% on both diff-pair setups but 0% on both same-pair setups, indicating systematic rejection rather than visual understanding. LLaVA shows the opposite pattern: 94–100% on same-pairs but only 12–44% on diff-pairs, reflecting a strong Yes-bias.

The Bias analysis (Table 4) confirms this: Llama has $d' = 0$ (no discriminability) with extreme No-bias ($c = 2.58$), while LLaVA has low sensitivity ($d' = 1.24$) with strong Yes-bias ($c = -1.20$). Molmo achieves the highest discriminability ($d' = 2.60$) but remains conservative. Only Qwen approaches balanced responding ($c = -0.35$), though its sensitivity is moderate ($d' = 1.52$).

Table 2 shows that attribute agreement is highest for SameSim pairs across all models (90.6–96.4%), confirming that visually similar same-identity pairs yield stable responses. Notably, LLaVA achieves the highest overall agreement (88.7%) despite its heavy Yes-bias, suggesting its visual encoder produces stable attribute representations even when identity judgments are unreliable.

Coherence analysis (Table 3) reveals that Setup 3 (Diff-Sim) is the primary source of incoherence across models. Llama produces 14 incoherent pairs and InternVL 12, *i.e.* models correctly judge "different" but their own attribute-level responses contradict this judgment. Overall coherence remains above 92% for all models, indicating that most incoherence is concentrated in the hardest condition.

Table 1. **Identity judgment accuracy (%)** ($N = 50$ pairs per setup).

Model	Same Diff.	Same Sim.	Diff Sim.	Diff Diff.	All
LLaVA	94.0	**100.0**	12.0	44.0	62.5
InternVL	32.0	**100.0**	**88.0**	**100.0**	**80.0**
Qwen	**74.0**	**100.0**	42.0	90.0	76.5
Molmo	10.0	92.0	**100.0**	**100.0**	75.5
Llama	0.0	0.0	**100.0**	**100.0**	50.0

Table 2. **Cross-image attribute agreement (%)** ($N = 50$ pairs per setup).

Model	Same Diff.	Same Sim.	Diff Sim.	Diff Diff.	All
LLaVA	87.7	**96.4**	86.4	84.4	**88.7**
InternVL	84.4	94.5	85.4	83.2	86.9
Qwen	83.6	93.5	83.8	82.2	85.8
Molmo	80.7	90.6	79.6	75.7	81.7
Llama	83.2	93.5	84.2	82.9	86.0

Table 3. **Identity-attribute coherence (%)**. Parentheses show incoherent pair count out of 50. Thresholds: $\tau_s = 0.80$, $\tau_d = 0.88$.

Model	Same Diff.	Same Sim.	Diff Sim.	Diff Diff.	All
LLaVA	90.0 (5)	100.0	94.0 (3)	86.0 (7)	92.5
InternVL	98.0 (1)	100.0	76.0 (12)	100.0	93.5
Qwen	90.0 (5)	100.0	82.0 (9)	98.0 (1)	92.5
Molmo	98.0 (1)	100.0	92.0 (4)	100.0	**97.5**
Llama	100.0	100.0	72.0 (14)	100.0	93.0

Table 4. **Bias analysis.** d': sensitivity (higher = better discrimination). c: criterion ($c < 0$: liberal/Yes-bias, $c > 0$: conservative/No-bias). H: hit rate, FA: false alarm rate.

Model	H	FA	d'	c	Bias
LLaVA	0.97	0.72	1.24	−1.20	Yes-bias
InternVL	0.66	0.06	1.93	0.56	No-bias
Qwen	0.87	0.34	1.52	−0.35	Yes-bias
Molmo	0.51	0.00	**2.60**	1.28	No-bias
Llama	0.00	0.00	0.00	2.58	No-bias

3.1. Discussion

Our four metrics collectively reveal that current VLMs do not perform compositional facial reasoning. The perception-reasoning gap rests on three concurrent findings: attribute agreement is high and stable across all models and setups (81.7–88.7%, Table 2), confirming genuine perceptual competence; identity judgment accuracy collapses where visual similarity conflicts with ground-truth identity (Table 1); and every model exhibits substantial criterion bias, with none approaching balanced responding (Table 4). Coherence localizes where this decoupling is most acute, concentrating in DiffSim, rather than serving as the primary evidence for the gap.

The most striking finding is that no model bases its identity decision on its own attribute-level evidence. Llama answers "different" to every pair; LLaVA answers "same" to almost every pair. Even InternVL, the top-accuracy model at 80.0%, achieves only 32% on SameDiff: its overall score is carried by No-bias succeeding on easy diff-pairs.

Performance degrades sharply when visual similarity conflicts with ground-truth identity. SameSim pairs are trivially easy, with all models achieving $\geq$92% accuracy. SameDiff and DiffSim pairs expose failures, revealing that VLMs rely on pixel-level resemblance as a proxy for identity rather than grounding decisions in discriminative features such as bone structure or facial proportions.

Finally, no single metric suffices to diagnose model behavior. InternVL leads on accuracy but produces 12 incoherent DiffSim pairs. LLaVA leads on attribute agreement but has the strongest Yes-bias and the second-lowest accuracy. Molmo leads on coherence but achieves only 10% on SameDiff. Only by examining all four metrics together does the full picture emerge: current VLMs have not learned to integrate attribute percepts in identity decisions. Per-model behavioral profiles are discussed in Appendix A.4.

4. Conclusion

In this work, we propose an evaluation framework that probes whether VLMs genuinely reason about facial identity by decomposing model behavior into four complementary dimensions: accuracy, attribute agreement, coherence, and response bias. Our analysis reveals a pervasive gap between perception and reasoning. Models describe fine-grained facial attributes with notable stability, yet their identity judgments remain largely decoupled from this evidence. No single dimension captures the full picture; only by examining all four dimensions together do the failures become apparent. We believe this perception-reasoning gap extends to other fine-grained visual reasoning tasks, and we advocate for evaluation protocols that explicitly assess model coherence.

References

[1] Anthropic. Claude ai model. https://www.anthropic.com, 2024. Accessed: 2026-03-30. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 3

[4] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Bhatt, Ethan Elber, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 3

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[6] Miriam Doh, Caroline Mazini Rodrigues, Nicolas Boutry, Laurent Najman, Matei Mancas, and Hugues Bersini. Bridging human concepts and computer vision for explainable face verification. *arXiv preprint arXiv:2403.08789*, 2024. 1

[7] Matthew C. Fysh and Markus Bindemann. Understanding face matching. *Quarterly Journal of Experimental Psychology*, 76(4):862–880, 2023. Epub 2022 Jun 17. 1

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[9] David M Green and John A Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966. 3

[10] Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob A Hansen, Sivan Doveh, Victor I Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, et al. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *Advances in Neural Information Processing Systems*, 37:22927–22946, 2024. 1

[11] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[13] Junwoo Lim and Ho-Sub Yoon. Vision-language model-based face verification for preventing unauthorized access. In *2025 16th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1906–1910, 2025. 1

[14] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025. 1

A. Additional Analysis and Design Considerations

A.1. Coherence Score: Formulation, Thresholds, and Interpretation

The coherence metric (Eq. 2) is intentionally self-referential: it evaluates whether a model's identity judgment is logically compatible with its own attribute-level responses, without reference to ground-truth labels. This design reflects a principled stance; we are not measuring whether the model is correct, but whether it is internally consistent. A model that always responds "different" and whose attribute agreement consistently falls below $\tau_d = 0.88$ will score high coherence, not because it reasons well, but because its bias happens to align with its attribute responses by default. This is a feature, not a flaw: such a model is coherent in the trivial sense, and our bias analysis (Table 4, $d' = 0.00$, $c = 2.58$ for Llama) makes this triviality explicit. Coherence and bias must therefore be read jointly; neither suffices alone.

The thresholds are set asymmetrically by design. Because demographically matched different-identity pairs can exhibit moderate attribute agreement even across distinct individuals, a stricter lower bound for "different" judgments would produce false incoherence on pairs that are genuinely similar in appearance. We set $\tau_s = 0.80$ and $\tau_d = 0.88$ based on the empirical attribute agreement distribution in Setup 3 (DiffSim), which represents the hardest and most diagnostic condition.

To clarify the Llama case specifically: Llama predicts "different" for every pair across all setups, so coherence is evaluated under the τ_d branch of Eq. 2 throughout. Incoherence fires only when Llama says "different" yet attribute agreement exceeds $\tau_d = 0.88$. This occurs in 14 DiffSim pairs, where Llama's own attribute responses describe two faces as nearly identical yet it still concludes they are different people. This is precisely the failure mode our framework is designed to surface: a judgment reached through reasoning that contradicts the model's own perceptual evidence, regardless of whether that judgment happens to be correct.

A.2. Interpretation of High Overall Coherence

Overall coherence above 92% across all models does not weaken the perception-reasoning gap claim; it reflects the structure of our evaluation design. Three of the four setups (SameSim, DiffDiff, and SameDiff) are conditions where even a biased model can produce judgments that happen to align with its attribute responses, because visual similarity and ground-truth identity are not in conflict. The diagnostic weight of our framework rests on Setup 3 (DiffSim), where the two dimensions conflict directly. Coherence in this setup drops to 72–94% across models, with the most incoherent models (Llama: 14 pairs, InternVL: 12 pairs) being precisely those that exhibit the strongest bias in Table 4.

More fundamentally, the perception-reasoning gap is not solely a coherence claim. It is established by the conjunction of three findings: (1) attribute agreement remains high and stable across all models and all setups (81.7–88.7%, Table 2), demonstrating genuine perceptual competence; (2) identity judgment accuracy collapses in setups where visual similarity conflicts with ground-truth identity (e.g., Molmo: 10% on SameDiff, Llama: 0% on both same-identity setups, Table 1); and (3) every model exhibits substantial criterion bias (Table 4), with no model approaching balanced responding. Taken together, these findings show that perceptual competence at the attribute level does not propagate into identity judgments, which are instead governed by response bias. Coherence localizes where this decoupling is most acute; it is not the sole carrier of the argument.

A.3. Diagnostic Strength of the Question Set

The 50 attribute questions used in our framework are drawn from the subset of CelebA attributes that are identity-defining, specifically structural and geometric facial properties such as nose shape, eye shape, jawline structure, and facial proportions. Attributes reflecting transient or controllable appearance factors, such as hair color, makeup, or accessories, were explicitly excluded, as described in Section 2.2. This distinction is central to our framework: we probe attributes that should, in principle, remain stable across images of the same person and differ across images of different people.

We note that using moderately general identity-defining attributes, rather than highly fine-grained ones, makes our coherence test conservative in the favorable direction. Generic attributes are more likely to yield high agreement across any pair of faces, which raises the baseline for attribute agreement regardless of identity. Violations of coherence under these conditions are therefore more meaningful, not less, since they occur despite the test being designed to make agreement easy. If attribute questions were more fine-grained and discriminative, agreement would be expected to drop for different-identity pairs, making incoherence in DiffSim even more pronounced. Our current design thus understates the severity of the perception-reasoning gap relative to what a stricter attribute set would reveal.

A.4. Per-Model Behavioral Profiles

The per-setup breakdowns in Tables 1–4 reveal distinct behavioral profiles that are obscured by averaged accuracy, and that directly bear on the perception-reasoning gap claim. We summarize each model in turn.

Llama-3.2-11B. Llama is the clearest case of bias dominating judgment. It answers "different" to every pair without exception ($H = 0.00$, $FA = 0.00$, $d' = 0.00$, $c = 2.58$), achieving 100% accuracy on both DiffDiff and DiffSim by chance alignment with its default response, and 0% on both same-identity setups. Despite this, its attribute agreement is stable across setups (82.9–93.5%, Table 2), confirming that its visual encoder extracts consistent facial representations. The 14 incoherent DiffSim pairs arise because Llama's attribute agreement occasionally exceeds $\tau_d = 0.88$ on visually similar lookalike pairs, yet it still responds "different." This model illustrates the ceiling case of judgment-attribute decoupling: maximum perceptual stability, zero identity discriminability.

LLaVA-OV-1.5-8B. LLaVA exhibits the mirror-image failure. It answers "same" to almost every pair ($H = 0.97$, $FA = 0.72$, $c = -1.20$), achieving 94–100% on same-identity setups but only 12–44% on different-identity setups. Yet it leads all models on overall attribute agreement (88.7%), indicating that its visual encoder is highly stable. Its 7 incoherent DiffDiff pairs arise when it says "same" for a visually dissimilar different-identity pair whose attribute agreement falls below $\tau_s = 0.80$. LLaVA illustrates that strong perceptual representations and strong response bias can coexist, and that accuracy on one identity condition reveals nothing about the other.

InternVL3-8B. InternVL is the top-accuracy model overall (80.0%) but achieves only 32% on SameDiff, the condition requiring recognition of identity across appearance variation. Its overall score is carried by No-bias succeeding on easy DiffDiff and DiffSim setups ($FA = 0.06$, $c = 0.56$). The 12 incoherent DiffSim pairs, the second-highest count, arise because InternVL correctly judges "different" on lookalike pairs while its own attribute responses describe the two faces as nearly identical. This model illustrates how accuracy can mislead: an 80% overall score masks catastrophic failure on one of the four setups and substantial attribute-judgment decoupling on the hardest condition.

Qwen2.5-VL-7B. Qwen is the closest to balanced responding among all models ($c = -0.35$), though its discriminability remains moderate ($d' = 1.52$). It achieves 74% on SameDiff, the best of any model on that condition, and 42% on DiffSim, reflecting genuine difficulty on lookalike pairs rather than pure bias. Its 9 incoherent DiffSim pairs confirm that the lookalike condition remains the critical stress test even for the least-biased model.

Molmo-O-7B. Molmo achieves the highest discriminability ($d' = 2.60$) and the highest overall coherence (97.5%), but scores only 10% on SameDiff. Its near-zero false alarm rate ($FA = 0.00$) indicates a strong conservative bias ($c = 1.28$) that happens to be beneficial on different-identity setups and detrimental on same-identity setups. Its high coherence is therefore partially a product of consistent No-bias rather than genuine attribute-grounded reasoning, a distinction that the bias analysis makes explicit.

Summary. Across all five models, no behavioral profile is consistent with attribute-grounded identity reasoning. The two models with the most stable attribute representations (LLaVA, Llama) have the worst discriminability. The model with the highest discriminability (Molmo) fails on SameDiff. The most balanced model (Qwen) still concentrates failures in DiffSim. This pattern is not coincidental; it reflects a systematic decoupling between the attribute-perception pathway and the identity-judgment pathway across architectures and parameter scales.