PAIRWISE AGREEMENT DOES NOT CERTIFY A SHARED WORLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can answer questions about a scene correctly while their answer probabilities fail to describe one shared world. We measure this failure by comparing the smallest change needed to make overlapping marginals agree with the smallest change needed to obtain them from one joint distribution. The difference isolates a global inconsistency that pairwise checks omit; a third term measures the excess discrepancy of the model's global readout. Probabilities are computed for complete autoregressive answers, including the native termination token. Across five models, all four with complete eight-world competence coverage exhibit a positive gap, and three reproduce it under both prompt forms. Testing three models on 512 GQA and CLEVR images yields 32 positive conditions on 30 distinct images with correct baseline single-question and joint answers, including 14 images for Llama-based Idefics3. Across four fixed readout variants, answer wording and order can change the global component even when every tested answer remains correct. These findings show why global compatibility must be tested at the probability level, with the answer interface explicitly specified.

1 INTRODUCTION

A model can answer local questions correctly while its answer probabilities admit no common joint distribution over the scene. Accuracy checks the most probable candidates; compatibility constrains the full probability tables. Figure 1 illustrates why agreement on shared marginals is insufficient.

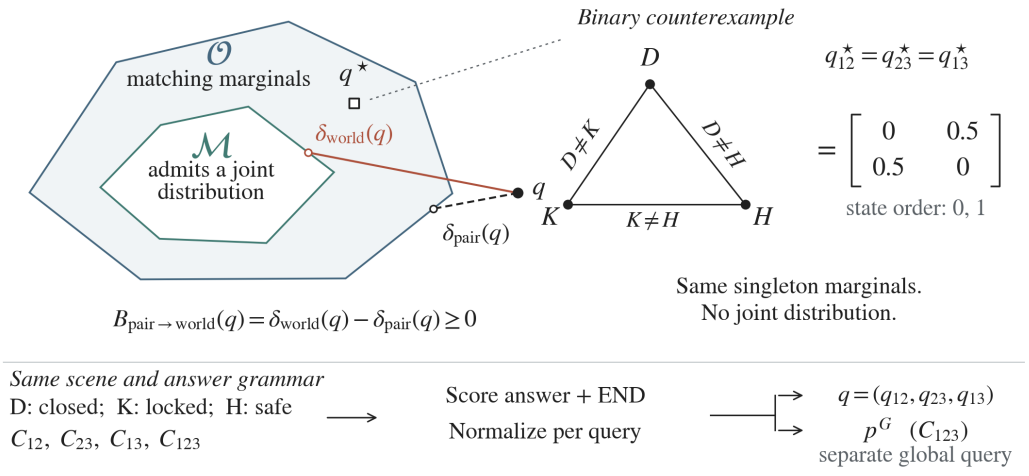


Figure 1: Pairwise overlap agreement can omit a global constraint. Right: a mathematical binary counterexample has matching singleton marginals but no joint distribution, placing $q^\star \in \mathcal{O} \setminus \mathcal{M}$ (left). The general point q shows the distances defining $B_{\text{pair}\to\text{world}}$. Below: complete-answer probabilities include the native terminator and are normalized per query; p^G is queried separately. Set geometry is schematic; distances use the ℓ_1 norm.

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.

Global compatibility matters when learned systems expose several predictions about a common environment. Latent world models support control by predicting future states or observations (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025). Generative simulators produce visual environments (Yang et al., 2023; Bruce et al., 2024), while vision-language models (VLMs) answer questions about objects and relations (Bai et al., 2025b). When several reports describe the same scene, their global compatibility becomes a basic evaluation question.

Current consistency tests usually measure a weaker property. Language-model work studies agreement across prompts, facts, explanations, or samples (Elazar et al., 2021; Kassner et al., 2021; Jung et al., 2022; Wang et al., 2023). Vision-language benchmarks test compositional reasoning, image-text ordering, and hallucination (Johnson et al., 2017; Hudson & Manning, 2019; Thrush et al., 2022; Li et al., 2023). These tests remain valuable, but local accuracy and matching overlap marginals do not imply global probabilistic realizability. Three pair distributions can agree on every shared singleton marginal while admitting no joint distribution over all variables. Figure 1 shows the smallest binary case.

The gap between local and global consistency has a precise mathematical basis. Marginal extension theory asks whether local measures have a common extension (Vorob'ev, 1962), and graphical-model theory separates local and marginal polytopes (Wainwright & Jordan, 2008). Local agreement is sufficient for particular acyclic structures, but not for general cyclic context families (Abramsky & Brandenburger, 2011; Atserias & Kolaitis, 2020). The empirical question is whether this distinction appears in learned-model probabilities after the measurement controls for model competence, prompt form, and complete autoregressive answers.

We address this question with a fixed probabilistic interface. A finite ontology defines valid worlds, local contexts, candidate statements, and shared conditioning information. Each model assigns teacher-forced probability to every statement followed by its native assistant terminator. We compare the resulting local laws with two nested sets: laws that satisfy pairwise overlap equalities and laws induced by a global world distribution. Their ℓ_1 distances isolate a nonnegative pair-to-world gap. A unique-maximum competence criterion establishes that the queried states are readable. In the controlled ontology, an eight-world coverage criterion selects the cells used for comparison; on natural question triples, we report both all-image and correctness-stratified results. These empirical criteria remain separate from mathematical feasibility.

The evidence separates controlled replication from visual generalization. Five VLMs are evaluated on 80 controlled stimuli, four evidence carriers, and two prompt forms. Every model with eight-world coverage has a positive gap; three reproduce it under both forms, and LLaVA-NeXT adds a Mistral-decoder result. We then test three models on 512 distinct GQA and CLEVR images. Qwen3-VL and Idefics3 both yield positive gaps with correct single-question and joint answers. Fixed changes to answer wording and order expose a further distinction: the answers can stay correct while the gap changes. We therefore report both the existence of the effect and its same-image stability.

Our contribution is an empirical diagnostic built on established marginal theory. It connects complete autoregressive sequence probabilities to finite local and global laws, separates three sources of disagreement in common ℓ_1 units, and identifies the omitted global component across qualified VLM architectures. The practical conclusion is direct: pairwise consistency should be reported as a local condition, not as a certificate of one shared probabilistic world.

2 Related Work

2.1 Marginal extension and local consistency

The existence of a global measure with specified marginals is a mature mathematical problem. Vorob'ev (1962) characterized consistent families of measures and their extensions, while Wainwright & Jordan (2008) developed the local and marginal polytopes used in graphical-model inference. Local consistency becomes sufficient under specific structural conditions, including acyclic context hypergraphs (Atserias & Kolaitis, 2020). Related local-to-global conditions arise in constraint satisfaction (Dechter, 1992). Cyclic contexts also appear in contextuality, where compatible local empirical models may lack a global section (Abramsky & Brandenburger, 2011). Related

problems include marginal-consistency message passing (Werner, 2015), compatibility of finite full conditionals (Slavkovic & Sullivant, 2004), and causal marginal problems (Gresele et al., 2022). We connect these ideas to a specified complete-answer interface and a learned-model measurement.

2.2 Consistency and probability in language models

Language-model consistency work studies whether prompts or related facts yield compatible outputs. Prior methods measure and regularize factual consistency (Elazar et al., 2021), maintain a memory of beliefs (Kassner et al., 2021), or search for logically consistent explanations (Jung et al., 2022). Other work corrects few-shot answer biases (Zhao et al., 2021) or estimates whether an answer is correct (Kadavath et al., 2022; Kuhn et al., 2023). Few-shot example order can change predictions (Lu et al., 2022), and several surface strings can compete for the same semantic answer (Holtzman et al., 2021). These lines motivate two choices in our protocol: prompt forms remain separate probability measurements, and answer probabilities are defined at the sequence level. The remaining distinction is global: even well-defined local answer laws may fail to be marginals of one world law.

Recent work brings probabilistic coherence directly to learned systems. Kotawala (2026) measures the distance from claims composed by multiple large language model (LLM) agents to a joint coherent polytope, while Andrews (2026) turns coherence axioms into label-free evaluation penalties. Choi et al. (2026) instead constructs a stationary distribution by repeatedly resampling variables from LLM conditionals. Our interface evaluates separately queried local and global VLM laws for one scene. The nested distances isolate the global component missed by overlap-only correction, and the complete-answer construction identifies the autoregressive event whose probability enters the test.

2.3 Diagnostics for vision-language models

Vision-language evaluation has moved from broad question answering toward controlled tests of objects, relations, and compositionality. Compositional Language and Elementary Visual Reasoning (CLEVR) uses a synthetic scene generator to isolate visual reasoning (Johnson et al., 2017), GQA adds compositional questions over real images (Hudson & Manning, 2019), and Massive Multi-discipline Multimodal Understanding (MMMU) tests broad multimodal reasoning (Yue et al., 2023). Winoground uses matched image-caption pairs to expose binding failures (Thrush et al., 2022), while Polling-based Object Probing Evaluation (POPE) measures object hallucination (Li et al., 2023). Our test addresses another level of evaluation. It asks whether several probability tables over overlapping scene variables have a common global realization after the model has demonstrated competence on the specified states.

2.4 World models and shared state

World models learn predictive state for control or generation (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025). Recent systems also learn interactive visual environments and closed-loop sensor simulation (Bruce et al., 2024; Yang et al., 2023). An architecture that explicitly samples one shared state and renders every view from that state induces a joint law by construction. Its external semantic readouts can still be audited against a specified ontology, but the raw generated tuple already has a joint distribution. Our diagnostic is most informative when local predictions are queried separately, when several models or agents contribute local laws, or when a shared latent state is not itself the object being certified.

3 Measuring Global Consistency

3.1 Variables and conditioning

The diagnostic begins with a finite ontology, because unconstrained outputs always admit a product joint distribution. Let V index the specified variables, $\mathcal{X}_v$ be the state space of variable v, and $\Omega \subseteq \prod_{v \in V} \mathcal{X}_v$ be the valid global worlds under a fixed condition. A local context $C \subseteq V$ observes the projection $\pi_C : \Omega \to \mathcal{X}_C$. For one model, scene, prompt, and output grammar, the model

induces a local law $q_C \in \Delta(\mathcal{X}_C)$. All local laws in one test share the same scene evidence and common prompt instructions; only the requested variables and their answer format differ.

Our controlled experiment uses three binary variables, door state D, key state K, and hazard state H. The valid set contains all eight assignments, and the local contexts are $C_{12} = (D, K)$, $C_{23} = (K, H)$, and $C_{13} = (D, H)$. The model also reports a global law $p^G \in \Delta_8$ for $C_{123} = (D, K, H)$. This complete binary cube is intentionally simple: global incompatibility cannot be attributed to an unreported physical constraint. The visual extension instead uses three original binary questions per image. There, the full cube is an outer approximation to any logical constraints among the answers, as detailed in Appendix I.2. Failure to admit a joint law on this full cube also rules out a joint law on any stricter semantic support.

3.2 What pairwise checks omit

Stack the three four-state pair laws as $q = (q_{12}, q_{23}, q_{13}) \in \mathcal{Q} = \Delta_4^3$. Let $\mathcal{O} \subseteq \mathcal{Q}$ be the overlap polytope in which the two reports for each singleton variable have the same marginal. Let $A \in \{0,1\}^{12\times 8}$ map a global eight-state law to its three pair marginals, and define the marginal polytope

$$\mathcal{M} = A\Delta_8. \tag{1}$$

Every global law has matching singleton marginals, so $\mathcal{M} \subseteq \mathcal{O} \subseteq \mathcal{Q}$. This inclusion is the only fact needed for the diagnostic.

We measure distance to both feasible sets in the same ℓ_1 units:

$$\delta_{\text{pair}}(q) = \min_{v\in\mathcal{O}} \|q - v\|_1, \qquad \delta_{\text{world}}(q) = \min_{v\in\mathcal{M}} \|q - v\|_1. \tag{2}$$

The set inclusion gives $0 \le \delta_{\text{pair}} \le \delta_{\text{world}}$ without an empirical threshold. We define the omitted component as

$$B_{\text{pair}\to\text{world}}(q) = \delta_{\text{world}}(q) - \delta_{\text{pair}}(q) \ge 0. \tag{3}$$

A positive value means that requiring a joint law costs more than matching the shared singleton marginals, with both costs measured from the original q. Both distances are finite linear programs.

The global distance is the compatibility test: $\delta_{\text{world}} = 0$ exactly when a joint law exists. The gap instead compares two minimum costs measured from the same q. It can vanish while both distances remain positive, so $B_{\text{pair}\to\text{world}} = 0$ alone does not certify compatibility.

The context structure determines whether the second program is necessary. If the local contexts satisfy the running-intersection property, consistent local laws can be joined into a global law when the specified support factorizes over those contexts, so overlap agreement is sufficient (Atserias & Kolaitis, 2020; Wainwright & Jordan, 2008). Cyclic context families can make the inclusion $\mathcal{M} \subseteq \mathcal{O}$ strict. The three pair contexts in our experiment form the smallest such binary cycle, and Figure 1 gives a point in $\mathcal{O} \setminus \mathcal{M}$. This yields a direct evaluation rule: inspect the context structure first, then solve the global program when overlap agreement does not guarantee an extension.

The model's global readout provides a third point of comparison. Its pair marginals are $Ap^G \in \mathcal{M}$, so

$$\begin{aligned} D_G(q, p^G) &= \|q - Ap^G\|_1 \\ &= \delta_{\text{pair}}(q) + B_{\text{pair}\to\text{world}}(q) + E_{\text{global readout}}(q, p^G), \end{aligned} \tag{4}$$

where $E_{\text{global readout}} = D_G - \delta_{\text{world}} \ge 0$. The three terms answer separate questions. Overlap distance measures unmatched shared marginals. The pair-to-world gap isolates the additional global-world constraint. Global-readout excess measures how much more local-law discrepancy the model's own global answer incurs than a nearest coherent explanation of its local answers. Table 1 summarizes the interpretation used throughout the experiment.

3.3 Complete autoregressive answer events

A candidate token string is not yet a complete answer event. For condition c, state s, candidate tokens $v_s = (v_{s,1}, \ldots, v_{s,L_s})$, and the model's native assistant terminator e, the autoregressive chain rule

gives

$$K^{\mathrm{pre}}(s \mid c) = \prod_{j=1}^{L_s} q(v_{s,j} \mid c, v_{s,<j}), \quad K^{\mathrm{complete}}(s \mid c) = K^{\mathrm{pre}}(s \mid c) q(e \mid c, v_s). \tag{5}$$

After normalization over the specified candidate family,

$$p^{\mathrm{complete}}(s \mid c) = \frac{p^{\mathrm{pre}}(s \mid c)\tau(s,c)}{\sum_r p^{\mathrm{pre}}(r \mid c)\tau(r,c)}, \qquad \tau(s,c) = q(e \mid c, v_s). \tag{6}$$

Equation (6) follows directly from the chain rule. It equals the prefix law exactly when τ is constant over candidates with positive prefix probability. We apply Equations (2), (3), and (4) to the complete-answer laws. Prefix laws are retained only as a matched measurement control.

3.4 Correctness and coverage

Correctness separates a probability-level failure from an incorrect answer. A condition is scene competent if the true state is the unique maximum-probability candidate in all three pair contexts and the global context. In the controlled study, a carrier-prompt cell is eligible if it contains a competent condition for every one of the eight worlds. Only competent conditions from eligible cells enter that study's primary gap analysis. This rule is fixed independently of the sign of $B_{\mathrm{pair}\to\mathrm{world}}$ and establishes coverage of its shared ontology.

The visual extension reports all images and then stratifies by correctness. Its stricter subset requires the three baseline single-question answers and all four joint-context answers to be correct. Eight-pattern coverage is reported separately because each image has different questions; it is not required to test a given probability table. The development-informed change to the execution rule is recorded in Appendix I.1. Both studies use $B_{\mathrm{pair}\to\mathrm{world}} > 10^{-8}$ as the fixed numerical resolution criterion, without fitting an effect threshold.

4 Experiments

4.1 Models, stimuli, and probability extraction

We evaluate five released VLM checkpoints with distinct visual and language stacks. The panel contains Qwen3-VL-4B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025a;b), Mage-VL with a from-scratch visual encoder and Qwen3 decoder (Yang et al., 2026), Idefics3 with a Llama 3.1 decoder (Laurençon et al., 2024), and LLaVA-NeXT with a Mistral decoder (Liu et al., 2024). It therefore covers Qwen, Llama, and Mistral decoder families while retaining one fixed task. The ontology contains the two states of each of D, K, and H. Pair prompts request one of four ordered statements. The global prompt requests one of eight. We use a direct grammar and a matched world-first grammar whose candidate strings begin with "WORLD:". No exemplar, numeric codebook, length normalization, or post-hoc calibration is used.

The 80 stimuli cover four carriers. Authoritative text states all three values, raster images print the same semantic words in three panels, icons depict the three states with four visual styles per world, and photographs compose six licensed real images into four variants per world. Each stimulus is queried under two prompt forms and four contexts. Across five models, this yields 3,200 laws, 800 local-global conditions, and 16,000 candidate sequences. We teacher-force every candidate and its native assistant terminator, then normalize sequence probabilities only within the appropriate four-state or eight-state candidate family.

Table 1: The decomposition separates three consistency failures. Each row states the meaning of a positive term.

Term	Compared quantities	Positive value means
δ_{pair}	q to $\mathcal{O}$	Shared overlap marginals disagree
$B_{\mathrm{pair}\to\mathrm{world}}$	$d_1(q,\mathcal{M}) - d_1(q,\mathcal{O})$	Global constraints add distance beyond overlap repair
$E_{\mathrm{global\ readout}}$	D_G versus δ_{world}	The model's global readout incurs excess local-law discrepancy

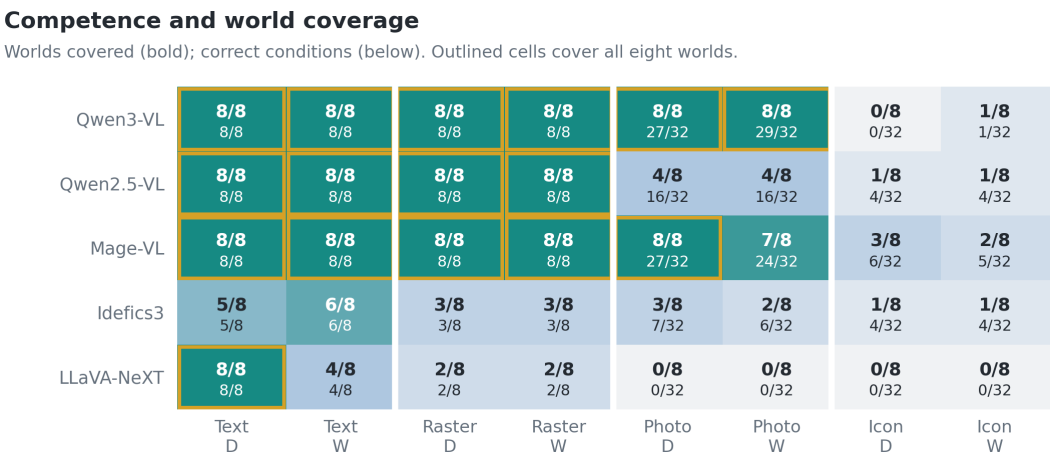


Figure 2: Four models provide complete eight-world evidence for the global test. Each cell reports competent worlds out of eight and competent conditions. Gold borders mark eligible cells. D and W denote direct and world-first prompts.

Table 2: Complete-answer outcomes across the model panel. D/W denotes direct and world-first prompts. Competent counts use all 160 conditions per model; Eligible counts carrier-prompt cells. Gap columns use competent conditions from eligible cells. A dash marks a prompt with no eligible cell.

Model	Competent	Eligible	$B_{\text{pair}\to\text{world}} > 10^{-8}$ D/W	Max $B_{\text{pair}\to\text{world}}$ D/W
Qwen3-VL-4B	89/160	6	9/2	1.05×10^{-4}/1.79×10^{-7}
Qwen2.5-VL-7B	72/160	4	7/7	6.35×10^{-3}/2.32×10^{-3}
Mage-VL-4B	94/160	5	6/3	3.25×10^{-1}/4.13×10^{-2}
Idefics3-8B	38/160	0	–/–	–/–
LLaVA-NeXT-7B	16/160	1	1/–	7.43×10^{-3}/–

4.2 Coverage in the controlled ontology

The coverage criterion separates a structural test from a visual recognition failure. Figure 2 shows the complete-answer competence counts. Qwen3-VL yields six eligible cells, Qwen2.5-VL yields four, and Mage-VL yields five. LLaVA-NeXT qualifies in the direct-text cell, giving an independent Mistral-decoder test. Idefics3 recognizes 38 of 160 conditions but no carrier-prompt cell covers all eight worlds, so it contributes no eligible condition to this comparison. In total, 187 competent conditions from 16 eligible cells enter the main analysis. All other conditions remain in the complete results.

4.3 Global gaps across qualified models

The main result is that all four models meeting the coverage criterion expose distance omitted by pairwise checks. Figure 3 reports all 187 competent conditions from eligible cells. Of these, 35 exceed the fixed 10^{-8} numerical criterion. Qwen3-VL, Qwen2.5-VL, and Mage-VL have resolved positives under both prompt forms. Mage-VL gives the largest gaps, 0.325 for direct prompts and 0.0413 for world-first prompts. LLaVA-NeXT adds a Mistral-decoder result: one of eight conditions in its eligible direct-text cell is positive, with a gap of 7.43×10^{-3}. Thus every model with eight-world coverage shows the omitted global component. Table 2 gives the model-level evidence.

The decomposition shows that the omitted component is distinct from two more common disagreements. All 187 eligible conditions have positive δ_{pair}, so unmatched overlap marginals are widespread. A positive $B_{\text{pair}\to\text{world}}$ appears in 35 conditions: requiring a joint law costs more than the minimum overlap correction. The model's own global answer also incurs positive excess dis-

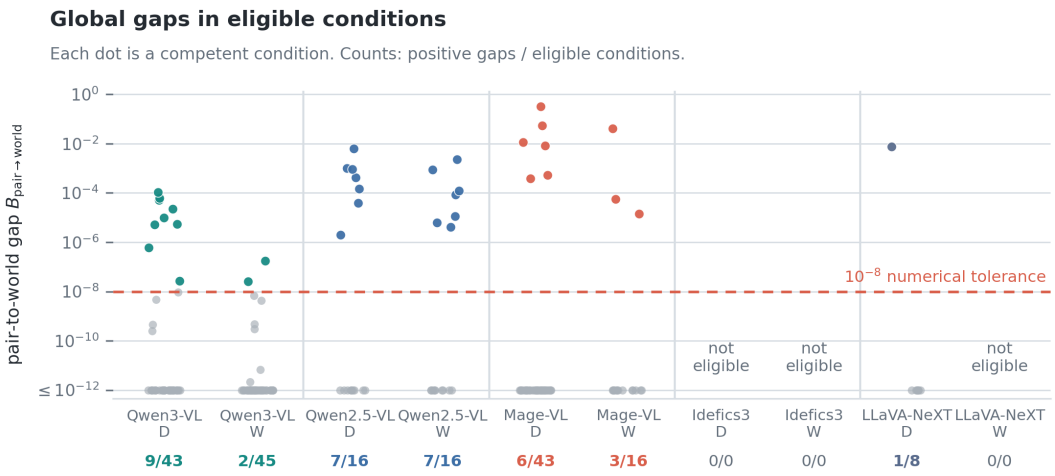


Figure 3: Every qualified model has a positive pair-to-world gap. Dots are scene-competent conditions from eligible cells. Colored dots exceed the fixed 10^{-8} numerical criterion. Counts give resolved positives over eligible conditions. Missing groups did not satisfy eight-world coverage.

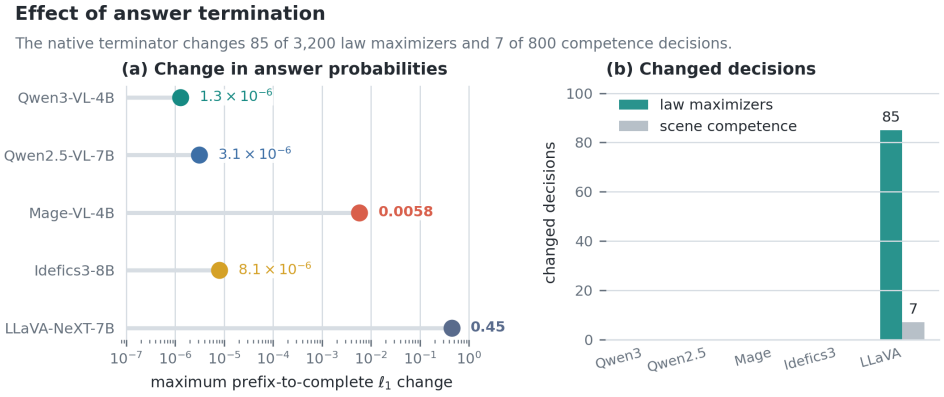


Figure 4: Complete-answer measurement is decision-relevant for LLaVA-NeXT. Left: maximum ℓ_1 transport for each model. Right: changes in law maximizers and scene competence. The eligible gap conclusion is unchanged.

crepancy in 185 conditions. These terms answer different questions and should not be combined into one generic consistency score.

The relative size of $B_{\text{pair}\to\text{world}}$ confirms that the global component can be substantive. Among the 35 resolved positive conditions, $B_{\text{pair}\to\text{world}}/\delta_{\text{world}}$ has a median of 1.95% and reaches 58.2%. Pairwise mismatch remains larger in most conditions, but it does not bound the unreported global component. This is why a small pairwise score cannot serve as a global certificate.

4.4 Effect of Answer Termination

Termination probability depends strongly on the model and candidate, so an unterminated candidate is not equivalent to a finished answer. The maximum prefix-to-complete ℓ_1 change ranges from 1.29×10^{-6} for Qwen3-VL to 0.451 for LLaVA-NeXT. Across the panel, exact termination changes 85 of 3,200 law maximizers, 7 of 800 scene-competence decisions, and 6 gap classifications before eligibility filtering. All of these decision changes occur in LLaVA-NeXT. The 187 eligible-condition gap classifications and the four qualified-model conclusions remain unchanged. Complete-event measurement therefore matters for the probability interface without creating the main finding.

4.5 CORRECT VISUAL ANSWERS CAN RETAIN A GLOBAL GAP

The gap also appears beyond the controlled ontology. We evaluate Qwen3-VL, Idefics3, and LLaVA-NeXT on 256 held-out images from each of GQA and CLEVR (Hudson & Manning, 2019; Johnson et al., 2017). Each image supplies three original binary questions, with yes/no or true/false answers in forward or reverse order. The 512 images yield 6,144 conditions. Among 877 conditions with correct baseline single-question and joint answers, 32 have positive gaps on 30 distinct images. Qwen3-VL contributes 16 images and Idefics3 contributes 14, with both models represented on both datasets. Their original-image accuracy exceeds a fixed shuffled-image control by 14.1 to 24.7 percentage points.

Existence across images does not imply stability across readouts. There are 119 model-image cases in which the baseline single-question answers and all four joint readout variants are correct. Eleven have a positive gap in at least one variant, but none has a positive gap in all four. For one Idefics3 CLEVR image, the gap is unresolved in forward order and reaches 0.1496 and 0.1202 in reverse order under the two answer vocabularies, with correctness unchanged. In the yes/no comparison, overlap distance falls from 0.8781 to 0.6182 even as the additional global component becomes positive and accounts for 19.5% of world distance. Thus improved overlap agreement can coexist with a larger omitted component on the same correctly answered image. Appendix I gives the full matrix, matched analysis, and numerical reconstruction; Figure 5 shows this image and its distances.

5 DISCUSSION

5.1 INTERPRETING THE GLOBAL GAP

Pairwise overlap agreement should be treated as one component of consistency, not as a global certificate. The distinction is operational: an evaluator can report δ_{pair}, $B_{\text{pair}\to\text{world}}$, and $E_{\text{global readout}}$ in the same units. A zero first term states that all specified singleton overlaps agree. It does not state that the local laws belong to the marginal polytope. A positive second term identifies exactly the additional distance omitted by that statement. This separation avoids choosing an arbitrary threshold at which a pairwise score is declared globally safe.

Joint-feasibility diagnostics are also emerging for learned systems. Kotawala (2026) evaluates claims composed by several LLM agents, Andrews (2026) evaluates coherence axioms, and Choi et al. (2026) constructs a stationary law from model conditionals. Our decomposition identifies the distance omitted specifically by overlap correction, while complete-answer measurement and correctness connect that geometric question to VLM probabilities. These approaches support a common reporting principle: a semantic consistency score should name the feasible set that it certifies.

The visual results distinguish recognition from probabilistic compatibility. Idefics3 contributes no eligible cell in the controlled ontology, yet gives correct-answer positive cases on both visual datasets. Its image-control gain also confirms that the original image matters for the measured task. Coverage and correctness therefore describe the empirical setting; they do not replace the global feasibility calculation or determine its answer.

The complete-answer derivation also clarifies what probability is being tested. Prior work documents sensitivity to few-shot example order, surface-form competition, and semantic aggregation (Lu et al., 2022; Holtzman et al., 2021; Kuhn et al., 2023). Our event is deliberately narrower than free-form semantic probability: each specified state has one complete statement followed by the model's assistant terminator. Its effect ranges from negligible in the two Qwen checkpoints to decision-relevant in LLaVA-NeXT. The native termination event is therefore part of the measurement, not an interchangeable formatting choice. Extending the diagnostic to free-form answers would require a specified parser and a sum over every accepted surface form.

5.2 RELATION TO WORLD-MODEL EVALUATION

The finding identifies a relation among probabilistic readouts and leaves the model's latent representation unspecified. A joint generator already induces a joint distribution over its raw outputs, and an explicit shared-state architecture induces a joint law by construction. The present test becomes nontrivial after an external ontology specifies which events across outputs denote the same

world variable. This makes the diagnostic relevant to separately queried views, distributed agents, heterogeneous predictors, and semantic audits of shared-state models.

Structural inconsistency also does not guarantee a planning loss. World-model research correctly evaluates predictions through downstream control (Hafner et al., 2020; 2025). Separate finite-state planning controls show task-dependent prediction gains from global-consistency features: the disagreement reward benefits, while the primary semantic reward has no resolved gain. These controls evaluate joint-feasibility features rather than $B_{\mathrm{pair}\to\mathrm{world}}$ alone. A nearest ℓ_1 repair can also have several optima with different task values (Appendix G.2). Structural diagnosis and decision improvement therefore require separate evidence.

5.3 Implications for consistency evaluation

A multi-view consistency benchmark should state the event, the conditioning information, and the level of consistency being tested. Factual consistency tests compare related beliefs (Elazar et al., 2021; Kassner et al., 2021), confidence calibration relates probabilities to correctness (Kadavath et al., 2022), and semantic uncertainty groups surface forms that express the same answer (Kuhn et al., 2023). Controlled vision-language benchmarks test recognition, binding, and hallucination (Johnson et al., 2017; Thrush et al., 2022; Li et al., 2023). Each measurement is valuable, but none implies a common extension unless the shared variables, valid worlds, and joint feasibility condition are explicit.

The matched visual analysis adds a practical reporting requirement: report the answer interface alongside the gap. Prior work establishes sensitivity to few-shot example order and surface-form competition (Lu et al., 2022; Holtzman et al., 2021). Here, fixed answer-vocabulary and question/output-order changes alter the global component even when all tested answers remain correct. This supports a diagnostic of specified answer laws, while a representation-level claim would require additional invariance evidence. Selecting a repaired global law remains a separate task-dependent decision.

The readout effect reflects changed model probabilities, not a permutation of table entries. Relabeling variables or binary states in fixed tables preserves both feasible sets and the ℓ_1 norm, and therefore preserves $B_{\mathrm{pair}\to\mathrm{world}}$ (Appendix I.2). The measured changes arise when the model is queried again under a different readout. Our reversal changes question and answer order together, so the comparison measures their combined effect.

The context structure determines the required consistency test. Acyclic contexts with factorized support need only overlap checks; a cyclic family can require the global program. Appendix B.5 gives the precise support condition. This structural distinction tells an evaluator when local checks are sufficient and when they omit a necessary calculation.

5.4 Scope

The evidence covers five VLM checkpoints, a controlled ontology, and two benchmark-derived visual subsets. The latter extend correct-answer positive cases to a Llama decoder but do not establish four-readout stability. Their balanced question triples are diagnostic subsets, not official benchmark scores. The experiments use still images and one answer string per state; video transitions, unrestricted semantic answers, and downstream control require their own specified events.

6 Conclusion

Local answers can agree on every shared marginal yet admit no global law. Our finite diagnostic separates pairwise inconsistency, the omitted global component, and global-readout excess. Controlled evidence spans four qualified models, while independent visual scenes add correct-answer positive cases for Qwen and Llama decoders. The matched analysis shows why both accuracy and readout stability must accompany probability-level claims.

A shared-world claim should specify complete answer events and conditioning information, report correctness and coverage, and test global feasibility. The complete reproduction code is provided in the Supplementary Material.

AI USE STATEMENT

Generative AI tools supported mathematical formalization, experimental code checking, and manuscript polishing. The authors developed the theoretical framework and experimental design, interpreted the results, and verified the citations, equations, code, figures, insights, claims, and prose. The authors take responsibility for all content.

REFERENCES

Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13(11):113036, 2011. doi: 10.1088/1367-2630/13/11/113036.

Isaiah Andrews. Revealed rationality: Label-free evaluation and regularization from representation theorems, 2026. URL https://arxiv.org/abs/2608.05015.

Albert Atserias and Phokion G. Kolaitis. Consistency, acyclicity, and positive semirings, 2020. URL https://arxiv.org/abs/2009.09488.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Sanghyeok Choi, Henry Gouk, and Esmeralda S. Whitammer. Structured inference with large language gibbs, 2026. URL https://arxiv.org/abs/2606.19264.

Rina Dechter. From local to global consistency. *Artificial Intelligence*, 55(1):87–107, 1992. doi: 10.1016/0004-3702(92)90043-W.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Luigi Gresele, Julius von Kügelgen, Jonas M. Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7793–7824, 2022. URL https://proceedings.mlr.press/v162/gresele22a.html.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019. doi: 10.1109/CVPR.2019.00686. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Johnson_CLEVR_A_Diagnostic_CVPR_2017_paper.html.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.82. URL https://aclanthology.org/2022.emnlp-main.82/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.697. URL https://aclanthology.org/2021.emnlp-main.697/.

Anany Kotawala. Locally coherent, globally incoherent: Bounding compositional incoherence in multi-component LLM agents, 2026. URL https://arxiv.org/abs/2605.30335.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Aleksandra B. Slavkovic and Seth Sullivant. The space of compatible full conditionals is a unimodular toric variety, 2004. URL https://arxiv.org/abs/math/0405046.

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*, 2022. doi: 10.1109/CVPR52688.2022.00517.

N. N. Vorob'ev. Consistent families of measures and their extensions. *Theory of Probability and Its Applications*, 7(2):147–163, 1962. doi: 10.1137/1107014.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Tomáš Werner. Marginal consistency: Upper-bounding partition functions over commutative semirings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1455–1468, 2015. doi: 10.1109/TPAMI.2014.2363833.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, Xinjie Zhang, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, et al. Mage-VL: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026. URL https://arxiv.org/abs/2607.24904.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. UniSim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023. doi: 10.1109/CVPR52729.2023.00140.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, 2023. URL https://arxiv.org/abs/2311.16502.

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

A Appendix Roadmap

The appendix follows the main argument in verification order. Appendix B states the finite probability results and their proofs. Appendix C gives the linear programs. Appendix D specifies prompts, candidates, carriers, and selection criteria. Appendix E reports the complete eligible-cell results. Appendix F details complete-answer transport. Appendix G records model qualification and fixed planning controls. Appendix H closes the controlled study with reproduction and numerical checks. Appendix I extends the evaluation to independent visual images, including the amended sampling protocol, complete result matrix, matched readouts, an illustrative image, and raw-score coverage.

B Finite Probability Results

B.1 Why an ontology is necessary

Proposition 1. Suppose only local output spaces $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ and probability measures $\nu_1, \ldots, \nu_n$ are given. Without shared variables, valid-world constraints, or specified cross-output maps, a joint distribution always exists.

Proof. Define $\Omega = \prod_i \mathcal{Y}_i$ and $P = \bigotimes_i \nu_i$. The ith marginal of P is ν_i. Therefore the question becomes nontrivial only after a shared structure rules out this unconstrained product construction. If a generator already outputs a tuple $Y_{1:n} = G(c, \xi)$, then the pushforward $G(c, \cdot)_{\#} P_\xi$ is also a joint law over that raw tuple. □

This proposition fixes the interpretation of our test. We do not ask whether separately named outputs can be placed in an arbitrary joint space. We ask whether their specified semantic events are marginals of a probability law on Ω.

B.2 Nested feasible sets and the decomposition

Proposition 2. For the three pair laws in Section 3, $\mathcal{M} \subseteq \mathcal{O} \subseteq \mathcal{Q}$ and

$$0 \leq \delta_{\text{pair}}(q) \leq \delta_{\text{world}}(q) \leq D_G(q, p^G). \tag{7}$$

Consequently Equation (4) has three nonnegative terms.

Proof. A point in $\mathcal{M}$ is Ap for some $p \in \Delta_8$. Any two pair marginals derived from p have identical marginals on their shared singleton, so $Ap \in \mathcal{O}$. Thus $\mathcal{M} \subseteq \mathcal{O}$. Distance to a superset cannot exceed distance to its subset, which proves the first inequality. Since $Ap^G \in \mathcal{M}$, distance to $\mathcal{M}$ cannot exceed distance to Ap^G. Defining $B_{\text{pair}\to\text{world}} = \delta_{\text{world}} - \delta_{\text{pair}}$ and $E_{\text{global readout}} = D_G - \delta_{\text{world}}$ gives the stated decomposition. □

Two boundary cases matter. First, $B_{\text{pair}\to\text{world}} = 0$ does not imply $q \in \mathcal{M}$ because the two distances can be equal and positive. Second, $E_{\text{global readout}} = 0$ means that Ap^G is one nearest point in $\mathcal{M}$. The nearest point need not be unique.

B.3 Exact pairwise agreement without a world law

Let each pair distribution place probability $1/2$ on $(0, 1)$ and $1/2$ on $(1, 0)$. Every singleton marginal is Bernoulli$(1/2)$, so the three tables lie in $\mathcal{O}$ and $\delta_{\text{pair}} = 0$. A global assignment with $D \neq K$, $K \neq H$, and $D \neq H$ cannot exist for binary variables. Hence no global law can have these three pair marginals, which places the family outside $\mathcal{M}$ and gives $\delta_{\text{world}} > 0$. This is the example shown in Figure 1.

B.4 Complete-answer transport

Proposition 3. Let v_s be the specified candidate for state s and e the native assistant terminator. The normalized complete-answer law is Equation (6). Moreover, $p^{\text{complete}} = p^{\text{pre}}$ if and only if $\tau(s, c)$ is constant over candidates with positive prefix probability.

Proof. The autoregressive chain rule gives $K(v_s, e \mid c) = K(v_s \mid c) q(e \mid c, v_s)$. Divide each side by the sum over specified candidates to obtain Equation (6). If τ is constant, it cancels during

normalization. Conversely, equality of the two normalized laws implies the same multiplicative factor for every state with positive prefix mass. □

The corresponding pairwise log-odds identity is

$$\log \frac{p_s^{\text{complete}}}{p_r^{\text{complete}}} = \log \frac{p_s^{\text{pre}}}{p_r^{\text{pre}}} + \log \frac{\tau_s}{\tau_r}. \tag{8}$$

No length penalty or calibration assumption enters this identity.

B.5 When overlap agreement suffices

Acyclic contexts admit a common extension when the support has no additional cross-context constraints. Specifically, let the contexts form a tree such that all contexts containing any variable form a connected subtree. This is the running-intersection property. Write $S_C = \pi_C(\Omega)$ and assume

$$\Omega = \{x \in \textstyle\prod_v \mathcal{X}_v : x_C \in S_C \text{ for every } C\}. \tag{9}$$

This is the required support factorization. For normalized laws supported on S_C and agreeing on overlaps, start at a root context and attach each child using its conditional law given its parent overlap. On zero-mass overlap events, choose the conditional arbitrarily because those events have zero probability in the extension. The running-intersection property preserves all previous marginals; support factorization keeps the resulting assignments in Ω. This is the standard tree extension argument (Wainwright & Jordan, 2008; Atserias & Kolaitis, 2020).

The support condition is essential even for an acyclic family. For example, two disjoint binary singleton contexts are acyclic, but requiring their values to be equal adds a constraint invisible to overlap checks. Deterministic singleton laws at zero and one then have zero overlap distance and no extension on that restricted support. The full binary cube in our controlled study avoids this issue. Its failure of overlap sufficiency comes from the cycle of three pair contexts.

C Linear Programs

C.1 Distance to the overlap polytope

Write $q \in \mathbb{R}^{12}$ for the stacked pair tables. The overlap polytope is specified by $v \geq 0$, one normalization equality per pair table, and three singleton-overlap equalities. Let these equalities be $Cv = d$. Introducing $t \geq 0$, the first distance is

$$\begin{aligned} \min_{v,t} \quad & \mathbf{1}^\top t \\ \text{s.t.} \quad & Cv = d, \quad v \geq 0, \\ & -t \leq v - q \leq t. \end{aligned} \tag{10}$$

The equality rows are fixed by the contexts and contain no learned parameter.

C.2 Distance to the marginal polytope

The global distance can be solved directly over $p \in \Delta_8$:

$$\begin{aligned} \min_{p,t} \quad & \mathbf{1}^\top t \\ \text{s.t.} \quad & p \geq 0, \quad \mathbf{1}^\top p = 1, \\ & -t \leq Ap - q \leq t, \quad t \geq 0. \end{aligned} \tag{11}$$

The programs use all four cells of every pair table. The reported ℓ_1 distance is therefore the sum over 12 probabilities, without division by two. The maximum possible distance between two stacked pair families is six. Both programs use primal and dual feasibility tolerances of 10^{-9}. The separate reporting criterion $B_{\text{pair}\to\text{world}} > 10^{-8}$ determines which gaps count as numerically resolved.

C.3 Complexity boundary

The eight-state program is exact and small. In a general ontology, $|\Omega|$ can grow exponentially with the number of variables or trajectory horizon. Acyclic context structures can admit direct extension, and low-treewidth systems can exploit factorization. Larger cyclic systems require relaxations, cut generation, or a separation oracle. Scaling the exact program to unrestricted video or long-horizon trajectories remains open.

D Experimental Protocol

D.1 Ontology and contexts

Table 3 gives the complete ontology. All eight combinations are valid worlds. The local contexts form a three-cycle because every variable appears in two pair reports.

Table 3: Finite ontology used in the controlled experiment.

Variable	State 0	State 1
D, door	CLOSED	OPEN
K, key	PADLOCKED	KEY
H, hazard	SAFE	WARNING

For a pair context, a candidate has the form "D is CLOSED; K is PADLOCKED." with the requested variables and values substituted. A global candidate has the form "D is CLOSED; K is PADLOCKED; H is SAFE." The world-first prompt adds "WORLD:" to the start of every candidate. Candidate families contain exactly four and eight statements, respectively. Each candidate is immediately followed by the model's native assistant-turn terminator.

D.2 Prompt specification

Every prompt first states the evidence source, then defines the ontology, requests one ordered context, prints the fixed output grammar, and asks for one statement with no other text. Authoritative-text evidence names all three states directly. Image evidence states that the three panels are views of the same world. The local and global prompts differ only in the requested context and its grammar. Direct and world-first prompts are never averaged because they define different autoregressive kernels.

D.3 Stimulus carriers

Table 4: Stimulus construction. The icon and photograph carriers use four visual variants for each world.

Carrier	Worlds	Stimuli	Evidence form
Authoritative text	8	8	State words in prompt
Raster	8	8	State words rendered in panels
Icon	8	32	Drawn state icons, four styles
Photograph	8	32	Licensed image composites
Total	8	80	

The photograph carrier uses one source image for each of the six atomic states. Five sources use the Creative Commons Attribution-ShareAlike license, and one uses Creative Commons Zero. Source links and licenses are included in the experiment artifact. The visual variants alter layout and presentation but not the specified world state.

D.4 Selection criteria and fixed decisions

The protocol fixes the following order before inspecting $B_{\text{pair}\to\text{world}}$.

1. Measure exact prefix and complete-answer probability for every specified candidate.
2. Mark a condition competent only when truth is the unique maximum-probability state in C_{12}, C_{23}, C_{13}, and C_{123}.
3. Mark a carrier-prompt cell eligible only when competent conditions cover all eight worlds.
4. Solve both distance programs for every condition and report competent conditions from eligible cells.
5. Call a gap numerically resolved when $B_{\text{pair}\to\text{world}} > 10^{-8}$.

Strict model-level replication requires at least one eligible cell and a resolved positive gap under both prompt forms. The panel also reports a descriptive qualified-model result: whether any eligible condition is positive. No carrier, prompt, candidate string, or threshold is selected from the gap values.

E Full Results

E.1 Competence and coverage

Table 5: Complete-answer competence by model, carrier, and prompt. Eligible cells cover all eight worlds. D is direct and W is world-first.

Model	Carrier	Competent conditions		Worlds covered		Eligible
		D	W	D	W	D/W
Qwen3-VL-4B	Text	8/8	8/8	8	8	yes/yes
	Raster	8/8	8/8	8	8	yes/yes
	Photo	27/32	29/32	8	8	yes/yes
	Icon	0/32	1/32	0	1	no/no
Qwen2.5-VL-7B	Text	8/8	8/8	8	8	yes/yes
	Raster	8/8	8/8	8	8	yes/yes
	Photo	16/32	16/32	4	4	no/no
	Icon	4/32	4/32	1	1	no/no
Mage-VL-4B	Text	8/8	8/8	8	8	yes/yes
	Raster	8/8	8/8	8	8	yes/yes
	Photo	27/32	24/32	8	7	yes/no
	Icon	6/32	5/32	3	2	no/no
Idefics3-8B	Text	5/8	6/8	5	6	no/no
	Raster	3/8	3/8	3	3	no/no
	Photo	7/32	6/32	3	2	no/no
	Icon	4/32	4/32	1	1	no/no
LLaVA-NeXT-7B	Text	8/8	4/8	8	4	yes/no
	Raster	2/8	2/8	2	2	no/no
	Photo	0/32	0/32	0	0	no/no
	Icon	0/32	0/32	0	0	no/no

E.2 Pair-to-world gap by eligible cell

Strict two-prompt replication holds for Qwen3-VL, Qwen2.5-VL, and Mage-VL. Authoritative text supplies qualified evidence for all three models, and raster supplies a second carrier. LLaVA-NeXT adds qualified direct-text evidence from a Mistral decoder. Eligible photograph cells broaden coverage for Qwen3-VL and Mage-VL, although only the Qwen3-VL direct cell contains a resolved positive.

Table 6: Pair-to-world gaps in eligible cells. Counts are resolved positives over scene-competent conditions.

Model	Carrier	Prompt	Positive / competent	Maximum $B_{\text{pair}\to\text{world}}$
Qwen3-VL-4B	Text	Direct	$2/8$	2.30×10^{-5}
Qwen3-VL-4B	Text	World-first	$2/8$	1.79×10^{-7}
Qwen3-VL-4B	Raster	Direct	$3/8$	1.00×10^{-5}
Qwen3-VL-4B	Raster	World-first	$0/8$	6.76×10^{-9}
Qwen3-VL-4B	Photo	Direct	$4/27$	1.05×10^{-4}
Qwen3-VL-4B	Photo	World-first	$0/29$	4.30×10^{-16}
Qwen2.5-VL-7B	Text	Direct	$4/8$	9.92×10^{-4}
Qwen2.5-VL-7B	Text	World-first	$6/8$	8.81×10^{-4}
Qwen2.5-VL-7B	Raster	Direct	$3/8$	6.35×10^{-3}
Qwen2.5-VL-7B	Raster	World-first	$1/8$	2.32×10^{-3}
Mage-VL-4B	Text	Direct	$3/8$	1.16×10^{-2}
Mage-VL-4B	Text	World-first	$2/8$	5.58×10^{-5}
Mage-VL-4B	Raster	Direct	$3/8$	3.25×10^{-1}
Mage-VL-4B	Raster	World-first	$1/8$	4.13×10^{-2}
Mage-VL-4B	Photo	Direct	$0/27$	3.33×10^{-16}
LLaVA-NeXT-7B	Text	Direct	$1/8$	7.43×10^{-3}

E.3 Scale of the resolved gaps

The resolved global component spans several orders of magnitude and is not concentrated in one model. Table 7 summarizes every positive value in eligible, scene-competent conditions. Mage-VL contributes nine positives and the largest absolute gap. LLaVA-NeXT has one positive condition, so its median equals that observation. All four qualified models contribute to the 35-condition distribution.

Table 7: Scale of positive pair-to-world gaps in eligible, scene-competent conditions. Here $\rho = B_{\text{pair}\to\text{world}}/\delta_{\text{world}}$. All rows use the fixed 10^{-8} numerical tolerance.

Model	n	min $B_{\text{pair}\to\text{world}}$	med $B_{\text{pair}\to\text{world}}$	max $B_{\text{pair}\to\text{world}}$	med ρ	max ρ
Qwen3-VL-4B	11	2.62×10^{-8}	5.33×10^{-6}	1.05×10^{-4}	0.32%	19.0%
Qwen2.5-VL-7B	14	1.99×10^{-6}	1.35×10^{-4}	6.35×10^{-3}	2.94%	7.2%
Mage-VL-4B	9	1.45×10^{-5}	8.45×10^{-3}	3.25×10^{-1}	4.44%	58.2%
LLaVA-NeXT-7B	1	7.43×10^{-3}	7.43×10^{-3}	7.43×10^{-3}	23.42%	23.4%
All qualified models	35	2.62×10^{-8}	8.56×10^{-5}	3.25×10^{-1}	1.95%	58.2%

E.4 All three decomposition terms

Table 8: Decomposition occupancy in competent conditions from eligible cells. A count reports values above 10^{-8}.

Model	Conditions	δ_{pair}	$B_{\text{pair}\to\text{world}}$	$E_{\text{global readout}}$
Qwen3-VL-4B	88	88	11	86
Qwen2.5-VL-7B	32	32	14	32
Mage-VL-4B	59	59	9	59
Idefics3-8B	0	0	0	0
LLaVA-NeXT-7B	8	8	1	8

Positive overlap distance shows that the learned local laws often disagree even at the singleton-overlap level. The pair-to-world gap remains informative because it measures the increase in optimal correction cost when a joint law is required. Global-readout excess is common and often larger,

showing that the explicit global readout is not generally an ℓ_1-nearest coherent point for the local laws.

F TERMINATION TRANSPORT

Table 9: Prefix-to-complete transport by model. Gap changes report all conditions and eligible conditions, respectively.

Model	Max ℓ_1	Maximizer	Competence	Gap all/eligible
Qwen3-VL-4B	1.29×10^{-6}	$0/640$	$0/160$	$0/0$
Qwen2.5-VL-7B	3.13×10^{-6}	$0/640$	$0/160$	$0/0$
Mage-VL-4B	5.83×10^{-3}	$0/640$	$0/160$	$0/0$
Idefics3-8B	8.06×10^{-6}	$0/640$	$0/160$	$0/0$
LLaVA-NeXT-7B	4.51×10^{-1}	$85/640$	$7/160$	$6/0$

The transport is exactly paired because prefix and termination log probabilities are reconstructed from the same forward token terms. The maximum change in $B_{\mathrm{pair}\to\mathrm{world}}$ between prefix and complete laws is 5.39×10^{-2} and occurs for LLaVA-NeXT. The six gap-classification changes occur outside eligible cells. The complete event therefore changes intermediate decisions while preserving the qualified-model result.

G BOUNDARY EXPERIMENTS

G.1 MODEL QUALIFICATION OUTSIDE STRICT REPLICATION

The complete model panel separates qualified evidence from unreadable cells. Idefics3 with a Llama 3.1 decoder is competent on 38 of 160 conditions, but its best cell covers six of eight worlds. It therefore supplies no eligible global test. LLaVA-NeXT with a Mistral decoder is competent on all eight direct-text conditions and exhibits a resolved positive gap, while its world-first text cell covers four worlds. This establishes a qualified direct-prompt extension without satisfying the stricter two-prompt replication rule. Reporting both outcomes preserves the distinction between absence of coverage and absence of the effect.

G.2 FIXED PLANNING CONTROLS

A separate synthetic study tests how consistency relates to decisions. It fits multinomial logistic predictors to an eight-state transition distribution over eight initial states and four actions. The joint predictor uses 11 features; each restricted local predictor uses seven, retaining only the state variables in its pair context. The primary setting uses interaction strength $\kappa = 4$, 1,024 samples per state-action condition, and 512 independent training replicates. The parameter κ penalizes the two all-equal next states in the true transition logits. Global-atom pseudocounts of $1/2$ and local-cell pseudocounts of one give matching smoothed training marginals. The artifact also reports $\kappa \in \{1.5, 4, 6\}$ and sample sizes $16, 64, 256, 1024$.

Two fixed rewards are linear in the three pair laws. The semantic reward is $4P(D = 1, K = 1) + 2P(K = 1, H = 0) - 6P(D = 1, H = 1)$; the disagreement reward sums the three pairwise inequality probabilities. Both subtract fixed action costs computed under the reference transition distribution. Each planner selects the action with highest predicted utility, and regret is evaluated under the known true transition law, then averaged over the eight initial states. In the primary semantic task, mean regret is 0.00949 for the restricted local predictor and 0.00531 for the joint predictor. On the disagreement task, their regrets are 0.000133 and 0.000375, respectively. Thus a joint predictor does not improve both tasks uniformly. These are newly fitted finite-state learners, separate from the VLM measurements.

The prediction test evaluates additional global-consistency features, not $B_{\mathrm{pair}\to\mathrm{world}}$ alone. Baseline features are pairwise probability error, excess cross-entropy, and maximum singleton disagreement

across actions. Four added features describe the world distance and the binary joint-feasibility violation after matching singleton marginals, each at the selected action and at its maximum across actions. The latter violation is the positive part of the lower minus upper feasible bound for $P(D = K = H = 1)$ given the pair tables. Logistic regression predicts positive regret using 409 training replicates and 103 held-out replicates; all eight states of a replicate remain in the same split. A positive log-loss reduction favors the augmented predictor. It is 0.00398 with a 95% interval $[-0.01972, 0.02768]$ for the semantic reward, and 0.04789 with interval $[0.02822, 0.06755]$ for the disagreement reward. The intervals use replicate-level Student t estimates. The semantic task therefore has no resolved incremental gain, whereas the disagreement task does.

Repair gains also depend on the reward. We compare four transformations of the restricted local laws: match singleton marginals while preserving pairwise odds ratios; then mix those tables toward uniform until jointly feasible; project the raw tables onto $\mathcal{M}$ in ℓ_1; or project the singleton-matched tables onto $\mathcal{M}$. Table 10 reports each method, including the two whose semantic-task outcomes coincide. Gain is raw-planner regret minus transformed-planner regret, averaged over the 512 replicates. All four gains are positive for the semantic task and negative for the disagreement task.

Table 10: Mean regret reduction in the primary finite-state setting. Positive values favor the transformation; brackets give replicate-level 95% Student t intervals.

Transformation	Semantic reward	Disagreement reward
Match singleton marginals	0.004527	-0.000333
	$[0.003420, 0.005634]$	$[-0.000342, -0.000324]$
Then mix toward uniform	0.004271	-0.000275
	$[0.003192, 0.005350]$	$[-0.000284, -0.000266]$
Project raw tables	0.001736	-0.000162
	$[0.000580, 0.002892]$	$[-0.000171, -0.000154]$
Project matched tables	0.004527	-0.000272
	$[0.003420, 0.005634]$	$[-0.000281, -0.000263]$

Nearest ℓ_1 projection need not select a unique decision. We optimize each action's utility over the full set of nearest distributions, rather than using only the solver-returned optimum. In the primary semantic setting, the mean utility range has width 0.117. Among 512 replicates, 308 (60.16%) allow both positive and negative regret gains, and 132 (25.78%) allow a negative gain or no change. Together, these 440 replicates (85.94%) lack a fixed gain sign. A task-independent rule for choosing among optima is therefore needed to define a single repair method.

H Reconstruction and Reproducibility Audit

H.1 Raw-to-result reconstruction

For each candidate token position, the artifact stores the selected-token logit a_j and the full-vocabulary log normalizer z_j. For the controlled study, the token-level audit reconstructs

$$\ell_s^{\mathrm{pre}} = \sum_{j=1}^{L_s} (a_j - z_j), \qquad \ell_s^{\mathrm{end}} = a_{L_s+1} - z_{L_s+1}, \qquad \ell_s^{\mathrm{complete}} = \ell_s^{\mathrm{pre}} + \ell_s^{\mathrm{end}}. \tag{12}$$

It then normalizes the log scores and verifies Equation (6). A separate implementation solves both linear programs for all 800 conditions across the five-model panel from the packaged laws. Neither audit imports the experiment's distance functions.

The largest numerical ordering violation between δ_{pair} and δ_{world} is 5.44×10^{-10}. One eligible direct-text condition for Qwen3-VL lies close enough to the fixed 10^{-8} criterion to change classification across solver formulations: its reported gap is 9.50×10^{-9}, compared with 1.04×10^{-8} in an independent solution. We retain the reported values throughout and do not count this condition among the 35 positives. All reported positives remain positive in the comparison, and every qualified-model conclusion is unchanged. The visual-test classifications agree for all 6,144 conditions.

Table 11: Maximum residuals from the independent reconstruction.

Check	Maximum residual
Selected-token sequence reconstruction, all models	0
Complete-answer transport identity, all models	3.33×10^{-16}
Reported controlled complete law	0
Controlled pair-to-world gap	1.21×10^{-9}
Visual-test pair-to-world gap	1.39×10^{-15}
Three-term decomposition	8.88×10^{-16}

H.2 Artifact contents

Inference uses unquantized weights in the 16-bit brain floating-point format (bfloat16), single-precision token logits and log normalizers, and double-precision score accumulation and distance optimization. The native assistant terminators are <|im_end|> (token 151645) for both Qwen models and Mage-VL, <end_of_utterance> (128258) for Idefics3, and </s> (2) for LLaVA-NeXT. Model revisions, candidate token sequences, and software versions are retained with the probability records.

The generated artifacts contain the specified law list, prompt and candidate records, stimulus registry, controlled-study images and licenses, source-image identifiers for the visual extension, selected-token logits, log normalizers, normalized laws, condition metrics, termination transport rows, and model-carrier-prompt summaries. The one-cell reproduction code downloads the source data and model weights, runs inference, and computes the reported tables and numerical checks. The main scientific result requires no sampled generation, semantic judge, or manual labeling after the ontology is fixed.

Table 12 summarizes the evidence for each claim.

Table 12: Evidence boundary for the submission.

Statement	Evidence status
Pairwise overlap agreement is not a general certificate of global realizability	Exact finite counterexample and Proposition 2
Every model with an eligible cell exhibits a positive pair-to-world gap	Four qualified models and 16,000 complete candidate sequences
The effect replicates under both prompts	Supported for Qwen3-VL, Qwen2.5-VL, and Mage-VL
The finding extends across decoder families	Controlled Mistral evidence; correct-answer visual cases for Idefics3-Llama
Positive cases persist under all four visual readouts	Not supported in the matched test
Prefix truncation creates the qualified-model finding	Not supported; eligible gap classifications are unchanged
$B_{\mathrm{pair}\to\mathrm{world}}$ alone improves planning	Not directly tested; auxiliary consistency features have task-dependent predictive value
The result covers video or free-form semantic answers	Not evaluated

I Independent Visual Images and Readout Sensitivity

The visual extension asks whether the diagnostic remains informative when the image and questions come from established visual reasoning datasets. It complements the controlled study, whose photograph carrier recombines six atomic source images. The controlled positive conditions include 20 on text, 11 on raster text, and four photograph variants of one Qwen3-VL scene. The extension instead uses different original images and different question triples, while preserving complete-answer

measurement and the three pair contexts. Its two questions are whether correct visual answers can retain a positive gap and whether that gap persists under fixed readout changes.

I.1 Data selection and the development-informed protocol

The sample contains original binary questions from GQA balanced testdev and CLEVR validation (Hudson & Manning, 2019; Johnson et al., 2017). We use the GQA question and image mirror maintained by `lmms-lab/GQA` and shard 000 of the CLEVR JPEG mirror `laion/clevr-webdataset`. Source image IDs, original question IDs, reference answers, question programs, and fixed dataset revisions are retained in the artifact. The CLEVR mirror stores an image repeatedly for different questions. We map each question index back to the original image and select each source image once.

Selection balances reference answers without using model predictions. For each image, questions with identical semantic programs are deduplicated and the remaining questions are placed in seeded random order. A bipartite matching assigns distinct images to the eight possible three-answer patterns, requiring enough yes and no questions for the assigned pattern. The first compatible questions in the shuffled lists form the triple. Each dataset supplies eight development images and 32 test images per pattern, giving 64 development and 256 test images per dataset. There is no image reuse across those splits. The test set is held out from this development analysis; the protocol does not establish that its images were absent from model pretraining.

Development results changed the execution rule before test inference. The initial rule required each single-question accuracy to have a Wilson 95% lower bound above $1/2$ and all eight answer patterns to contain a scene with correct pair and global answers. None of the six model-dataset groups passed, so no test inference ran under that rule. We retained that outcome and amended the protocol to measure every previously fixed test group, with correctness and pattern coverage reported separately. No test image, question, model, readout, or gap tolerance was changed or selected from test outputs.

The amendment separates empirical coverage from a pointwise feasibility test. The controlled ontology uses the same three variables in every stimulus; natural images use different questions. Missing a competent answer pattern in other images does not invalidate the distance calculation for the current image. We therefore report all test images first, followed by correctness-stratified descriptions. The original development qualification remains unsuccessful; the extension is not a retrospective pass of that rule.

I.2 Shared condition, answers, and support

Every local and global query sees the same full image and all three questions, denoted A, B, and C. Only the requested answer subset changes. The pair contexts are AB, BC, and AC, and the global context is ABC. Forward order lists the questions as A, B, C and reports each subset in that order. Reverse order lists them as C, B, A and reverses the reported subset. This factor changes question order and answer order together, so its effect cannot be assigned to either component alone.

Two fixed answer vocabularies yield four readout variants. The yes/no grammar uses strings such as `A=no; B=yes.`; the Boolean grammar uses `A=false; B=true.` and explicitly states that true means yes. The processor's native assistant-turn template supplies the continuation boundary. Each state is scored as the entire candidate followed by the native terminator, and its law is normalized over the four pair or eight global candidates. We do not average the four variants into one law. A variant defines one measured answer interface.

Logical relations among questions are handled conservatively. Let $W \subseteq \{0,1\}^3$ denote the unknown exact semantic support. The measured cube uses all eight assignments, so

$$\mathcal{M}(W) \subseteq \mathcal{M}(\{0,1\}^3) \subseteq \mathcal{O}, \qquad d_1(q, \mathcal{M}(W)) - d_1(q, \mathcal{O}) \geq B_{\text{pair}\to\text{world}}(q). \tag{13}$$

A positive gap on the outer support remains a lower bound on the additional distance for a stricter semantic support. A zero gap does not certify those unmodeled relations. Reference answers determine correctness, not a singleton support containing only the observed truth.

Relabeling the measured tables leaves the gap unchanged. A permutation of the three variables or a binary-label exchange induces permutations R of pair-table entries and S of world columns, with

$RA = AS$. The context family contains all three pairs, so $R\mathcal{O} = \mathcal{O}$ and $R\mathcal{M} = \mathcal{M}$. Because the ℓ_1 norm is permutation invariant,

$$B_{\text{pair}\to\text{world}}(Rq) = B_{\text{pair}\to\text{world}}(q). \tag{14}$$

This property concerns relabeling a fixed probability table. It does not require a language model to produce that same table after the prompt changes. The four measured readouts test the latter behavior empirically.

I.3 Models and workload

The visual study reuses three checkpoints from the controlled panel: `Qwen/Qwen3-VL-4B-Instruct`, `HuggingFaceM4/Idefics3-8B-Llama3`, and `llava-hf/llava-v1.6-mistral-7b-hf`. They represent Qwen, Llama, and Mistral decoders, respectively. Model-native processors and termination tokens are retained. Candidate scores use teacher forcing, unquantized model weights, single-precision logits, and double-precision accumulation and distance optimization.

Each test image supplies 16 joint laws and six single-question laws. The four joint laws per variant require $3 \times 4 + 8 = 20$ candidate sequences. Three baseline single-question queries and their three shuffled-image controls add 12 sequences, giving $4 \times 20 + 12 = 92$ per image and model. Thus each model has 47,104 test candidates, and the full extension has 141,312 candidates, 33,792 probability laws, and 6,144 joint conditions. The 12,288 development candidates are retained separately and never enter the test counts.

I.4 Correctness and image-level evidence

The strict correctness condition includes both single and joint questions. For a given image and readout, all three baseline single-question maximizers and the three pair and one global maximizers must uniquely equal the reference answers. The baseline single-question format is always yes/no in forward order; single-question answers under other formats were not measured. The weaker joint-correct condition requires only the three pairs and the global answer. Ties do not count as correct.

Table 13 reports the image-level evidence without pooling four measurements of an image as independent samples. Thirty distinct test images produce a strict-correct positive condition: 16 for Qwen3-VL and 14 for Idefics3, with no overlap between their positive-image sets. Both models have examples in both datasets. LLaVA-NeXT contributes no strict-correct positive case. The counts establish existence in these subsets, not high prevalence across arbitrary images or tasks.

Table 13: Visual positive cases and same-image persistence. The first count is the number of images with a strict-correct positive gap in at least one variant, out of all 256 images. The second is positive in all four variants, out of images that satisfy strict correctness in all four. A dash denotes an empty correctness subset. Maxima range over strict-correct conditions.

Model	Dataset	Any variant	All four variants	Max $B_{\text{pair}\to\text{world}}$
Qwen3-VL	GQA	11/256	0/61	0.0034
Qwen3-VL	CLEVR	5/256	0/31	0.0070
Idefics3	GQA	11/256	0/12	0.0930
Idefics3	CLEVR	3/256	0/12	0.1496
LLaVA-NeXT	GQA	0/256	0/3	$\leq 10^{-8}$
LLaVA-NeXT	CLEVR	0/256	–	$\leq 10^{-8}$

I.5 Complete readout matrix

Table 14 retains every model, dataset, and readout, including all-zero groups. Across the four variants there are 877 strict-correct conditions and 32 positive gaps. The corresponding 30 images are fewer because two images are positive under two variants. Conditional fractions describe the selected correctness subset; they are not estimates of unconditional image-level prevalence. The all-image column supplies the unfiltered denominator for each variant.

Table 14: Complete visual-test matrix. Entries are positive gaps divided by the specified condition count. All includes every image; Joint requires the four joint-context answers to be correct; Strict additionally requires the three baseline single-question answers to be correct. A $0/0$ entry denotes no qualifying condition, not a zero estimated rate. Values no greater than 10^{-8} are unresolved under the fixed numerical criterion.

Model	Dataset	Answers	Order	All	Joint	Strict	Max strict $B_{\mathrm{pair}\to\mathrm{world}}$
Qwen3-VL	GQA	yes/no	Forward	36/256	10/105	6/91	0.0030
Qwen3-VL	GQA	yes/no	Reverse	30/256	5/91	4/71	0.0034
Qwen3-VL	GQA	true/false	Forward	9/256	2/114	0/86	$\leq 10^{-8}$
Qwen3-VL	GQA	true/false	Reverse	6/256	2/105	1/72	2.89×10^{-5}
Qwen3-VL	CLEVR	yes/no	Forward	16/256	7/79	3/57	3.00×10^{-4}
Qwen3-VL	CLEVR	yes/no	Reverse	14/256	1/68	1/45	0.0013
Qwen3-VL	CLEVR	true/false	Forward	23/256	6/101	2/49	0.0070
Qwen3-VL	CLEVR	true/false	Reverse	5/256	1/88	0/43	$\leq 10^{-8}$
Idefics3	GQA	yes/no	Forward	22/256	2/48	2/40	0.0101
Idefics3	GQA	yes/no	Reverse	20/256	1/38	0/34	$\leq 10^{-8}$
Idefics3	GQA	true/false	Forward	22/256	4/64	4/48	0.0930
Idefics3	GQA	true/false	Reverse	43/256	5/49	5/35	0.0027
Idefics3	CLEVR	yes/no	Forward	8/256	0/62	0/51	$\leq 10^{-8}$
Idefics3	CLEVR	yes/no	Reverse	27/256	1/38	1/32	0.1496
Idefics3	CLEVR	true/false	Forward	15/256	1/50	1/41	0.0161
Idefics3	CLEVR	true/false	Reverse	64/256	2/29	2/24	0.1202
LLaVA-NeXT	GQA	yes/no	Forward	2/256	1/59	0/13	$\leq 10^{-8}$
LLaVA-NeXT	GQA	yes/no	Reverse	6/256	0/63	0/17	$\leq 10^{-8}$
LLaVA-NeXT	GQA	true/false	Forward	2/256	1/64	0/9	$\leq 10^{-8}$
LLaVA-NeXT	GQA	true/false	Reverse	0/256	0/63	0/10	$\leq 10^{-8}$
LLaVA-NeXT	CLEVR	yes/no	Forward	1/256	1/9	0/5	$\leq 10^{-8}$
LLaVA-NeXT	CLEVR	yes/no	Reverse	0/256	0/5	0/4	$\leq 10^{-8}$
LLaVA-NeXT	CLEVR	true/false	Forward	5/256	0/6	0/0	–
LLaVA-NeXT	CLEVR	true/false	Reverse	3/256	0/1	0/0	–

The positive component can be appreciable even though its frequency is low. The largest strict-correct gap is 0.1496, and the largest fraction $B_{\text{pair}\to\text{world}}/\delta_{\text{world}}$ among strict-correct positives is 27.4%. These are descriptive maxima over all fixed variants, not selected operating points. Most conditions have no resolved additional component, and a zero $B_{\text{pair}\to\text{world}}$ can coexist with positive overlap and world distances.

I.6 Does the model use the image?

A fixed shuffled-image control measures dependence on the original visual input. Within each dataset and split, images are placed in a seeded random cycle and replaced by their successor. The three questions and reference answers remain unchanged. Each image is used once as a replacement, and no image is paired with itself. Only single-question laws use this control; shuffled and original-image probabilities are never combined in the same joint-feasibility calculation.

Table 15 shows a positive original-image gain for Qwen3-VL and Idefics3 on both datasets and for LLaVA-NeXT on GQA. LLaVA-NeXT is near chance on CLEVR and shows no resolved gain. These controls support visual dependence at the model-dataset level. They do not establish that every individual positive case is caused by image content.

Table 15: Original versus shuffled images on the three baseline questions. Accuracy is a percentage; gain and its interval are percentage points. Intervals are descriptive image-level bootstrap intervals from 2,000 resamples of the per-image mean difference across three questions.

Model	Dataset	Original	Shuffled	Gain	95% interval
Qwen3-VL	GQA	75.5	50.8	24.7	[21.1, 28.5]
Qwen3-VL	CLEVR	64.3	50.3	14.1	[10.9, 17.1]
Idefics3	GQA	67.6	50.3	17.3	[13.2, 21.4]
Idefics3	CLEVR	73.7	50.9	22.8	[18.8, 27.1]
LLaVA-NeXT	GQA	58.9	50.4	8.5	[6.2, 10.7]
LLaVA-NeXT	CLEVR	50.1	50.0	0.1	[0.0, 0.4]

The intervals describe this fixed, truth-stratified diagnostic sample. Images are grouped across their three questions and repeated readouts, rather than counting answer sequences as independent observations. The shared image derangement can induce dependence between differences, and the intervals do not account for that dependence or model pretraining overlap. We use them as descriptive evidence of image reliance, not as population-level significance claims or official GQA/CLEVR benchmark scores.

I.7 Matched readouts with correctness held fixed

The strongest stability comparison keeps both the image and answer correctness fixed. Table 16 includes only images whose baseline single-question answers and all four joint readouts are correct. Among 119 model-image cases, eleven have at least one positive readout: nine have exactly one and two have exactly two. None has three or four. Thus the absence of four-way persistence is not explained entirely by a readout becoming incorrect.

Table 16: Number of positive readouts on images with all four joint readouts and baseline single-question answers correct. Columns 0 to 4 count images with that many positive gaps among the four variants.

Model	Dataset	Matched images	0	1	2	3	4
Qwen3-VL	GQA	61	54	7	0	0	0
Qwen3-VL	CLEVR	31	29	1	1	0	0
Idefics3	GQA	12	11	1	0	0	0
Idefics3	CLEVR	12	11	0	1	0	0
LLaVA-NeXT	GQA	3	3	0	0	0	0
LLaVA-NeXT	CLEVR	0	0	0	0	0	0

Some two-way comparisons do retain a positive gap. One Qwen3-VL CLEVR image is positive under both yes/no orders among 41 images correct in both. One Idefics3 CLEVR image is positive under both reverse-order vocabularies among 18 images correct in both. All other matched variant pairs have no shared strict-correct positive. These are descriptive comparisons of the same preselected variants, not new replication criteria. The test's declared four-variant support condition remains unmet.

I.8 An image with unchanged answers and changing gaps

Figure 5 shows the Idefics3 CLEVR image with the largest gap among that model's images correct under all four readouts. It is selected after measurement to illustrate the observed pattern, not to estimate its frequency. The original image is `CLEVR_val_000053`; the question IDs are 538, 535, and 534. Its three reference answers are yes, no, and no.

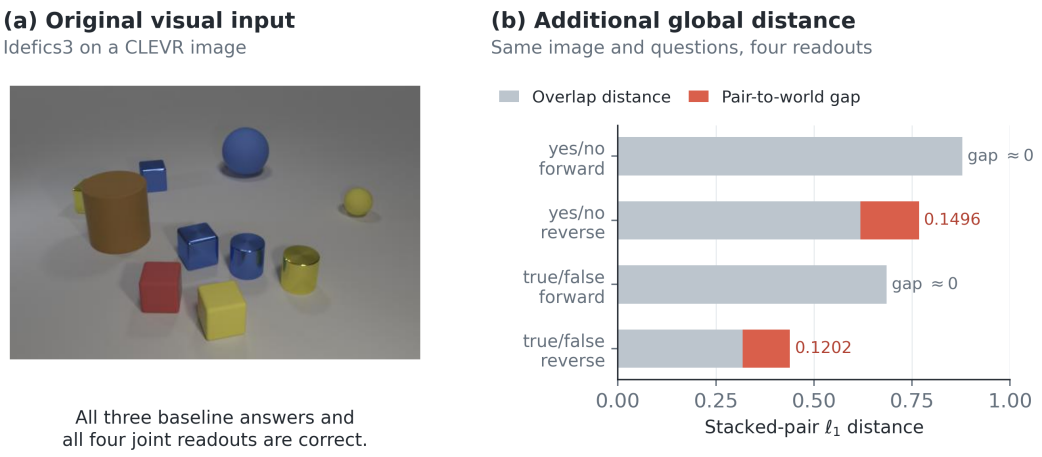


Figure 5: Correct answers persist while the additional global component changes. Idefics3 answers the three baseline questions and all four joint readouts correctly for this CLEVR image. Gray measures overlap distance and coral the additional pair-to-world gap. Reverse order produces a resolved extra component under both vocabularies; forward order does not. Gray remains positive, so an unresolved extra component does not mean the local laws are globally compatible. Image from CLEVR (Johnson et al., 2017).

The original questions make the shared condition concrete. Question A asks, "Is there anything else that is made of the same material as the red cube?" Question B asks, "Is the number of big brown objects that are on the right side of the blue cylinder less than the number of tiny red matte blocks that are behind the big rubber cylinder?" Question C asks, "Is there a big cyan object?" All questions appear in every query; only the requested subset and the fixed readout variant differ.

Table 17 separates the effect from overlap mismatch. In forward order, world distance equals overlap distance within numerical tolerance. In reverse order, the additional component accounts for 19.5% and 27.4% of world distance under yes/no and true/false, respectively. The model still selects the correct answers. Table 18 reports the underlying pair probabilities, with every column aligned back to the canonical A, B, C order before computing distances.

Table 17: Decomposition for the image in Figure 5. All rows have correct baseline single-question and joint answers. The ratio $\rho = B_{\text{pair}\to\text{world}}/\delta_{\text{world}}$ is reported for resolved positives.

Answers	Order	δ_{pair}	δ_{world}	$B_{\text{pair}\to\text{world}}$	ρ
yes/no	Forward	0.8781	0.8781	$\leq 10^{-8}$	–
yes/no	Reverse	0.6182	0.7678	0.1496	19.5%
true/false	Forward	0.6856	0.6856	$\leq 10^{-8}$	–
true/false	Reverse	0.3182	0.4384	0.1202	27.4%

Table 18: Complete pair laws for the illustrated Idefics3 image. Binary 0 means no or false, and 1 means yes or true. Columns refer to answer vocabulary and question/output order; each four-row pair block sums to one before rounding.

Context	State	yes/no forward	yes/no reverse	true/false forward	true/false reverse
AB	00	0.02804	0.12506	0.29877	0.16926
AB	01	0.00231	0.22361	0.04583	0.29785
AB	10	0.94122	0.63513	0.63837	0.52134
AB	11	0.02843	0.01620	0.01702	0.01155
BC	00	0.52977	0.59612	0.59353	0.53765
BC	01	0.00055	0.32293	0.00167	0.22934
BC	10	0.46951	0.07121	0.40458	0.22412
BC	11	0.00018	0.00975	0.00022	0.00889
AC	00	0.01758	0.08825	0.23938	0.06976
AC	01	0.00004	0.05938	0.00059	0.24404
AC	10	0.98196	0.83736	0.75980	0.63175
AC	11	0.00042	0.01501	0.00022	0.05445

I.9 Numerical reconstruction and evidence availability

Independent distance optimization preserves every visual-test gap decision. An implementation constructed separately from the experiment analyzer solves 12,288 overlap and world programs for all 6,144 conditions. The largest pair-to-world gap difference is 1.39×10^{-15}, with no change at the fixed 10^{-8} criterion. The reported laws and distances are used throughout the text, tables, and figures.

Candidate records cover the full visual test in one complete reproduction. Each of the three models contributes 47,104 test sequences, for a total of 141,312. The artifact retains the selected-token logits, log normalizers, candidate states, and normalized answer laws, with development records stored separately. Counts refer to distinct candidate identifiers. Where resumed journals repeat an identifier, reconstruction uses its last committed record, as does the inference loader. Token terms sum to the recorded sequence log probabilities without a residual. All tables and figures use this complete set of measurements.

The evidence supports a finite probability-level diagnostic with observed readout sensitivity. Independent images and a Llama-decoder model extend the existence result beyond the controlled text and composite-image study. The absence of four-way persistence limits an interface-invariant interpretation. Both findings are useful for evaluation: answer correctness, overlap distance, the additional global component, and readout stability answer different questions and should be reported separately.