SIMULVISION: DIAGNOSING AND RECOVERING MULTI-OBJECT ASSIGNMENT IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object errors in vision-language models (VLMs) are often interpreted as failures of visual binding or spatial reasoning, yet an incorrect global response does not reveal whether assignment-relevant information is genuinely unavailable. Inspired by the distinction between component access and multi-item integration, we study identity-location assignment under controlled evaluation. A component-preserving two-object diagnostic initially suggests a substantial mapping deficit: Qwen2.5-VL-7B fails on 31.4% of component-correct items under a fixed answer order. However, only 6.4% are incorrect under both answer orders, showing that many apparent assignment failures are response-unstable. We therefore introduce a closed-vocabulary multi-object assignment task that removes answer-position bias and free-form naming variability. On Qwen, local likelihood probing recovers substantial assignment-relevant signal even when monolithic response accuracy is low. Across both Qwen2.5-VL-7B and InternVL3.5-4B, enforcing one-to-one structure on the same local compatibility scores substantially improves exact assignment accuracy. We instantiate this intervention as Factorized Global Assignment (FGA), a training-free structured readout, and replicate the effect on an independently frozen 300-image confirmation set. These results show that multi-object evaluation should distinguish what assignment evidence can be elicited locally from how reliably several assignments are assembled into a globally consistent response.

1 INTRODUCTION

Modern vision-language models (VLMs) can recognize individual objects, describe complex scenes, and answer spatial questions with high accuracy. Yet these capabilities become fragile when several objects must be represented and related at once Thrush et al. (2022); Hsieh et al. (2023); Yuksekgonul et al. (2023); Zeng et al. (2024); Mitra et al. (2024); Alonso et al. (2025). A useful conceptual distinction comes from simultanagnosia, where access to individual elements can coexist with difficulty integrating several elements into a coherent visual whole Mazza (2017). We use this only as experimental motivation: access to visual components need not imply reliable multi-item integration. This motivates our central question for VLMs: when a multi-object response is wrong, is the assignment-relevant information actually unavailable, or does the failure arise when otherwise accessible evidence must be assembled into a joint response?

This question is difficult to answer from end-task accuracy alone. An incorrect multi-object response can reflect object recognition or localization errors, but it can also arise from answer-position preferences, non-canonical naming, or instability in converting available evidence into a final response. We therefore seek a diagnostic that separates missing component information from failures in object–location assignment, and then asks how much assignment-relevant evidence remains recoverable when the global response is incorrect.

We first introduce a component-preserving two-object swap assay. The image, object categories, box identities, and answer components are fixed while only the object-to-box assignment changes. On Qwen2.5-VL-7B Bai et al. (2025), 31.4% of component-correct items fail under a fixed answer order. However, after reversing the answer order for the same candidate assignments, only 6.4%

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.

remain incorrect under both orders. Thus, many errors that initially resemble failures of identity–location binding are response-unstable rather than order-invariant assignment failures.

To remove both answer-position bias and unrestricted naming variability, we next formulate multi-object perception as a closed-vocabulary one-to-one assignment task. Each image contains numbered boxes and an unordered set of valid object labels, and the model must assign every label to exactly one box. Under this controlled interface, a consistent local–global dissociation emerges: as the number of objects increases, exact accuracy of monolithic global generation deteriorates rapidly, while frozen VLM likelihoods evaluated independently for each box retain substantially stronger object–box assignment signal.

This dissociation motivates a direct intervention. We introduce **Factorized Global Assignment (FGA)**, a training-free structured readout that constructs an object–box compatibility matrix from the same frozen VLM likelihoods and enforces the one-to-one structure of the task. FGA introduces no learned parameters or additional supervision; it changes only how the local evidence is combined. Across Qwen2.5-VL-7B and InternVL3.5-4B Wang et al. (2025), FGA substantially improves exact assignment accuracy over independent local decoding using the same compatibility scores, and the effect replicates on an independently frozen confirmation set.

Taken together, our results show that assignment-relevant evidence can remain locally elicitable even when the model's final global response is incorrect. In the settings studied here, some apparent multi-object binding failures are therefore consistent with a gap between locally recoverable evidence and globally consistent response construction, rather than the complete absence of assignment-relevant information.

2 Related Work

Multi-object compositionality and visual binding. VLMs exhibit persistent weaknesses when multiple entities, attributes, or spatial relations must be composed correctly Thrush et al. (2022); Hsieh et al. (2023); Yuksekgonul et al. (2023); Zeng et al. (2024); Mitra et al. (2024); Alonso et al. (2025). These failures are closely related to the classical binding problem, in which representing individual entities does not guarantee that their identities, attributes, and locations are associated correctly Treisman (1996). Recent work has examined binding limitations in CLIP-like representations and large VLMs Kang et al. (2025); Campbell et al. (2024); Izadi et al. (2025), while mechanistic analyses have identified internal structures associated with spatial indexing and visual search Assouel et al. (2026); Kang et al. (2026). Our focus is complementary: rather than asking whether multi-object failures exist, we ask what an observed assignment error actually establishes about the availability of assignment-relevant evidence.

Evaluation interfaces and structured readout. Model behavior can depend strongly on how predictions are elicited. Multiple-choice evaluation is sensitive to option order in both textual and multimodal models Pezeshkpour & Hruschka (2024); Atabuzzaman et al. (2025); Wei et al. (2026), while task decomposition and representation-level analyses show that information absent from a final response may remain recoverable under alternative probes Yang et al. (2024); Huang et al. (2025); Li et al. (2025). Structured decoding further shows that external constraints can be imposed at inference time without retraining Park et al. (2024). We build on these observations but isolate a different question: whether the same locally elicitable object–box evidence can be assembled into a globally consistent one-to-one assignment. FGA serves as a controlled readout intervention for testing this distinction.

3 Diagnosing Multi Object Assignment Failures

We first examine whether conventional multi object evaluation can reliably distinguish missing identity and location information from instability in how otherwise available information is expressed and combined.

3.1 Component Preserving Errors Are Often Response Unstable

For each diagnostic item, we select two numbered objects from the same image and construct three related queries. An identity only query asks for the unordered set of object categories, a location only query probes the spatial relation between the numbered boxes, and a mapping query asks which object belongs to which box. The two mapping alternatives contain exactly the same object categories and box identifiers and differ only in their assignment. We call a mapping error component preserved only when both the identity and location controls are answered correctly. Figure 1 summarizes the component-preserving diagnostic and the answer-order reversal test.

This construction provides a stronger candidate for an integration error than ordinary end task failure. If the model recognizes the relevant identities and correctly resolves the spatial slots but nevertheless selects the wrong pairing, the remaining error cannot be straightforwardly attributed to missing object identity or location information.

We construct 262 diagnostic triplets from 129 COCO val2017 images Lin et al. (2014), yielding 786 component questions. Correct A/B choices are balanced. On Qwen2.5 VL 7B, identity, location, and mapping accuracies are 97.3%, 62.2%, and 66.4%, respectively. Both component controls are correct on 156 triplets. Among these, 49 fail the fixed order mapping query, corresponding to a component preserved conditional error rate of 31.4% (95% triplet CI: 24.4–38.5%; image cluster CI: 23.5–39.7%).

At first sight, this appears to indicate a substantial assignment deficit despite preserved access to the relevant components. However, this interpretation assumes that the mapping error reflects the semantic assignment rather than the way the alternatives are presented. We therefore evaluate every mapping item again after reversing the complete A/B answer strings while leaving the image, question, and candidate assignments unchanged.

The apparent deficit decreases sharply. For Qwen2.5 VL 7B, only 10 of the 156 component correct triplets remain incorrect under both answer orders, reducing the fixed order error from 31.4% to an order invariant residual of 6.4%. The same qualitative decomposition appears across model families. Phi 4 Multimodal decreases from a 34.9% fixed order conditional error to a 3.5% both orders wrong residual Abouelenin et al. (2025), while InternVL3.5 4B decreases from 15.2% to 8.1%.

The residual is therefore much smaller than the original fixed-order estimate. A prediction-blind GQA/Visual Genome screen shows similarly small both-orders-wrong residuals across model families Hudson & Manning (2019), while an additional frozen COCO holdout reproduces the fixed-order component-conditioned effect but lacks an answer-order-reversal evaluation. We therefore treat the two-object result as evidence that the measurement interface is sensitive, rather than as a prevalence estimate of a universal binding deficit.

An obvious alternative is to remove answer choices altogether. We tested an unrestricted generation protocol on the same 262 mapping items. Under exact COCO label scoring, however, accuracy falls to 6.1%. Within the 156 component correct subset, 141 responses are invalid under exact label parsing, whereas only two correspond to literal swapped mappings. Inspection shows that many invalid responses are visually reasonable but use noncanonical category descriptions, such as "tennis player" instead of person or "empty seat" instead of chair.

The two interfaces therefore introduce complementary confounds. Multiple choice queries mix assignment with response position, while unrestricted generation mixes assignment with lexical production. A cleaner test should eliminate both.

3.2 Closed Vocabulary Multi Object Assignment

We therefore formulate multi object perception as a closed vocabulary assignment problem. Each item contains numbered object boxes and an unordered set of valid object labels. The model must assign every supplied label to exactly one box using a canonical output format. Because the candidate identities are given explicitly, the model does not need to invent category names. Because no A/B alternatives are used, the semantic decision is not tied to answer position.

We construct nested k=2/3/4 items from COCO train2017. Eligible instances belong to distinct categories, are not marked as crowd objects, and have area greater than 500 pixels squared. When

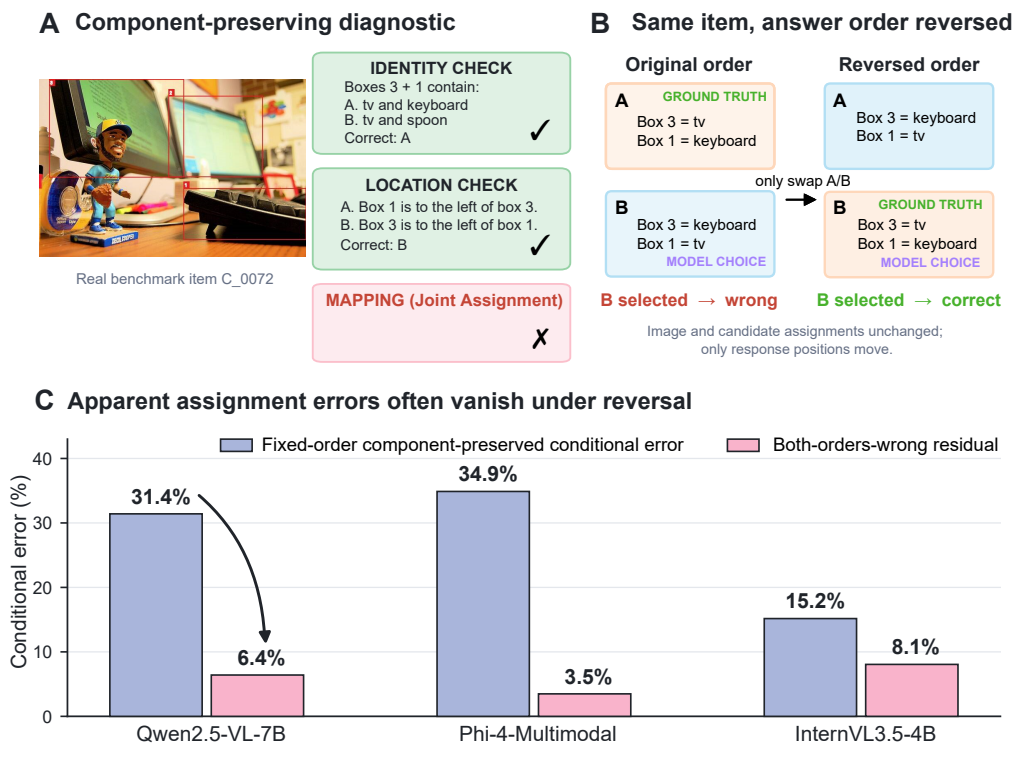


Figure 1: **Component-preserving diagnostic and answer-order reversal.** Identity and location controls can remain correct while the mapping query fails, but many such errors disappear when only the A/B response order is reversed. The residual persists across three model families.

the same category occurs more than once, the largest instance is retained. Box numbering and candidate label order are deterministically shuffled.

The development evaluation contains 152 items from 85 source images, with 85/43/24 items for k=2/3/4. All items pass deterministic integrity checks for unique identifiers, category uniqueness, vocabulary permutation, target consistency, bounding boxes, and rendered stimuli.

Because the development k=4 subset contains only 24 items, we additionally construct a larger confirmation set using the identical eligibility rule on images that were not used by earlier project artifacts. It contains 300 source images and 529 items, with 300/156/73 items for k=2/3/4. The confirmation manifest was frozen and SHA 256 hashed before any model output was collected; no prediction, difficulty, or performance signal entered item selection.

We evaluate complete assignment accuracy and individual box–label accuracy; formal metric definitions are given in Section 5.1.

3.3 Local Scoring Reveals Recoverable Assignment Evidence

With the evaluation interface controlled, we compare monolithic global generation with local compatibility scoring in the same frozen VLM. In the global condition, the model generates the complete mapping in one response. In the local condition, every candidate–box pair is scored using normalized teacher-forced likelihood. The independent Local decoder then assigns each candidate to its highest-scoring box without enforcing the one-to-one constraint.

The difference is substantial. On the development set, Qwen2.5 VL 7B achieves exact accuracies of 60.0%, 18.6%, and 4.2% under global generation for k=2/3/4, respectively. Local scoring reaches 88.2%, 62.8%, and 33.3%. The same separation appears on the independent confirmation set: global

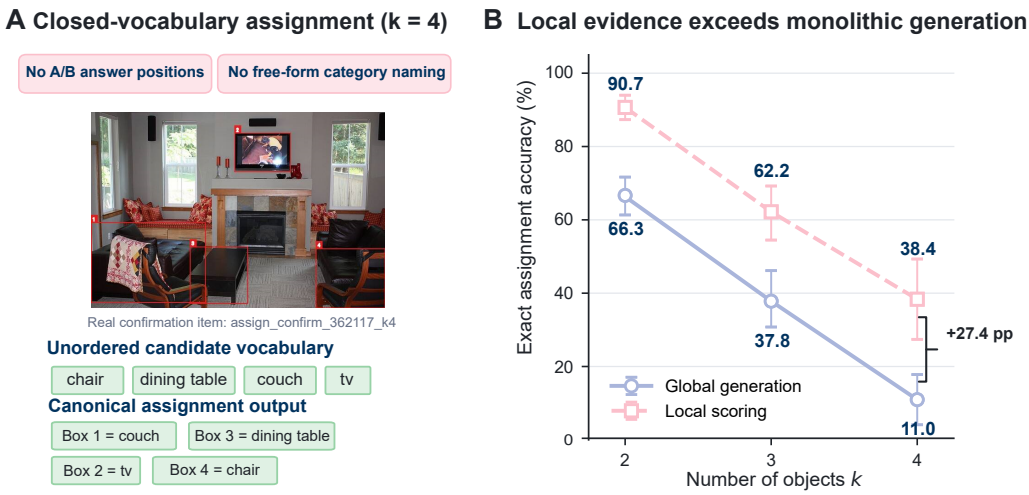


Figure 2: **Controlled assignment reveals a local-to-global gap.** On the Qwen confirmation set, independent local compatibility scoring retains substantially higher exact assignment accuracy than monolithic global generation as the number of objects increases.

generation reaches 66.3%, 37.8%, and 11.0%, whereas local scoring reaches 90.7%, 62.2%, and 38.4%. Figure 2 illustrates the controlled assignment interface and the resulting separation between global generation and local scoring on the confirmation set.

The key observation is not simply that local scoring performs better. Rather, substantial assignment evidence remains recoverable under the local scoring interface even in regimes where monolithic global generation performs poorly. This pattern shows that assignment-relevant signal remains recoverable under the local likelihood interface.

One concern is that images supporting larger k may simply be harder. We therefore construct a within image control using the 73 confirmation images with at least four eligible categories. For each image, we fix the same four object pool and evaluate all six two object subsets, all four three object subsets, and the full four object assignment. The source image and candidate pool are thus unchanged while only the number of simultaneously competing assignments varies.

Performance still decreases with k under this control for both global generation and local scoring. The increasing difficulty therefore cannot be explained solely by selecting harder images at larger object counts. Instead, the controlled task reveals a persistent separation between locally recoverable assignment evidence and joint-response accuracy. This local to global gap motivates the intervention in Section 4.

4 Factorized Global Assignment

Section 3 reveals a consistent gap between the information recoverable from local object–box scores and the accuracy of the final global response. This suggests a direct test of our interpretation. If useful assignment evidence is recoverable locally, we can test whether imposing the known one-to-one structure on the same scores improves exact assignment without modifying the VLM itself.

We therefore use **Factorized Global Assignment (FGA)** as a structured readout intervention. Instead of asking the model to generate the complete assignment as a single response, FGA first measures the compatibility between each candidate label and each numbered box, then combines these local scores under the one-to-one constraint implied by the task. The VLM remains frozen throughout, and FGA introduces no learned parameters or additional supervision. Its purpose is to test whether changing only the final organization of the same local evidence can recover globally consistent assignments.

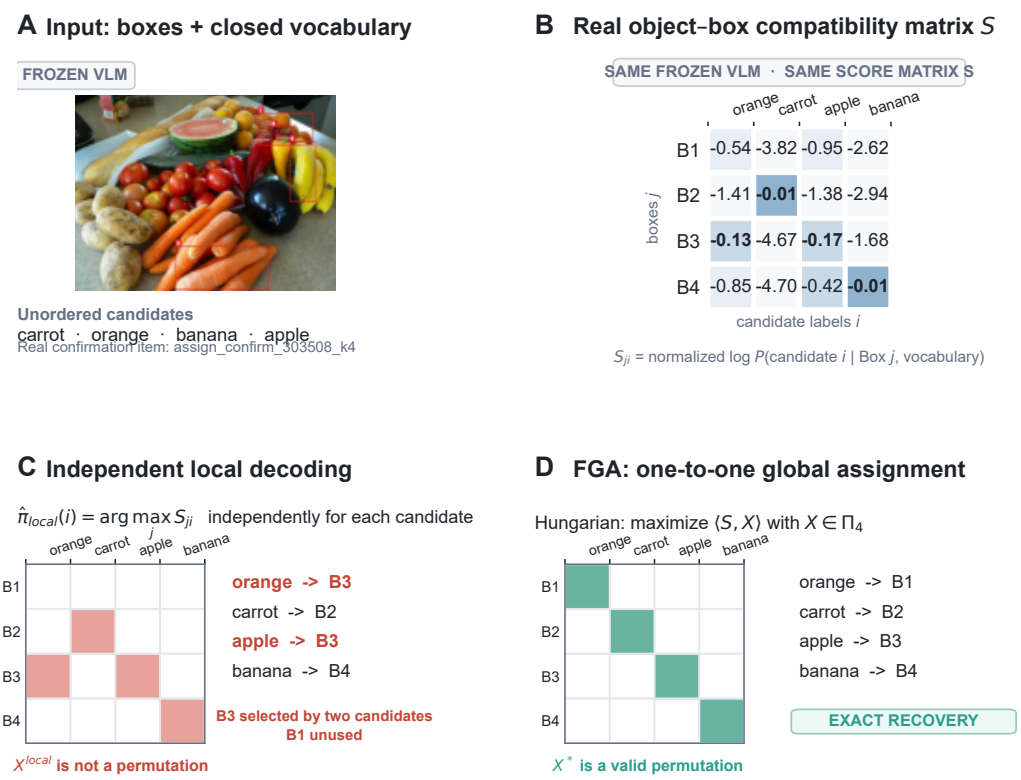


Figure 3: **Factorized Global Assignment from local compatibility scores.** Independent local decoding can produce mutually incompatible assignments from the same score matrix. FGA enforces the one-to-one structure of the task and recovers a valid global assignment in the illustrated real confirmation example.

4.1 Problem Formulation and Local Compatibility

Consider an image I with k numbered boxes $B = \{b_1, \ldots, b_k\}$ and a closed vocabulary of k candidate labels $O = \{o_1, \ldots, o_k\}$. The ground-truth assignment maps each candidate to its box: if candidate o_i is contained in box b_j, we write $\pi^\star(i) = j$.

For each target box b_j, we use the same image and candidate vocabulary as in Section 3 and evaluate every candidate label under the frozen VLM. Let $o_i = (w_{i,1}, \ldots, w_{i,T_i})$ denote the tokenization of candidate label o_i, where T_i is the number of label tokens. We define the local compatibility score as the token-normalized teacher-forced log likelihood

$$S_{ji} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P_\theta \left(w_{i,t} \mid I, b_j, O, w_{i,<t} \right), \tag{1}$$

where θ denotes the fixed model parameters. Normalizing by T_i makes scores more comparable across candidate labels with different token lengths.

Evaluating every candidate label against every box produces a compatibility matrix $S \in \mathbb{R}^{k \times k}$, where rows index boxes and columns index candidate labels, and S_{ji} measures the support that the frozen VLM assigns to candidate o_i for box b_j.

The independent local baseline decodes each candidate separately:

$$\hat{\pi}_{\mathrm{Local}}(i) = \arg \max_{j \in \{1,\ldots,k\}} S_{ji}. \tag{2}$$

Because candidates are decoded independently, multiple candidates may select the same box while another box remains unused. Such conflicts reveal cases in which individually plausible local decisions cannot simultaneously form a valid global assignment.

4.2 Global Assignment from Local Evidence

The local decoder in Eq. 2 assigns each candidate independently and therefore does not guarantee a valid one-to-one assignment. FGA instead uses the same compatibility matrix S while enforcing the global structure of the task.

FGA finds the globally valid permutation

$$\hat{\pi}_{\mathrm{FGA}} = \arg\max_{\pi \in \Pi_k} \sum_{i=1}^{k} S_{\pi(i),i}, \tag{3}$$

The constraints require every box to receive exactly one label and every label to be used exactly once. This is a standard linear assignment problem, which we solve exactly using the Hungarian algorithm Kuhn (1955). FGA introduces no new visual features, semantic information, learned parameters, or supervision; it only changes how the same local compatibility scores are combined. Because independent local decoding and FGA use the same score matrix, their comparison isolates the effect of enforcing one-to-one global structure.

5 Experiments

We evaluate the assignment readouts across models and data splits, and test whether enforcing one-to-one structure can recover globally consistent assignments from the same local evidence. We then examine what kinds of errors are corrected by FGA and test several alternative explanations for the observed gain.

5.1 Evaluation Protocol

We evaluate Qwen2.5-VL-7B and InternVL3.5-4B on the development and independent confirmation sets described in Section 3.2, using fixed model revisions and deterministic decoding. The two sets are source-image disjoint, and the confirmation manifest was frozen before any model output was collected.

Our primary metric is exact assignment accuracy, which requires the full object–box permutation to be correct. Pair accuracy measures individual box–label assignments. Confidence intervals are estimated with 2,000 paired bootstrap draws using source-image cluster resampling as the primary analysis. Greedy outputs are parsed deterministically against the supplied vocabulary, with unparseable responses counted as incorrect.

Because greedy output-format reliability differs across model families, we treat greedy generation as an end-to-end reference. The comparison between independent local scoring and FGA provides the cleaner test of structured assignment, since both use the same frozen compatibility matrix and differ only in the final readout.

5.2 Main Results

FGA improves globally consistent assignment from the same local scores. Development-set results show the same pattern and are reported in the Appendix; we focus here on the independently frozen confirmation set. Figure 4 summarizes the main confirmation-set results for both model families.

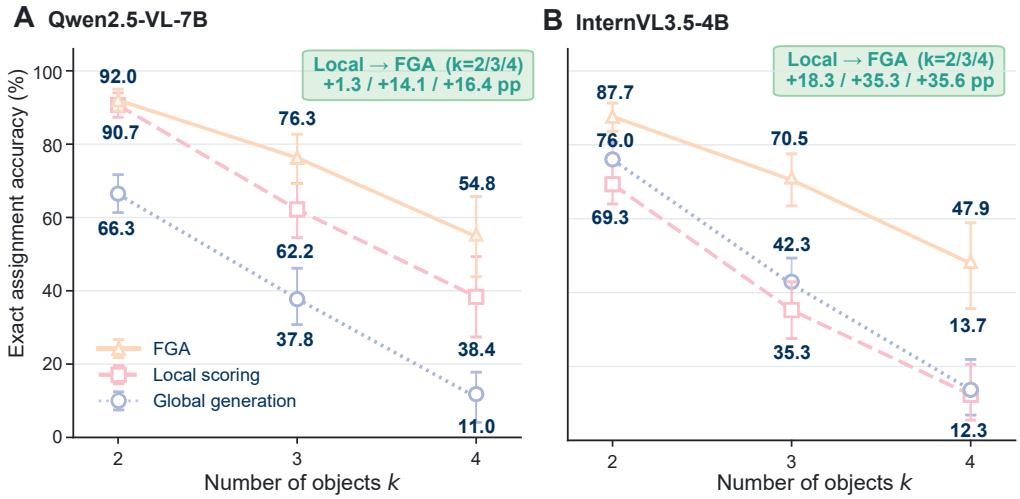


Figure 4: **Exact assignment accuracy on the independent confirmation set.** FGA improves over independent local decoding on both Qwen2.5-VL-7B and InternVL3.5-4B while using the same local compatibility scores; only the final assignment rule differs.

On the confirmation set, Qwen2.5-VL-7B achieves 66.3%, 37.8%, and 11.0% exact accuracy under greedy generation for $k = 2, 3, 4$, compared with 90.7%, 62.2%, and 38.4% for Local and 92.0%, 76.3%, and 54.8% for FGA. Relative to the same local score matrix, FGA improves over Local by +1.3, +14.1, and +16.4 percentage points. The same Local-to-FGA recovery replicates on InternVL3.5-4B, from 69.3%, 35.3%, and 12.3% to 87.7%, 70.5%, and 47.9%, corresponding to gains of +18.3, +35.3, and +35.6 points. Image-cluster confidence intervals exclude zero for all Local-to-FGA contrasts across both models.

The within-image control in Section 3.3 shows the same scaling trend: with the source image and four-object pool fixed, FGA accuracy decreases from 91.6% to 75.7% to 54.8% for Qwen and from 89.7% to 69.9% to 47.9% for InternVL as k increases from 2 to 4.

5.3 Why Does FGA Help?

The main results show that global structure improves exact assignment, but not why. We therefore examine the errors corrected by FGA and test two alternative explanations for the gain.

FGA primarily resolves incompatible local decisions. Independent local decoding can assign multiple candidate labels to the same box, leaving another box unused. On the confirmation set, Qwen produces duplicate assignments on 5, 42, and 33 items for $k = 2, 3, 4$, of which FGA converts 4, 22, and 12 into exact assignments. InternVL produces 81, 94, and 55 duplicate cases, of which FGA resolves 55, 55, and 26. Pair accuracy changes much less than exact accuracy, consistent with FGA improving the global structure of the assignment rather than uniformly improving individual box classification.

The effect is not specific to the Hungarian solver. Using the same Qwen confirmation score matrix, Greedy-Unique and Sinkhorn balancing with Hungarian rounding Sinkhorn & Knopp (1967); Cuturi (2013) both improve substantially over independent local argmax (Table 1). Sinkhorn matches FGA at $k = 2$ and $k = 4$ and is slightly higher at $k = 3$. These results support interpreting the gain as an effect of enforcing valid global assignment structure rather than of a uniquely privileged decoder.

Whole-sequence scoring does not explain the Qwen result. To test whether FGA benefits merely from evaluating more candidate assignments, we enumerate all $k!$ valid assignment strings and select the complete sequence with the highest teacher-forced likelihood under the same frozen Qwen model. On the confirmation set, whole-sequence scoring reaches 82.0%, 62.8%, and 34.2% for

Table 1: Comparison of alternative readouts on the Qwen2.5-VL-7B independent confirmation set. Structured readouts operate on the same frozen local compatibility scores unless otherwise specified.

Method	$k = 2$	$k = 3$	$k = 4$
Greedy generation	66.3	37.8	11.0
Whole-sequence scoring	82.0	62.8	34.2
Local argmax	90.7	62.2	38.4
Greedy-Unique	91.3	70.5	47.9
Sinkhorn	**92.0**	**76.9**	**54.8**
FGA (Hungarian)	**92.0**	76.3	**54.8**

$k = 2, 3, 4$, compared with 92.0%, 76.3%, and 54.8% for FGA. Thus, evaluating additional complete answers alone does not account for the gain; factorizing the response into local compatibility scores before imposing global structure remains substantially more effective.

Additional controls. The recovered compatibility signal remains strongly dependent on the visual input: replacing the correct image with a gray or unrelated image causes a large accuracy drop. Candidate-order permutations and blank-image language-prior calibration do not alter the main conclusion. The additional inference cost is dominated by compatibility scoring; the assignment step operates only on a $k \times k$ matrix. Full ablation results, prior-calibration results, and latency measurements are reported in the Appendix.

6 Discussion

Our results separate three sources of apparent multi-object failure that are often conflated in end-task evaluation. Answer-position sensitivity can create component-preserved errors in multiple-choice settings; unrestricted generation entangles assignment with lexical realization; and even after both confounds are controlled, independently plausible local decisions may fail to form a valid global assignment. The central empirical result is therefore not that a particular decoder solves visual binding, but that measured multi-object competence can depend on how assignment evidence is elicited and coordinated.

Importantly, local scoring repeatedly re-queries the frozen model under box-specific prompts. It therefore demonstrates that assignment-relevant evidence is locally elicitable, not that all such evidence is simultaneously represented or directly accessible during a single global forward pass.

The controlled design also defines the scope of the conclusion. The closed-vocabulary task assumes a known one-to-one candidate set with distinct categories and therefore isolates assignment from detection, open-vocabulary naming, missing candidates, and repeated-category instances. The full FGA analysis is currently restricted to COCO and two model families, and compatibility scoring incurs additional inference cost. FGA should therefore be viewed as a diagnostic structured readout rather than a general replacement for ordinary VLM inference.

7 Conclusion

We studied multi-object assignment by separating component access from global response construction. A component-preserving diagnostic shows that many apparent identity–location failures disappear under answer-order reversal, while a closed-vocabulary assignment task reveals that substantially stronger assignment evidence can be elicited under local scoring. Enforcing one-to-one structure on the same local scores recovers additional exact assignments across two VLM families without modifying model parameters. These results caution against equating end-task multi-object errors with missing visual information and motivate evaluation that separates local evidence elicitation from global response construction.

AI USE STATEMENT

Generative AI tools were used during this work for research brainstorming, refinement of experimental design, code assistance, organization and interpretation of experimental results, and manuscript drafting and language editing.

All AI-assisted research decisions, code, analyses, and written content were reviewed by the authors. Experimental procedures and numerical results were verified against the underlying code and stored outputs. AI-generated suggestions were treated as provisional and were retained only after author evaluation and validation.

The authors take full responsibility for the final content of this work, including all claims, analyses, code, figures, and text produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and does not involve human subjects, private data, or interventions involving people. The study focuses on evaluating and interpreting the behavior of vision-language models. We do not make clinical claims about artificial models or imply mechanistic equivalence between VLM behavior and neurological disorders. Simultanagnosia is used only as a conceptual source of experimental motivation.

REPRODUCIBILITY STATEMENT

We provide detailed information required to reproduce the experiments throughout the paper and appendix. Dataset construction rules, deterministic item generation, integrity checks, model revisions, prompting interfaces, likelihood scoring, evaluation metrics, and bootstrap procedures are documented in the supplementary material. The independent confirmation set was frozen and cryptographically hashed before any model outputs were collected. Random seeds and evaluation conventions are fixed and reported. We additionally provide code for dataset construction, local scoring, Factorized Global Assignment, structured readout baselines, and evaluation. An anonymous code repository will be provided with the submission.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, H. Awadalla, Nguyen Bach, Jianmin Bao, A. Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dongdong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Mingcheng Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yun-Meng Li, Chen Liang, Xihui Lin, Zeqi Lin, Meng-Jie Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, V. Mazalov, Ali Mousavi, Anh Nguyen, Jing Pan, D. Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xiaocheng Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, I. Zabir, Jianwen Zhang, L. Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *ArXiv*, March 2025. URL https://arxiv.org/abs/2503.01743.

Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez De Lacalle. Vision-language models struggle to align entities across modalities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18846–18862, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.965. URL https://aclanthology.org/2025.findings-acl.965/.

Rim Assouel, Declan Iain Campbell, Yoshua Bengio, and Taylor Whittington Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3RQ863cRbx.

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

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. *arXiv preprint arXiv:2502.13923*, 2025.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Webb. Understanding the limits of vision language models through the lens of the binding problem. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 113436–113460. Curran Associates, Inc., 2024. doi: 10.52202/079017-3604. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cdcc6d47c1627350014a3076112ab824-Paper-Conference.pdf.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116. Curran Associates, Inc., 2023. doi: 10.52202/075280-1355. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/63461de0b4cb760fc498e85b18a7fe81-Paper-Datasets_and_Benchmarks.pdf.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4830–4843, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.249. URL https://aclanthology.org/2025.findings-acl.249/.

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 (CVPR)*, June 2019.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Vahedi, Hosein Hasani, and Mahdieh Baghshah. Visual structures help visual reasoning: Addressing the binding problem in lvlms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 162930–162968. Curran Associates, Inc., 2025. doi: 10.52202/085713-5440. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ee8351f13c3015ac249ed5f767096966-Paper-Conference.pdf.

Raphaela Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 41245–41258, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/44956951349095f74492a5471128a7e0-Paper-Conference.pdf.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22436–22446, October 2025.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25ca.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Veronica Mazza. Simultanagnosia and object individuation. *Cognitive Neuropsychology*, 34(7-8): 430–439, 2017. doi: 10.1080/02643294.2017.1331212.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431, June 2024.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D' Antoni. Grammar-aligned decoding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 24547–24568. Curran Associates, Inc., 2024. doi: 10.52202/079017-0774. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2bdc2267c3d7d01523e2e17ac0a754f3-Paper-Conference.pdf.

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

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. doi: 10.2140/pjm.1967.21.343.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Anne Treisman. The binding problem. *Current Opinion in Neurobiology*, 6(2):171–178, 1996. ISSN 0959-4388. doi: https://doi.org/10.1016/S0959-4388(96)80070-5. URL https://www.sciencedirect.com/science/article/pii/S0959438896800705.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Sheng-Lun Wei, Yu-Ling Liao, Hen-Hsen Huang, and Hsin-Hsi Chen. Do multimodal LLMs understand order? measuring the fragility of multimodal reasoning under input order perturbations. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 9118–9128, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/4jtpgzks8pbr. URL https://aclanthology.org/2026.lrec-1.716/.

Qian Yang, Weixiang Yan, and Aishwarya Agrawal. Decompose and compare consistency: Measuring VLMs' answer reliability via task-decomposition consistency comparison. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3613–3627, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.211. URL https://aclanthology.org/2024.emnlp-main.211/.

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

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14141–14151, June 2024.

A Additional Diagnostic Experiments

The original component-preserving diagnostic uses three queries for each two-object image: identity-only, location-only, and the object–box mapping. A mapping failure is counted only when both component queries are correct. The mapping alternatives contain the same objects and box identifiers and differ only in their assignment. The answer-order test then swaps the full A/B strings while holding the image, semantics, and candidate assignments fixed. Table 2 reports the fixed-order conditional error and the stricter residual that is wrong under both answer orders.

A.1 Unrestricted generation

We also evaluated open-ended generation on the original 262 mapping items, without supplying the two candidate category names. An independently rerun exact-label parser confirms 16/262 exact COCO-label mappings (6.1%), six literal swaps, and 240 invalid mappings. Among the 156 items on which both component queries are correct, 141 are invalid under exact-label parsing and only two are literal swapped mappings. Table 3 illustrates how lexical realization is therefore entangled with assignment in this setting.

The GQA/Visual Genome screen was selected without model predictions. The eligible pool contained 706 images; 14 images were excluded by author quality control before inference, and 300 images were selected. All selected Visual Genome `coco_id` fields are null. The screen uses a different annotation source and is disjoint from the local COCO validation set under the recorded metadata, SHA-256, and near-duplicate checks. This screen tests whether the response-order effect transfers across annotation sources; it does not establish training-data provenance. No separate stricter-metadata GQA sensitivity experiment is available.

The additional fresh COCO holdout contains 300 triplets and has fixed-order outputs. Among 194 component-correct triplets, 58 fail the mapping query (29.9%). No answer-order-reversal artifact exists for this holdout, so a both-orders-wrong residual cannot be computed and is not reported.

B Dataset Construction and Quality Control

The closed-vocabulary task is constructed from COCO train2017. Eligible instances are non-crowd annotations with area strictly greater than $500\,\mathrm{px}^2$. We retain the largest instance of each category and require distinct categories within an item. Development source images are ordered by the hexadecimal SHA-256 digest of `assignment-20260901 || decimal(image_id)`. Confirmation first removes the frozen used-image registry, then orders the remaining local COCO images by the corresponding digest with salt `assignment-confirm-20260906`; the scan accepts the first 300 images with at least two eligible categories. For each $k \in \{2, 3, 4\}$, the k largest eligible categories are selected. Their box order is shuffled by a Python `random.Random` instance seeded with the hexadecimal image digest, independently for each k. Candidate order is shuffled with

Table 2: Component-preserved conditional errors (%). Parentheses give percentile 95% source-image bootstrap intervals from 2,000 draws. The fresh COCO holdout has fixed-order outputs only; no answer-order-reversal residual is available.

Dataset	Model	Eligible	Fixed-order error	Both-orders wrong
Original COCO	Qwen2.5-VL-7B	156	31.4 (23.5–39.7)	6.4 (2.7–10.4)
Original COCO	Phi-4-Multimodal	172	34.9 (27.9–42.1)	3.5 (1.1–6.5)
Original COCO	InternVL3.5-4B	211	15.2 (10.2–20.4)	8.1 (4.5–11.8)
GQA/VG screen	Qwen2.5-VL-7B	259	11.2 (7.5–15.3)	2.3 (0.8–4.3)
GQA/VG screen	Phi-4-Multimodal	226	11.9 (7.9–16.2)	4.4 (2.1–7.2)
GQA/VG screen	InternVL3.5-4B	265	5.3 (2.7–8.1)	1.9 (0.4–3.7)
Fresh COCO holdout	Qwen2.5-VL-7B	194	29.9	—

Table 3: Real unrestricted-generation responses from the original mapping diagnostic. The exact parser lowercases text and removes non-letter characters but requires the resulting wording to equal the COCO category label for each requested box.

Item	Ground-truth mapping	Raw response	Exact-parser outcome
C_0002	Box 2 = person; Box 4 = chair	Box 2 = Tennis player; Box 4 = Empty seat.	Invalid: Box 2: 'Tennis player' != 'person'; Box 4: 'Empty seat' != 'chair'
C_0014	Box 3 = person; Box 1 = vase	Box 3 = woman; Box 1 = vase.	Invalid: Box 3: 'woman' != 'person'
C_0047	Box 3 = airplane; Box 1 = truck	Box 3 = Airplane; Box 1 = Ground Support Equipment.	Invalid: Box 1: 'Ground Support Equipment' != 'truck'
C_0050	Box 4 = chair; Box 3 = tv	Box 4 = Armchair; Box 3 = Bookshelf	Invalid: Box 4: 'Armchair' != 'chair'; Box 3: 'Bookshelf' != 'tv'

the string seed `vocab:image_id:k` for development and `vocab-confirm:image_id:k` for confirmation. Thus the category sets are nested by construction, while box and vocabulary orders are deterministically regenerated for each k; no unrecorded RNG seed is involved.

The development evaluation contains 152 items from 85 represented source images: 85, 43, and 24 items for $k = 2, 3, 4$. The confirmation evaluation contains 529 items from 300 source images: 300, 156, and 73 items. Confirmation source images have zero overlap with the development train images, development evaluation images, and the wider used-image registry. The confirmation item file was frozen before model inference with SHA-256 `d5f6a1a12f0183d5d51353b0b94b3bca52f4da495205d8b5461b528cdd7d4788`.

Executed integrity checks verify unique question identifiers, distinct categories per item, candidate-vocabulary permutation, target/assignment agreement, expected output placeholders, nonnegative boxes, areas above the threshold, existence of all 529 rendered stimuli, render-count agreement, and zero source-image and question-identifier overlap with the development sets. Rendered-pixel uniqueness was not tested (the stored `render_unique` field is null), so no pixel-uniqueness claim is made.

C Evaluation and Statistical Details

Rows of the score matrix index boxes and columns index candidate labels: S_{ji} is the compatibility of candidate o_i with box b_j. In the serialized artifacts, the array is stored as $S[\text{box}][\text{candidate}]$.

The ground-truth candidate-to-box mapping is $\pi^\star(i) = j$.

$$\text{Exact} = \mathbb{I}\left[\forall i,\, \hat{\pi}(i) = \pi^\star(i)\right].$$

$$\text{Pair} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}\left[\hat{\pi}(i) = \pi^\star(i)\right].$$

The independent Local decoder is

$$\hat{\pi}_{\mathrm{Local}}(i) = \arg\max_j S_{ji}.$$

It can assign multiple candidates to the same box while leaving another box unused. FGA uses the same cells but restricts the result to a one-to-one assignment:

$$\hat{\pi}_{\mathrm{FGA}} = \arg\max_{\pi \in \Pi_k} \sum_{i=1}^{k} S_{\pi(i),i}.$$

where Π_k is the set of permutations of the k boxes. Uniform chance exact accuracy is $1/k!$.

Global responses are parsed as numbered `Box n = category` or `Box n: category` fields after lowercasing and retaining ASCII letters and spaces. A response is parseable only if all k fields are present. Invalid responses count as incorrect. Global pair accuracy is the fraction of ground-truth box fields parsed correctly. For a complete permutation this equals candidate-to-box pair accuracy after inversion.

Intervals use the 2.5th and 97.5th percentiles of 2,000 source-image cluster bootstrap draws. Sampling a source image retains all nested items belonging to that image. Confirmation analyses use seed 1235; the established within-image/development stream uses seed 1234. Paired contrasts are formed itemwise before cluster resampling.

D Prompts and Likelihood Scoring

The following are the exact stored prompts for confirmation item `assign_confirm_360880_k4`. The global prompt is:

```
The image contains numbered red boxes around candidate
objects.  The objects in the boxes are:  remote, chair,
person, couch (listed unordered).  Which object is assigned
to each box?  Answer exactly:  Box 1 = <object>; Box 2 =
<object>; Box 3 = <object>; Box 4 = <object>.
```

The corresponding local prompt for Box 1 is:

```
The image contains numbered red boxes around candidate
objects.  Which object is inside Box 1?  The objects are:
remote, chair, person, couch.  Answer exactly with the
object name.
```

Other items substitute the stored candidate vocabulary and numbered box while retaining these literal templates.

For candidate $o_i = (w_{i,1}, \ldots, w_{i,T_i})$ and box b_j, the teacher-forced compatibility is

$$S_{ji} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log p_\theta(w_{i,t} \mid I, b_j, O, w_{i,<t}).$$

The prompt tokens are masked from the loss. Completion token c is scored from logits at position $c-1$, and multi-token labels are averaged by the number of unmasked completion tokens. No backend-specific BOS/EOS behavior is asserted here.

The pinned revisions are `cc594898137f460bfe9f0759e9844b3ce807cfb5` for Qwen2.5-VL-7B, `93f923e1a7727d1c4f446756212d9d3e8fcc5d81` for Phi-4-Multimodal, and `481f6e32467eab4e922ccd7fd6cf420441a62331` for InternVL3.5-4B.

Implementation details. The confirmation greedy decoders use `do_sample=False`; Qwen uses `max_new_tokens=128` and downsizes inputs exceeding 112,896 pixels, whereas InternVL uses `max_new_tokens=64`, 448-pixel dynamic tiles, at most 12 tiles, and ImageNet mean/std normalization. Evaluation processes one item at a time, and each teacher-forced compatibility cell is a

Table 4: Global-generation parseability (count and %). All k numbered fields are required; invalid outputs count as incorrect.

Set	Model	$k = 2$	$k = 3$	$k = 4$
Development	Qwen2.5-VL-7B	85/85 (100.0)	43/43 (100.0)	24/24 (100.0)
Development	InternVL3.5-4B	74/85 (87.1)	34/43 (79.1)	10/24 (41.7)
Confirmation	Qwen2.5-VL-7B	300/300 (100.0)	156/156 (100.0)	73/73 (100.0)
Confirmation	InternVL3.5-4B	269/300 (89.7)	114/156 (73.1)	25/73 (34.2)

Table 5: Confirmation-set accuracy (%). Parentheses give percentile 95% source-image cluster bootstrap intervals from 2,000 draws. The paired FGA–Local rows report percentage-point contrasts computed itemwise before resampling.

Exact accuracy

Model	Method	$k = 2$	$k = 3$	$k = 4$
Qwen2.5-VL-7B	Greedy generation	66.3 (61.3–71.7)	37.8 (30.8–46.2)	11.0 (4.1–17.8)
Qwen2.5-VL-7B	Local	90.7 (87.3–94.0)	62.2 (54.5–69.2)	38.4 (27.4–49.3)
Qwen2.5-VL-7B	FGA	92.0 (89.0–95.0)	76.3 (69.2–82.7)	54.8 (43.8–65.8)
Qwen2.5-VL-7B	FGA - Local	+1.3 (+0.3–+3.0)	+14.1 (+9.0–+19.9)	+16.4 (+8.2–+26.0)
InternVL3.5-4B	Greedy generation	76.0 (70.7–80.7)	42.3 (34.6–49.4)	13.7 (6.8–21.9)
InternVL3.5-4B	Local	69.3 (64.0–74.7)	35.3 (27.6–42.9)	12.3 (5.5–20.5)
InternVL3.5-4B	FGA	87.7 (83.7–91.3)	70.5 (63.5–77.6)	47.9 (35.6–58.9)
InternVL3.5-4B	FGA - Local	+18.3 (+14.0–+22.7)	+35.3 (+27.6–+42.3)	+35.6 (+24.7–+46.6)

Pair accuracy

Model	Method	$k = 2$	$k = 3$	$k = 4$
Qwen2.5-VL-7B	Greedy generation	67.3 (62.3–72.7)	54.7 (48.7–61.1)	36.6 (30.1–43.5)
Qwen2.5-VL-7B	Local	91.5 (88.3–94.5)	81.8 (77.8–85.9)	69.5 (62.7–76.0)
Qwen2.5-VL-7B	FGA	92.0 (89.0–95.0)	83.3 (78.4–88.0)	72.9 (65.8–79.8)
Qwen2.5-VL-7B	FGA - Local	+0.5 (-0.2–+1.3)	+1.5 (-1.1–+4.1)	+3.4 (-0.7–+7.9)
InternVL3.5-4B	Greedy generation	78.0 (73.2–82.2)	54.1 (47.2–60.5)	21.9 (14.0–30.5)
InternVL3.5-4B	Local	82.8 (79.7–85.8)	70.5 (66.0–74.6)	61.3 (55.8–66.1)
InternVL3.5-4B	FGA	87.7 (83.7–91.3)	78.6 (73.3–83.8)	66.1 (58.2–74.0)
InternVL3.5-4B	FGA - Local	+4.8 (+1.8–+7.8)	+8.1 (+4.3–+12.2)	+4.8 (-1.0–+10.3)

singleton forward. Both assignment backends use 4-bit NF4 weights with double quantization and fp16 compute. The recorded confirmation runtime is Python 3.10.20, PyTorch 2.6.0+cu124, Transformers 5.14.1, Accelerate 1.14.0, bitsandbytes 0.49.2, SciPy 1.15.3, Pillow 12.2.0, and NumPy 2.2.6. Phi's diagnostic run instead records Transformers 4.48.2 and PEFT 0.13.2. No temperature value is passed to greedy generation, and no additional decoding seed is pinned because sampling is disabled.

E Additional Assignment Results

Development greedy generation is evaluated by reparsing the stored response against the ground-truth box assignment. This yields 60.0/18.6/4.2% for Qwen and 69.4/48.8/25.0% for InternVL at $k = 2/3/4$; older vocabulary-order-frame summaries are not used.

The within-image control fixes the same four-object source pool for each of 73 confirmation images and averages all six two-object and all four three-object subsets before comparison with the four-object item. Exact and pair accuracy continue to decline with k under this source-image-controlled construction.

F Alternative Readouts and Controls

Whole-sequence scoring evaluates every complete assignment sequence and is available for Qwen. Greedy-Unique sorts score cells and accepts a cell only when its box and candidate are both un-

Table 6: Development-set assignment accuracy (exact / pair, %). Greedy outputs are reparsed in the ground-truth box frame. Local is independent candidate-to-box decoding.

Model	Method	$k = 2$	$k = 3$	$k = 4$
Qwen2.5-VL-7B	Greedy generation	60.0 / 60.0	18.6 / 39.5	4.2 / 36.5
Qwen2.5-VL-7B	Local	88.2 / 90.0	62.8 / 82.9	33.3 / 70.8
Qwen2.5-VL-7B	FGA	90.6 / 90.6	79.1 / 86.0	41.7 / 67.7
InternVL3.5-4B	Greedy generation	69.4 / 70.6	48.8 / 61.2	25.0 / 36.5
InternVL3.5-4B	Local	63.5 / 78.2	41.9 / 72.9	29.2 / 62.5
InternVL3.5-4B	FGA	85.9 / 85.9	79.1 / 85.3	62.5 / 74.0

Table 7: Within-image control over the 73 confirmation images supporting four objects (%). Parentheses give percentile 95% source-image bootstrap intervals from 2,000 draws. For each image, $k = 2$ averages all six subsets and $k = 3$ all four subsets; $k = 4$ uses the full item.

Exact accuracy

Model	Method	$k = 2$	$k = 3$	$k = 4$
Qwen2.5-VL-7B	Greedy generation	61.0 (55.3–66.2)	26.0 (18.8–33.6)	11.0 (4.1–19.2)
Qwen2.5-VL-7B	Local	89.7 (86.1–92.9)	65.4 (57.9–72.3)	38.4 (27.4–49.3)
Qwen2.5-VL-7B	FGA	91.6 (87.9–94.5)	75.7 (68.8–81.8)	54.8 (43.8–65.8)
InternVL3.5-4B	Greedy generation	63.0 (56.2–69.4)	33.6 (27.1–40.1)	13.7 (5.5–21.9)
InternVL3.5-4B	Local	69.2 (64.2–73.7)	35.3 (28.8–42.1)	12.3 (5.5–20.5)
InternVL3.5-4B	FGA	89.7 (85.4–93.2)	69.9 (62.7–76.4)	47.9 (35.6–58.9)

Pair accuracy

Model	Method	$k = 2$	$k = 3$	$k = 4$
Qwen2.5-VL-7B	Greedy generation	62.8 (57.4–67.9)	43.9 (38.4–49.9)	36.6 (30.1–43.8)
Qwen2.5-VL-7B	Local	90.9 (87.3–93.9)	80.7 (75.8–85.2)	69.5 (62.3–76.0)
Qwen2.5-VL-7B	FGA	91.6 (87.9–94.5)	82.6 (77.7–87.0)	72.9 (65.4–80.1)
InternVL3.5-4B	Greedy generation	64.7 (58.0–71.2)	44.2 (37.6–50.6)	21.9 (13.7–29.8)
InternVL3.5-4B	Local	82.3 (79.0–85.3)	70.0 (65.6–74.0)	61.3 (56.2–66.4)
InternVL3.5-4B	FGA	89.7 (85.4–93.2)	77.9 (72.4–82.8)	66.1 (57.5–73.6)

used. The Sinkhorn baseline computes $\exp(S_{ji} - \max_{i'} S_{ji'})$ rowwise for numerical stability, uses fixed temperature $T = 1$, applies exactly 200 alternating row/column normalizations, and rounds the balanced matrix with Hungarian assignment. It has no learned parameters, temperature tuning, or convergence-based early stopping. This is fixed-temperature exponentiated-score Sinkhorn balancing followed by Hungarian rounding, connecting classical matrix scaling (Sinkhorn & Knopp, 1967) with the entropic optimal-transport formulation (Cuturi, 2013).

For the gray condition, code replaces every image pixel with RGB value 128 and then redraws the original item's red rectangles and white box numbers at their stored coordinates. For the mismatched condition, it keeps the prompt and candidate vocabulary fixed but supplies the first subsequent evaluation item with a different source-image identifier, including that item's already rendered overlays. Figure 5 shows the actual stored inputs for one development item. On Qwen development, correct-minus-gray FGA exact contrasts are 30.6 pp (95% CI: 18.8–42.4), 53.5 pp (95% CI: 37.2–72.1), and 33.3 pp (95% CI: 12.5–54.2) for $k = 2, 3, 4$; correct-minus-mismatched contrasts are 38.8 pp (95% CI: 27.1–50.6), 60.5 pp (95% CI: 46.5–74.4), and 41.7 pp (95% CI: 20.8–62.5). Candidate-order robustness uses three fixed deterministic permutations without selecting an order by outcome. The blank-score calibration subtracts each label's score on the corresponding gray image before structured readout.

Inference cost is dominated by construction of the $k \times k$ compatibility matrix. The following values describe the stored reference implementation, not optimized serving. Timing used one RTX 2080 Ti 11 GB from a two-GPU WSL2 host, one item per batch, and the first six development items for each k/model. Each item received two timed scoring passes; wall-clock intervals were synchronized with `torch.cuda.synchronize`, and model loading was excluded. Qwen recomputes image-conditioned scores for every cell, whereas InternVL reuses one visual encoding and performs k^2

Table 8: Alternative readouts and controls. Values are exact assignment accuracy (%). Whole-sequence scoring is exhaustive over complete outputs. Sinkhorn is fixed-temperature exponentiated-score balancing followed by Hungarian rounding.

Structured readouts, Qwen confirmation

Method	$k=2$	$k=3$	$k=4$
Greedy generation	66.3	37.8	11.0
Whole-sequence scoring	82.0	62.8	34.2
Local	90.7	62.2	38.4
Greedy-Unique	91.3	70.5	47.9
Sinkhorn + Hungarian	92.0	76.9	54.8
FGA-Hungarian	92.0	76.3	54.8

Visual dependence, Qwen development

Condition	$k=2$	$k=3$	$k=4$
Correct image	90.6	79.1	41.7
Gray image with boxes	60.0	25.6	8.3
Mismatched image	51.8	18.6	0.0

Candidate-order robustness

Summary	$k=2$	$k=3$	$k=4$
Mean of three permutations	91.4	81.4	50.0
Worst of three permutations	90.6	79.1	45.8

Blank-image calibration, Qwen confirmation

Scores	$k=2$	$k=3$	$k=4$
Raw	92.0	76.3	54.8
Blank-score calibrated	90.0	71.2	57.5

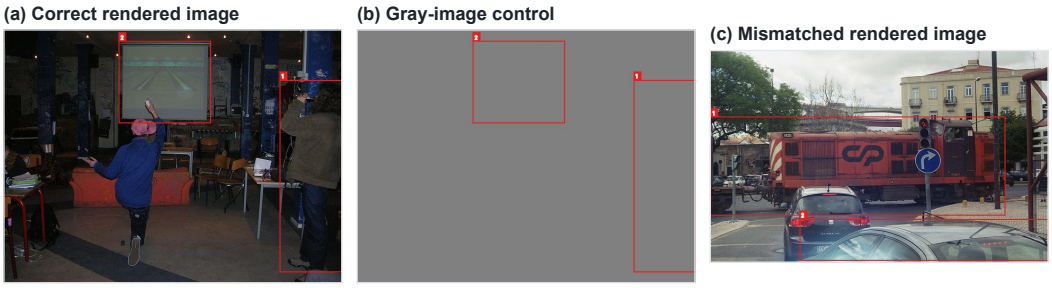


Figure 5: Actual visual-dependence inputs for development item `assign_eval_112065_k2`. (a) The stored correct rendered image. (b) The stored gray control, produced by filling the canvas with RGB 128 and redrawing the item's original red boxes and white numbers. (c) The stored mismatched input selected by the evaluation rule: the first subsequent item with a different `image_id`, here `assign_eval_297046_k2`. The query and candidate vocabulary remain those of the item in panel (a).

language-model scoring calls. No isolated Hungarian timing artifact exists, so no numerical claim about Hungarian latency is made.

G Structural Property of FGA

Proposition. Suppose every candidate column has a strict maximum and the independent choices

$$\hat{\pi}_{\mathrm{Local}}(i) = \arg\max_j S_{ji}$$

form a permutation of the boxes. Then this permutation is the unique optimum of the FGA objective. If column maxima contain ties and a selection of maxima forms a permutation, that permutation is an FGA optimum but need not be unique.

Proof. Let $\hat{\pi}$ denote the permutation formed by the candidate-wise maxima and let ρ be any feasible permutation. For every candidate i,

$$S_{\hat{\pi}(i),i} \geq S_{\rho(i),i}.$$

Table 9: Reference-implementation inference cost. Times are seconds per item (median [min, max]) over six items per k, with two synchronized timed passes; model loading is excluded. No isolated Hungarian timing was recorded.

Model	k	FGA forwards	FGA time	Greedy time
Qwen2.5-VL-7B	2	4 full	1.47 [1.39, 1.70]	1.18 [1.09, 1.29]
Qwen2.5-VL-7B	3	9 full	4.07 [3.46, 4.38]	1.57 [1.51, 1.67]
Qwen2.5-VL-7B	4	16 full	7.82 [6.75, 9.49]	2.04 [1.93, 2.16]
InternVL3.5-4B	2	1 vision + 4 cells	3.46 [1.29, 7.33]	2.46 [1.76, 3.82]
InternVL3.5-4B	3	1 vision + 9 cells	8.71 [8.14, 19.46]	3.17 [2.75, 4.90]
InternVL3.5-4B	4	1 vision + 16 cells	22.98 [2.19, 35.29]	4.31 [2.71, 5.38]

Summing these inequalities over candidate columns gives

$$\sum_i S_{\hat{\pi}(i),i} \geq \sum_i S_{\rho(i),i},$$

so $\hat{\pi}$ is globally optimal. Under strict maxima, any $\rho \neq \hat{\pi}$ differs on at least one candidate and the corresponding inequality is strict, giving a strictly smaller objective. With ties, the summed inequality remains non-strict and multiple optimal permutations may exist.