ONE FAILURE, MANY REPAIRS: MODULE-LEVEL REPAIR SPACES IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a vision–language model (VLM) gets a question wrong, is there only one internal way to make it right? From one LLaVA-OneVision-0.5B checkpoint (Base) we fine-tune, with each of two training seeds, all seven non-empty subsets of the vision encoder (V), projector (P), and language model (L) under identical data, exposure, and example order, and trace every error on 13 tasks of 2D geometry through the 2^3 design; since untrained modules stay bitwise equal to Base, trained modules can also be transplanted, reverted, and recombined. We ask three questions (values: seed 1101/1104). *RQ1: which repairs exist?* Of Base errors, 77.9/76.8% are repaired by some single-module adaptation and 45.2/43.5% by two or three different ones (44.0/50.0% when every model's task-level answer prior is removed); repairs that only a pair or the triple make are rare and form a different, non-additive regime. *RQ2: what does joint training build?* It mostly reorganizes available repairs instead of adding new ones (at most 7.0/8.5% of a joint set's repairs are made by no single set); V co-trained with L keeps only 38/39% of its stand-alone gain; the synergy of the analysed V+L pair-emergent repairs mostly survives recombination across independent runs, so it mostly does not require run-private code; and the training path can move which module carries a repair, at an accuracy cost when V trains first. *RQ3: why are several repairs available?* In the case analysed in depth (G02), Base already encodes the needed information, and V-only and L-only adaptation fix the failure along different causal routes, V-only through image keys and values that the frozen language model reads, L-only by changing L's own text computation; both meet in one late decision state, so where parameters adapt is not where the repair is computed. A behavioural failure need not correspond to a unique defect or repair location.

1 INTRODUCTION

Current VLMs fail simple 2D judgments, such as counting where two lines cross, deciding whether two circles touch, or comparing lengths (Rahmanzadehgervi et al., 2024; Kamoi et al., 2025; Xing et al., 2026). Such failures can often be removed by fine-tuning, and a line of work asks which module to tune (Liu et al., 2023; Karamcheti et al., 2024; Tong et al., 2024a; Li et al., 2025). This presupposes that a failure has a location and that the module to tune is where the defect sits.

We ask a more basic question: *when a VLM gets an item wrong, is there only one internal way to make it right?* A Base failure might be fixed by changing only the projector, only the vision encoder, or only the language model, or only when two or three modules change together. We call each such way a *module-level repair solution* and the set of solutions for a failure its *repair space*.

RQ1 (repair space). Which module-level repairs exist for the same failure: single, several alternative single-module repairs, pair- or triple-emergent, lost when more modules adapt, or none?

RQ2 (joint repair realization). When several repairs are available and modules learn together, what does joint training build (retention, winner-take-all reallocation, division of labour, complementary or co-adapted code, new capability), and why does it realize one implementation rather than another?

RQ3 (availability). Why does one failure admit several repairs? We treat the candidate explanations as non-exclusive families and test those the design reaches: (A) the information is already present but under-used; (B) several computational leverage points reach the same decision; (C) modules

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

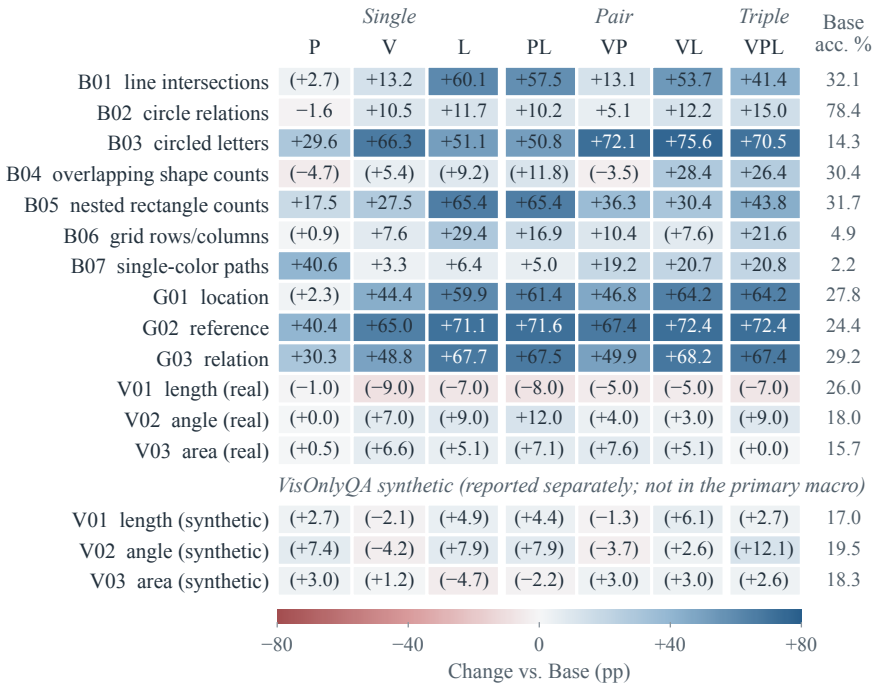

Figure 1: Factorial repair landscape: change in scene-first accuracy relative to the shared Base checkpoint (percentage points) for each trainable set, training seed 1101. Cells show net accuracy change, which mixes repaired Base errors with newly introduced ones; item-level repair is analysed in Section 4.1. Blue marks a gain and red a loss; the scale is symmetric about zero (± 80 pp) and every cell is annotated, so the sign is readable without colour. Parenthesized values: the 95% scene-family bootstrap interval of the change (2,000 replicates, scene families resampled within the task) includes zero (29 of the 91 primary cells). Right column: Base accuracy (%). Bottom block: VisOnlyQA synthetic splits, reported separately and excluded from the primary macro. Intervals are not adjusted for multiplicity.

compensate at their interfaces; (D) several parameter changes realize the same function; (E) different task strategies or cues; and (F) artifacts that mimic multiplicity.

We use a complete 2^3 factorial design: from one LLaVA-OneVision-0.5B checkpoint (Base) we train every non-empty subset of $\{V, P, L\}$ with the same data, loss, schedule, exposure, and example order. Every Base error can then be traced through all seven sets; a set leaves each module it does not train bitwise equal to Base, so trained modules can be transplanted into Base (sufficiency), reverted inside a joint model (necessity), and recombined across sets and training runs; and differences in which module carries a repair are not confounded with data or exposure. Our conclusion is that *a behavioural failure need not correspond to a unique internal defect or a unique repair location*.

Contributions. (i) An item-level vocabulary for module-level repair (emergence classes, multiplicity, lost repair, transplant sufficiency, necessity, recombination, and an outcome taxonomy of joint repair) that turns a factorial fine-tuning design into a map of repair solutions (Section 3.3). (ii) RQ1: repair spaces for 13 tasks in two seeds: most failures have one or several single-module repairs, multiplicity survives removing answer priors as a count (companion analysis), and multiple-single and pair-emergent repairs behave as different regimes (Section 4.1). (iii) RQ2: evidence that joint training reallocates more than it creates, that emergence and ownership differ, the synergy of the analysed V+L pair-emergent repairs mostly does not require run-private code, interface incompatibility explains the loss of stand-alone repairs only partly, and the training path steers ownership at an accuracy cost (Section 4.2). (iv) RQ3: causal evidence that on G02 V-only and L-only repairs take different routes to one late decision state, with probes, head ablations, margin decompositions, a pair-emergent case study, and cue controls that narrow, not exclude, alternatives (Section 4.3).

2 RELATED WORK

Which modules to adapt. VLM recipes differ in which modules they update: LLaVA kept the vision encoder frozen (Liu et al., 2023), LLaVA-OneVision trains the projector, then the full model (Li et al., 2025), and updating the vision encoder hurt, especially on localization, in Prismatic (Karamcheti et al., 2024) but generally helped in Cambrian-1 (Tong et al., 2024a). Tuning only selected language-model projections learns new skills with less forgetting (Zhu et al., 2025). For geometry, tuning only the language model performs comparably to full tuning on VisOnlyQA (Kamoi et al., 2025) and gives larger gains than tuning the vision encoder or projector on GePBench (Xing et al., 2026). Outside VLMs, the best layers to tune depend on the type of distribution shift (Lee et al., 2023). These studies ask where tuning helps most on average; we ask which failures each set repairs, whether different sets repair the same ones, and which module carries a repair under joint training.

Geometric perception in VLMs. BlindTest (Rahmanzadehgervi et al., 2024), VisOnlyQA (Kamoi et al., 2025), and GePBench (Xing et al., 2026), the sources of our tasks, report that VLMs struggle with judgments about overlapping or nearby primitives, lengths, angles, areas, and shape positions. Linear probes on features before and after the projector nearly solve two BlindTest tasks (Rahmanzadehgervi et al., 2024), and VLMs score well below direct readouts of their own vision encoders (Fu et al., 2025), whereas Tong et al. (2024b) relate VLM failures to visual patterns that challenge CLIP encoders (Radford et al., 2021). Whether a failure is limited by representation, by the interface, or by readout is therefore open; we test these hypotheses per task (Section 4.3). Probe accuracy alone does not show that a model uses the information (Hewitt & Liang, 2019; Belinkov, 2022).

Model stitching and module exchange. Model stitching connects the lower layers of one network to the upper layers of another through a trained layer to compare representations (Lenc & Vedaldi, 2015; Bansal et al., 2021). Our transplants need no adapter: every set starts from one checkpoint and leaves untrained modules unchanged, so a trained module can be placed into Base context directly, and the question is not representational similarity but which module is sufficient or necessary for a repair. Fine-tuning can enhance an existing circuit (Prakash et al., 2024) or learn a minimal wrapper over existing capabilities (Jain et al., 2024). Differences between fine-tuned and pretrained weights can be added and negated as task vectors (Ilharco et al., 2023); we analyse them per module.

Multiple solutions and loss landscapes. Minima of neural network losses are connected by low-loss paths (Garipov et al., 2018; Draxler et al., 2018), and networks trained from a shared state can become linearly connected (Frankle et al., 2020). Fine-tuned text classifiers from one pretrained model fall into distinct basins that correspond to different generalization strategies (Juneja et al., 2023), and underspecified pipelines return predictors that are equivalent in training-domain performance but behave differently elsewhere (D'Amour et al., 2022). Loss landscapes can be visualized along chosen directions (Li et al., 2018). We study multiplicity at an architectural level: solutions that differ in which module was allowed to change, compared item by item.

Mechanistic analyses of VLMs. Causal tracing has been adapted to BLIP and multimodal LLMs (Palit et al., 2023; Basu et al., 2024), and NOTICE corrupts images with semantic minimal pairs (Golovanevsky et al., 2025). In LLaVA models, object information concentrates in object-specific visual tokens read out at the last position (Neo et al., 2025), and visual information reaches question tokens in lower and middle layers (Zhang et al., 2025). Object localization, spatial-reasoning success, and spatial features that emerge during multimodal fine-tuning have been tied to small sets of attention heads or to attention on object locations (Schaumlöffel et al., 2026; Chen et al., 2025; Naghashyar et al., 2026). These analyses study released checkpoints or a single training pipeline; we compare checkpoints that differ by design in which modules adapted.

3 SETUP

3.1 MODEL, TRAINABLE SETS, AND TRAINING

The model is `llava-onevision-qwen2-0.5b-ov-hf` (Li et al., 2025): a SigLIP-family vision encoder (Zhai et al., 2023), an MLP projector, and a Qwen2 language model (Yang et al., 2024).

V comprises the vision tower; P the multimodal projector and the learned image-newline parameter; L the language model, including token embeddings and output head. Vision-tower parameters that do not feed the features the model uses (final layer norm, pooling head) are always frozen. The seven trainable sets are the *single* sets P, V, L, the *pair* sets PL, VP, VL, and the *triple* set VPL.[1] Gradients still flow through frozen modules, so in V, VP, and VL they reach V through frozen downstream modules. Trainable parameter counts and learning rates differ (L: 5×10^{-6}; V and P: 2×10^{-6}); contrasts estimate the effect of *allowing* a module to adapt under this one recipe, not a capacity- or learning-rate-matched effect of location (no learning-rate-ratio retrains were run).

Every run trains for 3 epochs over the same 26,000 rows (2,000 per task) with batch size 32, AdamW (Loshchilov & Hutter, 2019) without weight decay, learning rates 2×10^{-6} (V), 2×10^{-6} (P), and 5×10^{-6} (L), a cosine schedule after 74 warmup updates, FP32 weights with bfloat16 autocast, and gradient clipping at 1.0. The loss is the cross-entropy of the answer and end-of-sequence tokens. Within a seed all sets see the same examples in the same order (2,439 updates, 78,000 exposures); there is no checkpoint selection or hyperparameter search, and the final checkpoint is used. Checkpoints are kept at updates 0, 244, 610, 1,220, 1,830, 2,439 for the training-dynamics analysis. We train with two seeds, 1101 and 1104, whose 7 runs each completed with these counts (Section 3.4).

3.2 Tasks, evaluation, and scoring

The 13 tasks are seven BlindTest tasks (Rahmanzadehgervi et al., 2024) (B01 line intersections, B02 circle relations, B03 circled letters, B04 overlapping-shape counts, B05 nested-rectangle counts, B06 grid rows/columns, B07 single-colour paths), three VisOnlyQA geometry tasks (Kamoi et al., 2025) (V01 length, V02 angle, V03 area), and three GePBench tasks (Xing et al., 2026) (G01 location, G02 reference, G03 relation). Training rows for VisOnlyQA come from the authors' synthetic training splits; BlindTest and GePBench rows are generated from the authors' task definitions and generators, with scene families assigned to a split before question and style variants are generated. Evaluation uses the benchmarks' official test rows, which are never trained on. For each row the fixed set of legal answers is ranked by the summed conditional log likelihood of each answer string, so numbers are not comparable to generation-based leaderboards. An item is *correct* if the top-ranked answer is the gold answer. The primary metric is the scene-first 13-task macro accuracy: correctness is averaged within scene families, then over families within a task, then over tasks (9,452 official rows, with VisOnlyQA's real split); its synthetic split (300 rows) is reported separately.

3.3 Item-level repair

Let $c_S(i) \in \{0, 1\}$ be the correctness of official item i under set S, and $c_0(i)$ that of Base. The unit is a Base error, $i \in \mathcal{E} = \{i : c_0(i) = 0\}$. Singles are $\mathcal{S}_1 = \{P, V, L\}$, pairs $\mathcal{S}_2 = \{PL, VP, VL\}$.

Repair emergence. A Base error is *single-emergent* if $\max_{S \in \mathcal{S}_1} c_S(i) = 1$; *pair-emergent* if every single fails and $\max_{S \in \mathcal{S}_2} c_S(i) = 1$; *triple-emergent* if every single and every pair fails and $c_{VPL}(i) = 1$; and *unrepaired* if all seven sets fail. The four classes partition $\mathcal{E}$.

Pair-specific versus globally new repairs. For a pair XY with third single Z, the repair is *pair-specific* if $c_X(i) = c_Y(i) = 0$ and $c_{XY}(i) = 1$, and *globally new* if additionally $c_Z(i) = 0$. A pair-specific repair can thus be one the third single already reaches. Likewise, VPL has a *pair-relative gain* if every pair fails and $c_{VPL}(i) = 1$, and a *globally new* one if every single fails too.

Repair multiplicity. $n_1(i) = c_P(i) + c_V(i) + c_L(i)$ counts the alternative single-module repairs of a Base error (0: none, 1: unique, 2–3: alternatives); $n_2(i)$ is defined in the same way over pairs. Because overlap between two repair sets grows with their marginal repair rates, the observed overlap is compared with a null that preserves every set's repair count: within each task, one set's repairs are permuted over the Base errors; the expected overlap is computed in closed form and intervals come from a scene-family cluster bootstrap. The claim at stake is that the same failure can be repaired by different single-module adaptations, not that a given overlap is evidence of a shared mechanism.

Lost repair. For each of the nine edges $S \subset S'$ of the lattice where S' adds one module (e.g. V⊂VL), each item is *gained* ($c_S = 0, c_{S'} = 1$), *retained* (1,1), *lost* (1,0), or *still wrong* (0,0). The *lost-repair rate* of an edge is the share of Base errors repaired by S that S' no longer repairs.

[1] Run labels in the records: A=P, B=PL, C=VPL, D=VP, E=V, F=L, G=VL.

Capability emergence versus role reallocation. Joint adaptation creates new behavioural capability on pair- and triple-emergent items; on items a single set already repairs, it may still change which module carries the repair, which we measure with transplants. Write a model as a triple (V^a, P^b, L^c) of modules taken from the checkpoints of sets a, b, c (0 for Base). *Sufficiency* of module M trained in set S is the repair that M^S achieves in Base context, e.g. (V^{VL}, P^0, L^0); the 12 pairs of a module and a set that trains it form a transplant matrix. *Necessity* of M in a joint set is the repair lost when M alone is reverted to Base, e.g. (V^0, P^{VPL}, L^{VPL}). *Recombination* combines modules trained in different sets, e.g. (V^V, P^0, L^L). The *functional ownership* of a set's repairs by module M is the share of that set's repaired Base errors that the transplanted M^S also repairs, and its *retention* in a joint set S is the share of the single set M's repairs that M^S also makes. A repair of a joint set S is *new* if no single set repairs the error; *same-carrier* if, for some module M of S, both the single set M and M^S repair it; *reallocated* if single sets of S's modules repair it but none of those modules does as M^S; and *outside* otherwise (only an untrained module's single set repairs it).

3.4 Scope

Each seed is analysed separately with the same code and the shared Base. Values are seed 1101/seed 1104, with 95% percentile intervals of a scene-family cluster bootstrap within the seed (2,000 replicates; one draw for Base and every set), not adjusted for multiplicity; two seeds show which statements replicate. A statement without a seed holds in each seed; where they differ we say so. RQ1 and behavioural RQ2 also ran on a third seed, 1102: 84 of 95 two-seed checks hold (Appendix H); mechanistic analyses stay on two seeds. Item-level analyses, designed after seeing the official results and using models outside the original design, are exploratory.

4 Results

4.1 RQ1: which repairs exist for the same failure?

In both seeds every trainable set improves the primary macro over Base (25.8%; Appendix Tables 1 and 2), by +12.1 [+10.0, +14.5] and +12.5 [+10.3, +14.9] pp for P up to +34.3 [+31.6, +37.0] and +35.0 [+32.1, +37.9] pp for VPL, and every interval of a gain excludes zero. The four sets that train L reach 58.8–60.0/56.2–60.8%, above V and VP, which are above P. L alone is the best single set, and VPL exceeds it by +0.5 [−1.5, +2.4] and +4.5 [+2.5, +6.6] pp: within noise in seed 1101, where no difference among the four L-trained sets has an interval that excludes zero, but not in seed 1104, where L alone is below PL, VL, and VPL.

Module effects are strongly non-additive (Table 1). L has the largest main effect (+18.7/+18.5 pp), followed by V (+9.2/+10.0 pp) and P (+3.5/+4.1 pp); in both seeds both gaps in this ordering have intervals that exclude zero. All pairwise interactions are negative (INT_{VP} −4.3/−5.3, INT_{VL} −17.2/−12.5, INT_{PL} −7.1/−6.7 pp), the three-way interaction is positive (+11.5/+9.6 pp), and no contrast's interval includes zero. With one scene draw shared by both seeds, the only contrast whose seed difference excludes zero is INT_{VL} (+4.8 pp [+2.9, +6.7]), and L alone gains less in seed 1104 (−3.3 pp [−4.6, −2.0]). Once L is trained, also training V or P changes the macro by at most +1.3/+4.0 pp, less than the same module adds without L (Appendix A). Subadditivity admits two readings, which the item-level analyses separate: V and L may repair *the same* failures, or joint training may realize only one of several available repairs.

The landscape differs sharply across tasks (Figure 1; seed 1104: Appendix Figure 5); in both seeds the largest mean gains are on G02, B03, and G03. On B01 (line intersections), L gains +60.1/+53.8 pp, against +13.2/+13.0 pp for V and +2.7/+2.0 pp for P; on B01 and B05, L and PL gain more than VL and VPL, so opening V alongside L lowers accuracy there. On B07 (single-colour paths), P has the largest gain (+40.6/+41.4 pp), while V, L, and PL gain at most +6.4/+13.7 pp. On G02 (reference), from a Base accuracy of 24.4%, V gains +65.0/+66.3 pp, L +71.1/+71.7 pp, and P +40.4/+40.8 pp: each single module on its own closes more than half of the gap to perfect accuracy. On B03 (circled letters), every set that trains V gains at least +66.3/+64.6 pp, more than PL and L (at most +51.1/+49.4 pp). On V01 (length on real diagrams) no set improves (Appendix A).

Most of each gain requires the row's own image (Appendix B): with a blank image no set gains more than 7.4/7.7 pp over Base, and with another row's image (10 tasks) at most 17/18% of a set's gain survives, except for P (49/47%).

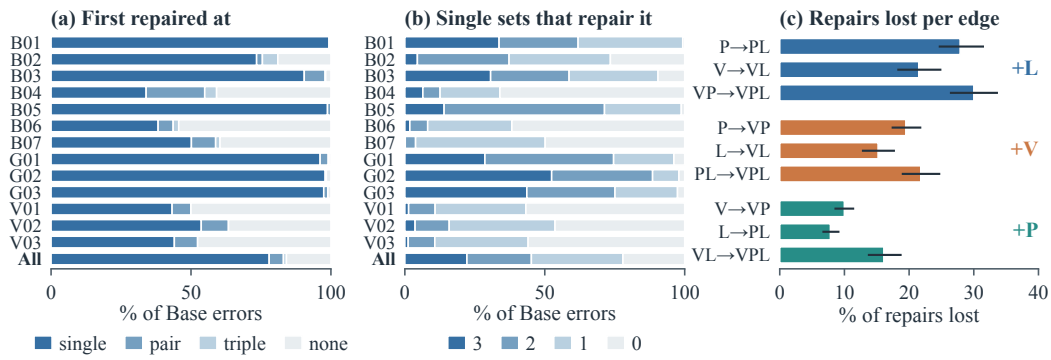

Figure 2: Repair spaces of the 13 tasks (RQ1), official predictions of training seed 1101. (a) Level at which each Base error is first repaired. (b) Number of single sets that repair it. (c) Share of a set's repairs lost when one module is added, coloured by the added module, with 95% scene-family cluster-bootstrap intervals.

Most failures have single-module repairs, and many have several. Of the shared Base's 6,390 errors (9,452 rows), 77.9 [75.2, 80.5] and 76.8 [74.2, 79.3]% are repaired by at least one of P, V, L, 5.1/5.8% first by a pair, 1.1/1.9% only by VPL, and 15.9/15.5% by no set (Figure 2a). Two or all three single sets repair 45.2/43.5%, more than the 32.7/33.3% with a unique one (Figure 2b; on G02, 52.5/53.7% by each of the three). Pair multiplicity is common ($n_2 \geq 2$: 64.1/61.1%), but a pair containing a repairing single is no new solution; counting *minimal* repairing sets (no proper subset repairs), 50.4 [47.7, 53.0] and 50.7 [48.0, 53.3]% have two or more (Appendix C), and the one triple, VPL, realizes its repairs in several ways (Appendix G.1). V and L co-repair about as often as chance given each set's repair count per task (observed/expected 1.020/0.997); most of the multiple-single count is thus task-level arithmetic (the per-task pigeonhole floor accounts for 75/69% of it; Appendix D.1), and item-level evidence rests on G02 (Section 4.3). Adding a module also loses 18.0/14.7% of the smaller set's repairs (nine edges), most when it is L (26.5/25.5%; Figure 2c). Tasks have distinct phenotypes (Appendix C, Table 38).

Multiplicity is not an answer-prior artifact at the level of counts (Appendix D.1). A single task-level answer bias can fix many errors of a skewed Base. Redefining Base errors and repairs with each model's own answer prior, cross-fitted on held-out scene families, leaves multiplicity at 44.0/50.0% (raw 45.2/43.5%; blank-image contextual calibration 37.6/36.5%). Of the raw multiple-single items, 31/30% are Base answer-prior errors that recalibrating Base alone fixes. As an upper bound on the prior share, dropping every repair that a pure answer-prior shift of Base reproduces leaves 18.1/16.6%, with 75/72% of the drop on B01; G02 keeps 60/60%.

Multiple-single and joint-only repairs are different regimes (Appendix P). Multiple-single items have several single-module routes before any joint training. For pair-emergent items we decompose each joint repair's answer margin into the stand-alone effects of its modules (each transplanted into Base) and their interaction: sub-threshold effects that add up to a repair are rare (1/1% of repaired instances); instead one jointly trained module repairs the item alone (48/42%) or the modules interact superadditively (46/52%). Triple-emergent items are essentially V×L: VPL's V and L without its P repair 99/98% of them. Outside B04, pair- and triple-level items replicate poorly across seeds (pair-level 26.2/20.4%, on B04 70.6/73.2%; Appendix D.7), so some may be single-set false negatives of one run. On unrepaired items the joint sets move the margin away from gold (80/76%), and the stand-alone effects would add up to a repair in 2/3%: these failures are not repaired under the current adaptation procedure (data, budget, optimizer), not shown to be unrepairable.

4.2 RQ2: WHAT DOES JOINT TRAINING BUILD FROM THE AVAILABLE REPAIRS?

All 26 models of each seed (sets, transplants, reversions, recombinations) were re-scored on 9,452 rows (6,383 Base errors; Appendix F); M^S denotes the model with M from S and Base elsewhere.

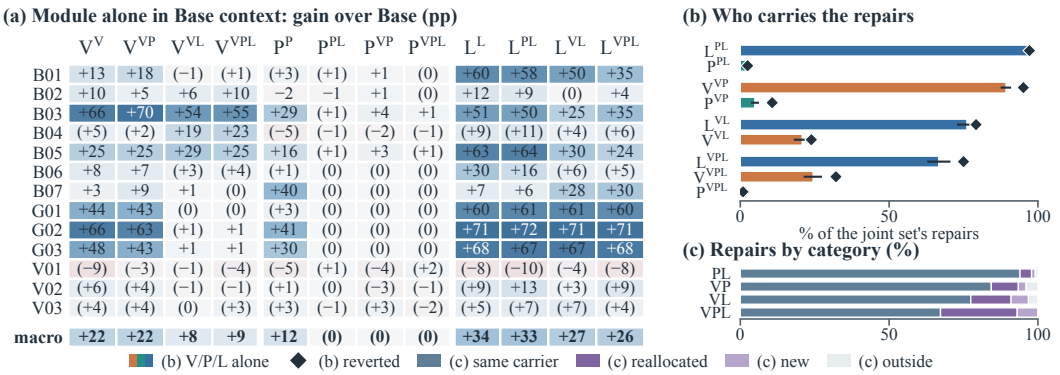

Figure 3: What joint adaptation changes (RQ2; seed 1101, re-scored, 13 tasks). (a) Gain over Base (pp) of M^S, module M trained in set S alone in Base context; colours as in Figure 1; parentheses: the 95% interval includes zero. (b) Share of a joint set's repairs that the module makes alone (bar, ownership, 95% interval) or loses on reversion (diamond, necessity). (c) Categories of Section 3.3.

Joint training reallocates more than it creates (Figure 3). Co-trained with L, V loses most of its stand-alone function (V^{VL} gains +8.5/+8.4 pp, 38/39% of V^V's gain), P loses it with any partner, and L keeps most of its own. Each joint set's repairs are carried mostly by one module: L^S alone makes 97.5/98.1, 76.1/70.0, and 66.7/73.4% of the repairs of PL, VL, and VPL, and V^{VP} 89.3/88.0% of those of VP. New repairs are 1.3–7.0/2.1–8.5% of a joint set's repairs and reallocated ones 3.9–25.8/4.3–17.7%; in both seeds and every joint set, reallocation outnumbers new capability (after answer-prior adjustment it still does for VL and VPL, not always for PL; Appendix D.1).

An outcome taxonomy classifies most joint repairs (Appendix G.1). In both seeds it sorts each joint repair by which of its modules repaired it as single sets and which repair it alone after joint training. In VL, 41.7/36.3% of the repairs are winner-take-all (several single-module repairs existed; one module carries the joint repair), 24.4/20.8% keep their sole single-module carrier, 6.3/9.4% keep two independent carriers, and 5.0/5.6% gain a redundant duplicate; a soft division of labour, in which stand-alone effects add up to the repair, accounts for 0.8/0.9%, and globally new capability for 5.8/7.2%; lost repairs and antagonism are identified too. *Emergence is not ownership:* no single set of the same run learned a pair-emergent repair, yet the solution can belong to one module, as PL's L alone repairs 88/80% of PL's pair-emergent items and VP's V 57/44% of VP's.

Jointly trained but not run-private code, for V+L. Neither V nor L repairs the 372/421 pair-specific items of VL (Base, V, and L wrong, VL right; Section 3.3). Combining VL's V with the L of VL trained with the other seed keeps 83/80% of them (the reverse 85/82%), against 39/41% with the single-trained L of the same seed (Appendix G.2). The synergy of the analysed V+L pair-emergent repairs thus mostly does not require run-private code; it needs a partner trained in the joint regime, also from the other seed's run, though the two runs' modules are not exactly interchangeable (13-task accuracy −3.0/+3.2 pp relative to VL). For PL, VP, and VPL no such recombination exists, and complementary and co-adapted repairs stay inseparable (18.5/9.7% of VPL's repairs).

Interface incompatibility is partial on G02. V^{VL} in Base context barely repairs G02 (25.2/25.2% against 90.7/90.7% for V^V). Retraining only the projector between V^{VL} and Base L for 244 updates beats an equally trained projector with Base V by +8/+10 pp but recovers only 0.20/0.23 of the gap on G02 (0.58/0.56 of the row-level 13-task gap; Appendix G.3).

Joint training settles on the L-side implementation early. In both seeds the L update of VL points along the L-only update ($\cos = 0.72/0.71$), its V update does not (0.23/0.23), and co-training L shrinks the V and P updates (norm ratios 0.72/0.72, 0.18/0.18; Appendix I). A control that keeps a frozen L inside the global gradient clip leaves V's update aligned with the V-only update (cosine 0.90/0.91, against 0.27/0.26 for VL at update 244; not a clipping artifact in either seed; Appendix G.4), but why V's stand-alone repairs are lost remains unresolved (item-level gradient starvation is not supported; Appendix G.5). The carrier is fixed early: L_t^{VL} alone makes 74.2–83.0/68.9–77.0% of VL's repairs at every stored update from the first (10% of training), and V does not first build a repair that L takes over (Appendix J).

The training path steers ownership, at an accuracy cost (Appendix K). Retraining VL with only V (*V first*) or only L (*L first*) trained for the first 1,220 updates moves the carrier in both seeds: V alone makes 88.1/71.7% of V first's final repairs and L alone 96.6/96.3% of L first's (unchanged replicate: V 22.3/21.9%, L 73.8/75.1%). Against the replicate, V first changes accuracy by −9.6 [−11.6, −7.6] and −4.8 [−7.1, −2.4] pp, beyond the run-to-run floor, and L first by −1.6/−2.6 pp: not detectably in seed 1101, but with an interval below zero in seed 1104 ([−4.6, −0.6]). The path can causally move which module owns an available repair, but not every ownership is reached at equal accuracy (possibly because the later module trains only in the schedule's decaying half).

4.3 RQ3: WHY DOES ONE FAILURE ADMIT SEVERAL REPAIRS?

We separate four levels: *failure locus* (what Base lacks or misuses), *reach* (which modules can reach a compensating change), *realization* (which reachable repair training builds; RQ2), and *measurement* (whether an apparent repair is an artifact); families A, B–D, and E–F address locus, reach, and measurement. Locus and reach need not coincide: on G02 V's repair reroutes the frozen language model; on B04 some repairs need two modules. Appendix E gives each family's status; task cards relate phenotypes to task demands only descriptively (Appendix L).

(A) The information is present but under-used (supported on G02, partial on B02). Probes fit on training rows and tested on official rows (Appendix M) show that neither G02 nor B02 is representation-limited: in Base, the G02 target's type is decodable from its own image tokens (balanced accuracy 0.66, chance 0.07) and the B02 signed gap from the circles' contact region (R^2 0.92), and Base's final prompt token carries the G02 answer (0.58 over layers 12–23, chance 0.25) that its own answer misses (23.5% correct); a pre-set cloze test of a binding failure was not confirmed (Appendix D.3). Decodability does not make a module a carrier: V^{VL} raises the target's decodability (+0.14/+0.12) yet alone barely repairs G02.

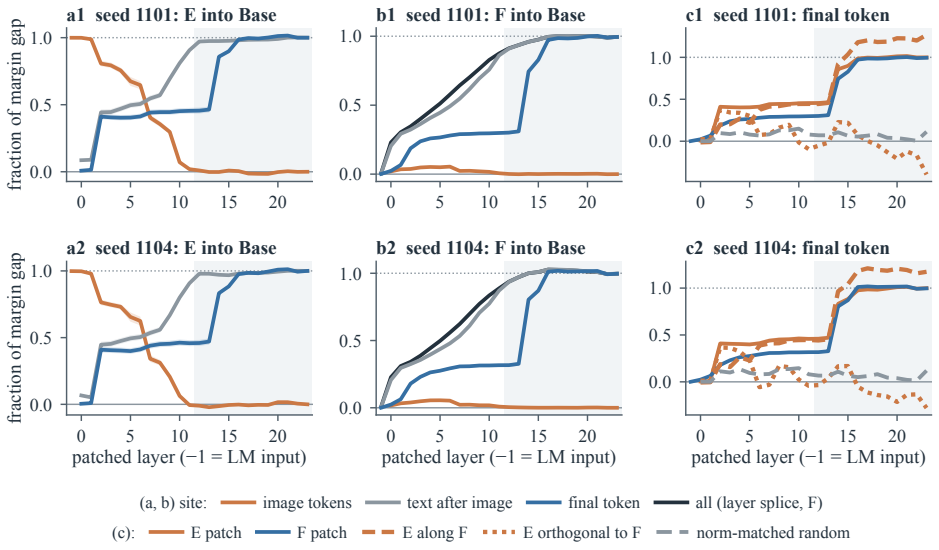

Figure 4: Causal patching on G02 (RQ3; top: seed 1101, bottom: seed 1104), items that Base fails and both V (run E) and L (run F) repair. Fraction of a model's answer-margin advantage over Base recovered when its states after layer l (−1: language-model input) at the named positions are put into Base. (a) V into Base: image tokens, text tokens after the image, final token. (b) L into Base: all positions (layer splice), image tokens, text after the image, final token. (c) Final token only: V's and L's patches, V's final-token change projected on L's same-item change and its orthogonal rest, and a norm-matched random control. Lines: means with 95% scene-family intervals; shaded: layers 12–23.

(B, C) Different causal routes; on G02 they meet in one decision state (supported on G02). On the 207 (seed 1101) and 212 (seed 1104) G02 items that Base fails and both V and L repair, we patch states between Base and each model and measure the fraction of the answer-margin gap recovered

(Figure 4; Appendix O). The V-only repair sits in the visual token field but not at the target: V's projector features at non-object positions give Base 0.85/0.88 of V's advantage, those at the target's tokens 0.08/0.08. The frozen language model reads it through image keys and values jointly (keys alone 0.30/0.29, values alone 0.34/0.34, both 1.00/1.00), half of it by layer 6/6. The L-only repair's changes at image-token positions carry none of its effect (at most 0.05/0.06); it is computed in the text positions. Both hand the decision to the final token at the same layer (14/14 for V, 14/14 for L), where V's change lies along L's (projecting it on L's recovers 1.04/1.03 of V's effect). The V-only change read from image keys and values in the lower layers thus surfaces on the frozen language model's late final-token query (Q/K/V decomposition, layers 12–23: both repairs raise the target-over-distractor attention score through that query, V +0.14/+0.13, L +0.25/+0.28; no late key side is detectable; Appendix N). The two repairs boost largely the same heads (4/4 of each top five shared, seed 1101/1104), but ablating L's top ten removes only 0.19/0.21 of V's and 0.14/0.14 of L's margin gain: the repair is distributed. On B02 the repairs also take different routes and reach the final token at the same layer but, unlike on G02, in different states (Appendix O.4).

(D) Coexisting implementations (supported on G02), but no first-order prediction (partial). On G02 the V-only and L-only repairs show no accuracy dip between them and together are no better than L alone (additivity $-67.6/-69.9$ pp; Appendix Q): geometry evidence that several implementations coexist, not an explanation. A first-order test at Base (575 items; per block, margin gradient times realized update) does not separate the regimes: L is judged locally capable in every class, unrepaired items included (100/100% of those sampled), so the Base Jacobian does not predict the repair that training realizes (Appendix P).

Pair-emergent V+L repairs: superadditive, mostly not run-private (supported for VL in both seeds; B04 is one case). Without a stand-alone carrier, VL's pair-emergent repairs are superadditive (100/100%) with a fixed sign pattern: VL's L alone moves the margin away from gold (92/97%), its V alone helps little, and together they repair; a superadditive interaction (not always a repair) persists with a single-trained partner of the same seed in 97/96% of them (partner-specific: 3/4%). In the one task with enough pair-emergent items for a case study, B04 with VL (141/135 items; Appendix P), odd gold counts are repaired by the single-trained V and L together (82/73%), two separable deficits that compose, whereas even counts are not (0/0%) but are by VL's V with the co-trained L of the other seed (97/77%): co-adaptation to a jointly trained partner, not to one run (an L trained without V rarely answers even counts; Appendix P.3).

(E, F) Cues and artifacts are narrowed, not excluded (partial; Appendix D.3). On G01–G03 Base answers (A) on every row, and V's and L's repairs follow the option content under cyclic option permutations (on G02, 92/90% and 100/99% of co-repaired items stay right under all four orders); under template-matched image swaps on G01 and G02 (13/15 G02 recipients, seed 1101/1104) they follow the image, not the text. On G03, however, L keeps its answer under a random image swap on 84/90% of co-repaired rows, so an image-light text route stays open there; counterfactual presence and symmetry interventions are untested, and the repairs are bound to the trained answer format. Alternative strategies and shortcuts are therefore not ruled out in general.

Synthesis. A behavioural failure need not identify a unique defect, repair location, or implementation: failure locus, parameter-adaptation location, computation location, and ownership can all differ (G02 dissociates the first three; in RQ2 ownership moves with the training path). Some failures admit interchangeable single-module routes to one decision state (G02); others need complementary cross-module changes (B04). Joint training selects, reallocates, or coordinates among these routes.

5 LIMITATIONS

We study one 0.5B model with one fixed recipe and no learning-rate or epoch search, so repair availability is relative to this recipe; route steering uses one unlock point. Two seeds do not estimate seed variance, the item-level analyses are exploratory, and the causal analysis covers one task in depth. Correctness is the top-ranked legal answer, so items near the decision boundary can flip for reasons unrelated to repair; a failed transplant shows only that a module is not sufficient in Base context. All numbers come from official test rows under likelihood ranking, whose pretraining exposure is unknown; evaluations ran on two GPU types, and RQ2 uses re-scored predictions (Appendix F).

ETHICS STATEMENT

The study uses public benchmarks and a public model checkpoint, collects no data from people, and involves no human subjects. BlindTest (MIT), VisOnlyQA (GPL-3.0; its real-split diagrams come from MathVista, CC BY-SA 4.0), GePBench (GPL-3.0; we use its TestMini split), and llava-onevision-qwen2-0.5b-ov-hf (Apache-2.0) are used under their licences; no images are redistributed. Our results concern a small model on synthetic and diagrammatic 2D geometry and should not be taken as evidence about the reliability of VLMs in consequential settings.

REPRODUCIBILITY STATEMENT

Section 3 gives the model, recipe, training seeds, data sizes, scoring, and every item-level definition. Every number in the tables, figures, and text of Section 4 and the task-card table is generated by scripts from hash-checked evaluation records, per-item predictions, the locked task cards, and statistics of the checkpoint weights and activations; the generators re-check each qualitative statement in the text against these records in each seed and fail if one no longer holds where the text states it. Scoring excludes the end-of-sequence token, breaks ties by the first-listed option, and runs under bfloat16 autocast; AdamW uses $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, and a batch of 32 is accumulated from 32 micro-batches of one row; runs interrupted by the scheduler resumed from checkpoints with restored RNG state. Software: PyTorch 2.11.0 (CUDA 12.8) and Transformers 4.51.3; evaluations ran on NVIDIA RTX PRO 6000 Blackwell and RTX 6000 Ada GPUs. An anonymized release of code, configuration, per-item predictions, and the trained checkpoints from which transplants are composed will accompany the paper. Intervals resample test items (scene families) within a seed and do not include run-to-run training variance; a same-recipe rerun of VL differs from the original by +2.2/+0.2 pp in the primary macro. The governed training runs used NVIDIA RTX PRO 6000 Blackwell GPUs, and re-scoring and retrains used NVIDIA RTX 6000 Ada GPUs, with PyTorch 2.11.0+cu128 and Transformers 4.51.3. The training pipeline trained the seven sets for eight seeds (1101–1108) and computed aggregate official evaluations for part of them; this paper analyses seeds 1101 and 1104 at item level, and 1102 for the behavioural checks of Appendix H; the other seeds were not analysed at item level and are not reported. BlindTest and GePBench training rows come from the same generators as their official test rows (disjoint scene families), so held-out does not mean out-of-distribution; VisOnlyQA trains on synthetic and tests on real diagrams.

LLM USE STATEMENT

In this work, the authors used generative AI tools to aid in polishing the writing and improving the readability of the manuscript, and to assist with writing and debugging the experimental code. The authors carefully reviewed, verified, and edited all AI-generated content and code, and take full responsibility for the content and correctness of this final work.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/0dfe31d6e703e138d46a7d2fced38b7c-Abstract-Conference.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? An attention mechanism perspective on focus areas. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9910–9932, 2025. URL https://openreview.net/forum?id=k7vcuqLK4X.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1309–1318, 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3259–3269, 2020. URL https://proceedings.mlr.press/v119/frankle20a.html.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=qQb1JLrwol.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, 2018. URL https://papers.nips.cc/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? A mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, 2025. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024. URL https://arxiv.org/abs/2404.15255.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=hY6M0JHl3uL.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. VisOnlyQA: Large vision language models still struggle with visual perception of geometric information. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=PYHwlyu2fa.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23123–23144, 2024. URL https://proceedings.mlr.press/v235/karamcheti24a.html.

Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=APuPRxjHvZ.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=zKv8qULV6n. arXiv:2408.03326.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Lachin Naghashyar, Hunar Batra, Ashkan Khakzar, Philip Torr, Ronald Clark, Christian Schroeder de Witt, and Constantin Venhoff. Towards understanding multimodal fine-tuning: Spatial features. arXiv:2602.08713, 2026. URL https://arxiv.org/abs/2602.08713.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2856–2861, 2023.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, 2024. URL https://openaccess.thecvf.com/content/ACCV2024/html/Rahmanzadehgervi_Vision_language_models_are_blind_ACCV_2024_paper.html. Linear-probing experiments appear in the extended version, arXiv:2407.06581v6, 2025.

Timothy Schaumlöffel, Martina G. Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31356–31365, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024a. URL https://papers.nips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024b.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to Makelov et al. (2023)'s "interpretability illusion" arguments. arXiv:2401.12631, 2024. URL https://arxiv.org/abs/2401.12631.

Shangyu Xing, Changhao Xiang, Xinyu Liu, Zhangtai Wu, Zhen Wu, Yifan Yue, Yuteng Han, Fei Zhao, and Xinyu Dai. GePBench: Evaluating fundamental geometric perception for multimodal large language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/64463. arXiv:2412.21036.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671. arXiv:2407.10671.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025.

Zhen Zhu, Yiming Gong, Yao Xiao, Yaoyao Liu, and Derek Hoiem. How to teach large multimodal models new skills. arXiv:2510.08564, 2025. URL https://arxiv.org/abs/2510.08564.

A FACTORIAL CONTRASTS AND PER-TASK RESULTS

Table 1 gives the gains and factorial contrasts of Section 4.1 per seed, their mean over the two seeds, and the seed difference, all with one scene-family draw shared by both seeds. Table 2 gives the primary macro accuracy of Base and every set with 95% scene-family intervals, Table 4 the factorial contrasts, and Table 6 per-task scene-first accuracy with the official row and scene-family denominators, for seed 1101; Tables 3, 5, and 7 and Figure 5 give the same for seed 1104.

Simple effects. Without L, training V adds +12.7–+22.8/+11.2–+21.3 pp and training P adds +2.1–+12.1/+2.4–+12.5 pp (seeds 1101/1104), every interval above zero. Once L is trained, also training V or P changes the macro by only -0.8 to +1.3 pp in seed 1101, where none of these four changes has an interval that excludes zero, and by +0.5 to +4.0 pp in seed 1104, where three of the four intervals lie above zero. In both seeds each with-L change is smaller than every without-L change of the same module.

V01. On V01 (length on real diagrams) no set improves: every change is negative (-9.0 to -1.0 and -13.0 to -3.0 pp); with 100 rows, no interval excludes zero in seed 1101 and only that of VL does in seed 1104.

Table 1: Change of the primary scene-first macro accuracy over Base (pp, top) and factorial contrasts (pp, bottom) per training seed, pooled over the seeds (mean of the seeds' estimates), and the seed difference, each with its 95% scene-family bootstrap interval (one draw shared by both seeds; 2000 replicates). Intervals reflect sampling of evaluation scenes, not training-seed variation.

	Seed 1101	Seed 1104	Pooled	1104−1101
P	+12.1 [+10.0, +14.5]	+12.5 [+10.3, +14.9]	+12.3 [+10.1, +14.7]	+0.4 [−0.2, +1.0]
V	+22.8 [+20.1, +25.5]	+21.3 [+18.7, +24.1]	+22.1 [+19.5, +24.7]	−1.5 [−3.1, +0.1]
L	+33.8 [+31.3, +36.4]	+30.5 [+28.0, +33.0]	+32.1 [+29.8, +34.6]	−3.3 [−4.6, −2.0]
PL	+33.0 [+30.5, +35.6]	+31.5 [+29.0, +34.0]	+32.2 [+29.8, +34.7]	−1.5 [−2.8, −0.2]
VP	+24.9 [+22.2, +27.5]	+23.7 [+21.1, +26.4]	+24.3 [+21.8, +26.8]	−1.1 [−2.7, +0.4]
VL	+33.6 [+30.7, +36.2]	+34.5 [+31.9, +37.2]	+34.0 [+31.5, +36.6]	+0.9 [−0.9, +2.7]
VPL	+34.3 [+31.6, +37.0]	+35.0 [+32.1, +37.9]	+34.6 [+32.0, +37.3]	+0.7 [−0.9, +2.4]
ME_V	+9.2 [+7.7, +10.6]	+10.0 [+8.7, +11.4]	+9.6 [+8.3, +10.8]	+0.9 [−0.2, +1.9]
ME_P	+3.5 [+2.7, +4.4]	+4.1 [+3.3, +4.9]	+3.8 [+3.1, +4.6]	+0.6 [−0.2, +1.4]
ME_L	+18.7 [+16.8, +20.5]	+18.5 [+16.7, +20.3]	+18.6 [+16.8, +20.3]	−0.2 [−1.3, +0.8]
INT_{VP}	−4.3 [−5.9, −2.7]	−5.3 [−7.0, −3.7]	−4.8 [−6.2, −3.4]	−1.0 [−2.8, +0.8]
INT_{VL}	−17.2 [−20.0, −14.6]	−12.5 [−15.1, −10.0]	−14.9 [−17.3, −12.5]	+4.8 [+2.9, +6.7]
INT_{PL}	−7.1 [−8.8, −5.6]	−6.7 [−8.3, −5.2]	−6.9 [−8.3, −5.6]	+0.4 [−1.2, +2.1]
INT_{VPL}	+11.5 [+8.5, +14.8]	+9.6 [+6.6, +12.8]	+10.6 [+8.0, +13.3]	−1.9 [−5.1, +1.3]

Table 2: Official behavioral evaluation of training seed 1101 (Base is the single shared pretrained checkpoint). Primary: 13-task scene-first macro accuracy (%): rows are averaged within scene families, scene families within tasks, then tasks (VisOnlyQA real split only). Secondary: row-weighted accuracy over the same 9,452 rows. Δ is computed from unrounded values. 95% intervals are percentile intervals of a scene-family cluster bootstrap within the seed (2,000 replicates; scene families resampled within each task, the same draw for Base and every set); they reflect sampling of evaluation scenes, not variation across training seeds, and are not adjusted for multiplicity.

		Single			Pair			Triple
Trainable set	Base	P	V	L	PL	VP	VL	VPL
run label	–	A	E	F	B	D	G	C
Primary scene-first macro	25.8	37.9	48.6	59.5	58.8	50.6	59.4	60.0
95% CI, lower	23.8	35.8	46.5	57.4	56.7	48.6	57.2	57.8
95% CI, upper	27.8	40.0	50.7	61.7	60.8	52.8	61.3	62.2
Δ vs. Base	–	+12.1	+22.8	+33.8	+33.0	+24.9	+33.6	+34.3
95% CI, lower	–	+10.0	+20.1	+31.3	+30.5	+22.2	+30.7	+31.6
95% CI, upper	–	+14.5	+25.5	+36.4	+35.6	+27.5	+36.2	+37.0
Row-weighted accuracy	32.4	40.8	51.4	73.0	71.3	51.9	73.3	69.5

Table 3: Official behavioral evaluation of training seed 1104 (Base is the single shared pretrained checkpoint). Primary: 13-task scene-first macro accuracy (%): rows are averaged within scene families, scene families within tasks, then tasks (VisOnlyQA real split only). Secondary: row-weighted accuracy over the same 9,452 rows. Δ is computed from unrounded values. 95% intervals are percentile intervals of a scene-family cluster bootstrap within the seed (2,000 replicates; scene families resampled within each task, the same draw for Base and every set); they reflect sampling of evaluation scenes, not variation across training seeds, and are not adjusted for multiplicity.

		Single			Pair			Triple
Trainable set	Base	P	V	L	PL	VP	VL	VPL
run label	–	A	E	F	B	D	G	C
Primary scene-first macro	25.8	38.3	47.1	56.2	57.3	49.5	60.3	60.8
95% CI, lower	23.8	36.2	45.1	54.2	55.2	47.5	58.2	58.5
95% CI, upper	27.8	40.4	49.2	58.2	59.2	51.6	62.4	63.0
Δ vs. Base	–	+12.5	+21.3	+30.5	+31.5	+23.7	+34.5	+35.0
95% CI, lower	–	+10.3	+18.7	+28.0	+29.0	+21.1	+31.9	+32.1
95% CI, upper	–	+14.9	+24.1	+33.0	+34.0	+26.4	+37.2	+37.9
Row-weighted accuracy	32.4	40.7	51.0	68.9	69.5	53.4	72.9	73.3

Table 4: Factorial contrasts of the primary scene-first macro (percentage points, seed 1101) in difference-in-differences units; interactions are full differences-in-differences, not ± 1-coded regression coefficients. Each contrast combines the seven configurations with the shared Base value. 95% intervals are percentile intervals of a scene-family cluster bootstrap within the seed (2,000 replicates; scene families resampled within each task, the same draw for Base and every set); they reflect sampling of evaluation scenes, not variation across training seeds, and are not adjusted for multiplicity.

	Main effects			Interactions			
Seed 1101	ME_V	ME_P	ME_L	INT_{VP}	INT_{VL}	INT_{PL}	INT_{VPL}
Estimate	+9.2	+3.5	+18.7	−4.3	−17.2	−7.1	+11.5
95% CI, lower	+7.7	+2.7	+16.8	−5.9	−20.0	−8.8	+8.5
95% CI, upper	+10.6	+4.4	+20.5	−2.7	−14.6	−5.6	+14.8

Definitions (letters are run labels and denote macro accuracies; A=P, B=PL, C=VPL, D=VP, E=V, F=L, G=VL): $\mathrm{ME}_V = [(E - \mathrm{Base}) + (D - A) + (G - F) + (C - B)]/4$; $\mathrm{ME}_P = [(A - \mathrm{Base}) + (D - E) + (B - F) + (C - G)]/4$; $\mathrm{ME}_L = [(F - \mathrm{Base}) + (B - A) + (G - E) + (C - D)]/4$; $\mathrm{INT}_{VP} = [(D - E - A + \mathrm{Base}) + (C - G - B + F)]/2$; $\mathrm{INT}_{VL} = [(G - E - F + \mathrm{Base}) + (C - D - B + A)]/2$; $\mathrm{INT}_{PL} = [(B - A - F + \mathrm{Base}) + (C - D - G + E)]/2$; $\mathrm{INT}_{VPL} = C - G - D - B + E + A + F - \mathrm{Base}$.

Table 5: Factorial contrasts of the primary scene-first macro (percentage points, seed 1104) in difference-in-differences units; interactions are full differences-in-differences, not ± 1-coded regression coefficients. Each contrast combines the seven configurations with the shared Base value. 95% intervals are percentile intervals of a scene-family cluster bootstrap within the seed (2,000 replicates; scene families resampled within each task, the same draw for Base and every set); they reflect sampling of evaluation scenes, not variation across training seeds, and are not adjusted for multiplicity.

	Main effects			Interactions			
Seed 1104	ME_V	ME_P	ME_L	INT_{VP}	INT_{VL}	INT_{PL}	INT_{VPL}
Estimate	+10.0	+4.1	+18.5	−5.3	−12.5	−6.7	+9.6
95% CI, lower	+8.7	+3.3	+16.7	−7.0	−15.1	−8.3	+6.6
95% CI, upper	+11.4	+4.9	+20.3	−3.7	−10.0	−5.2	+12.8

Definitions (letters are run labels and denote macro accuracies; A=P, B=PL, C=VPL, D=VP, E=V, F=L, G=VL): $\mathrm{ME}_V = [(E - \mathrm{Base}) + (D - A) + (G - F) + (C - B)]/4$; $\mathrm{ME}_P = [(A - \mathrm{Base}) + (D - E) + (B - F) + (C - G)]/4$; $\mathrm{ME}_L = [(F - \mathrm{Base}) + (B - A) + (G - E) + (C - D)]/4$; $\mathrm{INT}_{VP} = [(D - E - A + \mathrm{Base}) + (C - G - B + F)]/2$; $\mathrm{INT}_{VL} = [(G - E - F + \mathrm{Base}) + (C - D - B + A)]/2$; $\mathrm{INT}_{PL} = [(B - A - F + \mathrm{Base}) + (C - D - G + E)]/2$; $\mathrm{INT}_{VPL} = C - G - D - B + E + A + F - \mathrm{Base}$.

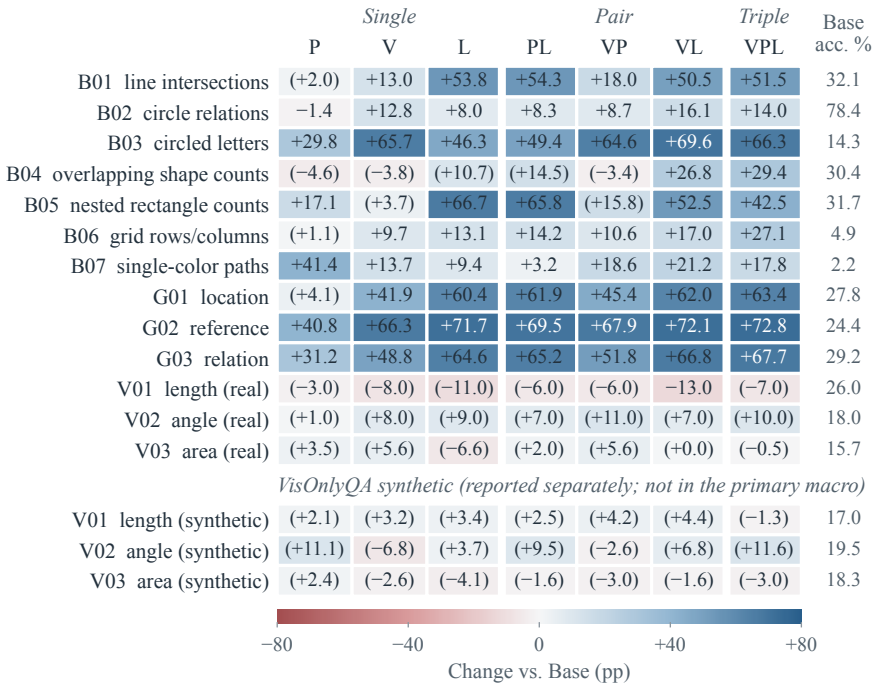

Figure 5: Factorial repair landscape: change in scene-first accuracy relative to the shared Base checkpoint (percentage points) for each trainable set, training seed 1104. Cells show net accuracy change, which mixes repaired Base errors with newly introduced ones; item-level repair is analysed in Section 4.1. Blue marks a gain and red a loss; the scale is symmetric about zero (± 80 pp) and every cell is annotated, so the sign is readable without colour. Parenthesized values: the 95% scene-family bootstrap interval of the change (2,000 replicates, scene families resampled within the task) includes zero (30 of the 91 primary cells). Right column: Base accuracy (%). Bottom block: VisOnlyQA synthetic splits, reported separately and excluded from the primary macro. Intervals are not adjusted for multiplicity.

Table 6: Per-task scene-first accuracy (%), seed 1101; Base is the shared checkpoint. Rows/scenes are the official denominators. †VisOnlyQA synthetic splits are reported separately and excluded from the primary macro. Scene-family bootstrap intervals of every per-task change accompany the released records.

Task	Source	Rows/scenes	Base	P	V	L	PL	VP	VL	VPL
B01 line intersections	BlindTest	3600/300	32.1	34.8	45.3	92.2	89.6	45.2	85.8	73.5
B02 circle relations	BlindTest	1344/224	78.4	76.9	88.9	90.1	88.6	83.5	90.6	93.5
B03 circled letters	BlindTest	624/52	14.3	43.9	80.6	65.4	65.1	86.4	89.9	84.8
B04 overlapping shape counts	BlindTest	960/38	30.4	25.7	35.7	39.6	42.2	26.9	58.8	56.8
B05 nested rectangle counts	BlindTest	240/40	31.7	49.2	59.2	97.1	97.1	67.9	62.1	75.4
B06 grid rows/columns	BlindTest	528/22	4.9	5.9	12.5	34.3	21.8	15.3	12.5	26.5
B07 single-color paths	BlindTest	720/180	2.2	42.8	5.6	8.6	7.2	21.4	22.9	23.1
G01 location	GePBench	437/186	27.8	30.1	72.1	87.7	89.2	74.6	92.0	91.9
G02 reference	GePBench	319/136	24.4	64.8	89.3	95.5	96.0	91.8	96.8	96.8
G03 relation	GePBench	380/311	29.2	59.5	78.0	96.9	96.7	79.1	97.4	96.6
V01 length (real)	VisOnlyQA	100/100	26.0	25.0	17.0	19.0	18.0	21.0	21.0	19.0
V02 angle (real)	VisOnlyQA	100/100	18.0	18.0	25.0	27.0	30.0	22.0	21.0	27.0
V03 area (real)	VisOnlyQA	100/99	15.7	16.2	22.2	20.7	22.7	23.2	20.7	15.7
V01 length (synth.)†	VisOnlyQA	100/88	17.0	19.7	15.0	22.0	21.4	15.7	23.1	19.7
V02 angle (synth.)†	VisOnlyQA	100/95	19.5	26.8	15.3	27.4	27.4	15.8	22.1	31.6
V03 area (synth.)†	VisOnlyQA	100/82	18.3	21.3	19.5	13.6	16.1	21.3	21.3	20.9

Table 7: Per-task scene-first accuracy (%), seed 1104; Base is the shared checkpoint. Rows/scenes are the official denominators. †VisOnlyQA synthetic splits are reported separately and excluded from the primary macro. Scene-family bootstrap intervals of every per-task change accompany the released records.

Task	Source	Rows/scenes	Base	P	V	L	PL	VP	VL	VPL
B01 line intersections	BlindTest	3600/300	32.1	34.1	45.1	85.9	86.4	50.1	82.6	83.6
B02 circle relations	BlindTest	1344/224	78.4	77.0	91.2	86.5	86.8	87.1	94.6	92.4
B03 circled letters	BlindTest	624/52	14.3	44.1	80.0	60.6	63.6	78.8	83.8	80.6
B04 overlapping shape counts	BlindTest	960/38	30.4	25.8	26.5	41.1	44.8	27.0	57.1	59.8
B05 nested rectangle counts	BlindTest	240/40	31.7	48.8	35.4	98.3	97.5	47.5	84.2	74.2
B06 grid rows/columns	BlindTest	528/22	4.9	6.1	14.6	18.0	19.1	15.5	22.0	32.0
B07 single-color paths	BlindTest	720/180	2.2	43.6	16.0	11.7	5.4	20.8	23.5	20.0
G01 location	GePBench	437/186	27.8	31.9	69.7	88.2	89.7	73.2	89.8	91.2
G02 reference	GePBench	319/136	24.4	65.2	90.7	96.1	93.9	92.3	96.4	97.2
G03 relation	GePBench	380/311	29.2	60.5	78.0	93.8	94.4	81.0	96.0	96.9
V01 length (real)	VisOnlyQA	100/100	26.0	23.0	18.0	15.0	20.0	20.0	13.0	19.0
V02 angle (real)	VisOnlyQA	100/100	18.0	19.0	26.0	27.0	25.0	29.0	25.0	28.0
V03 area (real)	VisOnlyQA	100/99	15.7	19.2	21.2	9.1	17.7	21.2	15.7	15.2
V01 length (synth.)†	VisOnlyQA	100/88	17.0	19.1	20.3	20.5	19.5	21.2	21.4	15.7
V02 angle (synth.)†	VisOnlyQA	100/95	19.5	30.5	12.6	23.2	28.9	16.8	26.3	31.1
V03 area (synth.)†	VisOnlyQA	100/82	18.3	20.7	15.7	14.2	16.7	15.2	16.7	15.2

B IMAGE DEPENDENCE OF THE GAINS

Table 8 reports each set under controls that remove the image (blank image), exchange it for the image of another row of the same task (image swap), append an answer-code codebook (option remap), or append the scene's geometric facts as text (G02 only). They bound how much of each gain can be obtained without the row's own image. Under option remap every set scores below Base and the sets that train L score lowest, so the gains are tied to the trained answer format (both seeds; seed 1104: Table 9).

Table 8: Control conditions: scene-first macro accuracy (%) under each control, seed 1101 (Base is the shared checkpoint). Blank image replaces the pixels with a same-size black image and keeps the question; option remap keeps the prompt and appends an explicit answer-code codebook, a diagnostic prompt extension rather than a relabeling. Image swap substitutes the image of a different row of the same task with identical dimensions and different pixels (deterministic cyclic donor). Such a donor exists for every row of the 10 BlindTest and GePBench tasks, but only for 2 of the 100 V02 real rows and 4 of the 100 V03 real rows, and for no V01 real row; both swap columns are therefore scene-first macros over those 10 tasks (9,152 rows), so clean and swapped accuracy cover the same rows. The official summary's image-swap macro, which averages every task with at least one swap row at equal weight and is not row-matched to the clean macro, is not used. The text oracle keeps the image and appends the scene's geometric facts, but not the answer, to the prompt, and is defined only for G02 (319 rows), so that column is G02 accuracy. Target-highlight and OCR-only controls were unavailable for every row.

	13 tasks			10 swap tasks		G02
Condition	Clean	Blank image	Option remap	Clean	Image swap	Text oracle
Base	25.8	22.6	23.3	27.5	24.3	24.4
P	37.9	23.4	19.3	43.3	32.1	27.0
V	48.6	25.1	18.2	56.7	28.4	46.1
L	59.5	29.7	12.3	70.7	31.3	90.8
PL	58.8	28.2	12.5	69.3	30.8	90.4
VP	50.6	25.5	18.9	59.2	29.5	43.6
VL	59.4	29.9	10.8	70.9	30.8	92.4
VPL	60.0	28.8	13.3	71.9	30.7	92.6

Table 9: Control conditions: scene-first macro accuracy (%) under each control, seed 1104 (Base is the shared checkpoint). Blank image replaces the pixels with a same-size black image and keeps the question; option remap keeps the prompt and appends an explicit answer-code codebook, a diagnostic prompt extension rather than a relabeling. Image swap substitutes the image of a different row of the same task with identical dimensions and different pixels (deterministic cyclic donor). Such a donor exists for every row of the 10 BlindTest and GePBench tasks, but only for 2 of the 100 V02 real rows and 4 of the 100 V03 real rows, and for no V01 real row; both swap columns are therefore scene-first macros over those 10 tasks (9,152 rows), so clean and swapped accuracy cover the same rows. The official summary's image-swap macro, which averages every task with at least one swap row at equal weight and is not row-matched to the clean macro, is not used. The text oracle keeps the image and appends the scene's geometric facts, but not the answer, to the prompt, and is defined only for G02 (319 rows), so that column is G02 accuracy. Target-highlight and OCR-only controls were unavailable for every row.

	13 tasks			10 swap tasks		G02
Condition	Clean	Blank image	Option remap	Clean	Image swap	Text oracle
Base	25.8	22.6	23.3	27.5	24.3	24.4
P	38.3	23.1	19.4	43.7	31.9	26.7
V	47.1	24.8	19.8	54.7	28.3	39.3
L	56.2	30.2	15.1	68.0	31.5	87.7
PL	57.3	28.9	14.4	68.2	31.4	88.6
VP	49.5	25.0	20.1	57.3	29.1	37.7
VL	60.3	28.2	13.8	73.0	30.9	91.8
VPL	60.8	28.3	12.1	72.8	31.0	93.4

C REPAIR SPACES: PAIR-RELATIVE COMPARISONS AND SEED 1104

Figure 6 repeats Figure 2 for seed 1104. Joint adaptation adds coverage for 6.2 [4.8, 7.5] and 7.7 [6.2, 9.2]% of Base errors (seeds 1101 and 1104; singles cover 77.9/76.8%, adding the pairs 83.0/82.6%, adding VPL 84.1/84.5%). Comparing a pair only with its own members overstates this: PL, VP, and VL repair 86, 193, and 377 (seed 1101) and 123, 299, and 418 (seed 1104) Base errors that both of their members miss, but only 49, 76, and 244 and 79, 79, and 290 of them are also missed by the third single set. Likewise, 41 of the 109 (seed 1101) and 45 of the 165 (seed 1104) errors that VPL repairs and no pair does have a single-set repair. Along the lattice, the edge that loses the most is VP→VPL in seed 1101 (30.0% [26.3, 33.7]) and P→PL in seed 1104 (33.6% [29.8, 37.7]); L→PL loses 7.8/6.0%. Of the Base errors that a single set repairs, 5.3/5.8% are repaired by none of the four joint sets. The overlap of P and L relative to chance is 1.008 [0.978, 1.036] and 0.964 [0.924, 0.999] and that of P and V 1.41 [1.35, 1.47] and 1.45 [1.38, 1.51]. On B07 the rates of V and L are 3.4/14.9% and 7.2/10.5%; pair-first repairs are next most frequent on V02 (9.8%, seed 1101) and V03 (17.9%, seed 1104).

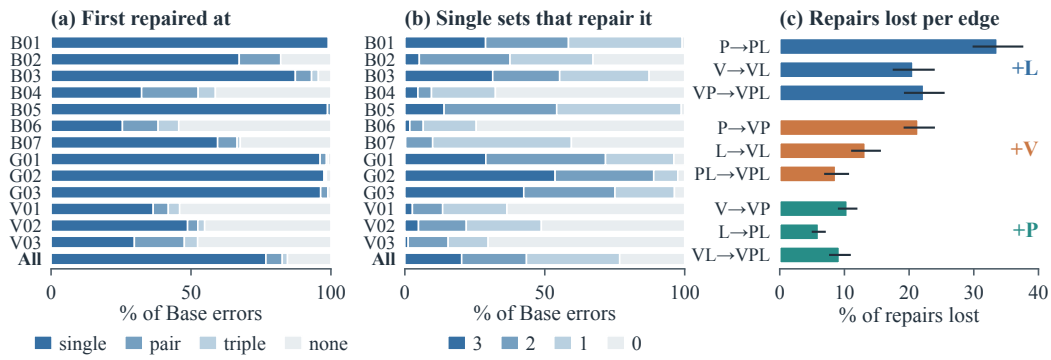

Figure 6: Repair spaces of the 13 tasks (RQ1), official predictions of training seed 1104. (a) Level at which each Base error is first repaired. (b) Number of single sets that repair it. (c) Share of a set's repairs lost when one module is added, coloured by the added module, with 95% scene-family cluster-bootstrap intervals.

Opening more modules also removes repairs. Along the nine edges that add one module, 18.0 [16.5, 19.7] and 14.7 [13.4, 16.2]% of the smaller set's repairs are lost (Figure 2c). Adding L loses the most (26.5/25.5% averaged over its three edges) and adding P the least (11.3/8.6%); in both seeds every edge that adds L loses more than every edge that adds P. By task, lost repair ranges from 1.8/1.7% (G02/G03) to 66.5/75.4% (B07), where PL loses 91.8/90.1% of the repairs of P.

Pair and minimal multiplicity (Figures 7 and 8; values seed 1101/1104, 95% scene-family cluster-bootstrap intervals). Over pairs, n_2 is 0, 1, 2, and 3 for 21.7/22.5, 14.2/16.4, 31.8/27.9, and 32.2/33.1% of Base errors, so 64.1 [61.5, 66.5] and 61.1 [58.3, 63.8]% have two or three repairing pairs. A pair that contains a repairing single, however, is not a new solution. A set S among P, V, L, PL, VP, VL, VPL is a *minimal* repair of a Base error if S repairs it and no proper subset of S does; the minimal sets of an error form an antichain, so there are at most three. The number of minimal repairs is 0 (exactly the unrepaired errors), 1, 2, and 3 for 15.9 [13.3, 18.6] and 15.5 [13.0, 18.2], 33.7 [31.1, 36.3] and 33.8 [31.2, 36.4], 28.2 [26.3, 30.0] and 30.2 [28.3, 32.3], and 22.2 [19.8, 24.7] and 20.4 [18.0, 22.8]%, so 50.4 [47.7, 53.0] and 50.7 [48.0, 53.3]% have two or more minimal repairs, fewer than have two or more repairing pairs but more than have two or more single-set repairs (45.2/43.5%): on 4.5/6.1% of Base errors a pair or VPL is a minimal repair next to a single-set one. At the triple level the count cannot separate solutions, since VPL is one set; its repairs are realized in several internal ways (Appendix G.1).

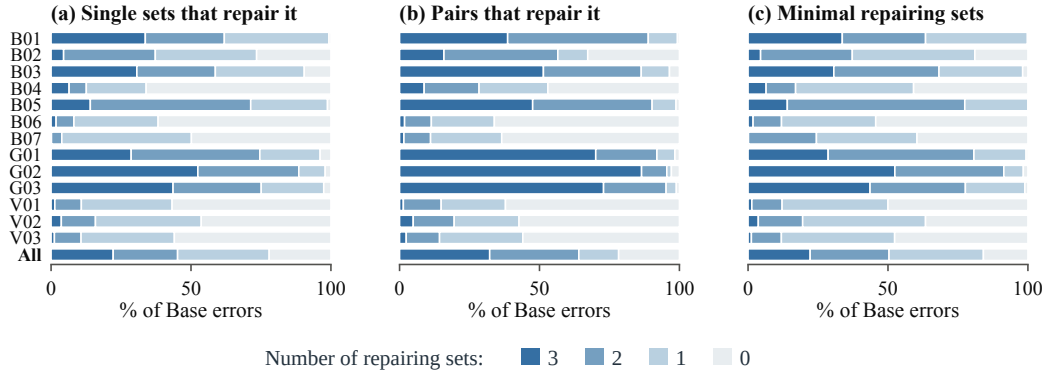

Figure 7: Repair multiplicity per task, official predictions of training seed 1101. (a) Single sets that repair a Base error (n_1). (b) Pairs that repair it (n_2). (c) Minimal repairing sets among P, V, L, PL, VP, VL, VPL (the set repairs it and no proper subset does); a pair that contains a repairing single is not counted.

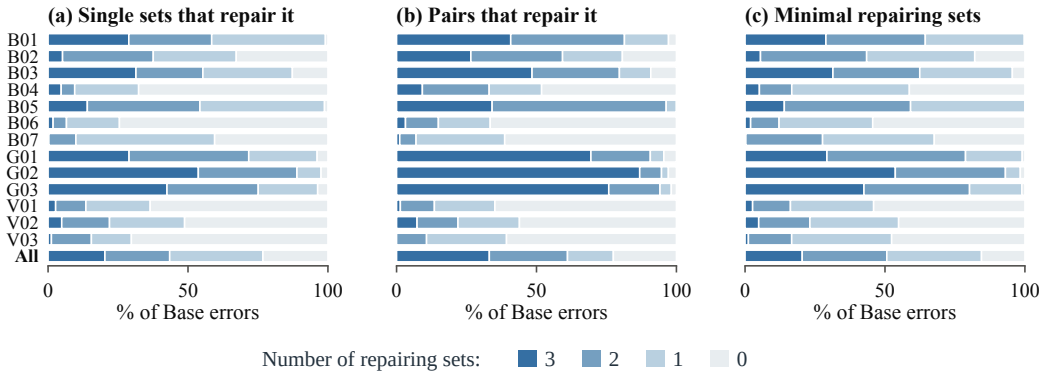

Figure 8: Repair multiplicity per task, official predictions of training seed 1104. (a) Single sets that repair a Base error (n_1). (b) Pairs that repair it (n_2). (c) Minimal repairing sets among P, V, L, PL, VP, VL, VPL (the set repairs it and no proper subset does); a pair that contains a repairing single is not counted.

Tasks have distinct repair phenotypes (Appendix Table 38). In both seeds, more than 95% of Base errors have a single-set repair on B01, B05, G01, G02, and G03, and at least 35% are repaired by no set on B04, B06, and V01–V03; B07 reaches this level in seed 1101 only (39.6/32.4%). Pair-first repairs are most frequent on B04 (20.9/20.2%). B07 is the one task on which the projector dominates: P repairs 43.5/44.3% of its Base errors, 6.0/3.0 times the rate of the better of V and L. For RQ1, then, in both seeds most failures have one or several single-module repairs, joint-only repairs are a small minority concentrated on a few tasks, and enlarging the trainable set both creates and removes repairs.

D Robustness of the repair space (RQ1)

This appendix asks whether the repair-space statements of Section 4.1 (multiplicity, pair- and triple-emergent repairs, unrepaired failures) survive answer priors, scorer ties, alternative cues, label conventions, and a change of training seed. Every analysis uses the official predictions of Section 4.1 (or re-scores them with the scorer of Section 4.2), treats each seed separately (values seed 1101/seed 1104), and reports 95% scene-family bootstrap intervals (2,000 replicates, Section 3.4). The audits were designed after the repair space was known and are exploratory.

D.1 Answer priors and prior-adjusted multiplicity

Question. Base's outputs are nearly constant on several tasks (on G01–G03 it answers (A) on every row; Appendix D.3). A task-level shift of answer preference is a change any module could make, and it "repairs" every error whose gold answer is the newly favoured class. How much of the multiplicity of Section 4.1, the share of Base errors repaired by at least two single sets (45.2/43.5%), is such a shift?

Design. Three corrections, each fitted by a 2-fold scene-family cross-fit so that no row's own family enters its correction. *XF* gives every model (Base and the seven sets) its own per-task answer-class bias and redefines the Base errors as the errors of recalibrated Base (tasks whose held-out gold classes are mostly absent from the fitting fold are left uncorrected: B03, B06). *CC* calibrates each model with its own blank-image scores, without gold labels. *MIM* keeps the Base errors and drops a set's repair whenever an item-independent per-task bias added to Base, fitted to imitate that set's own answers, already answers the item correctly; it is an upper bound on the share of multiplicity that an answer-prior shift can explain.

Result (Table 10). With every model at its own cross-fitted answer prior, multiplicity is 44.0 [41.8, 46.0] and 50.0 [47.5, 52.2]% (XF) and 37.6 [35.1, 40.0] and 36.5 [34.0, 39.2]% (CC), against 45.2 [42.7, 47.7] and 43.5 [40.8, 46.1]% raw: in both seeds more than a third of the Base errors keep two or more single-module repairs. About a third of the raw multiple-single items are Base answer-prior

errors that recalibrated Base already answers (31/30% under XF, 32/31% under CC), and most of the others stay multiple-single (55/55% and 62/60% of all raw multiple-single items). Dropping every prior-reproducible repair (MIM) leaves 18.1 [16.5, 19.7] and 16.6 [15.1, 18.1]%, less than half of the raw value; B01 supplies 75/72% of this drop. MIM is a loose bound: only 14 [11, 16] and 19 [16, 23]% of the raw multiple-single items are prior-reproducible in every set that repairs them. By module, 74/75% of P's repairs are prior-reproducible, 48/51% of V's, and 19/19% of L's.

The same correction explains the item-level co-repair excess between P and V: they co-repair 1.41 [1.35, 1.47] and 1.45 [1.38, 1.51] times as often as within-task independence predicts, but only 1.04/1.03 times within task and gold class, whereas V and L stay at 1.02/1.00 (within task) and 1.03/1.04 (within task and gold class). Much of the multiplicity is also forced by the marginal repair rates: summed over tasks, the pigeonhole floor $\max(0, r_X + r_Y - 1)$ accounts for 75/69% of the multiple-single items, and their count is at or below within-task independence (observed/expected 0.91/0.92). On G02, multiplicity is 89/89% raw, 90/89% under XF, 88/89% under CC, and 60/60% under MIM. On B02, MIM keeps most of it (37/38% raw, 33/33% MIM; 8/7% of V's and 7/10% of L's B02 repairs are prior-reproducible), and on the errors of recalibrated Base it is higher (70/71% XF, 69/77% CC).

The RQ2 statements of Section 4.2 were recomputed under each correction. L remains the module with the highest stand-alone sufficiency in PL, VL, and VPL (V in VP), and reverting L removes 79/75% of VL's repairs raw, 71/67% under XF, 64/59% under CC, and 85/82% under MIM. Real-located repairs outnumber new ones in VL and VPL under every correction; in PL it holds without correction but not under XF or MIM.

Reading. Multiplicity survives prior adjustment as a count, not as a count of item-specific mechanisms. A large part of P's and a smaller part of V's single-module repairs are answer-prior shifts, whereas L's mostly are not, and on the tasks where each single module repairs most errors the co-repair is largely forced by the repair rates. "Several single modules repair the same failure" is therefore primarily a task-level statement; the item-level reading (different routes to the same item) rests on the tasks where content repairs remain after MIM, above all G02, analysed in Section 4.3.

Table 10: Prior-adjusted multiplicity (share of Base errors repaired by at least two of P, V, L) under the three corrections, per seed, with 95% scene-family intervals. Base err.: Base errors under the estimator (XF redefines them as recalibrated Base's). MS: multiple-single items. *Base right*: raw multiple-single items that recalibrated Base answers correctly; *still MS*: raw multiple-single items that remain multiple-single. Last three columns: per-task multiplicity.

seed	estimator	Base err.	multiplicity % [95% CI]	MS	raw MS: Base right %	still MS %	B01 %	B02 %	G02 %
1101	raw	6390	45.2 [42.7, 47.7]	2889	–	–	62	37	89
1101	XF	5988	44.0 [41.8, 46.0]	2632	31	55	43	70	90
1101	CC	6354	37.6 [35.1, 40.0]	2386	32	62	39	69	88
1101	MIM	6390	18.1 [16.5, 19.7]	1159	–	40	9	33	60
1104	raw	6390	43.5 [40.8, 46.1]	2777	–	–	58	38	89
1104	XF	5988	50.0 [47.5, 52.2]	2992	30	55	57	71	89
1104	CC	6354	36.5 [34.0, 39.2]	2322	31	60	37	77	89
1104	MIM	6390	16.6 [15.1, 18.1]	1061	–	38	8	33	60

D.2 Scorer ties and candidate length

Question. Correctness is the top-ranked legal answer, so items near the decision boundary can flip for reasons unrelated to repair (Limitations), and summed log likelihood can favour short candidates. Do near-ties or candidate length inflate multiplicity? *Design.* All seven sets and Base are re-decided (i) with a 0.25-nat dead zone around the decision threshold, in which a margin too small to clear the zone counts as wrong, (ii) with random instead of first-option tie-breaking, and (iii) by mean per-token instead of summed log likelihood. *Result.* With the dead zone, multiplicity falls from 45 [42, 48] and 43 [40, 47]% to 38 [35, 41] and 35 [32, 39]% and most multiple-single items keep their class; random tie-breaking changes the multiple-single count from 2889/2777 to 2914/2785. Multiple-single items are over-represented among near-tie Base errors: 57 [50, 64] and 54 [47, 60]% of the near-tie Base errors are multiple-single, against 44 [41, 47] and 42 [39, 46]% of the other Base errors. Only B06 has candidates of unequal token length (its 528 rows); length normalization changes no item class outside B06, and any of the seven sets answers only 7.0 [0.0, 19.0] and 8.8 [0.0, 23.9]% of B06's unrepaired items. *Reading.* Near-ties add to multiplicity but do not create it,

and no task's repair space is a candidate-length artifact. Re-scoring is bitwise identical on the GPU type that produced the predictions and agrees closely across GPU types (Appendix F).

D.3 Alternative cues on G01–G03: the cue battery

Question. On the three GePBench tasks, where multiplicity is highest, are the repairs of V and L a letter-position shortcut, an image-independent text cue (a text prior only L can learn), or a repair of a content-to-letter binding failure of Base? *Design.* Base, V, L, and VL of each seed are scored through the official input path and scorer, which reproduces the re-scored records of Section 4.2 bitwise on unchanged rows (maximum difference 0.0e+00 nats), on four variants of every official row: the option contents cyclically shifted over the fixed letters; a cloze format (question stem only, the four option texts as candidates, decided by per-token mean log likelihood as fixed in advance, and by summed log likelihood); a template-matched image swap (the prompt of one row with the image of a row that has the same stem and option set but a different gold content; *follows the image* = picks the donor's gold); and the official random same-task image swap. Swap statistics are over the items that both V and L repair (co-repaired).

Result (Table 11). In both seeds Base answers (A) on every row of G01–G03, so its letter accuracy is the share of gold-(A) rows. *Option shifts:* the repairs follow the option content, not a letter position. On G02, V and L average 90/90% and 95/96% over the four shifts, and 92/90% and 100/99% of the co-repaired items are right under all four; on G01 and G03 the repairs follow the content as well. *Template-matched swaps:* on G01 (187/187 co-repaired recipients) L picks the donor image's gold on 94/94% and keeps the recipient's answer on 4/3%, and V follows the image on 77/75%; on G02 (13/15 recipients) V and L follow the image on 100/100% and 100/100%. *Random swaps:* on G02, V and L keep their clean answer on 39/37% and 42/42% of the co-repaired rows (Base on every row), but on G03 L keeps it on 84/90% against 64/59% for V, and the template-matched test that would decide G03 has only 1/1 usable recipient. *Cloze:* Base's per-token cloze accuracy is not above chance on any of the three tasks (G02: 19/19%), so the binding-failure test fixed in advance is not confirmed; with summed log likelihood, which favours short option texts, Base reaches 42/42% on G02, so the G02 reading depends on the normalization. On all three tasks the advantage of V and L over Base shrinks in the cloze format (G02: V from +66.1/+66.1 to +31.3/+24.1 points, L from +71.2/+71.8 to +25.4/+21.3).

Reading. A letter-position shortcut is ruled out on G01–G03, and an image-independent text cue is argued against for L's repairs on G01 and G02 (G02 with few items) but not on G03, where L's repairs survive a random image swap far more often than V's. The repairs are tied to the trained multiple-choice format rather than to a transferable readout of the option content. Presence cues are examined next; symmetry-breaking (mirrored) rows and an all-present counterfactual set were not run, so alternative strategies are narrowed, not ruled out.

Table 11: Cue battery on G01–G03 (%, seed 1101/seed 1104). *Letter*: official format; *shifts*: mean over the four cyclic option shifts; *all shifts*: co-repaired items (Base wrong, V and L right) that stay right under all four shifts; *cloze*: option texts as candidates, per-token (tok) or summed (sum) log likelihood; *tmpl. swap*: share of co-repaired recipients on which the model picks the donor image's gold; *rand. swap*: share of co-repaired rows whose answer is unchanged under the official random image swap (Base keeps its answer on every row). –: not defined.

task	set	letter	shifts	all shifts	cloze (tok)	cloze (sum)	tmpl. swap	rand. swap
G01	Base	29/29	25/25	–	28/28	28/28	–	–
	V	72/69	71/70	78/80	62/55	62/55	77/75	21/21
	L	89/88	89/89	98/98	61/56	61/56	94/94	25/24
	VL	93/90	92/90	–	46/56	46/56	96/94	22/24
G02	Base	24/24	25/25	–	19/19	42/42	–	–
	V	90/90	90/90	92/90	51/44	53/53	100/100	39/37
	L	95/95	95/96	100/99	45/41	54/52	100/100	42/42
	VL	96/96	96/96	–	40/46	52/54	100/100	39/45
G03	Base	28/28	25/25	–	23/23	33/33	–	–
	V	77/78	73/72	79/78	28/42	29/34	–	64/59
	L	97/94	96/94	98/98	46/48	41/41	–	84/90
	VL	97/96	96/96	–	45/46	38/36	–	76/78

D.4 A PRESENCE SHORTCUT ON G02

Question. On G02 most distractor options name shape types that are absent from the image, so detecting which option types are present answers most rows without evaluating the question's predicate. Is this how the single-module repairs work? *Design.* The official rows are split by whether exactly one option type is present in the image (70.5% of the 319 rows; the gold option is present on every row) or several (94 rows), where a presence-only strategy scores about the reciprocal of the number of present options (46% on average). *Result.* On multi-present rows V (81.9 [74.3, 89.0] and 84.0 [75.7, 91.6]%) and L (91.5 [84.9, 97.0] and 91.5 [84.9, 97.0]%) are above the presence-only rate, whereas P is not (44.7 [35.0, 53.9] and 44.7 [34.9, 54.0]%, down from 69.3/69.8% on single-present rows). When V and L err on multi-present rows they pick a present distractor (94/93% and 100/88% of their errors), so every set uses presence as a filter. Multiplicity is not confined to single-present rows: 83 [72, 92] and 82 [71, 91]% of the multi-present Base errors are multiple-single, against 91/92% of the single-present ones (the difference's interval includes zero). *Reading.* A presence shortcut fits P's G02 repairs, not V's or L's; this is observational, and the all-present counterfactual rows that would test it directly were not generated.

D.5 WHAT THE UNREPAIRED FAILURES ARE

Question. No set repairs 15.9/15.5% of Base errors. Are they failures no module can fix, or do they reflect label conventions, candidate length, or image resolution? *Design.* Base, V, L, and VL of each seed are scored with the scorer of Section 4.2 on three variants, each with a rule fixed in advance: on V01–V03, rows whose gold is option (e) are re-scored with the option contents cyclically shifted so that the gold content sits at (a) or (c) (a *label-position artifact* if some trained model answers at least half of the unrepaired gold-(e) items); on B06 every model is re-decided by per-token log likelihood (a *length artifact* under the same rule); on B02 the Base-error images are cropped around the contact region of the two circles to a half or a quarter of their side and resized back (a *resolution ceiling* if some trained model answers at least half of the near-tangent unrepaired items). *Result.* On V01–V03, gold-(e) rows are 21/21% of the rows but 46/44% of the unrepaired items, and no trained model answers these items in the clean scoring; with the gold content at (a), one of V, L, VL answers 63 [49, 76] and 56 [42, 69]% of them, and L (41 [27, 55] and 50 [37, 64]%) and VL (45 [31, 59] and 56 [43, 69]%) are each above chance: the label-position rule is met in both seeds. Moving the gold content to (c) instead rescues fewer (20/13%), while Base itself rises to 47/44% there. On B06, length normalization lets a trained model answer 2 [0, 7] and 4 [0, 12]% of the 273/273 unrepaired items, so they are not a length artifact. On B02, a quarter-side crop lets a trained model answer 27 [16, 40] and 52 [37, 67]% of the unrepaired items, none of which any set answers at native resolution, but only 28 [10, 50] and 46 [22, 76]% of the 25/24 near-tangent ones: the resolution-ceiling rule is not met. Across tasks, unrepaired items are not characterized by an image that fails to help Base: the image raises Base's gold log likelihood (positive pointwise information, next subsection) on 51 [42, 60] and 53 [43, 63]% of them. Among Base errors with at least three candidates, the gold answer is Base's runner-up for 18/22% of the unrepaired items, against 56/56% of the multiple-single ones. Label support explains a small part of the unrepaired set. Only B06 has gold classes that no training row carries (answers with a count outside the training range 3–8; 144 official rows, all Base errors). These are unrepaired in 99.3/98.6% of cases, against 14.0/13.6% of the Base errors whose gold class has training support, and no set that trains L ever outputs such a class; but they make up only 14.1/14.3% of all unrepaired items. *Reading.* "Unrepaired" is not one thing. On V01–V03 it is largely a label-position effect of the trained sets (tasks on which every model is near chance, Appendix D.7); on B02 a minority is resolution-limited for the trained models; on B06 it is not a length artifact, and part of it lacks any training support for the gold class. Unrepaired therefore means not observed under the current adaptation procedure (data, budget, optimizer), not impossible, and for the V01–V03 items and part of the B02 items not "information absent".

D.6 PER-ITEM CORRELATES OF MULTIPLICITY

Question. Beyond task-level rates, what distinguishes a Base error with several single-module repairs from one with a single repair or none? *Design.* Four item-level measures, each regressed or compared by class within seed: (i) pointwise information of the image for Base, $\mathrm{PVI}(i) = \log p_{\mathrm{Base}}(\text{gold} \mid \text{image}) - \log p_{\mathrm{Base}}(\text{gold} \mid \text{blank image})$, from the stored blank-image scores; (ii) the per-item decodability of the task variable by the Base probes of Appendix M (G02 answer letter,

B02 relation); (iii) whether the gold answer is Base's runner-up; and (iv) whether a set's repair raises the gold answer's log likelihood more than it lowers Base's answer (*booster*) or the reverse. *Result.* Multiple-single Base errors have positive PVI more often than errors with one single-module repair (51 [48, 53] and 52 [50, 55]% vs. 33 [29, 37] and 30 [26, 34]%), but given task and Base margin PVI predicts multiple-single membership only weakly (coefficient per standard deviation 0.016 [−0.006, 0.041] and 0.033 [0.008, 0.060]), less than the Base margin does (0.101 [0.067, 0.136] and 0.090 [0.052, 0.124]). On the items each set repairs, L's own PVI exceeds V's (5.41/5.20 vs. 1.37/1.40), so L's repairs are not image-independent by this measure. On G02 the Base probe's per-item decodability of the answer predicts the number of single-module repairs (slope 0.88 [0.67, 1.10] and 0.93 [0.71, 1.17], also within gold class), whereas on B02 it does not (−0.70 [−2.42, 0.93] and −0.39 [−2.13, 1.29]). The gold answer is Base's runner-up for 56/56% of the multiple-single items, 28/29% of the single-repair items, and 18/22% of the unrepaired items. The single sets repair by different margin routes: 93/93% of P's repairs are boosters, 76/75% of V's, and 28/28% of L's, which mostly lower Base's answer instead. *Reading.* Multiplicity is concentrated where Base already ranks the gold answer close and, on G02, where Base's own states decode the answer, consistent with information that is present but under-used there (not on B02); the image's pointwise information adds little beyond the Base margin. At the output, the same failure is repaired along different routes: P mainly raises the gold answer, L mainly suppresses Base's answer.

D.7 Cross-seed stability of the item classes

Question. Are an item's repair class and its multiplicity properties of the item, or of one training run? *Design.* The two seeds share Base and every official Base error. For each class of seed S we count the items that have the same class in the other seed and compare with the share expected under within-task independence; we compute the between-seed κ of $n_1(i)$ (Section 3.3); and for every task and class cell a family-level and an item-level permutation test (20,000 permutations) are Benjamini–Hochberg-corrected over the 57 cells. *Result* (Table 12). Every set's repair set replicates at 87.4–96.8/82.3–96.5%, and multiple-single items stay multiple-single in the other seed (88.7/92.3%; quadratic-weighted κ of n_1 0.90 [0.88, 0.91], unweighted 0.72). Unrepaired items stay unrepaired (81.3/83.4%). Pair-level items are fragile: 45.7/39.9% are pair-level again, above independence (observed/expected 3.55, the same from either seed) but far below the multiple-single rate; 24.2/31.4% of them get exactly one single-module repair and 19.6/16.6% none in the other seed. They replicate at 70.6/73.2% on B04 against 26.2/20.4% on the other tasks; PL and VP pair-emergent items at 12.2/7.6% and 9.2/8.9%; and triple-level items at 17.6/10.0%, not detectably above independence. After correction, the only pair-level cell that survives both permutation tests is B04 (1 of 8 cells), the VL pair-emergent cells that survive are B04, B07, no PL or VP pair-emergent cell survives, and every multiple-single cell does (13 of 13). The near-chance tasks are V01, V02, V03; with them excluded the multiple-single share is 46.5/44.5% (all tasks: 45.2/43.5%). *Reading.* Multiple-single repair is a reproducible property of the item in both seeds, and so is non-repair; pair- and triple-emergent counts outside B04 are largely run-specific and may be false negatives of the single sets in one run. This supports treating multiple-single and pair-emergent repairs as different regimes, with the pair-emergent regime characterized on B04.

Table 12: Cross-seed replication of the item classes: for the items of each class in seed S, the share (%) that have the same class in the other seed, with 95% scene-family intervals and the number of items in parentheses, and observed/expected co-repair under within-task independence (o/e; it is the same from either seed). MS: multiple-single; S1: exactly one single-module repair; pair-level: first repaired by a pair; PE: pair-emergent items that the named pair repairs.

class	seed 1101	seed 1104	o/e
MS	88.7 [86.7, 90.6] (2,889)	92.3 [91.0, 93.5] (2,777)	1.51
S1	79.3 [76.7, 81.7] (2,091)	77.9 [75.4, 80.2] (2,129)	2.21
pair-level	45.7 [35.8, 54.5] (326)	39.9 [29.9, 49.0] (373)	3.55
VL PE	55.7 [44.9, 66.1] (244)	46.9 [35.5, 56.9] (290)	4.16
PL PE	12.2 [4.8, 20.5] (49)	7.6 [2.8, 13.9] (79)	4.76
VP PE	9.2 [2.7, 17.5] (76)	8.9 [2.6, 16.7] (79)	2.42
triple-level	17.6 [2.7, 29.6] (68)	10.0 [1.3, 21.5] (120)	4.48
unrepaired	81.3 [76.5, 85.4] (1,016)	83.4 [78.0, 87.8] (991)	1.98

D.8 SHARED EXAMPLE ORDER WITHIN A SEED

Question. Within a seed all seven sets see the same examples in the same order. Could order effects make different sets flip the same items, inflating co-repair? *Design.* Both seeds share Base and its errors but see the data in different orders. For two sets X, Y we compute co-repair relative to within-task independence with both taken from one seed (same order) and from different seeds (different order); the *same-order excess* averages the within-seed minus the cross-seed value over the two seeds, which cancels each seed's own differences. Multiplicity is also recomputed with P, V, and L taken from any combination of the two seeds. *Result.* For every pair of single sets the same-order excess is negligible: 0.000 [−0.004, 0.004] for V and L, 0.001 [−0.003, 0.005] for P and V, and 0.003 [0.000, 0.005] for P and L, every interval inside ±0.05 (a materiality margin stated after a first run had flagged the P and L interval as lying just above zero). A same-order excess appears only between sets that share a trained module (V and VP: 0.035; V and VL: 0.015; L and PL: 0.030). With single sets drawn from any combination of seeds, multiplicity ranges from 43.4% to 45.3%, against 45.2/43.5% within seed. *Reading.* The shared example order does not inflate the co-repair of different single modules; the multiple-single count is not an order artifact.

E CANDIDATE EXPLANATIONS OF REPAIR AVAILABILITY (RQ3) AND ARTIFACT CHECKLIST

RQ3 asks why one failure admits several repairs. The candidate explanations below are not mutually exclusive, and one failure can be repaired by different modules for different reasons. They separate two questions: what computation Base lacks or misuses (failure locus) and which trainable modules can reach a compensating functional change (reach); the two need not coincide (on G02 V's repair reroutes the frozen language model; on B04 some repairs need two modules). Tables 13 and 14 list each candidate family (A information, B computation, C interface, D function and parameter space, E task) with the evidence in this paper, its status, and the experiment that would decide it; Table 15 gives the checklist of artifacts that could mimic or inflate multiple repair (family F). A status holds in both seeds unless stated; *supported* means supported on the task named, not in general; *partial* means consistent but not decisive or confounded; *unresolved* means untested or open; *against* means the evidence contradicts the candidate where tested. Route steering, update geometry, and training dynamics concern which available repair joint training realizes (RQ2) and are not RQ3 evidence. E, F, G label the runs of V, L, VL, and M^S is module M of set S alone in Base context.

Table 13: RQ3 candidate explanations, families A (information) and B (computation): evidence in this paper, status, and the experiment that decides it.

Candidate	Evidence in this paper	Status	Deciding experiment
A1 information present but under-used	On G02 and B02 Base's image tokens decode the task variable, and Base's final token carries the G02 answer that its own output misses (Section 4.3). Per item, Base-probe decodability predicts the number of single-module repairs on G02, not on B02 (App. D.6). Causal patching (App. O.1): the G02 target option is decodable at Base's option tokens although Base answers wrongly. The cloze test of a content-to-letter binding failure is not confirmed (App. D.3).	supported (G02); partial (B02)	amnesic removal of the probe subspaces in Base, E, F, G
A2 redundant copies (views, register-like tokens)	Causal patching: on G02 V's repair enters the frozen LM through the background image tokens, not the target-object tokens (App. O.1); on B02 it needs the contact-region features, though no image region suffices alone (App. O.4). On both tasks L's image-token states carry none of L's repair. The repairs take different routes, not two copies of the same visual evidence.	against (two visual copies; G02, B02); unresolved (register-like or distributed field)	base-view vs. tile patching
A3/E alternative strategies and cues (shortcut)	Letter-position shortcut ruled out on G01–G03; image-independent text cue argued against on G01/G02 but live on G03 (App. D.3); presence heuristic fits P, not V or L, on G02 (App. D.4). On G02 the final-token changes of E and F converge item by item; on B02 only partly (App. O.3). No counterfactual cue intervention.	partial; not ruled out	all-present, cue-conflict, and mirrored G02/B02 rows
A4 dormant capability behind a small wrapper	Low-rank truncations of F's update keep most of its repairs on G02 and B02 (the lowest rank tested does not on G02); on G02, norm-matched random V and L directions repair almost nothing (App. P.5); Base does not answer G02 from appended text facts (App. B).	partial	rank sweep by item class; text-only scoring of the original LM

Table 13: RQ3 candidate explanations, families A (information) and B (computation): evidence in this paper, status, and the experiment that decides it. (continued)

Candidate	Evidence in this paper	Status	Deciding experiment
B1/B5 several leverage points on one late decision state	G02: both single-module repairs act through the final-token query side (Section 4.3). Causal patching (App. O.1): E's image-token states stop mattering by the middle layers and its text states carry the repair; E and F hand the decision to the final token at the same layer; E's final-token change lies along F's same-item change. E and F boost largely the same heads, but ablating them removes little of either repair (distributed; App. O.2). On B02 the two repairs also take different routes and reach the final token at the same layer, but their final-token changes do not align (App. O.4, O.3).	supported (G02); B02: several leverage points, no convergence	the same causal design on further tasks
B3 serial two-stage composition (pair-only)	VL pair-emergent repairs without a single carrier are superadditive in the gold margin, and the synergy transfers to independently trained and cross-seed partners; sub-threshold additive combinations are rare (App. P.1, G.2). B04: odd counts are repaired by E.V+F.L although neither single-module set repairs them; even counts need co-trained modules, and co-trained partners from another joint set or seed also work (App. P.3). V or L trained on B04 alone repairs few of B04's VL-only items (7/5% and 0/7%, against 43/49% for E.V+F.L; App. P.4).	supported (VL, B04) as description; PL, VP, VPL not separable	stage-specific per-item probes
B3 sign epistasis	On superadditive VL repairs L^{VL} alone moves the margin away from gold; confounded with interface incompatibility. A retrained projector between V^{VL} and Base L (P-stitch) recovers only part of the lost G02 repair (App. G.3).	partial	margin trajectories over checkpoints
B distributed or compositional limitation (no single stage suffices)	5.1/5.8% of Base errors are first repaired by a pair and 1.1/1.9% only by VPL; B04 has the largest pair-first share (20.9/20.2%; Appendix L). Minimal repairing sets (Appendix C) count such a pair or triple as a solution only when none of its subsets repairs the item. Failure locus and reach differ: a two-module repair need not mean a two-stage defect (B3 above).	supported (B04) as description; mechanism as B3	stage-specific per-item probes
B4 parallel or backup pathways	G02: no direct readout of the object tokens' values; both repairs act through the late query side; head ablation finds no single pathway (distributed). B02: the two repairs' final-token changes do not align (App. O.4).	unresolved (G02: many heads or compensation); partial (B02); unresolved (B07)	path blocking on B02
B late overwrite of an answer Base computes	Logit lens (App. O.3): Base's G02 answer is at chance at every layer, so no lens-readable answer is overwritten; on B02 Base's lens accuracy peaks in middle layers above its output layer.	against (G02); partial (B02)	patching B02's middle-layer final-token state
B position bias of rotary attention	Base's G02 target-minus-distractor attention logit falls with the target tokens' distance in the sequence, but the repair rates of E, F, and G are unrelated to it (per-item analysis).	against (G02)	none needed

Table 14: RQ3 candidate explanations (continued): families C (interface), D (function and parameter space), and E (task).

Candidate	Evidence in this paper	Status	Deciding experiment
C1 upstream compensation	V-only changes the frozen LM's final-token query with Base LM weights; only L raises per-token target attention (Section 4.3); with the scene's geometric facts appended to the G02 prompt, V reaches far lower accuracy than L (App. B).	supported (G02)	soft visual prompt with V, P, L frozen
C2 downstream compensation	Not tested.	unresolved	direct vs. total logit effect on VL pair-emergent items
C3 projector compensation	P is inert in margin space and 74/75% of its repairs are reproduced by an answer-prior shift (App. D.1); on G02 it fits a presence shortcut (App. D.4).	partial (mostly prior)	rank-1 truncation of the P update
C4/C5 reparameterization; run-private interface code	The P&V co-repair excess is an answer-prior effect (1.41/1.45 within task, 1.04/1.03 within task and gold class). The synergy of the analysed VL pair-emergent repairs mostly does not require run-private code (cross-partner and cross-seed recombination; App. G.2). PL, VP, VPL have no cross-partner recombination.	against (VL private code); unresolved (PL, VP, VPL)	cross-partner recombinations for PL and VP
C tied embedding (a readout handle only L has)	Over all 13 tasks F's transformer blocks alone keep nearly all of its repairs; its output head alone repairs a small, mostly image-independent (prior-type) subset, and F without its head keeps its blank-image advantage (App. P.6).	against as main handle	none needed

Table 14: RQ3 candidate explanations (continued): families C (interface), D (function and parameter space), and E (task). (continued)

Candidate	Evidence in this paper	Status	Deciding experiment
D1 degeneracy	E and F repair the same G02 items with different mechanisms and image dependence (Sections 4.2–4.3); their answers share little information beyond gold and task ($I(E; F \mid \text{gold, task})/H(E \mid \text{gold, task})$ 0.11/0.15); on G01–G03 neither's repairs are image-coherent within scene families.	supported (G02)	per-task decomposition of E and F predictions
D2/D3 first-order reach; V×L coupling	Base-Jacobian test: L is judged locally capable in every class, so first-order reach does not predict which repair training realizes (negative; App. P.2). VL pair-emergent repairs have a large positive V×L interaction; triple-emergent repairs are carried by the V×L term.	partial (first-order: negative)	(α, β) grid on B04 by item class
D4 connected repair manifold	No barrier between the V-only and L-only repairs on G02 and B02 (Section 4.3, App. Q); geometry, not an explanation of availability.	supported (geometry only)	B04 grids by item class
D sub-threshold additivity	Additive combinations of sub-threshold single-module effects are rare among pair-emergent repairs (App. P.1).	against	none needed
D generic multiplicity (partitions)	On G02 each half of F's block update (lower or upper LM layers) keeps most of F's repairs; on B02 only the early half. Neither half of E's update does on G02; each half does on B02 and B04 (App. P.5).	partial (task-dependent)	sub-block training at matched updates
E1 location of the train–eval shift	Descriptive only: label-deciding computations relate to repair phenotypes over 13 tasks (Section 4.3).	partial	rank correlation with locked shift labels
E2 multi-task conflict	Single-task control on B04 (App. P.4): V or L trained on B04 alone does not repair B04's VL-only items (verdict: consistent with V×L synergy in both seeds; the Adam step budget is not matched and memorization is not excluded). The trained sets move most unrepaired items away from gold (App. P.1).	partial (B04 pair-emergent); unresolved (unrepaired items)	single-task runs on B06 and V01–V03
E symmetry	Not tested (mirrored rows need the rule generator).	unresolved	mirrored and rotated rows

Table 15: Artifact checklist: artifacts that could mimic or inflate multiple repair, how this paper controls them, and what remains.

Artifact	Control in this paper	Status	Remaining test
Seed stochasticity	Two seeds, analysed separately. Multiple-single items replicate (88.7/92.3%); pair-level items do not (45.7/39.9%) except on B04; triple-level items rarely (17.6/10.0%) (App. D.7).	controlled (MS); partial (pair, triple)	more seeds (seed 1102 covers the behavioural checks, App. H)
Scorer noise, ties	Bitwise re-scoring gates on the producing GPU type and cross-hardware agreement (App. F); a 0.25-nat dead zone lowers multiplicity but does not remove it, and random tie-breaking changes it little (App. D.2).	partial	–
Near-chance tasks	Only V01, V02, V03 are near chance; without them the multiple-single share is 46.5/44.5% (all tasks 45.2/43.5%; App. D.7).	controlled	–
Probe decodability ≠ sufficiency	V^{VL} raises G02 target decodability yet barely repairs (Section 4.3); V's target-object tokens decode better than Base's but transplanting them repairs little (App. O.1). Probes are read as decodability only.	guarded	amnesic removal
Transplant incompatibility	Reversions and recombinations (Section 4.2, App. F); the P-stitch recovers only part of the lost G02 repair (App. G.3); cross-seed recombinations keep VL pair-emergent repairs (App. G.2).	partly controlled	stitches for PL, VP
Memorization	Every trained set saw the training rows for 3 epochs; the official rows are held out, but for BlindTest and GePBench they come from the same generators, so they are not counterfactual.	unresolved	render-style-shifted rows
Shortcut	Letter position excluded; text cue argued against on G01/G02, live on G03 (App. D.3); presence explains P only (App. D.4).	not ruled out	counterfactual cue rows
Answer prior and aliasing	Base answers (A) on every G01–G03 row. Multiplicity survives cross-fitted and blank-image calibration as a count (44.0/50.0% and 37.6/36.5%, raw 45.2/43.5%); the prior-reproducible upper bound leaves 18.1/16.6%; P's and part of V's repairs are prior shifts, and much co-repair is forced by the repair rates (App. D.1).	partial	oracle re-derivation of gold on flagged items
Label position, convention, support	V01–V03 unrepaired items are largely a label-position effect; B06's are not a length artifact; part of B02's are resolution-limited. Gold classes without training support (B06 only) are almost always unrepaired but are 14.1/14.3% of all unrepaired items (App. D.5).	partial	task-oracle audit of unrepaired items
Shared example order	Same-order excess of co-repair between single sets inside ±0.05; multiplicity almost unchanged with single sets from different seeds (App. D.8).	controlled	–
Optimization budget and procedure	One fixed recipe without search, so unrepaired is written as "not observed under the current adaptation procedure". A clip-only control (V trained inside the joint clipping set of VL, L not updated) shows that joint gradient clipping does not explain the rotation of V's update in VL (App. G.4).	partial	longer or re-tuned training of the unrepaired tasks

F TRANSPLANTS, REVERSIONS, AND RECOMBINATIONS

Section 4.2 compares 26 models per seed on all 13 tasks (9,452 official rows, 1,788 scene families, 6,383 Base errors); values below are seed 1101/seed 1104. All of them, Base and the seven sets included, were scored in one campaign per seed with a re-implementation of the official scorer (same legal answers, summed log likelihood of each answer string, and tie rule), which reproduced official predictions bitwise on the GPU type that produced them; the official predictions of seed 1101 were produced on another GPU type. Tables 16 and 17 compare the re-scored and official predictions of Base and the seven sets: correctness agrees on 98.2–99.7/98.2–100.0% of items (the predicted answer on at least 97.0/97.0%) and 92.9–99.8/99.8–100.0% under the blank image; the lowest per-task agreement is P on B07 (92.9%) in seed 1101 and Base on V01 (95.0%) in seed 1104, so these numbers can differ slightly from Section 4.1. Every comparison in Section 4.2 is between re-scored models of the same seed.

Table 16: Re-scored versus official predictions of Base and the seven sets (seed 1101, 13 tasks, 9,452 rows). Macro: scene-first macro accuracy (%). Agreement: share of items whose correctness (or predicted answer) is the same; lowest task: the task with the lowest correctness agreement. Blank: correctness agreement under the blank-image control.

	Macro %			Agreement %			
Set	official	re-scored	diff (pp)	correct	answer	lowest task	Blank %
Base	25.8	25.8	+0.02	98.17	97.04	V01 95.0	99.82
P	37.9	37.7	−0.18	98.36	96.97	B07 92.9	92.91
V	48.6	48.1	−0.46	99.39	97.91	V02 96.0	99.38
L	59.5	59.4	−0.13	99.59	99.28	B03 98.4	99.79
PL	58.8	58.1	−0.65	99.48	99.29	V03 93.0	99.78
VP	50.6	50.4	−0.21	98.98	98.00	V03 94.0	98.76
VL	59.4	58.9	−0.42	99.68	99.50	V01 97.0	99.16
VPL	60.0	60.2	+0.19	99.57	99.39	V02 97.0	99.75

Table 17: Re-scored versus official predictions of Base and the seven sets (seed 1104, 13 tasks, 9,452 rows). Macro: scene-first macro accuracy (%). Agreement: share of items whose correctness (or predicted answer) is the same; lowest task: the task with the lowest correctness agreement. Blank: correctness agreement under the blank-image control.

	Macro %			Agreement %			
Set	official	re-scored	diff (pp)	correct	answer	lowest task	Blank %
Base	25.8	25.8	+0.02	98.17	97.04	V01 95.0	99.82
P	38.3	38.3	0.00	100.00	100.00	B01 100.0	100.00
V	47.1	47.1	0.00	100.00	100.00	B01 100.0	100.00
L	56.2	56.2	0.00	100.00	100.00	B01 100.0	100.00
PL	57.3	57.3	0.00	100.00	100.00	B01 100.0	100.00
VP	49.5	49.5	0.00	100.00	100.00	B01 100.0	100.00
VL	60.3	60.3	0.00	100.00	100.00	B01 100.0	100.00
VPL	60.8	60.8	0.00	100.00	100.00	B01 100.0	100.00

Co-trained with the language model, the vision encoder loses most of its stand-alone function, and the projector loses it with any partner (Figure 3a). In Base context, V^{VP} gains as much as V^{V} (+22.3/+22.1 vs +22.3/+21.3 pp; both intervals of the difference include zero), but V^{VL} and V^{VPL} gain only +8.5/+8.4 and +9.1/+8.9 pp, 38/39 and 41/42% of it. P^{PL}, P^{VP}, and P^{VPL} gain 0.0/0.0, +0.3/+0.5, and 0.0/−0.1 pp (every interval includes zero), against +11.9/+12.5 pp for P^{P}. The language model keeps its gain when co-trained with P (+32.6/+31.3 vs +33.6/+30.4 pp) and most of it with V (+26.8/+26.5 and +26.4/+23.8 pp).

Each joint set's repairs are carried mostly by one module (Figure 3b). The transplanted module alone makes 97.5/98.1% of the repairs of PL (L^{PL}), 89.3/88.0% of VP (V^{VP}), and 76.1/70.0% and 66.7/73.4% of VL and VPL (L^{VL}, L^{VPL}), and reverting it loses 97.2/97.2, 95.1/95.2, 79.2/75.0, and 75.0/76.8% of them. P alone makes at most 4.9/4.8% of any joint set's repairs; in VL and VPL, V alone makes 20.8/25.0% and 24.5/22.1%.

Joint adaptation mostly reassigns repairs that single modules already make (Figure 3c). New repairs are 1.3/2.1, 2.6/2.6, 5.8/7.2, and 7.0/8.5% of the repairs of PL, VP, VL, and VPL, and reallo-

cated ones 3.9/4.3, 9.1/7.6, 13.5/17.7, and 25.8/17.0%; in both seeds and every joint set reallocation outnumbers new capability, and most repairs keep their single-module carrier. Seen from the single modules, V co-trained with L still makes 24.7–24.9/27.0–27.9% of the repairs of V^V, against 88.8/89.3% when co-trained with P; P keeps at most 4.3/4.5% of its repairs with any partner; L keeps 91.2/94.0% with P and 65.3–76.5/71.3–76.8% with V.

On G02, V^V reaches 90.7/90.7% but V^{VL} only 25.2/25.2% (Base 24.4%), whereas L^{VL} reaches 95.2/92.8% (VL: 96.6/96.4%). On B07, co-training with V raises L's stand-alone gain (+27.5/+25.8 vs +6.5/+9.4 pp for L^L). Tables 18 and 19 give the transplant matrix with retention, ownership, necessity, and blank-image gain, Tables 20 and 21 each joint set's repairs by category, and Tables 22 and 23 the recombinations (seeds 1101 and 1104; Figure 9 repeats Figure 3 for seed 1104). For a pair, reverting one module leaves the other module's transplant, so the necessity of one module is one minus the ownership of the other; reverting P loses 2.5/1.9% and 0.9/0.4% of the repairs of PL and VPL but 10.7/12.0% in VP. Pooled over the joint sets, most reallocated repairs have no single carrier in the joint set rather than a carrier that moved to another module (1,469 vs 488 and 1,292 vs 541); in seed 1101 this holds in every joint set, in seed 1104 in all but PL. The mean absolute change of a module's stand-alone repair rate between single and joint training (over the nine joint cells) is 17.4 [16.3, 18.4] and 16.4 [15.5, 17.5] pp averaged over tasks (20.0/18.8 pp pooled), largest on G02 (39.5/39.3 pp). Pooled over the four joint sets, 397/498 Base errors (6.2/7.8%) are repaired by a joint set and by no single set, and 263/285 (4.1/4.5%) by a single set and by no joint set.

Table 18: The 12-cell transplant matrix (seed 1101, re-scored predictions, 13 tasks). M^S: module M trained in set S, alone in Base context (the diagonal cells M^M are the single sets). Gain: scene-first macro over the re-scored Base (pp). Repair: share of the 6,383 Base errors it repairs (pooled). Retention: share of M^M's repairs that M^S also makes. Ownership: share of S's repairs that M^S also makes. Necessity: share of S's repairs lost when M alone is reverted to Base. Blank: gain over Base under the blank-image control (pp). Brackets: 95% scene-family cluster-bootstrap intervals (2,000 replicates).

Model	Gain (pp)	Repair %	Retention %	Ownership %	Necessity %	Blank (pp)
V^V	+22.3 [+19.7, +25.0]	47.4	–	–	–	+2.1
V^{VP}	+22.3 [+19.6, +25.0]	49.0	88.8 [87.0, 90.4]	89.3 [87.8, 90.6]	95.1 [94.1, 96.1]	+2.2
V^{VL}	+8.5 [+6.9, +10.2]	17.3	24.7 [21.8, 27.3]	20.8 [18.3, 23.2]	23.9 [20.9, 26.8]	+0.2
V^{VPL}	+9.1 [+7.4, +10.8]	18.4	24.9 [22.0, 27.4]	24.5 [21.7, 27.1]	32.1 [28.7, 35.5]	+0.1
P^P	+11.9 [+9.7, +14.2]	33.4	–	–	–	−0.1
P^{PL}	0.0 [−0.6, +0.7]	2.0	3.0 [2.2, 4.0]	2.8 [2.2, 3.5]	2.5 [1.8, 3.2]	+0.1
P^{VP}	+0.3 [−0.7, +1.4]	3.6	4.3 [3.1, 5.6]	4.9 [3.9, 5.9]	10.7 [9.4, 12.2]	−0.1
P^{VPL}	0.0 [−0.6, +0.6]	1.5	1.7 [1.2, 2.4]	2.2 [1.8, 2.7]	0.9 [0.6, 1.3]	−0.2
L^L	+33.6 [+31.0, +36.4]	64.5	–	–	–	+6.5
L^{PL}	+32.6 [+30.1, +35.2]	61.9	91.2 [89.8, 92.6]	97.5 [96.8, 98.2]	97.2 [96.5, 97.8]	+5.1
L^{VL}	+26.8 [+24.0, +29.4]	56.3	76.5 [73.4, 79.4]	76.1 [73.2, 79.1]	79.2 [76.8, 81.7]	+6.6
L^{VPL}	+26.4 [+23.8, +29.1]	49.1	65.3 [61.6, 69.1]	66.7 [63.1, 70.3]	75.0 [72.3, 77.7]	+6.1

Table 19: The 12-cell transplant matrix (seed 1104, re-scored predictions, 13 tasks). M^S: module M trained in set S, alone in Base context (the diagonal cells M^M are the single sets). Gain: scene-first macro over the re-scored Base (pp). Repair: share of the 6,383 Base errors it repairs (pooled). Retention: share of M^M's repairs that M^S also makes. Ownership: share of S's repairs that M^S also makes. Necessity: share of S's repairs lost when M alone is reverted to Base. Blank: gain over Base under the blank-image control (pp). Brackets: 95% scene-family cluster-bootstrap intervals (2,000 replicates).

Model	Gain (pp)	Repair %	Retention %	Ownership %	Necessity %	Blank (pp)
V^V	+21.3 [+18.6, +24.1]	47.9	–	–	–	+1.7
V^{VP}	+22.1 [+19.5, +24.8]	48.5	89.3 [87.7, 90.8]	88.0 [86.4, 89.6]	95.2 [94.2, 96.1]	+2.0
V^{VL}	+8.4 [+6.7, +10.0]	19.3	27.9 [25.1, 30.5]	25.0 [22.4, 27.6]	30.0 [26.9, 33.2]	−0.1
V^{VPL}	+8.9 [+7.3, +10.6]	19.1	27.0 [24.2, 29.5]	22.1 [19.5, 24.6]	26.5 [23.5, 29.8]	−0.4
P^P	+12.5 [+10.3, +14.8]	33.6	–	–	–	0.0
P^{PL}	0.0 [−0.6, +0.5]	1.9	2.7 [1.9, 3.7]	2.8 [2.2, 3.6]	1.9 [1.3, 2.6]	−0.3
P^{VP}	+0.5 [−0.5, +1.6]	3.5	4.5 [3.4, 5.8]	4.8 [3.9, 5.8]	12.0 [10.4, 13.6]	−0.5
P^{VPL}	−0.1 [−0.7, +0.4]	1.5	2.0 [1.3, 2.7]	1.9 [1.4, 2.5]	0.4 [0.2, 0.7]	0.0
L^L	+30.4 [+27.9, +33.1]	59.1	–	–	–	+7.1
L^{PL}	+31.3 [+28.8, +33.9]	59.7	94.0 [92.8, 95.1]	98.1 [97.4, 98.7]	97.2 [96.4, 97.8]	+6.0
L^{VL}	+26.5 [+24.0, +29.1]	48.6	71.3 [68.0, 74.7]	70.0 [66.8, 73.1]	75.0 [72.4, 77.6]	+4.8
L^{VPL}	+23.8 [+21.0, +26.6]	53.1	76.8 [73.4, 79.7]	73.4 [70.2, 76.4]	76.8 [74.2, 79.4]	+5.0

Table 20: Each joint set's repaired Base errors by category (count and % of the set's repairs, pooled over 13 tasks; seed 1101, re-scored). New: no single set repairs the error; carried if some module of S alone (M^S) repairs it, joint-only if none does. Same carrier: the single set of a module M of S repairs it and so does M^S. Moved: single sets of S's modules repair it, but only another module of S does alone after joint training. No carrier: single sets of S's modules repair it, but no M^S does. Outside: only the single set of the module S does not train repairs it. Reallocated = moved + no carrier. Last two columns: 95% scene-family intervals.

		New		Same		No			
Set	Repairs	carried	joint-only	carrier	Moved	carrier	Outside	New %	Reallocated %
PL	3,941	44 (1.1)	6 (0.2)	3,704 (94.0)	64 (1.6)	89 (2.3)	34 (0.9)	1.3 [0.8, 1.8]	3.9 [3.0, 4.9]
VP	3,071	47 (1.5)	33 (1.1)	2,590 (84.3)	63 (2.1)	216 (7.0)	122 (4.0)	2.6 [2.0, 3.2]	9.1 [7.9, 10.4]
VL	4,127	88 (2.1)	153 (3.7)	3,196 (77.4)	121 (2.9)	438 (10.6)	131 (3.2)	5.8 [4.1, 7.6]	13.5 [11.2, 15.9]
VPL	3,749	103 (2.7)	158 (4.2)	2,522 (67.3)	240 (6.4)	726 (19.4)	0 (0.0)	7.0 [5.0, 8.8]	25.8 [22.7, 28.7]

Table 21: Each joint set's repaired Base errors by category (count and % of the set's repairs, pooled over 13 tasks; seed 1104, re-scored). New: no single set repairs the error; carried if some module of S alone (M^S) repairs it, joint-only if none does. Same carrier: the single set of a module M of S repairs it and so does M^S. Moved: single sets of S's modules repair it, but only another module of S does alone after joint training. No carrier: single sets of S's modules repair it, but no M^S does. Outside: only the single set of the module S does not train repairs it. Reallocated = moved + no carrier. Last two columns: 95% scene-family intervals.

		New		Same		No			
Set	Repairs	carried	joint-only	carrier	Moved	carrier	Outside	New %	Reallocated %
PL	3,790	65 (1.7)	16 (0.4)	3,501 (92.4)	116 (3.1)	47 (1.2)	45 (1.2)	2.1 [1.3, 3.1]	4.3 [3.5, 5.2]
VP	3,103	37 (1.2)	44 (1.4)	2,571 (82.9)	40 (1.3)	195 (6.3)	216 (7.0)	2.6 [2.0, 3.2]	7.6 [6.5, 8.7]
VL	4,075	91 (2.2)	202 (5.0)	2,934 (72.0)	114 (2.8)	606 (14.9)	128 (3.1)	7.2 [5.4, 8.9]	17.7 [15.1, 20.3]
VPL	4,214	170 (4.0)	189 (4.5)	3,140 (74.5)	271 (6.4)	444 (10.5)	0 (0.0)	8.5 [6.4, 10.5]	17.0 [14.4, 19.8]

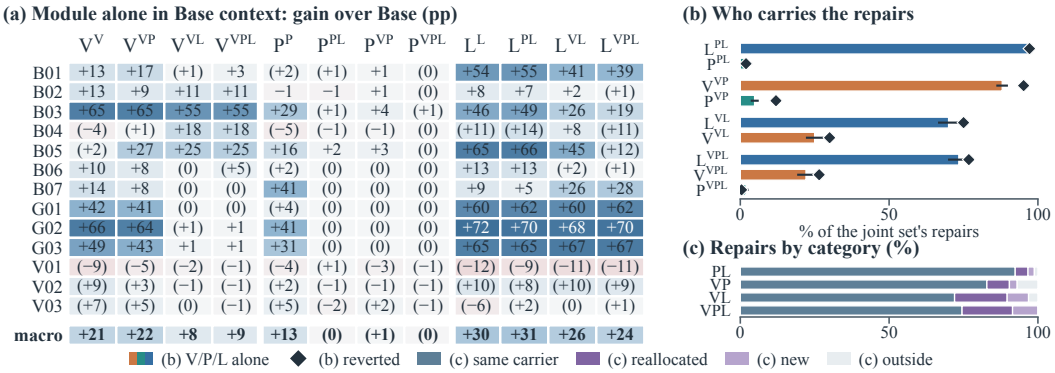

Figure 9: What joint adaptation changes (RQ2; seed 1104, re-scored, 13 tasks). (a) Gain over Base (pp) of M^S, module M trained in set S alone in Base context; colours as in Figure 1; parentheses: the 95% interval includes zero. (b) Share of a joint set's repairs that the module makes alone (bar, ownership, 95% interval) or loses on reversion (diamond, necessity). (c) Categories of Section 3.3.

Separately trained modules combine, but not as well as jointly trained ones (Table 22; $V^V{+}L^L$ denotes the recombination (V^V, P^0, L^L)). $V^V{+}L^L$ gains +30.1/+32.2 pp, −3.0 [−5.1, −1.0] and −2.3 [−4.0, −0.5] pp relative to VL, and −3.5 [−5.7, −1.5] and +1.8 [−0.3, +3.8] pp relative to L^L alone: below L^L in seed 1101 but not detectably different from it in seed 1104; in both seeds it loses more member repairs than it gains. With the co-trained vision encoder, $V^{VL}{+}L^L$ matches VL (+1.1 [−1.0, +3.1] and +0.3 [−1.8, +2.3] pp). Every recombination with V^V falls 3.0–12.7/2.3–12.2 pp below its jointly trained set, $P^P{+}L^L$ differs from PL by −1.3 [−2.9, +0.4] and −1.6 [−3.3, 0.0] pp, and every recombination gains 7.9–46.1/4.1–41.5 pp less than its members' summed gains.

Table 22: Recombinations of modules trained in different sets, alone in Base context otherwise (seed 1101, re-scored, 13 tasks). Joint: the jointly trained set with the same trained modules. Gain and differences: scene-first macro (pp) with 95% scene-family intervals; − joint, − best member, − sum: gain minus the joint set's gain, minus the larger member gain, and minus the sum of the member gains (each member alone in Base context). Kept: share of the Base errors repaired by at least one member that the recombination also repairs. Emerg./lost: Base errors repaired by the recombination and no member / by some member but not the recombination. Blank: blank-image gain over Base.

Model	Joint	Gain (pp)	− joint	− best member	− sum	Kept %	Emerg./lost	Blank
$V^V{+}L^L$	VL	+30.1 [+27.1, +33.2]	−3.0 [−5.1, −1.0]	−3.5 [−5.7, −1.5]	−25.8	84.3	217 / 731	+6.1
$V^V{+}L^{VL}$	VL	+27.9 [+25.0, +30.8]	−5.3 [−7.1, −3.3]	+1.1 [−1.2, +3.3]	−21.2	83.1	278 / 755	+5.1
$V^{VL}{+}L^L$	VL	+34.2 [+31.6, +36.9]	+1.1 [−1.0, +3.1]	+0.6 [−1.1, +2.2]	−7.9	84.4	437 / 685	+6.6
$V^V{+}P^P$	VP	+14.9 [+12.1, +17.7]	−9.8 [−12.3, −7.2]	−7.5 [−10.0, −4.9]	−19.4	53.7	518 / 1,622	+2.4
$P^P{+}L^L$	PL	+31.1 [+28.5, +33.7]	−1.3 [−2.9, +0.4]	−2.5 [−4.0, −1.1]	−14.5	73.2	176 / 1,251	+6.6
$V^V{+}P^P{+}L^L$	VPL	+21.7 [+18.8, +24.8]	−12.7 [−15.0, −10.4]	−11.9 [−14.3, −9.6]	−46.1	58.8	212 / 2,046	+5.8

Table 23: Recombinations of modules trained in different sets, alone in Base context otherwise (seed 1104, re-scored, 13 tasks). Joint: the jointly trained set with the same trained modules. Gain and differences: scene-first macro (pp) with 95% scene-family intervals; − joint, − best member, − sum: gain minus the joint set's gain, minus the larger member gain, and minus the sum of the member gains (each member alone in Base context). Kept: share of the Base errors repaired by at least one member that the recombination also repairs. Emerg./lost: Base errors repaired by the recombination and no member / by some member but not the recombination. Blank: blank-image gain over Base.

Model	Joint	Gain (pp)	− joint	− best member	− sum	Kept %	Emerg./lost	Blank
$V^V{+}L^L$	VL	+32.2 [+29.2, +35.5]	−2.3 [−4.0, −0.5]	+1.8 [−0.3, +3.8]	−19.5	85.2	314 / 682	+6.3
$V^V{+}L^{VL}$	VL	+29.1 [+26.2, +32.1]	−5.4 [−7.2, −3.6]	+2.6 [+0.4, +4.6]	−18.7	82.6	430 / 746	+4.4
$V^{VL}{+}L^L$	VL	+34.7 [+32.1, +37.6]	+0.3 [−1.8, +2.3]	+4.3 [+2.7, +5.9]	−4.1	90.0	549 / 417	+6.4
$V^V{+}P^P$	VP	+15.9 [+13.1, +18.9]	−7.9 [−10.8, −5.0]	−5.4 [−8.5, −2.6]	−18.0	50.9	503 / 1,707	+1.6
$P^P{+}L^L$	PL	+29.8 [+27.3, +32.5]	−1.6 [−3.3, 0.0]	−0.6 [−2.1, +0.8]	−13.2	70.4	227 / 1,333	+6.8
$V^V{+}P^P{+}L^L$	VPL	+22.8 [+20.0, +25.8]	−12.2 [−14.4, −10.1]	−7.7 [−9.9, −5.4]	−41.5	55.8	219 / 2,167	+6.7

Transplants inherit the image dependence of the module they carry. With a blank image, every model with a trained L gains +4.8 to +6.6 and +4.4 to +7.1 pp over Base (same control), including L^{PL}, L^{VL}, and L^{VPL}, and every model with the Base L at most +2.4/+2.0 pp (Table 18).

G JOINT REPAIR REALIZATION: OUTCOME TAXONOMY, RECOMBINATION, STITCHING, AND CONTROLS

This appendix extends the RQ2 analysis of Section 4.2. It first classifies every repair and every lost repair of a joint set by how joint training realized it (Appendix G.1), and then tests three alternatives to the reading that joint training reorganizes available repairs: code private to one training run between co-trained modules (cross-seed recombination, Appendix G.2), interface incompatibility of the co-trained vision encoder (P-stitch, Appendix G.3), and a procedural coupling of V to L through joint gradient clipping (Appendix G.4); a regression asks whether L's early repair of an item predicts V's loss of it (Appendix G.5). Unless stated otherwise the analyses reuse the re-scored models of the transplant analysis (Appendix F; 9,452 official rows, 6,383 Base errors), analyse each seed separately with 95% scene-family intervals (2,000 replicates, the draw of Appendix F), and give values as seed 1101/seed 1104. They are exploratory (Section 3.4).

G.1 OUTCOME TAXONOMY OF JOINT REPAIR

Question. When a joint set S repairs a Base error, did it keep a single-module solution, duplicate it, keep one of several, move it to another module, or build a repair that no module makes alone; and when S fails where a single set succeeds, what happened to that solution?

Design. For a Base error i and a joint set S, let C_1 be the modules of S whose single set repairs i and K the modules M of S whose transplant M^S repairs i in Base context (the *stand-alone carriers*). Two further tests enter only where no module carries i alone: whether the stand-alone effects, added in log-likelihood space over the legal answers ($\sum_M \ell_{M^S}(c) - (|S| - 1)\,\ell_0(c)$ for each candidate c), already rank the gold answer first; and, for VL, the within-seed recombinations $V^{VL}{+}L^L$, $V^V{+}L^{VL}$, and $V^V{+}L^L$ of Appendix F. Table 24 gives each leaf's criterion and what separates it from its neighbour. Leaves A–G partition the repairs of each joint set in every task; H classifies lost repairs; I–M are module-level or descriptive labels. The leaves refine the categories of Section 3.3 (same carrier = A1 + A2 + winner-take-all + redundant duplication with a retained carrier; no single carrier = D + E + F + E/F-ambiguous; the globally new share equals the new share of Appendix F), and every quantity the two analyses share was recomputed and agrees exactly in both seeds.

Table 24: Leaves of the joint-repair taxonomy. J: the joint set S repairs the Base error; C_1: modules of S whose single set repairs it; K: modules M of S whose transplant M^S repairs it in Base context. The right column names the neighbouring leaf and the test that separates them; the status says what the available models identify.

leaf	criterion	separated from its neighbour by; status
G new, globally new	J; no single set (P, V, L) repairs it	constituent-new only: whether the single set of the module S does not train repairs it; identified
G new, constituent-new only	J; $C_1 = \emptyset$; the third single set repairs it (pairs only)	globally new: the third single set repairs it; for VPL (no third module) the two coincide; identified
A1 sole-carrier retention	J; $\lvert C_1\rvert = 1$; $K = C_1$	winner-take-all: only one constituent had a single-set solution; identified
A2 independent retention	J; $\lvert C_1\rvert \geq 2$; $C_1 \subseteq K$	redundant duplication: every carrier is a retained single-set solution; identified
B redundant duplication	J; $C_1 \neq \emptyset$; $\lvert K\rvert \geq 2$; not A2	A2: at least one carrier is new to its module; identified (VPL: reversions reported as a flag)
C winner-take-all	J; $\lvert C_1\rvert \geq 2$; $\lvert K\rvert = 1$; $K \subseteq C_1$	A2: only one of several single-set solutions still carries alone; identified at stand-alone sufficiency (graded help of the other module not excluded)
C carrier moved	J; $C_1 \neq \emptyset$; $\lvert K\rvert = 1$; $K \cap C_1 = \emptyset$	winner-take-all: the surviving carrier had no single-set solution; identified
D soft division (additive)	J; $C_1 \neq \emptyset$; $K = \emptyset$; stand-alone effects add up to the repair	E/F: additivity at the answer log-likelihoods (first order); identified at that level
E complementary	VL: as D but not additive; a co-trained module repairs it with a single-trained partner that fails alone	F: the contribution transfers to a foreign partner; VL only

Table 24: Leaves of the joint-repair taxonomy. J: the joint set S repairs the Base error; C_1: modules of S whose single set repairs it; K: modules M of S whose transplant M^S repairs it in Base context. The right column names the neighbouring leaf and the test that separates them; the status says what the available models identify. (continued)

leaf	criterion	separated from its neighbour by; status
F co-adaptation (private code)	VL: not additive, no transfer, and a co-trained module breaks a single-trained partner's own repair	E: no transfer; candidate only (the break may be a generic shift, not shared code)
E/F ambiguous	not additive; recombinations uninformative or conflicting	E vs. F not identifiable; all such repairs of PL, VP, VPL (no cross-recombination of their co-trained modules)
H lost repair	not J; $C_1 \neq \emptyset$; $K = \emptyset$	antagonism: no module of S still repairs it alone; identified
H antagonism	not J; $K \neq \emptyset$	lost repair: a module of S alone repairs it, S does not; identified
I passenger (module level)	the module moved, but alone, removed from S, and (VPL) added to each sub-model its effect stays within ± 2 pp	contributor or antagonist: some effect's interval leaves the band; ambiguous when intervals are wide
J representation without function	a probe of the task variable improves over Base; no stand-alone repair	probed module–task cells only (G02, B02, B07)
K parameter vs. computation ownership	V-only parameters; the change of target selection sits on the frozen LM's query side	G02 only (Q/K/V, Appendix N)
L item routing	the stand-alone carrier set of each repaired item	descriptive; the leaves above are functions of it
M pair → triple	a pair's repairs in VPL: same carriers, changed carriers, or lost	descriptive

Results (Table 25, Figure 10). *Most joint repairs have exactly one stand-alone carrier, and where several constituents had a solution, joint training usually keeps one.* In every joint set and both seeds, sole-carrier retention, winner-take-all, and moved carriers together make up more than half of the repairs. Winner-take-all is the largest leaf outside retention (PL 35.7/33.1, VP 48.6/48.9, VL 41.7/36.3, VPL 39.3/37.6%) and outnumbers independent retention of two solutions, which stays below a tenth of every set's repairs (1.4/1.3, 2.7/2.9, 6.3/9.4, and 3.3/2.3%). Redundant duplication created by joint training is 1.4/1.5, 1.6/1.0, 5.0/5.6, and 11.9/8.8%.

Table 25: Leaves of the joint-repair taxonomy (Table 24) as a share (%) of each joint set's repaired Base errors, items pooled over the 13 tasks; each cell is seed 1101/seed 1104. Per-task shares and intervals are released with the item tables.

leaf (share of the arm's repairs, %)	PL	VP	VL	VPL
new, globally new (G)	1.3 / 2.1	2.6 / 2.6	5.8 / 7.2	7.0 / 8.5
new, constituent-new only (G)	0.9 / 1.2	4.0 / 7.0	3.2 / 3.1	0.0 / 0.0
independent retention (A)	1.4 / 1.3	2.7 / 2.9	6.3 / 9.4	3.3 / 2.3
sole-carrier retention (A)	55.5 / 56.4	31.4 / 30.1	24.4 / 20.8	12.8 / 26.0
redundant duplication (B)	1.4 / 1.5	1.6 / 1.0	5.0 / 5.6	11.9 / 8.8
winner-take-all (C)	35.7 / 33.1	48.6 / 48.9	41.7 / 36.3	39.3 / 37.6
carrier moved (C)	1.6 / 3.1	2.1 / 1.3	2.9 / 2.8	6.2 / 6.2
soft, additive (D)	0.0 / 0.0	0.7 / 0.5	0.8 / 0.9	0.8 / 0.8
complementary, transfers (E)	0.0 / 0.0	0.0 / 0.0	3.8 / 5.3	0.0 / 0.0
co-adaptation cand. (F)	0.0 / 0.0	0.0 / 0.0	0.8 / 1.1	0.0 / 0.0
E vs F ambiguous	2.3 / 1.2	6.4 / 5.8	5.2 / 7.5	18.5 / 9.7
repaired Base errors	3,941 / 3,790	3,071 / 3,103	4,127 / 4,075	3,749 / 4,214

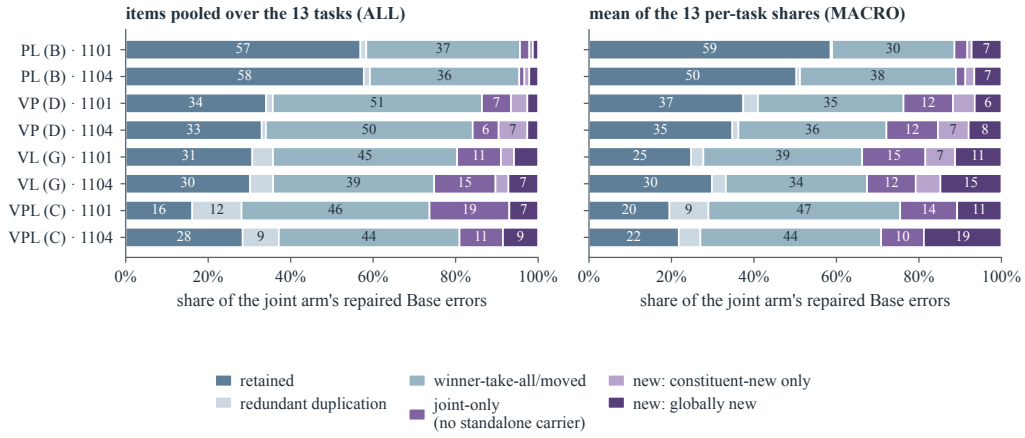

Figure 10: Outcome families of joint repair per joint set and seed (run labels: B = PL, D = VP, G = VL, C = VPL); left: items pooled over the 13 tasks, right: mean of the per-task shares. Retained: A1 + A2; redundant duplication: B; winner-take-all / moved: C; joint-only (no stand-alone carrier): D + E + F + E/F-ambiguous; new: G, split into constituent-new only and globally new.

New capability is the smaller share, and emergence is not ownership. Globally new repairs are 1.3/2.1, 2.6/2.6, 5.8/7.2, and 7.0/8.5% of the repairs of PL, VP, VL, and VPL; counting also the repairs that only the third single set makes (constituent-new), 2.1/3.3, 6.6/9.6, 9.0/10.3, and 7.0/8.5%. In every joint set and both seeds, reallocated repairs (Section 3.3) outnumber the globally new ones (paired differences +7.7 [+4.6, +10.8] and +10.5 [+7.3, +13.8] pp for VL, every interval of the four sets above zero). Of the new (constituent-new) repairs, one co-trained module alone in Base context makes 90.5/79.4, 51.5/48.1, 54.0/47.7, and 39.5/47.4%: no single set of the joint set's own modules makes these repairs, yet once joint training has learned them, the solution can belong to one module. Of the rest, only a few are additive (1.2/2.4, 1.0/2.4, 0.0/1.7, and 2.3/3.6% of the new repairs, against 8.3/18.3, 47.5/49.5, 46.0/50.6, and 58.2/49.0% synergistic), and among repairs with a single-set solution the additive leaf D is at most 1% of each set's repairs (0.8/0.9% in VL) and smaller than the synergistic leaves. The synergy is strongest where V and L are co-trained: the median interaction of the joint-only repairs (gold margin of S minus the sum of its modules' stand-alone changes) is 10.5/7.2 and 6.4/11.5 nats in VL and VPL, against 1.9/2.5 and 1.2/1.1 in PL and VP.

Complementary versus co-adapted repairs can be told apart only in VL. There, the synergistic repairs whose contribution transfers to a single-trained partner (E, 3.8/5.3%) outnumber the private-code candidates (F, 0.8/1.1%), and 5.2/7.5% remain ambiguous. In PL, VP, and VPL every synergistic joint-only repair is E/F-ambiguous (2.3/1.2, 6.4/5.8, and 18.5/9.7%), because no recombination of their co-trained modules with foreign partners exists; for them, private code is neither supported nor excluded. The F leaf is a candidate only; Appendix G.2 tests it for VL across seeds.

Lost repairs and antagonism (H). Every joint set loses between a tenth and a fifth of the Base errors that its constituents' single sets repair, with no module left carrying them (15.2/16.7, 13.3/13.9, 10.3/15.6, and 17.1/14.2%). On a further 2.3/2.0, 4.8/5.3, 8.9/4.9, and 12.6/7.1% of them a module of the joint set still repairs the item alone but the joint set does not, most often in VPL.

Passenger updates (I). Averaged over tasks, P is a passenger in PL and VPL: it moves (update norm 0.13/0.13 and 0.10/0.10 times that of P), but alone, reverted, or added to each sub-model its effect stays within ±2 pp. In VP (norm ratio 0.23/0.23) removing P costs accuracy, so P contributes there.

Representation without function (J) and parameter versus computation ownership (K), G02. V^{VL} raises the decodability of the G02 target's type at its image tokens (+0.14/+0.12; Appendix M), but its stand-alone G02 gain has an interval inside ±2 pp (+0.9/+0.9 pp, upper bounds +2.0/+2.0): the representation changed without a stand-alone repair. Conversely, V, whose trained parameters are all in the vision encoder, changes late-layer target selection only on the query side of the frozen language model (key side +0.005/+0.013, query side +0.143/+0.134; Appendix N): the module that adapted is not where the computation changes.

Pair to triple (M). Most repairs of VL keep their stand-alone carriers in VPL (64.2/74.3%), whereas few repairs of VP do (21.1/23.7%) and L joins as a stand-alone carrier on 55.7/57.6% of them. PL's repairs keep their carriers in VPL in 55.5/73.1% and are lost in 21.7/8.7%. Repairs of VPL that no pair and no single set makes are 1.9/2.9% of its repairs.

Reading. In both seeds joint training mostly realizes one of the available single-module solutions, or moves a repair to one carrier, rather than keeping several side by side or summing partial contributions; the joint-only repairs it adds are rarely additive and mostly synergistic, strongly so when V and L are co-trained, and part of what emerges only under joint training is carried by one module afterwards. The leaves are defined by stand-alone sufficiency in Base context: a module that is not a stand-alone carrier may still help inside S, and additivity is tested at the answer log-likelihoods only. Status: the leaf shares are supported in both seeds; co-adaptation (F) is a candidate, and complementary versus co-adapted is unresolved outside VL.

G.2 Cross-seed recombination: is the VL synergy private to one run?

Question. Within a seed, the repairs that VL makes and neither V nor L makes survive poorly when a co-trained module is paired with a single-trained partner (Table 26). Is that because the co-trained V and L share code that only their own run produces (the F leaf), or do co-trained partners from an independent run also work?

Design. We combine V^{VL} of one seed with L^{VL} of the other seed (P at Base), in both directions, and score the two models on all 13 tasks like the models of Appendix F (settings of the seed-1104 campaign, bitwise-gated). For each seed s the references are s's VL repairs (4,127/4,075 Base errors) and its VL pair-specific items (Base, V, and L wrong, VL right; 372/421); *retention* is the share of them a model repairs. Rules fixed before scoring: the recombination is *equivalent* to VL if its 13-task accuracy is within 2 pp of s's VL and it retains at least 80% of the pair-emergent items; the synergy is *run-private* if the recombination retains fewer of them than the within-seed $V^{VL}{+}L^{L}$ (paired interval below zero) while the latter retains at least half. The cross-seed $V^{V}{+}L^{L}$ control was not scored.

Results (Table 26). In both seeds and both directions the cross-seed recombinations retain most of the pair-emergent items: with s's V and the other seed's L, 83 [76, 88] and 80 [73, 85]%; with the other seed's V and s's L, 85 [78, 93] and 82 [76, 88]%. A same-run co-trained module with a single-trained partner retains far fewer (39/41% for $V^{VL}{+}L^{L}$, 40/47% for $V^{V}{+}L^{VL}$); every paired difference between a cross-seed recombination and these two lies above zero (against $V^{VL}{+}L^{L}$: +44.1 and +46.8 pp in seed 1101, +38.7 and +41.6 pp in seed 1104). The run-private rule is therefore met in neither direction. Nor is the equivalence rule: the recombinations' 13-task accuracy differs from s's VL by -3.0 [-4.2, -1.8] and +3.0 [+1.8, +4.1] pp in seed 1101 and by +3.2 [+1.9, +4.4] and -2.8 [-4.1, -1.5] pp in seed 1104, in opposite directions.

Reading. The synergy of the analysed VL pair-emergent repairs mostly does not require run-private code: a co-trained partner from an independent run substitutes for the run's own partner on most of them, whereas a single-trained partner from the same run does not. What VL co-training builds between V and L is thus reproduced by independent co-trainings, although the accuracy offsets show that the two runs' modules are not exactly interchangeable. This concerns VL only: PL, VP, and VPL have no recombinations of their co-trained modules and remain not separable (Appendix G.1). Status: supported for VL in both seeds; unresolved for PL, VP, and VPL.

Table 26: Cross-seed recombination of VL's modules (P at Base). For reference seed s: 13-task accuracy, and the share of s's VL repairs and of its VL pair-emergent items that each model repairs (%, 95% scene-family intervals in brackets for the recombinations). "s" and "other" name the seed a module comes from; the other rows are models of seed s.

	seed 1101			seed 1104		
model	acc.	VL rep.	pair-em.	acc.	VL rep.	pair-em.
VL of seed s	73	100	100	73	100	100
VL of the other seed	73	88	72	73	89	73
V^{VL} (s) + L^{VL} (other)	70	88	83 [76, 88]	76	95	80 [73, 85]
V^{VL} (other) + L^{VL} (s)	76	97	85 [78, 93]	70	88	82 [76, 88]
V^{VL}+L^{L}	74	88	39	75	88	41
V^{V}+L^{VL}	67	84	40	68	85	47
V^{V}+L^{L}	69	84	31	72	87	41
V^{VL} alone	39	21	12	40	25	11
L^{VL} alone	65	76	46	62	70	40
V	51	58	0	51	60	0
L	73	84	0	69	80	0
Base	32	–	–	32	–	–

G.3 P-STITCH: IS THE LOST STAND-ALONE REPAIR OF THE CO-TRAINED VISION ENCODER INTERFACE INCOMPATIBILITY?

Question. On G02, V^{VL} in Base context scores like Base (24/24% vs. 24/24%), although V alone reaches 90/90% (Appendix F). Does the co-trained vision encoder lack a usable repair, or are its outputs only off-interface for the Base projector and the frozen Base language model?

Design. We train only P, from Base, between the frozen V^{VL} and the frozen Base L, for 244 updates (the first stored update of the sets) with the recipe of P (learning rate 2×10^{-6}, 74 warmup updates, the cosine schedule of the full run stopped at update 244) on exactly the examples and order of P's first 244 updates. The update-matched control is P at update 244, which differs from the stitch only in using Base V. Rule fixed before scoring, on G02: if the stitch exceeds the control (interval above zero) and its interval lies above P's final accuracy, the lost repair is interface incompatibility; if it exceeds the control only, the incompatibility is *partial*; otherwise it is not supported. We also report the share of the gap between V^{VL} and V that the stitch recovers.

Results (Table 27). On G02 the stitch reaches 38/39%: above V^{VL} (+13 [+10, +17] and +15 [+11, +19] pp) and above the update-matched control (+8 [+5, +11] and +10 [+6, +13] pp), but below P's full-budget accuracy (-25 [-32, -17] and -23 [-30, -16] pp) and far below V (+52 [+47, +58] and +50 [+44, +56] pp). It recovers a share 0.20 [0.15, 0.26] and 0.23 [0.17, 0.29] of the gap to V, less than half, and the pre-declared verdict is *partial* in both seeds. The gain needs the image (blank image: stitch 24/24%, V^{VL} 24/24%), and 98/98% of the Base errors the stitch repairs on G02 are also repaired by V. Over the 13 tasks the stitch also exceeds the update-matched control (+10 [+9, +12] and +9 [+8, +11] pp) and recovers a share 0.58 [0.48, 0.69] and 0.56 [0.45, 0.67] of the macro gap to V.

Reading. The co-trained vision encoder's lost stand-alone G02 repair is partly interface incompatibility: a retrained projector lets the frozen Base language model use part of V^{VL}'s change, beyond what the same projector training achieves with Base V, and what it recovers are V's repairs. It is not mostly incompatibility at this budget, so "V co-trained with L loses its stand-alone repair" should be read as "loses most of it through the Base interface; a short projector-only stitch recovers part of it". Limits: one budget (244 updates, not converged; a longer stitch could recover more) and one cell; the L side (the co-trained L behind a stitched interface) was not run, so for L incompatibility is not separated from a change of its computation. Model stitching usually inserts a trained layer to compare representations (Lenc & Vedaldi, 2015; Bansal et al., 2021); here the retrained projector is the model's own interface and the question is functional. Status: partial, in both seeds.

Table 27: P-stitch: accuracy (%) on G02 and the 13-task macro. V|VL: the co-trained vision encoder with Base P and L (V^{VL}); A@244: P at update 244 (update-matched control); A: P at the end; P-stitch: V^{VL}, P trained for 244 updates, Base L; E = V; G = VL.

	seed 1101		seed 1104	
model	G02	macro	G02	macro
Base	23.5	25.5	23.5	25.5
V|VL (G.V + Base P)	24.5	34.4	24.5	34.2
A@244 (P only, 244 upd.)	30.1	31.9	29.8	31.8
A (P only)	62.4	37.5	62.4	38.0
P-stitch (G.V + P*)	37.6	42.2	39.5	40.9
E (V only)	89.7	47.7	89.7	46.3
G (VL)	95.6	58.9	95.6	60.4

G.4 Clipping control for the update geometry

Question. Co-training L rotates V's update away from its single-set direction ($\cos(\Delta V_{VL}, \Delta V_V) = 0.23/0.23$ at the end, 0.27/0.26 at update 244) and shrinks it (Section 4.2, Appendix I). Training clips the global gradient norm over all trainable parameters, so in VL L's larger gradient sets V's clip factor. Does the joint clip alone produce the rotation and the shrinkage?

Design. A control retrains VL's recipe with V and L trainable from the first update (L has gradients, enters the global clip norm, and has optimizer state) but with L's learning rate zero at every update, so L never moves: the model is that of V, the clip set that of VL. One run per seed trains 244 updates on the full schedule, and its V update ΔV_{ctl} is compared with those of V and VL at update 244. Rule fixed before the runs finished: the control reproduces a share $f = (c - \cos(\Delta V_{\mathrm{ctl}}, \Delta V_V))/(c - \cos(\Delta V_{VL}, \Delta V_V))$ of the gap below a run-to-run ceiling c; $f \geq 2/3$ means a clipping artifact, $f \leq 1/3$ that clipping does not explain the rotation. The primary ceiling is the cosine between VL and its replicate, the secondary that between V and the V-first retrain, which trains V alone until update 1,220 (both from Appendix K, on another GPU type, the same hardware split as between the control and V).

Results (Table 28). The control's L equals Base in every one of its tensors in both seeds. Its V update points along V's ($\cos(\Delta V_{\mathrm{ctl}}, \Delta V_V) = 0.90/0.91$, against 0.27/0.26 for VL), between the two run-to-run ceilings (0.87/0.76 for the VL rerun, 0.96/0.96 for the V rerun), so $f = -0.04/-0.29$ (0.09/0.07 with the secondary ceiling): not clipping, in both seeds. It is not shrunk (norm ratio to ΔV_V 1.09/1.08, against 0.56/0.56 for VL), and in each of the 26 encoder layers its cosine with ΔV_V exceeds that of VL (medians over layers 0.90/0.91 vs. 0.28/0.27). V's clip factor is even smaller in the control than in VL (L, held at Base, keeps a larger pre-clip gradient), and the control's training loss tracks that of V, not that of VL.

Reading. Global-norm clipping does not explain the rotation or the shrinkage of V's update: they come from L moving during co-training, so the update geometry of Section 4.2 needs no "under joint clipping" qualifier. The control covers update 244 (10% of training) only, where the rotation is already present (Appendix I). Status: a clipping artifact is not supported in either seed.

Table 28: Clipping control at update 244: cosines and norm ratios of V updates ($\Delta V = V_t - V_0$) of the control (ctl: VL's recipe with L's learning rate zero), V (E), and VL (G); ceilings: the VL replicate and the V-first retrain (V alone until update 1,220) against their counterparts; f: share of the gap between the ceiling and $\cos(\Delta V_G, \Delta V_E)$ that the control reproduces.

step 244	seed 1101	seed 1104
$\cos(\Delta V_{\text{ctl}}, \Delta V_E)$	0.90	0.91
$\cos(\Delta V_G, \Delta V_E)$	0.27	0.26
$\cos(\Delta V_{\text{ctl}}, \Delta V_G)$	0.28	0.27
ceiling: G rerun	0.87	0.76
ceiling: E rerun	0.96	0.96
$\|\Delta V_{\text{ctl}}\|/\|\Delta V_E\|$	1.09	1.08
$\|\Delta V_G\|/\|\Delta V_E\|$	0.56	0.56
gap reproduced f (G / E rerun)	-0.04 / 0.09	-0.29 / 0.07

G.5 Retention regression: does L's early repair predict V's loss of an item?

Question. Co-trained with L, V keeps only 24.7–24.9/27.0–27.9% of V's repairs, against 88.8/89.3% with P (Section 4.2, Appendix F). If the module that reduces an item's loss first starves the other of gradient on that item, the earlier L^{VL} repairs an item, the less V^{VL} should keep V's repair of it.

Design. The population is each seed's Base errors that V repairs; the outcome is whether V^{VL} also repairs the item, and a placebo outcome whether V^{VP} does. The predictor is the stable onset of L_t^{VL}'s repair (the first of the stored updates 244, 610, 1,220, 1,830, and 2,439 from which L alone repairs the item at every later one, ordered, with "never" last; Appendix J), in a ridge-penalized logistic regression with task fixed effects, Base's gold margin, and the number of the other single sets that repair the item, with intervals from scene-family bootstrap refits. Rule fixed in advance: starvation is supported in a seed if the onset coefficient for V^{VL} lies above zero and above the placebo's (paired interval above zero).

Results. Of V's repairs (3,026/3,057 Base errors), V^{VL} keeps 24.7/27.9% and the placebo V^{VP} 88.8/89.3%. Descriptively, V^{VL} keeps fewer of the repairs that L_t^{VL} makes from update 244 on (15.8 [12.8, 18.9] and 22.2 [18.5, 25.8]%; 1,603/1,363 items) than of those it never makes (39.8 [33.5, 46.3] and 34.5 [30.1, 39.0]%; 871/1,179 items). Once task, Base margin, and the other single sets are held fixed, the onset-rank coefficient's interval includes zero in both seeds (0.04 [−0.05, 0.14] and 0.03 [−0.05, 0.11]), its difference from the placebo's −0.01 [−0.15, 0.14] and 0.01 [−0.12, 0.13], and the adjusted retention is 23.6/27.2% for items L repairs from update 244 and 25.9/28.7% for items it never repairs. With the onset replaced by an indicator of repair from update 244 the coefficient is −0.50 [−0.84, −0.19] and −0.07 [−0.44, 0.24], and its difference from the placebo's (−0.35 [−0.88, 0.19] and −0.05 [−0.57, 0.49]) includes zero in both seeds. The pre-declared verdict is *not supported* in both seeds.

Reading. V co-trained with L loses most of V's stand-alone repairs, and the loss is specific to co-training with L (the placebo keeps most of them). It is not graded by when L learns the item, so gradient starvation in this item-timing form is not supported. Joint clipping alone leaves V's update close to V's (Appendix G.4), which makes the procedural coupling an unlikely source; the rotation of V's update comes from L moving during co-training, which is consistent with, but does not show, the same source for the lost repairs. The regression is observational: onset, the Base margin, and the task are adjusted for, but items that L repairs early may differ from late ones in ways the covariates miss. Status: item-level gradient starvation is not supported in either seed; the source of the lost repairs remains unresolved.

H A third training seed for the behavioural results (seed 1102)

What was run. Seed 1102 trains the seven sets of Section 3 with the same data, recipe, and update and exposure counts, and its own example order. Its models enter only the behavioural analyses: the factorial landscape and the repair space of RQ1, the transplants, reversions, and recombinations of RQ2 (Appendix F), and the outcome taxonomy (Appendix G.1). The update geometry, training dynamics, route steering, and every mechanistic analysis of RQ3 stay on seeds 1101 and 1104. All

247 scoring jobs that these analyses read ran on one GPU type (NVIDIA RTX 6000 Ada) with the scorer settings of seed 1104. A pool of RTX A6000 GPUs was tested first and not used for these jobs: re-scoring 864 rows there gave identical correctness on 99.19% of them, below the pre-set acceptance threshold (all 28 differences are ties at bfloat16 precision); those GPUs scored only seed 1102's training-dynamics and loss-landscape models, which enter no result of this paper.

Source of the numbers. For seeds 1101 and 1104, the RQ1 landscape and repair space use the official, receipt-verified predictions. For seed 1102 the official predictions of P, PL, VPL, and VP are incomplete, so the RQ1 landscape and repair space are computed from the re-scored predictions of the same models, which RQ2 uses in every seed, with the same estimators (13 tasks, 2,000 scene-family replicates). Two checks support this substitution. The same pipeline applied to seed 1104 reproduces the official outcome of 89 of the 89 claim checks of RQ1 and the behavioural RQ2 analyses that both sources evaluate (values in column "1104 (re-scored)" of Table 29). Because seed 1104's official predictions of the seven sets agree with their re-scored ones on every item (Table 17), this check isolates the replacement of the official Base. On the four models with official predictions (Base, V, L, VL), seed 1102's re-scored predictions pass 3 of the 3 agreement checks of Appendix F.

Claims. We re-evaluated every binding claim check of RQ1, of the behavioural RQ2 analyses, and of the outcome taxonomy with the generators' own code and thresholds (Tables 30–33). Of the 95 checks that the paper states for both seeds, 84 hold in seed 1102: the RQ1 landscape 14/19, the RQ1 repair space 25/28, RQ2 28/30, and taxonomy 17/18. The holding checks include all 10 checks of these results on which the abstract, introduction, or contributions rely: for example, every set is above Base; most Base errors are repaired by some single-module adaptation, and more by two or three than by exactly one; V co-trained with L keeps only a quarter to under a third of its stand-alone repairs; and reallocated repairs outnumber new ones in every joint set. For seed 1102, 6 checks cannot be evaluated: the image-swap and option-remap controls were not scored, the G02 plan numbers exist only for seed 1101, and three taxonomy checks need update geometry, probes, or causal patching. Of the 9 checks that the paper states per seed, seed 1102 has seed 1101's outcome on 4 and seed 1104's on 5. On the L-trained sets it matches neither: L is not below all three of PL, VL, and VPL, but not every difference among them includes zero either (rows 4–5).

What does not replicate. Seed 1102 fails 11 checks (rows 10, 11, 15, 16, 18 (RQ1 landscape); 34, 43, 56 (RQ1 repair space); 66, 89 (RQ2); 103 (taxonomy)).

- RQ1 landscape: the ordering of simple effects (every without-L simple effect of V and P above zero, and every with-L simple effect smaller than every without-L one), and the task patterns of B01 (L and PL gain the most; L without V gains more than L with V) and B07 (P gains the most; V, L, and PL gain the least).
- RQ1 repair space: B04 as the task with the largest pair-emergent share; the V–L overlap lying within the pre-set band around chance; and P's lead being specific to B07.
- RQ2: L co-trained with V keeping more than the pre-set share of its stand-alone gain; and $P^P{+}L^L$ matching PL within the pre-set margin.
- Taxonomy: additive joint-only repairs staying below the pre-set bound in every joint set.

Most of these are task-level patterns or orderings of secondary contrasts. Three lie close to their pre-set thresholds: the V–L overlap, 0.973 [0.946, 0.996] (observed/expected), stays near chance but its interval ends just below the band; L co-trained with V keeps 74% of L-alone's gain; and one joint set's additive share is 1.2%. One is a clear difference: $P^P{+}L^L$ differs from PL by -4.2 [-5.8, -2.6] pp in seed 1102, so the statement that it matches PL does not replicate. The checks on which the abstract rests hold in all three seeds. Three seeds still do not estimate seed variance, and the main text states without a seed only what holds in seeds 1101 and 1104.

Table 29: Seed 1102: headline values (95% scene-family intervals). Columns 1101 and 1104 are the paper's values (RQ1 landscape and repair space from official predictions); "1104 (re-scored)" applies seed 1102's pipeline to seed 1104; RQ2 uses re-scored predictions in every column.

	quantity	1101	1104	1104 (re-scored)	1102 (re-scored)
R1	gain vs Base, P (pp)	+12.1 [+10.0, +14.5]	+12.5 [+10.3, +14.9]	+12.5 [+10.3, +14.8]	+12.2 [+10.0, +14.5]
R1	gain vs Base, V (pp)	+22.8 [+20.1, +25.5]	+21.3 [+18.7, +24.1]	+21.3 [+18.6, +24.1]	+22.0 [+19.5, +24.7]
R1	gain vs Base, L (pp)	+33.8 [+31.3, +36.4]	+30.5 [+28.0, +33.0]	+30.4 [+27.9, +33.1]	+31.4 [+28.8, +33.9]
R1	gain vs Base, PL (pp)	+33.0 [+30.5, +35.6]	+31.5 [+29.0, +34.0]	+31.5 [+28.9, +34.1]	+32.9 [+30.4, +35.6]
R1	gain vs Base, VP (pp)	+24.9 [+22.2, +27.5]	+23.7 [+21.1, +26.4]	+23.7 [+21.2, +26.4]	+22.9 [+20.5, +25.3]
R1	gain vs Base, VL (pp)	+33.6 [+30.7, +36.2]	+34.5 [+31.9, +37.2]	+34.5 [+31.8, +37.2]	+33.2 [+30.7, +36.0]
R1	gain vs Base, VPL (pp)	+34.3 [+31.6, +37.0]	+35.0 [+32.1, +37.9]	+35.0 [+32.1, +38.0]	+35.3 [+32.7, +37.9]
R1	main effect V (pp)	+9.2 [+7.7, +10.6]	+10.0 [+8.7, +11.4]	+10.0 [+8.7, +11.4]	+9.2 [+8.1, +10.4]
R1	main effect P (pp)	+3.5 [+2.7, +4.4]	+4.1 [+3.3, +4.9]	+4.1 [+3.3, +4.9]	+4.2 [+3.3, +5.1]
R1	main effect L (pp)	+18.7 [+16.8, +20.5]	+18.5 [+16.7, +20.3]	+18.5 [+16.6, +20.3]	+18.9 [+17.0, +20.7]
R1	interaction VL (pp)	−17.2 [−20.0, −14.6]	−12.5 [−15.1, −10.0]	−12.5 [−15.1, −10.0]	−14.2 [−16.7, −11.8]
R1	interaction VPL (pp)	+11.5 [+8.5, +14.8]	+9.6 [+6.6, +12.8]	+9.6 [+6.5, +12.9]	+11.9 [+8.6, +15.2]
R2	single-emergent share of Base errors (%)	77.9 [75.2, 80.5]	76.8 [74.2, 79.3]	76.8 [74.2, 79.3]	77.4 [74.8, 79.8]
R2	pair-emergent share (%)	5.1 [3.9, 6.4]	5.8 [4.6, 7.1]	5.9 [4.7, 7.1]	4.9 [4.0, 5.9]
R2	triple-emergent share (%)	1.1 [0.6, 1.6]	1.9 [1.2, 2.6]	1.9 [1.2, 2.6]	2.1 [1.3, 2.9]
R2	unrepaired share (%)	15.9 [13.3, 18.6]	15.5 [13.0, 18.2]	15.4 [13.0, 18.2]	15.6 [13.2, 18.3]
R2	repaired by $\geq$2 single sets (%)	45.2	43.5	43.6	41.6
R2	V&L overlap, observed/expected	1.020 [1.002, 1.036]	0.997 [0.974, 1.018]	0.999 [0.977, 1.021]	0.973 [0.946, 0.996]
R3	V alone (E) in Base context (pp)	+22.3 [+19.7, +25.0]	+21.3 [+18.6, +24.1]	+21.3 [+18.6, +24.1]	+22.0 [+19.5, +24.7]
R3	V of VL in Base context (pp)	+8.5 [+6.9, +10.2]	+8.4 [+6.7, +10.0]	+8.4 [+6.7, +10.0]	+6.9 [+5.4, +8.6]
R3	L alone (F) (pp)	+33.6 [+31.0, +36.4]	+30.4 [+27.9, +33.1]	+30.4 [+27.9, +33.1]	+31.4 [+28.8, +33.9]
R3	L of VL in Base context (pp)	+26.8 [+24.0, +29.4]	+26.5 [+24.0, +29.1]	+26.5 [+24.0, +29.1]	+23.1 [+20.3, +25.9]
R3	VL repairs carried by L alone (%)	76.1 [73.2, 79.1]	70.0 [66.8, 73.1]	70.0 [66.8, 73.1]	74.0 [71.3, 76.6]
R3	V retained in VL (% of V-alone repairs)	24.7 [21.8, 27.3]	27.9 [25.1, 30.5]	27.9 [25.1, 30.5]	23.5 [20.9, 26.0]
R3	VL: new repairs (%)	5.8 [4.1, 7.6]	7.2 [5.4, 8.9]	7.2 [5.4, 8.9]	5.2 [4.0, 6.5]
R3	VL: reallocated repairs (%)	13.5 [11.2, 15.9]	17.7 [15.1, 20.3]	17.7 [15.1, 20.3]	19.4 [17.0, 21.8]
R3	recombination $V^V + L^L$ gain (pp)	+30.1 [+27.1, +33.2]	+32.2 [+29.2, +35.5]	+32.2 [+29.2, +35.5]	+30.8 [+27.8, +33.8]

Table 30: Seed 1102: claim checks of the RQ1 landscape (✓ holds, × fails, – not evaluable). Status: "both seeds" = stated for seeds 1101 and 1104; "per seed" = stated for one seed only; "n/e" = not evaluable for seed 1102.

#	claim (where)	1101	1104	1102	status
1	every trainable set is above Base and every 95% interval of the change excludes zero (Abstract / R1)	✓	✓	✓	both seeds
2	P has the smallest and VPL the largest gain (point estimates) (R1)	✓	✓	✓	both seeds
3	every L-trained set (L, PL, VL, VPL) is above V and VP, which are above P, each difference's interval excluding zero (R1)	✓	✓	✓	both seeds
4	no difference among the four L-trained sets has a 95% interval that excludes zero (R1)	✓	×	×	per seed
5	L alone is below PL, VL, and VPL, each difference's 95% interval excluding zero (R1)	×	✓	×	per seed
6	L alone is the best single set and VPL is above it in the point estimate (R1)	✓	✓	✓	both seeds
7	every factorial contrast's 95% interval excludes zero (R1)	✓	✓	✓	both seeds
8	main effects ordered ME_L $>$ ME_V $>$ ME_P $>$ 0, and the intervals of ME_L-ME_V and ME_V-ME_P exclude zero (R1)	✓	✓	✓	both seeds
9	all pairwise interactions negative, three-way positive, and |INT_VL| is the largest interaction (Abstract / R1)	✓	✓	✓	both seeds
10	every without-L simple effect of V and P has a 95% interval above zero (R1)	✓	✓	×	both seeds
11	for V and P, every with-L simple effect is smaller than every without-L simple effect (R1)	✓	✓	×	both seeds
12	none of the four with-L simple effects has a 95% interval that excludes zero (R1)	✓	×	×	per seed
13	three of the four with-L simple effects have a 95% interval above zero and the fourth includes zero (R1)	×	✓	✓	per seed
14	the three largest mean gains are on G02, B03, and G03 (in some order; the text names them) (R1)	✓	✓	✓	both seeds
15	B01: L and PL gain more than every other set, and L alone gains more than twice V alone (R1)	✓	✓	×	both seeds
16	B01: L and PL (L without V) gain more than VL and VPL (L with V) (R1)	✓	✓	×	both seeds
17	B05: L and PL (L without V) gain more than VL and VPL (L with V) (R1)	✓	✓	✓	both seeds
18	B07: P alone has the largest gain; V, L, and PL have the three smallest (R1)	✓	✓	×	both seeds
19	G02: V alone and L alone each gain more than 60 pp, P alone more than 30 pp (R1)	✓	✓	✓	both seeds
20	G02: each single set alone closes more than half of the gap between Base and 100% (Abstract / Intro / R1)	✓	✓	✓	both seeds
21	B03: every V-trained set gains more than PL and L (R1)	✓	✓	✓	both seeds
22	V01 (length, real): no set improves (every change is negative) (Abstract / R1)	✓	✓	✓	both seeds
23	V01 (length, real): no single set's 95% interval excludes zero (the text says so) (R1)	✓	×	✓	per seed
24	V01 (length, real): only VL's 95% interval excludes zero (below zero) (R1)	×	✓	×	per seed
25	blank-image gain is less than 30% of each set's clean gain (most of each gain needs the row's image) (R1)	✓	✓	✓	both seeds
26	under image swap at most a fifth of each set's clean gain survives, except P ($<$ 0.6) (R1)	✓	✓	–	n/e
27	under option remap every set is below Base and every L-trained set is below every L-frozen set (App. B)	✓	✓	–	n/e

Table 31: Seed 1102: claim checks of the RQ1 repair space (notation as in Table 30).

#	claim (where)	1101	1104	1102	status
28	most Base errors have a single-module repair; fewer than one in ten first appear at a pair or VPL (Abstract / R2)	✓	✓	✓	both seeds
29	the single-repairable share's 95% interval lies above 70% (R2)	✓	✓	✓	both seeds
30	task-mean single share equals the MACRO row (R2)	✓	✓	✓	both seeds
31	the >95%-single tasks are B01, B05, G01, G02, and G03 (the text names them) (R2)	✓	✓	✓	both seeds
32	B07 has at least 35% unrepaired Base errors (R2)	✓	×	✓	per seed
33	the two task groups (>95% single; ≥35% unrepaired) are disjoint and non-empty (R2)	✓	✓	✓	both seeds
34	B04 has the largest pair-first share (the text names B04) (R2)	✓	✓	×	both seeds
35	new capability equals pair-first plus triple-first (R2 / R3)	✓	✓	✓	both seeds
36	PL: pair-specific repairs outnumber globally new ones (R2)	✓	✓	✓	both seeds
37	VP: pair-specific repairs outnumber globally new ones (R2)	✓	✓	✓	both seeds
38	VL: pair-specific repairs outnumber globally new ones (R2)	✓	✓	✓	both seeds
39	more than a quarter of VPL's pair-relative gains is already reached by a single set ('Likewise') (R2)	✓	✓	✓	both seeds
40	more Base errors have two or three single-module repairs than exactly one (Abstract / R2)	✓	✓	✓	both seeds
41	G02: more than half of the Base errors are repaired by all three single sets (R2)	✓	✓	✓	both seeds
42	L is the unique single repair far more often than V or P (> 3x each) (R2)	✓	✓	✓	both seeds
43	V and L repair the same errors about as often as chance: 95% interval within [0.95, 1.05] (R2)	✓	✓	×	both seeds
44	P&L overlap is not above chance (interval does not lie above 1) (R2)	✓	✓	✓	both seeds
45	P&V overlap is clearly above chance (CI lower bound > 1.2) (R2)	✓	✓	✓	both seeds
46	G02 V&L overlap is also near chance (within 5%) (R2)	✓	✓	✓	both seeds
47	pair-level multiplicity (n_2≥2) overstates the minimal count: more Base errors have ≥2 repairing pairs than ≥2 minimal repairs (R1 (main.tex))	✓	✓	✓	both seeds
48	more Base errors have two or three minimal repairs than exactly one (R1 (main.tex))	✓	✓	✓	both seeds
49	minimal multiplicity ≥2 exceeds n_1≥2 (singles are always minimal; App. repair spaces) (R1 (main.tex))	✓	✓	✓	both seeds
50	every one of the nine edges loses some repairs (> 5%) (R2)	✓	✓	✓	both seeds
51	the edge with the largest lost-repair rate adds L, and L->PL has the smallest (R2)	✓	✓	✓	both seeds
52	every edge that adds L loses more repairs than every edge that adds P (R2)	✓	✓	✓	both seeds
53	mean lost-repair rate ordered by the added module: L > V > P (R2)	✓	✓	✓	both seeds
54	B07 has the largest lost-repair rate and P->PL is among its largest edges (R2)	✓	✓	✓	both seeds
55	lost repair on G02/G03 (lowest tasks) is below 5% (R2)	✓	✓	✓	both seeds
56	B07: P's repair rate is over 2.5x both V's and L's; on every other task P exceeds neither by 5 points (R2 / R4)	✓	✓	×	both seeds

Table 32: Seed 1102: claim checks of RQ2 (notation as in Table 30; "scorer" = agreement of re-scored and official predictions, for seed 1102 on Base, V, L, VL only).

#	claim (where)	1101	1104	1102	status
57	26 models: Base + 7 sets, 9 transplants, 3 reversions, 6 recombinations (R3)	✓	✓	✓	both seeds
58	re-scored correctness agrees with the official predictions on ≥ 98% of items for every real set (R3 / App)	✓	✓	✓	scorer
59	re-scored macros differ from the official ones by less than 1 pp (R3 / App)	✓	✓	✓	scorer
60	the lowest per-task agreement is still high (≥ 90%), so that set's numbers on that task 'differ slightly' (App)	✓	✓	✓	scorer
61	V trained with P gains as much as V alone (interval of the difference includes 0) (R3)	✓	✓	✓	both seeds
62	V trained with L (VL, VPL) gains clearly less than V alone (interval below 0) (R3)	✓	✓	✓	both seeds
63	V trained with L keeps well under half of V-alone's macro gain (< 45%) (R3)	✓	✓	✓	both seeds
64	P trained with any partner: every interval of the stand-alone gain includes 0; P alone's excludes 0 (R3)	✓	✓	✓	both seeds
65	L trained with P keeps L-alone's gain (difference interval includes 0) (R3)	✓	✓	✓	both seeds
66	L trained with V keeps most of L-alone's gain: difference interval below 0 but gain > 75% of L alone (R3)	✓	✓	×	both seeds
67	on G01–G03, V alone gains more than 40 pp but V trained with L (VL, VPL) less than 2 pp (Fig. 3a)	✓	✓	✓	both seeds
68	on B07, L trained with V gains more than L alone (interval of the difference above 0) (R3)	✓	✓	✓	both seeds
69	G02: V alone ≥ 85%, V trained with L within 2 pp of Base, L trained with V ≥ 90% and within 5 pp of VL (R3)	✓	✓	✓	both seeds
70	G02 numbers reproduce the plan (sections 17, 20; seed 1101 only) (R3)	✓	–	–	n/e
71	the module that carries most of each joint set's repairs: L in PL, V in VP, L in VL, L in VPL (R3)	✓	✓	✓	both seeds
72	that module alone repairs > 65% of its set's repairs, > 2x any other module, and its reversion loses > 70% (R3)	✓	✓	✓	both seeds
73	P carries at most 5% of any joint set's repairs; its reversion loses at most 5% of PL's and VPL's repairs but more than twice that in VP (R3)	✓	✓	✓	both seeds
74	in VL and VPL, V alone carries between 15% and 30% of the set's repairs (R3)	✓	✓	✓	both seeds
75	V co-trained with L keeps only a quarter to under a third of V-alone's repairs (every interval within 20–33%); V co-trained with P > 85% (R3 / Contrib.)	✓	✓	✓	both seeds
76	P retains at most 5% of P-alone's repairs with any partner (R3)	✓	✓	✓	both seeds
77	L retains > 90% with P and 60–80% with V (VL, VPL), less than with P (R3)	✓	✓	✓	both seeds
78	the role-reallocation magnitude is largest on G02 (R3)	✓	✓	✓	both seeds
79	in every joint set, reallocated repairs outnumber new ones (R3 / Abstract)	✓	✓	✓	both seeds

Table 32: Seed 1102: claim checks of RQ2 (notation as in Table 30; "scorer" = agreement of re-scored and official predictions, for seed 1102 on Base, V, L, VL only). (continued)

#	claim (where)	1101	1104	1102	status
80	the largest reallocated share exceeds the largest new share (range quoted in the contributions) (Contrib.)	✓	✓	✓	both seeds
81	new repairs are under 10% of every joint set's repairs ('rarely') (Abstract / R3)	✓	✓	✓	both seeds
82	most of every joint set's repairs keep the carrier of single-module training (same carrier > 65%) (R3)	✓	✓	✓	both seeds
83	most reallocated repairs have no single carrier rather than a carrier that moved (every joint set) (App)	✓	×	×	per seed
84	pooled over the four joint sets, most reallocated repairs have no single carrier rather than a carrier that moved (App)	✓	✓	✓	both seeds
85	every recombination with V trained alone falls below its jointly trained set (interval below 0) (R3)	✓	✓	✓	both seeds
86	VˆV+LˆL is below its better member LˆL (interval below 0) (R3 / App)	✓	×	×	per seed
87	VˆV+LˆL is below VL (interval below 0) and loses more member repairs than it gains (R3 / App)	✓	✓	✓	both seeds
88	VˆVL+LˆL matches VL (interval of the difference includes 0) (R3)	✓	✓	✓	both seeds
89	PˆP+LˆL is within 2 pp of PL (interval includes 0) (R3)	✓	✓	×	both seeds
90	every recombination gains less than the sum of its members' gains (interval below 0) (R3)	✓	✓	✓	both seeds
91	each recombination is compared with the jointly trained set of the same modules (R3)	✓	✓	✓	both seeds
92	blank image: every model with a trained L gains more than every model with the Base L; all L-trained intervals exclude 0 (R3)	✓	✓	✓	both seeds

Table 33: Seed 1102: claim checks of the outcome taxonomy (Appendix G.1; notation as in Table 30).

#	claim (where)	1101	1104	1102	status
93	R1a: reallocated (RQ2) > globally new in every joint arm (paired CI of the difference > 0, ALL)	✓	✓	✓	both seeds
94	R1c: reorganised (C1 != {} and not retained: B, C, D, E, F, E/F) > constituent-new in every joint arm (paired CI > 0, ALL)	✓	✓	✓	both seeds
95	R2a: V's standalone repair drops when L is trainable (V|VL, V|VPL minus V|V: CI < 0) and not when only P co-trains (V|VP minus V|V: CI contains 0)	✓	✓	✓	both seeds
96	R2b: P trained with any partner has no standalone gain (P|PL, P|VP, P|VPL: CI contains 0) while P|P > 0	✓	✓	✓	both seeds
97	R2c: the L standalone drop is smaller than the V drop when V and L co-train (VL and VPL: L|S - L|L > V|S - V|V, point estimates)	✓	✓	✓	both seeds
98	R3: the joint L update keeps the single-L direction (cos(dL_VL, dL_L) and cos(dL_VPL, dL_L) > 0.5, each above the V counterpart cos(dV_VL, dV_V), cos(dV_VPL, dV_V))	✓	✓	–	n/e
99	most repairs of every joint arm are carried by exactly one module alone (A1 + C_wta + C_moved > 50%, ALL)	✓	✓	✓	both seeds
100	winner-take-all is the largest leaf outside retention in every joint arm (C_wta > every other non-A leaf, ALL)	✓	✓	✓	both seeds
101	winner-take-all outnumbers independent retention in every joint arm (C_wta > A2, ALL)	✓	✓	✓	both seeds
102	independent retention (A2) is below 10% of every joint arm's repairs (ALL)	✓	✓	✓	both seeds
103	additive (soft) joint-only repairs are at most 1% of each arm's repairs and fewer than the synergistic ones (D ≤ 1%; D < E + F + E/F, ALL)	✓	✓	×	both seeds
104	VL synergistic repairs with informative recombinations: transfer (E) outnumbers private code (F)	✓	✓	✓	both seeds
105	PL, VP, and VPL have no E/F separation (all synergistic joint-only repairs are E/F-ambiguous)	✓	✓	✓	both seeds
106	constituent-new exceeds globally new in the pairs (the third module's single arm repairs some pair repairs), equal in VPL	✓	✓	✓	both seeds
107	every joint arm loses 10–20% of the Base errors its constituents' single arms repair, with no module left carrying them (H_lost, ALL)	✓	✓	✓	both seeds
108	antagonism (a module of S alone repairs, S does not) is largest in VPL (ALL)	✓	✓	✓	both seeds
109	P is a passenger in PL and VPL (moved; every effect's CI within ±2 pp, MACRO)	✓	✓	✓	both seeds
110	P in VP is not a passenger (removing P from VP costs accuracy, MACRO)	✓	✓	✓	both seeds
111	G02: V of VL is representation without function (probe gain, standalone gain CI within ±2 pp)	✓	✓	–	n/e
112	G02: E (V-only parameters) changes late-layer target selection on the frozen LM's query side only (K: parameter ownership != computation ownership)	✓	✓	–	n/e
113	pair -> triple: most VL repairs keep their standalone carriers in VPL (> 60%), while most VP repairs change carriers or are lost (same < 30%) and L joins as a carrier for more than half of them	✓	✓	✓	both seeds

I UPDATE GEOMETRY

Figure 11 shows the update geometry of Section 4.2 for both seeds; every statement below holds in each (values seed 1101/seed 1104; in both seeds every trained tensor at step 0 equals Base).

Over transformer layers only, without embeddings, $\cos(\Delta L_{VL}, \Delta L_L) = 0.54/0.53$. Each module's updates split into two groups by whether one other module is co-trained: V updates by L (cosine 0.80/0.82 within {V, VP}, 0.62/0.59 within {VL, VPL}, 0.24/0.24 across), P updates by L (0.46/0.46, 0.48/0.48, 0.16/0.16), and L updates by V (0.86/0.86, 0.85/0.83, 0.72/0.71); co-training V shrinks the L update less (norm ratio 0.94/0.92) than co-training L shrinks the upstream updates, and ΔL_{PL} is closer to ΔL_L (0.86/0.86) than ΔL_{VL} is. Along depth, the V-side cosine peaks at 0.35/0.34 in the first half of the encoder and falls to 0.11/0.10 in its last third, where the VL update is also relatively small (per-layer norm ratio 0.96/0.96 in the first, 0.57/0.61 in the last layer); the L-side cosine is flat (0.52–0.63/0.49–0.61). The split is present at step 244 (10% of training; cosines 0.82/0.78 for L and 0.27/0.26 for V), and from step 1,220 every update has cosine $\geq 0.977/0.980$ with its final direction, so VL does not first form a V-only-like update and then leave it. Co-training L thus redirects and shrinks the upstream updates, while the L update stays closer to its single-module solution.

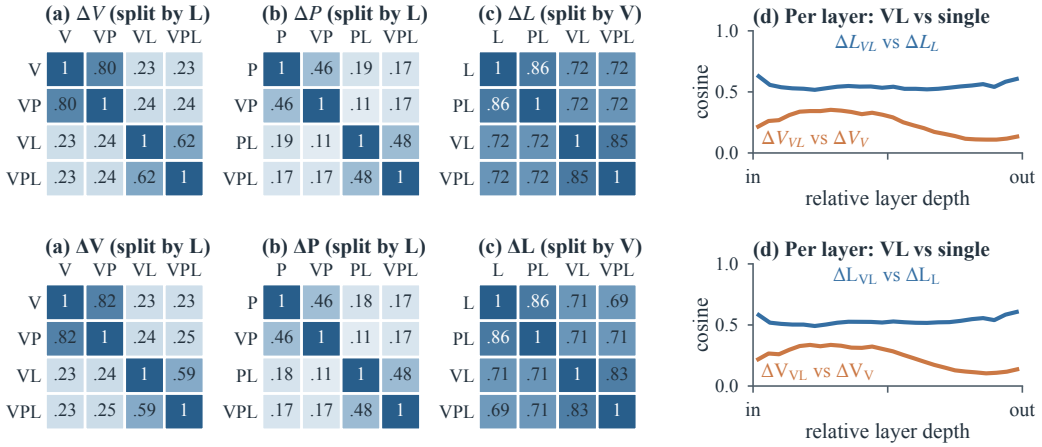

Figure 11: Update geometry (RQ2; top: seed 1101, bottom: seed 1104). (a–c) Cosine between the final updates $\Delta X_S = X_S - X_0$ of one module trained in different sets, ordered so that the blocks differ in whether the named partner is co-trained. (d) Per-layer cosine of the VL update with the single-module update.

J Training dynamics

The training dynamics of Section 4.2 use 81 models per seed scored with the scorer of Section 4.2 (Appendix F): Base, the seven sets at updates 244, 610, 1,220, 1,830, and 2,439, and every module of a joint set at the same updates alone in Base context; the final-update models are those of Section 4.2. They cover all 13 tasks (9,452 official rows, 1,788 scene families, 6,383 Base errors). Figures 12 and 13 show seeds 1101 and 1104; Tables 34 and 35 give, for every joint set and update, the share of the set's repairs that each module alone makes at the same update. Values below are seed 1101/seed 1104. The *stable onset* of a final repair is the first stored update from which the set repairs the item at every later stored update; its *carriers* at onset and at the end are the modules that repair it alone at these two updates. A final repair's route (Figure 12c) is *same* if the two carrier sets are equal and non-empty, *added* or *dropped* if the carrier set at the end strictly contains, or is strictly contained in, the one at onset (including none→some and some→none), *switched* if both are non-empty and neither contains the other, and *never* if no module alone repairs the item at either update. In both seeds, L carries 66.0–81.1/73.3–78.6% of the repairs of VPL at every stored update, as it does for VL, and every set reaches 54–80/51–86% of its final gain by update 244. On G02, V reaches 68.6/70.2% by update 244, whereas V_t^{VL} stays within 1.1/1.1 pp of Base at every stored update. The seeds differ in VPL: in seed 1101, unlike in VL, the share of its repairs that no module makes alone rises from 12.9% at update 244 to 22.1% at update 610 and then stays (23.6% at the end); in seed 1104 it does not (13.7, 17.2, and 15.0%), so we do not interpret this difference between VPL and VL.

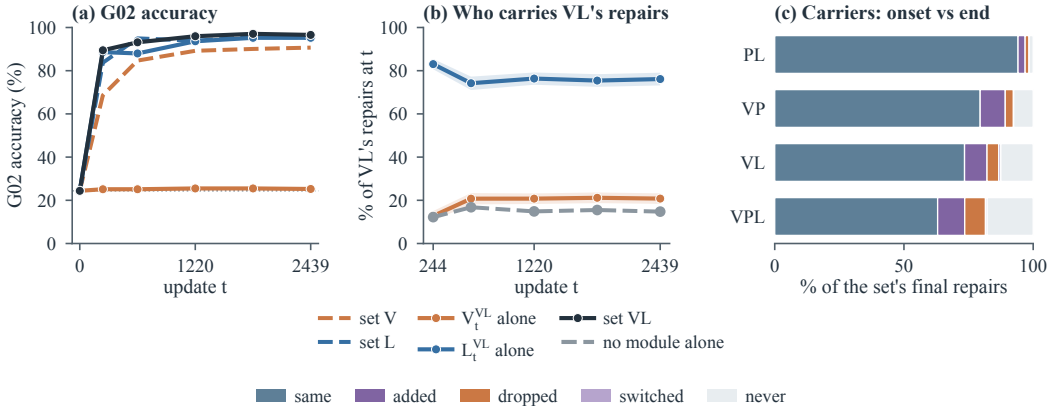

Figure 12: Training dynamics of repair (seed 1101, 13 tasks). (a) G02 accuracy over updates of VL, of its modules alone in Base context (V_t^{VL}, L_t^{VL}), and of the single sets V and L (dashed); dotted: Base. (b) Share of VL's repairs at update t that each module makes alone (pooled; bands: 95% scene-family intervals). (c) Each joint set's final repairs by the modules that make them alone at the set's stable onset and at the end.

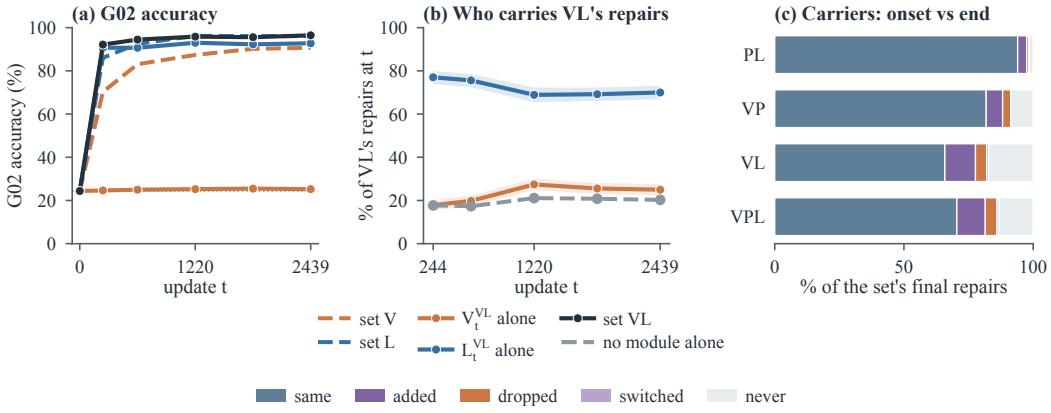

Figure 13: Training dynamics of repair (seed 1104, 13 tasks). (a) G02 accuracy over updates of VL, of its modules alone in Base context (V_t^{VL}, L_t^{VL}), and of the single sets V and L (dashed); dotted: Base. (b) Share of VL's repairs at update t that each module makes alone (pooled; bands: 95% scene-family intervals). (c) Each joint set's final repairs by the modules that make them alone at the set's stable onset and at the end.

Table 34: Who carries a joint set's repairs during training (seed 1101, 13 tasks, pooled). Share (%) of the Base errors that set S repairs at update t which module M of S, taken at the same update and placed alone in Base context (M_t^S), also repairs; last column: repaired by no module alone. Brackets: 95% scene-family intervals.

Set	Update	V alone	P alone	L alone	no module alone
PL	244	–	2.3 [1.8, 2.9]	99.1 [98.7, 99.4]	0.9 [0.6, 1.3]
	610	–	2.8 [2.2, 3.5]	98.3 [97.7, 98.8]	1.7 [1.2, 2.3]
	1220	–	2.3 [1.7, 3.0]	98.0 [97.4, 98.6]	2.0 [1.4, 2.6]
	1830	–	2.5 [1.9, 3.2]	97.3 [96.5, 98.0]	2.7 [2.0, 3.5]
	2439	–	2.8 [2.2, 3.5]	97.5 [96.8, 98.2]	2.5 [1.8, 3.2]
VP	244	91.1 [89.5, 92.6]	3.9 [2.9, 5.0]	–	8.5 [7.0, 10.0]
	610	89.3 [87.8, 90.7]	4.5 [3.5, 5.6]	–	10.5 [9.1, 11.9]
	1220	88.6 [87.1, 90.1]	4.3 [3.3, 5.3]	–	11.0 [9.6, 12.5]
	1830	89.5 [88.1, 90.8]	4.7 [3.6, 5.7]	–	10.1 [8.8, 11.5]
	2439	89.3 [87.8, 90.6]	4.9 [3.9, 5.9]	–	10.2 [8.9, 11.6]
VL	244	12.8 [10.5, 15.0]	–	83.0 [80.6, 85.4]	12.2 [9.9, 14.5]
	610	20.8 [18.1, 23.4]	–	74.2 [71.0, 77.3]	16.8 [14.1, 19.4]
	1220	20.8 [18.3, 23.2]	–	76.4 [73.5, 79.2]	14.9 [12.1, 17.5]
	1830	21.2 [18.7, 23.6]	–	75.4 [72.5, 78.4]	15.5 [12.7, 18.3]
	2439	20.8 [18.3, 23.2]	–	76.1 [73.2, 79.1]	14.8 [12.0, 17.5]
VPL	244	18.2 [15.6, 20.8]	2.0 [1.5, 2.5]	81.1 [78.7, 83.6]	12.9 [10.9, 14.9]
	610	24.4 [21.6, 27.2]	2.1 [1.7, 2.7]	66.7 [63.4, 70.3]	22.1 [19.1, 25.0]
	1220	25.0 [22.1, 27.7]	2.3 [1.8, 2.8]	66.3 [62.9, 69.8]	24.1 [20.9, 27.3]
	1830	24.8 [22.0, 27.4]	2.2 [1.7, 2.7]	66.0 [62.6, 69.5]	24.0 [20.8, 27.3]
	2439	24.5 [21.7, 27.1]	2.2 [1.8, 2.7]	66.7 [63.1, 70.3]	23.6 [20.2, 26.9]

Table 35: Who carries a joint set's repairs during training (seed 1104, 13 tasks, pooled). Share (%) of the Base errors that set S repairs at update t which module M of S, taken at the same update and placed alone in Base context (M_t^S), also repairs; last column: repaired by no module alone. Brackets: 95% scene-family intervals.

Set	Update	V alone	P alone	L alone	no module alone
PL	244	–	2.0 [1.4, 2.6]	98.8 [98.4, 99.2]	1.1 [0.7, 1.6]
	610	–	2.8 [2.2, 3.4]	98.2 [97.6, 98.7]	1.8 [1.3, 2.4]
	1220	–	3.0 [2.4, 3.6]	97.7 [97.1, 98.3]	2.3 [1.7, 2.9]
	1830	–	2.8 [2.1, 3.5]	97.6 [96.9, 98.2]	2.4 [1.8, 3.1]
	2439	–	2.8 [2.2, 3.6]	98.1 [97.4, 98.7]	1.9 [1.3, 2.6]
VP	244	91.9 [90.4, 93.2]	3.4 [2.5, 4.4]	–	8.0 [6.7, 9.4]
	610	88.8 [87.2, 90.2]	4.4 [3.5, 5.4]	–	10.6 [9.2, 12.1]
	1220	89.9 [88.4, 91.4]	4.3 [3.4, 5.3]	–	9.9 [8.5, 11.4]
	1830	89.2 [87.7, 90.6]	4.7 [3.8, 5.8]	–	10.1 [8.8, 11.6]
	2439	88.0 [86.4, 89.6]	4.8 [3.9, 5.8]	–	11.2 [9.8, 12.8]
VL	244	17.8 [15.5, 20.3]	–	77.0 [73.9, 79.9]	17.7 [14.9, 20.8]
	610	19.8 [17.3, 22.4]	–	75.6 [72.4, 78.6]	17.3 [14.8, 19.9]
	1220	27.4 [24.7, 30.1]	–	68.9 [65.5, 72.3]	21.1 [18.1, 24.1]
	1830	25.5 [22.9, 28.1]	–	69.2 [66.0, 72.3]	20.8 [17.9, 23.7]
	2439	25.0 [22.4, 27.6]	–	70.0 [66.8, 73.1]	20.3 [17.4, 23.3]
VPL	244	18.6 [16.4, 20.9]	2.0 [1.5, 2.4]	78.6 [75.5, 81.3]	13.7 [11.4, 16.1]
	610	20.2 [17.7, 22.6]	1.9 [1.5, 2.4]	73.5 [70.0, 76.8]	17.2 [14.6, 20.2]
	1220	21.3 [18.7, 23.8]	1.9 [1.5, 2.3]	73.5 [70.0, 76.5]	16.0 [13.3, 19.0]
	1830	22.1 [19.5, 24.5]	2.0 [1.6, 2.5]	73.3 [70.0, 76.3]	15.4 [12.7, 18.1]
	2439	22.1 [19.5, 24.6]	1.9 [1.4, 2.5]	73.4 [70.2, 76.4]	15.0 [12.4, 17.8]

K ROUTE STEERING

Route steering (Section 4.2) retrains VL (P frozen) in each seed with the recipe of Section 3 and the same examples in the same order as that seed's sets. In *V first* and *L first* one module trains alone for the first 1,220 updates; the other then joins with a fresh optimizer state, ramped in on the shared learning-rate schedule, so it trains only during the decaying second half of that schedule. The *replicate* changes nothing. All three ran on another GPU type than the sets of Sections 4.1–4.2 and were scored with the scorer of Section 4.2 under the same settings (Appendix F; for B06 the two campaigns partly used different code paths for multi-token answers, each verified to give bitwise-identical log likelihoods); VL in Tables 36 and 37 and Figures 14 and 15 (seeds 1101 and 1104) is the model of Section 4.2. Values below are seed 1101/seed 1104. In both seeds, at update 1,220 the locked module is bitwise equal to Base, and the module trained alone behaves like the single set at the same update: V first's V agrees with V on 94.8/94.5% of items (accuracy difference -0.3

[−1.5, +0.8] and −0.5 [−1.8, +0.7] pp) and L first's L with L on 93.8/92.6% (+0.9 [−0.1, +1.8] and −0.5 [−2.1, +1.2] pp), against 60.0/62.4% and 82.2/77.2% for the modules of VL. At the final update the module trained first alone still makes 96.5/97.3% (V first) and 96.1/96.0% (L first) of its solo-phase repairs and behaves like its single-module set: V first's V agrees with V on 93.1/94.0% of items and L first's L with L on 94.4/93.2%, against 61.6/61.9% and 80.3/81.0% for the modules of VL. The analysis covers all 13 tasks (9,452 official rows, 1,788 scene families, 6,383 Base errors) with 2,000 scene-family bootstrap replicates. Against the replicate, V first loses accuracy on B01 (−38.1 [−43.2, −32.7] and −15.0 [−19.2, −10.4] pp), where V is the weaker single-module repair (Section 4.1). L first loses on B03, where L is the weaker one, in seed 1101 (−16.7 pp [−29.2, −5.1]) but not detectably in seed 1104 (−8.3 pp [−18.8, +1.8]); so a cost where the module trained first is the weaker repair is not a rule. On G02, V first and L first do not differ detectably (−0.9 [−3.8, +1.8] and −2.1 [−5.3, +0.6] pp) despite opposite carriers (V alone 91.7/88.5%, L alone 99.1/99.1% of the repairs), both slightly below the replicate (V first −2.9/−3.7, L first −2.1/−1.6 pp).

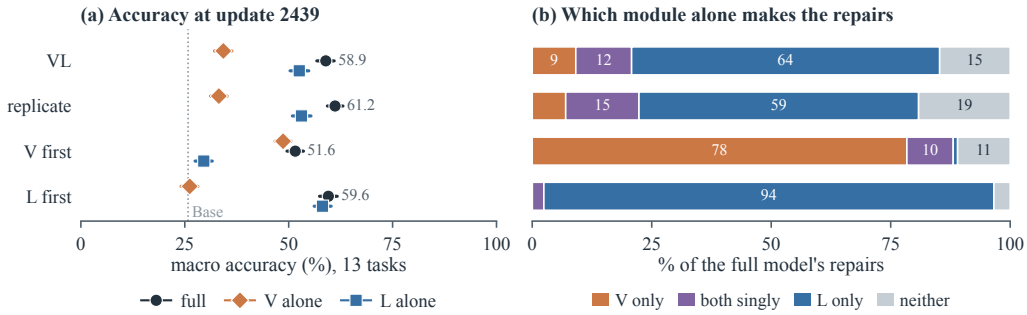

Figure 14: Route steering of VL (seed 1101, 13 tasks, update 2,439). (a) Scene-first macro accuracy of the full model and of its V or L alone in Base context (bars: 95% scene-family intervals; dotted: Base). (b) The full model's repairs of Base errors by which module alone also repairs them (pooled).

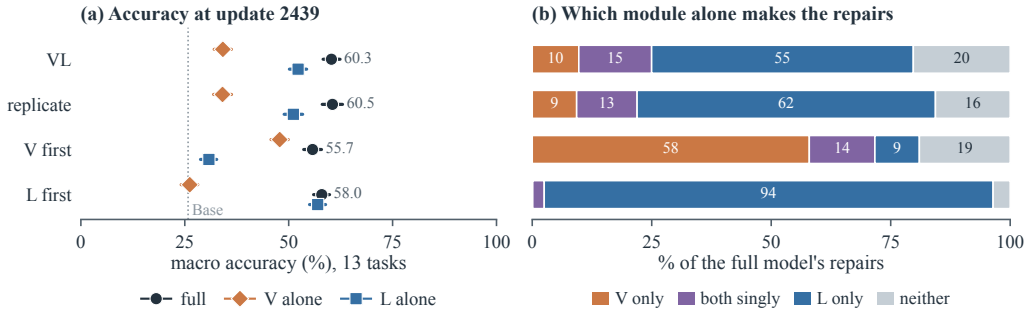

Figure 15: Route steering of VL (seed 1104, 13 tasks, update 2,439). (a) Scene-first macro accuracy of the full model and of its V or L alone in Base context (bars: 95% scene-family intervals; dotted: Base). (b) The full model's repairs of Base errors by which module alone also repairs them (pooled).

Table 36: Route steering (seed 1101, 13 tasks). VL (Section 4.2) and three retrains with its data and order: *replicate* (its recipe unchanged), *V first* (only V trained for updates 1–1,220, then V and L), *L first* (the reverse). Accuracy: scene-first, 13-task macro (%) of the full model and of its V or L alone in Base context; repairs: share (%) of the full model's repairs of Base errors that the module alone also repairs, and that neither module alone repairs (pooled). Brackets: 95% scene-family intervals. At update 1,220 the locked module is still Base, so V first is its V alone and L first its L alone.

		Accuracy %			Repairs of the full model %		
Run	Update	full	V alone	L alone	V alone	L alone	neither
VL	2,439	58.9 [56.8, 61.0]	34.3	52.5	20.8 [18.3, 23.2]	76.1 [73.2, 79.1]	14.8
replicate	2,439	61.2 [59.3, 63.1]	33.2	53.1	22.3 [19.9, 24.6]	73.8 [71.0, 76.7]	19.2
V first	2,439	51.6 [49.6, 53.6]	48.7	29.6	88.1 [86.3, 89.8]	10.6 [9.0, 12.4]	11.0
L first	2,439	59.6 [57.4, 61.7]	26.2	58.1	2.4 [1.9, 3.0]	96.6 [95.6, 97.4]	3.4
VL	1,220	57.7 [55.6, 60.0]	34.4	51.3	20.8 [18.3, 23.2]	76.4 [73.5, 79.2]	14.9
replicate	1,220	58.9 [56.8, 61.0]	33.3	52.7	22.2 [19.8, 24.5]	75.4 [72.6, 78.2]	17.8
V first	1,220	47.9 [45.9, 50.0]	47.9	25.8	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]	0.0
L first	1,220	56.2 [54.1, 58.3]	25.8	56.2	0.0 [0.0, 0.0]	100.0 [100.0, 100.0]	0.0

Table 37: Route steering (seed 1104, 13 tasks). VL (Section 4.2) and three retrains with its data and order: *replicate* (its recipe unchanged), *V first* (only V trained for updates 1–1,220, then V and L), *L first* (the reverse). Accuracy: scene-first, 13-task macro (%) of the full model and of its V or L alone in Base context; repairs: share (%) of the full model's repairs of Base errors that the module alone also repairs, and that neither module alone repairs (pooled). Brackets: 95% scene-family intervals. At update 1,220 the locked module is still Base, so V first is its V alone and L first its L alone.

		Accuracy %			Repairs of the full model %		
Run	Update	full	V alone	L alone	V alone	L alone	neither
VL	2,439	60.3 [58.3, 62.4]	34.2	52.3	25.0 [22.4, 27.6]	70.0 [66.8, 73.1]	20.3
replicate	2,439	60.5 [58.1, 63.0]	34.1	51.1	21.9 [19.4, 24.4]	75.1 [72.0, 78.0]	15.6
V first	2,439	55.7 [53.6, 57.9]	47.8	30.8	71.7 [68.8, 74.5]	23.0 [20.1, 25.8]	19.0
L first	2,439	58.0 [56.0, 59.8]	26.2	57.0	2.5 [1.9, 3.1]	96.3 [95.4, 97.1]	3.6
VL	1,220	58.3 [56.3, 60.4]	34.2	51.1	27.4 [24.7, 30.1]	68.9 [65.5, 72.3]	21.1
replicate	1,220	59.0 [56.6, 61.5]	34.4	48.9	23.6 [20.9, 26.2]	74.1 [71.0, 77.1]	16.0
V first	1,220	46.5 [44.5, 48.6]	46.5	25.8	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]	0.0
L first	1,220	54.6 [52.5, 56.6]	25.8	54.6	0.0 [0.0, 0.0]	100.0 [100.0, 100.0]	0.0

L TASK COMPUTATION DESCRIPTORS AND REPAIR PHENOTYPES

Each task's computation card records the required semantic state, the selection or binding of the target, the core operation, the aggregation scope, and the precision regime; the six descriptors are scored from task definitions, generators, and official rows. Table 38 gives the locked descriptors of Section 4.3 next to the per-task repair phenotypes of Section 4.1 for seed 1101, and Table 39 for seed 1104.

Section 4.3 relates repair availability to task demands along the chain *task computation* → *Base bottleneck* → *available repairs*. Each task receives a computation card from which six descriptors are scored 0 (not needed), 1 (auxiliary), or 2 (decides the label), locked before any accuracy or repair number was consulted (Appendix L). We test three Base-bottleneck hypotheses with probes and interventions whose targets the cards define: *representation-limited* failures (information missing upstream; V repair matters, multiplicity is low), *interface-limited* failures (V carries the signal but P does not deliver it usably; P repair is strong), and *selection- or readout-limited* failures (information present, downstream decision wrong; repairs at several modules, high multiplicity). A fourth, *distributed or compositional* limitation has no single stage whose change suffices; its repairs are pair- or triple-emergent (B04, the task with the largest pair-first share). The six descriptors describe the computation a task needs and were locked before any result; they are not a map from task demands to modules: G02 is selection-heavy, yet P, V, and L each repair it, and V's repair acts through the frozen language model (Section 4.3).

With 13 tasks, the descriptors support descriptive statements only. (i) In both seeds, none of the 5 tasks on which almost every Base error has a single-set repair needs label-deciding metric precision

or connectivity, whereas each task with many unrepaired errors has a label-deciding continuous estimate (V01–V03), global count (B04, B06), or, in seed 1101, connectivity (B07); the converse fails for B02 and B05 (19.0/17.9% and 0.0/0.0% unrepaired) and, in seed 1104, for B07. (ii) Query-conditioned selection alone does not predict multiplicity: where the selected target's category is the answer (B03 and G02), 59–89/55–89% of Base errors have two or more single-set repairs, but where selection is combined with a continuous estimate (V01, V02, and V03), only 11–16/14–22%. (iii) B07, the only connectivity-heavy task, is the only task on which the projector dominates repair, as an interface-limited bottleneck would predict, though over all tasks an answer-prior shift reproduces 74/75% of P's repairs (Appendix D.1).

Table 38: Task computation descriptors (left; Pr metric precision, Se query-conditioned selection, Re relation, Ag global aggregation, Co connectivity, Id discrete identity; 0 not needed, 1 auxiliary, 2 decides the label; locked before any accuracy or repair number was consulted) and repair phenotypes (right; official predictions of training seed 1101, % of Base errors; Base err. = number of Base errors). First repair: smallest level with a repairing set. Single-set repair: repaired by P, V, L. $n_1 \geq 2$: two or three single-set repairs. Lost: repairs lost when one module is added, pooled over the nine edges.

	Computation descriptors						Base	First repair				Single-set repair				
Task	Pr	Se	Re	Ag	Co	Id	err.	single	pair	triple	none	P	V	L	$n_1 \geq 2$	Lost
B01	1	0	2	1	0	0	2,444	99	1	0	0	44	60	91	62	20
B02	2	0	2	0	0	0	290	73	2	6	19	4	50	60	37	14
B03	0	2	0	0	0	2	535	90	7	0	2	39	79	61	59	9
B04	0	0	0	2	0	0	683	34	21	4	41	7	18	28	13	15
B05	1	0	0	2	0	0	164	99	1	0	0	26	62	96	71	18
B06	0	0	0	2	0	0	502	38	5	2	54	4	11	33	8	52
B07	0	1	0	1	2	0	704	50	9	2	40	43	3	7	4	67
G01	1	1	1	0	0	0	311	96	3	0	1	37	72	90	75	4
G02	1	2	1	1	0	2	244	98	0	0	2	57	88	94	89	2
G03	0	1	2	0	0	0	273	97	1	0	1	45	74	97	75	2
V01	2	2	1	0	0	0	74	43	7	0	50	22	18	16	11	47
V02	2	2	1	0	0	0	82	54	10	0	37	20	28	26	16	44
V03	2	2	1	0	0	0	84	44	8	0	48	17	21	18	11	50
All							6,390	78	5	1	16	34	47	64	45	18

Table 39: Task computation descriptors (left; Pr metric precision, Se query-conditioned selection, Re relation, Ag global aggregation, Co connectivity, Id discrete identity; 0 not needed, 1 auxiliary, 2 decides the label; locked before any accuracy or repair number was consulted) and repair phenotypes (right; official predictions of training seed 1104, % of Base errors; Base err. = number of Base errors). First repair: smallest level with a repairing set. Single-set repair: repaired by P, V, L. $n_1 \geq 2$: two or three single-set repairs. Lost: repairs lost when one module is added, pooled over the nine edges.

	Computation descriptors						Base	First repair				Single-set repair				
Task	Pr	Se	Re	Ag	Co	Id	err.	single	pair	triple	none	P	V	L	$n_1 \geq 2$	Lost
B01	1	0	2	1	0	0	2,444	99	0	0	0	44	60	83	58	14
B02	2	0	2	0	0	0	290	67	15	0	18	5	61	43	38	11
B03	0	2	0	0	0	2	535	87	6	2	4	40	78	56	55	9
B04	0	0	0	2	0	0	683	32	20	6	41	7	9	30	10	8
B05	1	0	0	2	0	0	164	99	1	0	0	26	44	98	54	16
B06	0	0	0	2	0	0	502	25	13	7	54	4	13	17	7	38
B07	0	1	0	1	2	0	704	60	7	1	32	44	15	11	10	75
G01	1	1	1	0	0	0	311	96	2	1	1	38	69	89	72	5
G02	1	2	1	1	0	2	244	98	0	0	2	57	88	95	89	2
G03	0	1	2	0	0	0	273	96	3	0	1	46	74	94	75	2
V01	2	2	1	0	0	0	74	36	5	4	54	19	16	18	14	46
V02	2	2	1	0	0	0	82	49	4	2	45	22	29	24	22	35
V03	2	2	1	0	0	0	84	30	18	5	48	19	21	6	15	62
All							6,390	77	6	2	16	34	48	59	43	15

M TASK-SPECIFIC PROBES

The probes of Section 4.3 ask where a linear readout finds the variable that each task's computation card names: the target's identity and its selection on G02, the signed gap and the relation on B02, and the path count and component membership on B07. They compare Base, V, L, VL, and the transplants V^{VL} and L^{VL} of each seed (final update; E, F, G label the runs as in Appendix N; values are seed 1101/seed 1104, and Base, shared by both seeds, has one value), captured with a prefix-only forward pass in the numerical regime of the official evaluation. *Image sites* are projector outputs: the mean over all tokens, a 3×3 grid of the global tile, and object- or region-specific means over the tokens whose source cell touches the region (G02: the target objects, and each object as a unit; B02: the two disks, and a disk of half the radius at the midpoint of the centres). P is Base in every model, so an image site depends only on V: Base, L, and L^{VL} share one set of features, VL and V^{VL} another. *Language-model sites* are the residual stream after each decoder layer at the final prompt token and, for G02, at the last token of each option line. Classifiers are L2-regularized logistic regressions and continuous targets ridge regressions on standardized features, with the penalty chosen by 5-fold cross-validation grouped by scene family on the training rows (2,000 per task; 10,994 G02 objects), refit, and tested once on the official rows (G02 319, B02 1,344, B07 720; 1,631 G02 objects and 1,276 options), whose scene families are disjoint from training. Scores are balanced accuracy (chance in parentheses) or R^2, with 2,000 scene-family bootstrap replicates shared by all models, so model contrasts are paired. The summaries were fixed in advance: every image site is reported (Tables 40 and 41; Figures 16 and 17), and language-model layers are summarized over layers 12–23 (the band of Appendix N) by the mean test score, the number of layers whose interval lies above chance, and the number whose paired contrast with Base lies above or below zero (Tables 42 and 43). We do not select the best cross-validated layer, because every trained model saw the training rows and its cross-validation score saturates. The models' own first-token readout on the captures agrees with their official predictions on at least 96.0/98.1% of rows.

Summary of the probe results. Task-specific linear probes (fit on training rows, tested on official rows; B07's do not transfer) show that neither G02 nor B02 is representation-limited: in Base (shared by both seeds) the target's type is decodable from its own image tokens (balanced accuracy 0.66 [0.59, 0.76], chance 0.07) and the signed gap from the circles' contact region (R^2 0.92 [0.90, 0.93]). Base's final prompt token carries the G02 answer (0.58 over layers 12–23, chance 0.25) that its own answer misses (23.5% correct), a selection- or readout-limited pattern, but barely the B02 relation (0.36; contact region 0.69; chance 0.33). In both seeds, L raises final-token decodability (G02 +0.33/+0.33, B02 +0.23/+0.17; above Base at 12/12 and 12/11 of the 12 layers), V that of the G02 target's tokens (+0.12/+0.14; every interval above zero), and so does V^{VL} (+0.14/+0.12), which alone barely repairs G02 (Section 4.2): decodability does not make a module a carrier. On G02, where Base queries at the answer position do not prefer target over distractor image tokens, we split each repair's change in their score contrast (layers 12–23) into a key side (Base query, repaired keys) and a query side (repaired query, Base keys; Appendix N). In both seeds, both single-module repairs act through the query side (V +0.14/+0.13, L +0.25/+0.28; every interval above zero) and neither through the keys, although V changes the keys' target–distractor contrast more than L: V reaches target selection indirectly, through what its image tokens feed into lower layers, and only L also raises per-token attention to the target, so the expected split (V on keys, L on query) holds only for L.

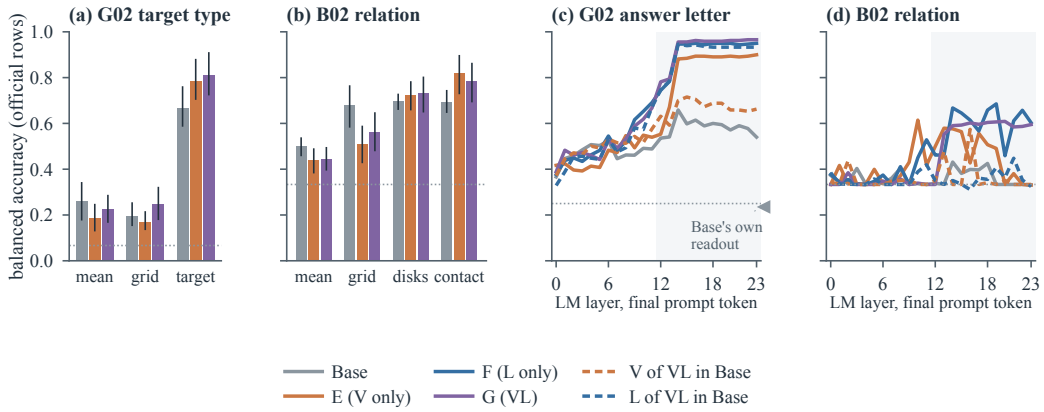

Figure 16: Task-specific probes on the official rows (seed 1101). (a, b) Balanced accuracy at the image sites of G02 (target type; *target*: the target objects' tokens) and B02 (relation; *disks*, *contact*: region-specific sites) for the three vision encoders, with 95% scene-family intervals. (c, d) Balanced accuracy at the final prompt token by layer for the six models (shaded: layers 12–23; dotted: chance); the marker in (c) is Base's own first-token readout accuracy.

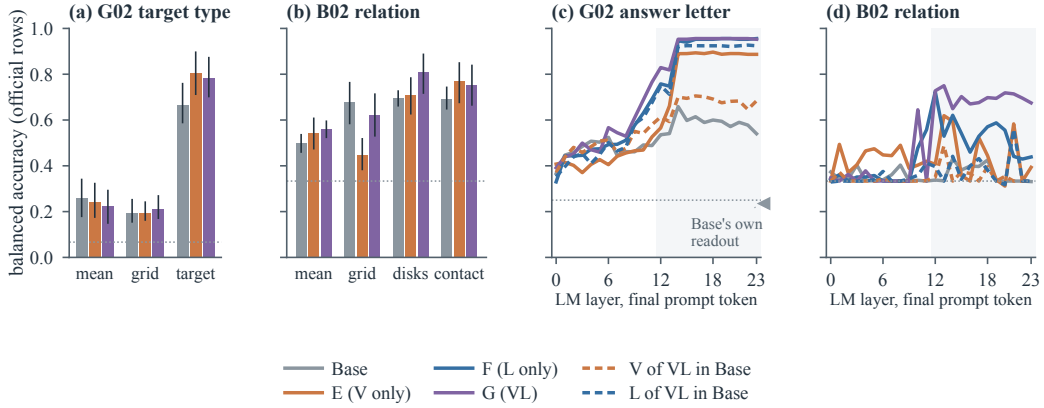

Figure 17: As Figure 16, seed 1104.

G02. Each object's type is decodable from its own tokens in Base (0.65 [0.59, 0.71]), and V and V^{VL} change this little (+0.05 [−0.01, +0.10] and +0.09 [+0.02, +0.14] and +0.05 [−0.01, +0.11] and +0.03 [−0.02, +0.08]): undetectably in seed 1101, whereas in seed 1104 V raises it slightly. Whether an object is the target is near chance at its image tokens (0.53–0.55/0.53–0.54), which see no query. Target selection appears in the language model: at the option tokens Base already separates the target, a present distractor, and an absent one (0.66, chance 0.33, above chance at every layer), and every trained model raises this (L +0.20/+0.21, V +0.14/+0.13, V^{VL} +0.09/+0.08). At the final token V raises the decodability of the answer (+0.26/+0.26; above Base at 11/11 of 12 layers) but not of the target's type (+0.01/+0.04; 3/2 layers), whereas L raises both (+0.33/+0.33 and +0.14/+0.15). V^{VL} raises the answer's decodability over Base (+0.08/+0.09) but stays below V (−0.18/−0.17; below at 11/11 of 12 layers), and it answers like Base (24.5/24.5% vs 23.5% correct at the captured position).

B02. The two disks' tokens also carry the gap and the relation in Base (R^2 0.68, balanced accuracy 0.70), whereas the mean over all tokens does not carry the gap (R^2 −1.60). The official gaps extend beyond the training range, so R^2 also measures extrapolation. In both seeds V fits the continuous gap worse at the contact region (R^2 −0.12 [−0.15, −0.10] and −0.19 [−0.23, −0.15]) and at the disks (−0.30 [−0.40, −0.21] and −0.27 [−0.38, −0.18]); it makes the relation more decodable at the

contact region in seed 1101 (+0.13 [+0.04, +0.21]) but not detectably in seed 1104 (+0.08 [−0.01, +0.16]). V^{VL} does the same, less strongly, in seed 1101 (+0.09 [+0.01, +0.17] and −0.04 [−0.07, −0.02]); in seed 1104 neither change is detectable. At Base's final token the gap is not linearly recoverable ($R^2 < 0$ at every layer 12–23) and the relation is above chance at only 6 of 12 layers. There, L and VL raise the relation (+0.23/+0.17 and +0.21/+0.33) more than V (+0.10/+0.07), which raises it at 8 of 12 layers and lowers it at none in seed 1101, but at 7 and 3 in seed 1104; L^{VL} and V^{VL} leave it about unchanged on average (−0.01/+0.02 and 0.00/0.00). L^{VL} answers like Base in seed 1101 (78.2% vs 78.4%; VL 90.6%) and slightly better in seed 1104 (80.8%; VL 94.6%), matching Appendix Q, where the B02 repair of VL needs both of its updates in seed 1101 and ΔL_{VL} alone makes a small part of it in seed 1104.

Table 40: Probe scores at the image (projector) sites on the official rows (seed 1101): balanced accuracy (chance in parentheses) or R^2, with 95% scene-family intervals. P is Base in every model, so each column is one vision encoder: Base's (shared by Base, L, and L^{VL}), V's, or VL's (shared by VL and V^{VL}). †: targets that depend on the query, controls at image-only sites. B07 *pair joined*: range of the point estimates over the station pairs.

task	target (chance)	site	Base V (Base, F, L^{VL})	V of E	V of G (G, V^{VL})
G02	target type (0.07)	mean	0.26 [0.18, 0.34]	0.18 [0.13, 0.25]	0.23 [0.17, 0.28]
		grid 3×3	0.19 [0.16, 0.25]	0.17 [0.14, 0.21]	0.25 [0.18, 0.32]
		target obj.	0.66 [0.59, 0.76]	0.78 [0.71, 0.88]	0.81 [0.73, 0.91]
G02	object type (0.07)	each object	0.65 [0.59, 0.71]	0.70 [0.64, 0.76]	0.70 [0.64, 0.76]
G02	object is target† (0.50)	each object	0.53 [0.51, 0.56]	0.54 [0.52, 0.57]	0.55 [0.53, 0.58]
G02	answer letter† (0.25)	mean	0.24 [0.20, 0.28]	0.26 [0.22, 0.30]	0.26 [0.21, 0.30]
		grid 3×3	0.22 [0.17, 0.26]	0.24 [0.19, 0.29]	0.25 [0.20, 0.30]
		target obj.	0.25 [0.20, 0.29]	0.23 [0.19, 0.27]	0.26 [0.21, 0.31]
G02	template† (0.25)	mean	0.33 [0.26, 0.40]	0.30 [0.27, 0.36]	0.37 [0.30, 0.45]
		grid 3×3	0.30 [0.27, 0.33]	0.32 [0.29, 0.35]	0.31 [0.29, 0.34]
		target obj.	0.36 [0.33, 0.40]	0.38 [0.31, 0.46]	0.37 [0.34, 0.41]
B02	relation (0.33)	mean	0.50 [0.46, 0.54]	0.44 [0.38, 0.49]	0.45 [0.40, 0.49]
		grid 3×3	0.68 [0.58, 0.76]	0.51 [0.43, 0.59]	0.56 [0.48, 0.65]
		two disks	0.70 [0.66, 0.73]	0.72 [0.66, 0.78]	0.73 [0.65, 0.80]
		contact region	0.69 [0.65, 0.74]	0.82 [0.73, 0.90]	0.78 [0.70, 0.86]
B02	gap (R^2)	mean	−1.60 [−2.03, −1.29]	−0.94 [−1.24, −0.73]	−1.04 [−1.32, −0.84]
		grid 3×3	−0.44 [−0.69, −0.25]	0.66 [0.60, 0.71]	0.91 [0.89, 0.93]
		two disks	0.68 [0.64, 0.71]	0.38 [0.27, 0.46]	0.68 [0.64, 0.72]
		contact region	0.92 [0.90, 0.93]	0.79 [0.77, 0.81]	0.88 [0.85, 0.90]
B02	radius (R^2)	mean	0.04 [−0.07, 0.11]	−1.34 [−1.66, −1.13]	−1.11 [−1.40, −0.93]
		grid 3×3	0.71 [0.67, 0.74]	0.94 [0.93, 0.95]	0.95 [0.94, 0.96]
		two disks	0.68 [0.63, 0.71]	0.37 [0.26, 0.43]	0.57 [0.51, 0.62]
		contact region	0.49 [0.37, 0.58]	−0.62 [−0.90, −0.42]	0.15 [−0.03, 0.27]
B07	path count (0.33)	mean	0.01 [0.00, 0.02]	0.00 [0.00, 0.00]	0.01 [0.00, 0.02]
		grid 3×3	0.05 [0.03, 0.07]	0.00 [0.00, 0.00]	0.03 [0.01, 0.05]
B07	routes (0.33)	mean	0.50 [0.46, 0.54]	0.44 [0.40, 0.48]	0.49 [0.45, 0.54]
		grid 3×3	0.36 [0.31, 0.41]	0.33 [0.33, 0.33]	0.33 [0.31, 0.33]
B07	pair joined (0.50)	mean	0.47–0.63	0.47–0.73	0.43–0.71
		grid 3×3	0.49–0.68	0.52–0.75	0.42–0.68

Table 41: As Table 40, seed 1104.

task	target (chance)	site	Base V (Base, F, L^{VL})	V of E	V of G (G, V^{VL})
G02	target type (0.07)	mean	0.26 [0.18, 0.34]	0.24 [0.18, 0.32]	0.22 [0.15, 0.29]
		grid 3×3	0.19 [0.16, 0.25]	0.20 [0.16, 0.24]	0.21 [0.17, 0.27]
		target obj.	0.66 [0.59, 0.76]	0.80 [0.71, 0.90]	0.78 [0.70, 0.87]
G02	object type (0.07)	each object	0.65 [0.59, 0.71]	0.74 [0.68, 0.79]	0.68 [0.62, 0.75]
G02	object is target† (0.50)	each object	0.53 [0.51, 0.56]	0.53 [0.51, 0.56]	0.54 [0.52, 0.57]
G02	answer letter† (0.25)	mean	0.24 [0.20, 0.28]	0.25 [0.21, 0.30]	0.26 [0.22, 0.31]
		grid 3×3	0.22 [0.17, 0.26]	0.22 [0.18, 0.27]	0.25 [0.20, 0.29]
		target obj.	0.25 [0.20, 0.29]	0.21 [0.16, 0.26]	0.23 [0.18, 0.27]
G02	template† (0.25)	mean	0.33 [0.26, 0.40]	0.32 [0.28, 0.38]	0.32 [0.28, 0.37]
		grid 3×3	0.30 [0.27, 0.33]	0.31 [0.28, 0.33]	0.31 [0.28, 0.34]
		target obj.	0.36 [0.33, 0.40]	0.35 [0.31, 0.38]	0.37 [0.34, 0.41]
B02	relation (0.33)	mean	0.50 [0.46, 0.54]	0.54 [0.47, 0.61]	0.56 [0.52, 0.59]
		grid 3×3	0.68 [0.58, 0.76]	0.45 [0.38, 0.52]	0.62 [0.53, 0.71]
		two disks	0.70 [0.66, 0.73]	0.71 [0.63, 0.78]	0.81 [0.72, 0.89]
		contact region	0.69 [0.65, 0.74]	0.77 [0.68, 0.85]	0.75 [0.67, 0.84]
B02	gap (R^2)	mean	−1.60 [−2.03, −1.29]	−1.01 [−1.31, −0.81]	−0.47 [−0.71, −0.31]
		grid 3×3	−0.44 [−0.69, −0.25]	0.19 [0.04, 0.31]	0.41 [0.31, 0.48]
		two disks	0.68 [0.64, 0.71]	0.41 [0.29, 0.50]	0.77 [0.72, 0.81]
		contact region	0.92 [0.90, 0.93]	0.73 [0.69, 0.76]	0.90 [0.88, 0.91]
B02	radius (R^2)	mean	0.04 [−0.07, 0.11]	−1.37 [−1.69, −1.14]	−1.55 [−1.89, −1.33]
		grid 3×3	0.71 [0.67, 0.74]	0.98 [0.98, 0.98]	0.94 [0.93, 0.95]
		two disks	0.68 [0.63, 0.71]	0.45 [0.36, 0.51]	0.71 [0.67, 0.74]
		contact region	0.49 [0.37, 0.58]	−1.06 [−1.40, −0.82]	0.31 [0.16, 0.41]
B07	path count (0.33)	mean	0.01 [0.00, 0.02]	0.00 [0.00, 0.00]	0.00 [0.00, 0.01]
		grid 3×3	0.05 [0.03, 0.07]	0.00 [0.00, 0.00]	0.04 [0.02, 0.07]
B07	routes (0.33)	mean	0.50 [0.46, 0.54]	0.57 [0.53, 0.61]	0.49 [0.45, 0.53]
		grid 3×3	0.36 [0.31, 0.41]	0.33 [0.33, 0.33]	0.32 [0.27, 0.36]
B07	pair joined (0.50)	mean	0.47–0.63	0.44–0.63	0.45–0.58
		grid 3×3	0.49–0.68	0.49–0.68	0.48–0.66

B07. B07 has no object site, because its official rows carry no station or route geometry. Every B07 probe separates the training families almost perfectly (cross-validated balanced accuracy at least 0.93/0.93), yet the path count does not transfer to the official rows: its interval reaches or lies below chance at every image site and every layer 12–23 in every model, except in 0/1 of the 84 cells (largest point estimate 0.38/0.37, chance 0.33). We therefore do not interpret B07; Base answers 2.2% of its official rows correctly.

Controls and caveats. Image-only sites see no query. The answer letter, which depends on option order, stays at chance there in every model (0.22–0.26/0.21–0.26, every interval containing 0.25), whereas the template is above chance at every image site (0.30–0.38/0.30–0.37), so the scenes differ by template and it is only a partial control; at Base's final token the template is read from the prompt (1.00). Every trained model saw the training rows for 3 epochs, so its probe is fit on familiar inputs and tested on new ones, whereas Base's is not. A probe score measures decodability, not use (Hewitt & Liang, 2019; Belinkov, 2022): a higher score does not show that the model reads the variable, and a lower one does not show that the information is lost. Final-token decodability can draw on the question and option text as well as on the image. Each seed contributes one training run per set. The seed-1104 probes were fit from a working copy with uncommitted changes to the analysis code (its tree hash is recorded), so their provenance check fails; a refit from a committed tree has not been run.

Table 42: Probe scores in the language model on the official rows (seed 1101): mean over layers 12–23 of the per-layer test score (balanced accuracy, chance in parentheses, or R^2) at the final prompt token or (‡) at the last token of each option line. Small numbers: for Base, the layers whose 95% interval lies above chance (above zero for R^2); for the other models, the layers whose paired contrast with Base lies above/below zero (of 12; B07 *pair joined*: mean over the station pairs, counts of 72).

task	target (chance)	Base	E (V)	F (L)	G (VL)	V^{VL}	L^{VL}
G02	answer letter (0.25)	0.58 12	0.84 11/0	0.92 12/0	0.93 12/0	0.67 11/0	0.91 12/0
G02	target type (0.07)	0.47 12	0.49 3/1	0.62 10/0	0.63 9/0	0.55 10/0	0.62 9/0
G02	template (0.25)	1.00 12	0.97 0/1	0.96 0/2	0.97 0/2	1.00 0/0	0.98 0/3
G02	option status‡ (0.33)	0.66 12	0.80 12/0	0.87 12/0	0.90 12/0	0.75 12/0	0.86 12/0
G02	option is target‡ (0.50)	0.80 12	0.88 12/0	0.93 12/0	0.95 12/0	0.85 10/0	0.94 12/0
B02	relation (0.33)	0.36 6	0.46 8/0	0.59 12/0	0.57 11/1	0.37 3/4	0.36 5/5
B02	gap (R^2)	−38.2 0	−5.0 12/0	−1.4 12/0	−8.0 10/1	−21.9 10/2	−7.5 10/2
B02	radius (R^2)	−124.3 0	−3.1 11/0	−5.0 11/1	−37.3 9/3	−69.8 9/3	−9.9 11/1
B07	path count (0.33)	0.12 0	0.00 0/9	0.26 11/0	0.25 10/0	0.06 1/7	0.19 6/0
B07	routes (0.33)	0.29 2	0.41 10/2	0.64 11/1	0.38 11/0	0.36 9/3	0.61 11/1
B07	pair joined (0.50)	0.50 7	0.51 16/9	0.55 36/5	0.55 37/9	0.50 7/6	0.55 36/8

Table 43: As Table 42, seed 1104.

task	target (chance)	Base	E (V)	F (L)	G (VL)	V^{VL}	L^{VL}
G02	answer letter (0.25)	0.58 12	0.84 11/0	0.92 12/0	0.93 12/0	0.67 11/0	0.89 12/0
G02	target type (0.07)	0.47 12	0.52 2/1	0.62 10/0	0.66 11/0	0.55 9/0	0.65 11/0
G02	template (0.25)	1.00 12	0.96 0/3	0.95 0/1	0.97 0/2	1.00 0/0	0.97 0/0
G02	option status‡ (0.33)	0.66 12	0.79 12/0	0.87 12/0	0.89 12/0	0.74 12/0	0.87 12/0
G02	option is target‡ (0.50)	0.80 12	0.86 10/0	0.94 12/0	0.94 12/0	0.85 12/0	0.92 12/0
B02	relation (0.33)	0.36 6	0.44 7/3	0.54 11/0	0.70 12/0	0.36 2/3	0.38 2/2
B02	gap (R^2)	−38.2 0	−10.2 9/3	−9.5 11/1	−0.8 12/0	−88.0 3/9	−2.4 12/0
B02	radius (R^2)	−124.3 0	−12.5 9/3	−6.7 12/0	−13.7 10/2	−31.4 8/4	−7.6 11/0
B07	path count (0.33)	0.12 0	0.01 0/8	0.25 9/0	0.26 11/0	0.11 4/2	0.19 5/0
B07	routes (0.33)	0.29 2	0.42 10/1	0.71 11/1	0.41 11/0	0.33 6/6	0.58 10/2
B07	pair joined (0.50)	0.50 7	0.51 17/6	0.54 31/5	0.53 24/9	0.50 10/12	0.54 29/3

N Q/K/V ANALYSIS ON G02

The Q/K/V analysis of Section 4.3 asks at which stage of late-layer attention the single-module repairs of G02 act. It compares Base, V, L, VL, and the transplants V^{VL} and L^{VL} of each seed (final update) on the 319 official G02 rows, in the numerical regime of the official evaluation; values are seed 1101/seed 1104, and every statement holds in both seeds unless noted. Tables 44 and 45 and Figures 18 and 19 label the sets by their runs (E = V, F = L, G = VL). A forward pre-hook on every attention module of the language model recomputes, from the module's own input with its own projections and rotary embedding, the query q at the final prompt token, whose logits score the answer, and the keys k and values at all positions. *Target* tokens are the image tokens whose source cell contains the drawn outline of an object of the target type, *distractor* tokens those that contain the outline of another object, and tokens on both are dropped (medians 108 and 249 of 3,699 image tokens; 2 rows have no distractor token). With $s = q^\top k/\sqrt{d}$, the *score contrast* is the mean of s over target tokens minus that over distractor tokens, and the per-token attention *log ratio* is the log of the mean attention weight on target tokens minus that on distractor tokens. Both are averaged over heads and over layers 12–23 (the upper half, fixed before looking), with 2,000 scene-family bootstrap replicates. For a model M, the *key side* is the change over Base when Base's query reads M's keys, the *query side* the change when M's query reads Base's keys, and the *interaction* the rest of M's own change. Each query and key comes from its model's own forward pass on the row, so M's query also reflects M's image tokens as read in lower layers. Attention rebuilt from the tapped q, k, and values matches each attention module's output within 2.8/2.5% (median 0.6/0.6%), and the readout at the captured position agrees with the model's prediction on at least 98.1/100.0% of rows.

Base queries do not prefer target over distractor tokens (score contrast 0.00 [−0.04, +0.05]). V, L, VL, and L^{VL} all raise the contrast, and all through the query side (Table 44): no key side is detectable, and each is below the query side. At every layer the key side stays within ±0.044/0.047, whereas the query side is small in the lower half and largest in layers 12–23 (Figure 18, left; peak at layer 21 for L). V^{VL} changes nothing. V changes the target-minus-distractor contrast of the keys more than L (relative L2 change 75/74% vs 42/40%), yet read by Base's query these keys give no gain (0.00 [−0.06, +0.07] and +0.01 [−0.05, +0.07]), and they add at most a little to L's query (interaction of $q_{\mathrm{F}} \cdot k_{\mathrm{E}}$: +0.02 [−0.03, +0.07] and +0.05 [+0.01, +0.09]; nothing detectable in seed 1101, a small positive interaction in seed 1104). V's gain comes from its final-token query, which moves by 48/47% (L: 52/50%) although its language model is Base, and from the interaction of that query with its own keys (+0.08 [+0.04, +0.13] and +0.07 [+0.03, +0.12]).

In attention weights, L raises the per-token target-versus-distractor log ratio (+0.19 [+0.12, +0.25] and +0.24 [+0.18, +0.30]) and V does not (+0.02 [−0.06, +0.10] and −0.01 [−0.10, +0.08]): it puts more of the final token's attention on the image (8.5/10.3% vs 5.5% for Base) but a smaller share of that on target tokens (1.8/1.5% vs 3.8%). No model reads the answer from the object tokens' values directly: their target-minus-distractor direct contribution to the answer-letter logit margin (attention times value times output projection, taken through the final norm onto the unembedding of the answer letter minus the mean of the other letters, summed over heads and layers 12–23) is at most 0.007/0.006 logits in magnitude, interval included, in every model. The query-side change is not specific to repaired items: L's query side is also positive on the 75 rows Base already answers correctly (+0.26 [+0.19, +0.33] and +0.26 [+0.19, +0.33]). The expected split, V changing the key/value side and L the query/readout side, therefore holds for L only; V reaches the query stage indirectly, through what its image tokens feed into lower layers.

Caveats. Only the final prompt token is analysed, not binding at the option tokens; averaging over heads dilutes any small set of selection heads (head-level rankings are exploratory and not used); the token sets rest on stroke geometry; the value readout is a linear direct attribution through a frozen final norm; and each seed contributes one training run per set.

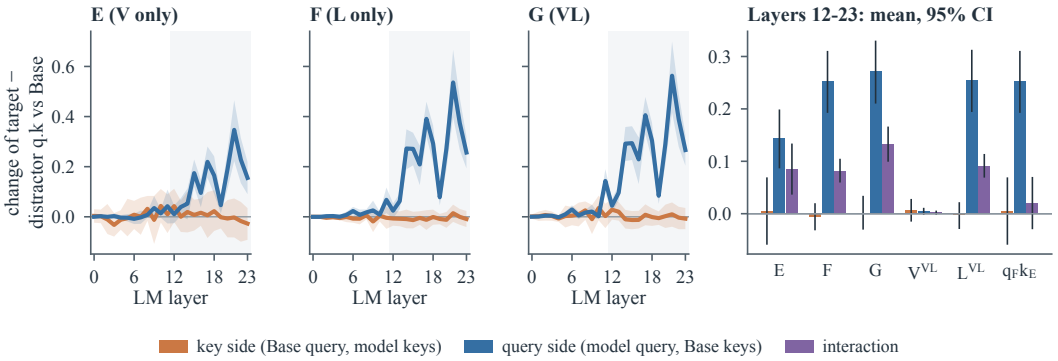

Figure 18: Q/K/V decomposition on G02 (seed 1101; E = V, F = L, G = VL). Left: per-layer change over Base of the target-minus-distractor score contrast at the final prompt token, key and query sides (bands: 95% scene-family intervals; shaded: layers 12–23). Right: mean over layers 12–23 with 95% intervals; V^{VL}, L^{VL}: transplants of VL's modules into Base; $q_F k_E$: L's query reading V's keys.

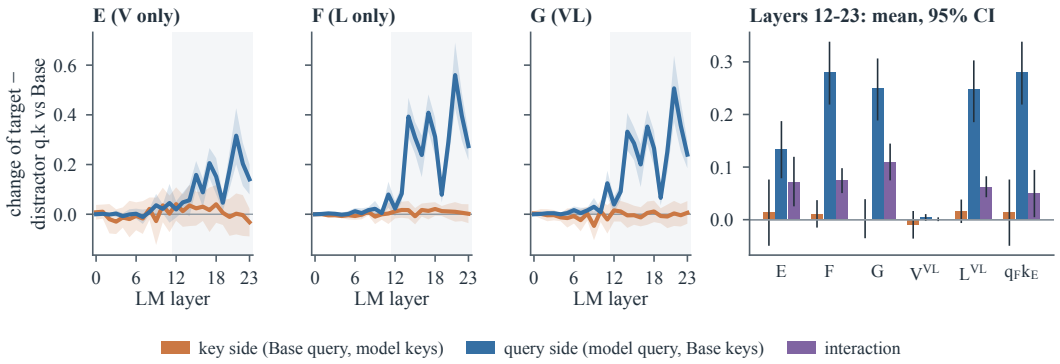

Figure 19: As Figure 18, seed 1104.

Table 44: Change over Base of the target-minus-distractor score contrast at the final prompt token on G02 (layers 12–23, mean over heads, seed 1101), split into key side (Base query, model keys), query side (model query, Base keys), and interaction; brackets: 95% scene-family intervals. V^{VL}, L^{VL}: a module of VL alone in Base context; $q_F \cdot k_E$: the query of L reading the keys of V.

model	total	key side	query side	interaction
E (V only)	+0.23 [+0.16, +0.31]	0.00 [−0.06, +0.07]	+0.14 [+0.09, +0.20]	+0.08 [+0.04, +0.13]
F (L only)	+0.33 [+0.26, +0.39]	−0.01 [−0.03, +0.02]	+0.25 [+0.19, +0.31]	+0.08 [+0.06, +0.10]
G (VL)	+0.41 [+0.33, +0.48]	0.00 [−0.03, +0.03]	+0.27 [+0.21, +0.33]	+0.13 [+0.10, +0.16]
V^{VL}	+0.01 [−0.01, +0.03]	+0.01 [−0.01, +0.03]	0.00 [0.00, +0.01]	0.00 [0.00, 0.00]
L^{VL}	+0.34 [+0.27, +0.41]	0.00 [−0.03, +0.02]	+0.26 [+0.20, +0.31]	+0.09 [+0.07, +0.11]
$q_F \cdot k_E$	+0.28 [+0.20, +0.36]	0.00 [−0.06, +0.07]	+0.25 [+0.19, +0.31]	+0.02 [−0.03, +0.07]

Table 45: As Table 44, seed 1104.

model	total	key side	query side	interaction
E (V only)	+0.22 [+0.14, +0.30]	+0.01 [−0.05, +0.07]	+0.13 [+0.08, +0.19]	+0.07 [+0.03, +0.12]
F (L only)	+0.37 [+0.30, +0.43]	+0.01 [−0.01, +0.04]	+0.28 [+0.22, +0.34]	+0.07 [+0.05, +0.10]
G (VL)	+0.36 [+0.29, +0.43]	0.00 [−0.03, +0.04]	+0.25 [+0.19, +0.30]	+0.11 [+0.08, +0.14]
V^{VL}	−0.01 [−0.03, +0.02]	−0.01 [−0.03, +0.01]	0.00 [0.00, +0.01]	0.00 [0.00, 0.00]
L^{VL}	+0.33 [+0.26, +0.39]	+0.02 [0.00, +0.04]	+0.25 [+0.19, +0.30]	+0.06 [+0.04, +0.08]
$q_F \cdot k_E$	+0.34 [+0.26, +0.43]	+0.01 [−0.05, +0.07]	+0.28 [+0.22, +0.34]	+0.05 [+0.01, +0.09]

O CAUSAL ROUTES OF THE SINGLE-MODULE REPAIRS

The probes and the Q/K/V decomposition of Section 4.3 (Appendices M and N) show that on G02 Base's tokens already carry task-relevant information and that the V-only and the L-only repair both act through the late query side at the answer position. Neither analysis shows which states the repaired models *use*. This appendix intervenes on the forward pass. The models of each seed are taken at the final update and labelled by their runs as in Appendix N: Base; E = V (trained V, Base P and L); F = L (Base V and P, trained L); G = VL. Values are seed 1101/seed 1104, brackets are 95% scene-family cluster bootstrap intervals (2,000 replicates, computed within each seed), and every statement holds in both seeds unless a seed is named. We report decodability (probes, logit lens), causal use (necessity: Base's state put back into an adapted model), and sufficiency (an adapted model's state put into Base) separately. These analyses locate the routes of repairs that training produced; they do not test whether the repaired models use an alternative strategy or a shortcut cue, for which no counterfactual cue intervention exists yet, so such explanations are not ruled out.

O.1 G02: WHICH STATES CARRY THE V-ONLY AND THE L-ONLY REPAIR

Question. Which states of E and of F carry their G02 repairs, and do the two repairs reach the answer by the same computation?

Design. On the 319 official G02 rows, in the numerical regime of the official scorer (whose margins the clean pass reproduces bitwise for Base, E, and F), we replace a receiver's hidden states at a set of positions after one layer (or at the language-model input) with a donor's states from its own clean forward pass on the same item, and let the receiver compute the remaining layers (Meng et al., 2022; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024); the outcome is the fraction of the donor–receiver gap in the answer margin (log-probability of the answer letter minus the best other letter) recovered or removed, as a ratio of means over the primary items (Base wrong, E and F right: 207/212 rows). Sites are image positions (target-object strokes, other-object strokes, all object strokes, the non-object *background* positions, size-matched random background positions), text positions after the image (all of them, the question and options, the final prompt token), and every position; controls are norm-matched random perturbations of the same states, and a second stage recomputes Base's text tokens over a key/value cache in which only the image positions' keys and/or values come from E. The intervention grid was fixed before any patched outcome was computed; the second stage was declared after a two-row smoke run of the first, and the claim thresholds were written after the first seed-1101 chunks and before any seed-1104 result, so seed 1104 is the out-of-sample check.

Result: the V-only repair. At the language-model input E's repair sits in the background positions of the visual token field, not in the objects (Table 46): E's projector outputs at the background positions alone recover 0.85 [0.83, 0.87] and 0.88 [0.86, 0.90] of the gap in Base (accuracy 92/89%), those on all object strokes 0.20/0.19, and those on the target objects 0.08/0.08, less than as many random background positions (target minus random: -0.14 [-0.16, -0.12] and -0.17 [-0.19, -0.15]). Necessity mirrors it: Base's background features put into E remove 0.80 [0.77, 0.83] and 0.81 [0.78, 0.83] of E's advantage (E's accuracy falls to 8/5%), Base's object-stroke features 0.15/0.12. A norm-matched random perturbation of all image features recovers 0.11/0.10. Inside the frozen language model, E's image-token states fall below half of the repair at layer 7/7 and carry at most 0.02/0.02 of it from layer 11 on, while its text-token states reach half at layer 6/6 and carry 0.99/0.99 of it in layers 12–23; the question and option tokens alone carry 0.01/0.00 (Figure 20). The second stage shows how the text reads the image: E's image keys and values at every layer reproduce E in Base's text computation (1.00/1.00), keys alone give 0.30 [0.29, 0.32] and 0.29 [0.28, 0.31] and values alone 0.34 [0.32, 0.35] and 0.34 [0.33, 0.36], so both are needed; restricted to the background positions they give 0.90/0.91, to the object strokes 0.16/0.16. These reads are spread over the lower layers: accumulated from layer 0 they reach half of the repair at layer 6/6, and no single layer gives more than 0.43/0.44.

Result: the L-only repair. F's image-token states carry none of its repair (at most 0.05/0.06 in magnitude at any layer), and its text-token states do as well as all of its states (difference at most 0.10/0.09). The splice of F's layers up to l with Base's above reaches half of the repair at layer 5/6 and 0.98/1.00 in layers 12–23, and the reverse splice loses 0.82/0.82 of F's advantage there: F computes its repair gradually in the text positions of the lower and middle layers.

Result: where the routes meet. Both repairs hand the decision to the final prompt token at the same layer: the final-token state alone crosses half of the repair at layer 14/14 for E and 14/14 for F (0.89/0.88 and 0.84/0.87 in layers 12–23), whereas one layer earlier the final token alone carries 0.46/0.47 (E) and 0.31/0.33 (F) and all text tokens together 0.97/0.98 and 0.94/0.97. At the final token E's change lies along F's: projected on F's change on the same item it recovers 1.04 [1.01, 1.09] and 1.03 [1.00, 1.07] of E's effect in layers 12–23, and its orthogonal rest $-0.05/-0.07$. F's change contains more than E's direction (along E's change 0.62/0.63, orthogonal rest 0.51 [0.49, 0.53] and 0.53 [0.51, 0.55]). Norm-matched random final-token perturbations recover 0.06/0.06 (E) and 0.06/0.06 (F), and the same-item cosine of the two changes at the final token, 0.80/0.78, exceeds the cosine between different items, 0.62/0.62. Base's text-token states put into E remove its repair (1.00/1.00 in layers 12–23), and Base's final-token state removes it from layer 10/10 on: E's final-token state is necessary from there but sufficient only from layer 14/14.

Table 46: G02 activation patching (E = V, F = L; primary items, 207/212 rows in seed 1101/seed 1104): fraction of the margin gap between receiver and donor recovered in the receiver (E or F → Base, sufficiency) or removed from it (Base → E, necessity), with 95% scene-family intervals. Rows "layers 12–23" are the mean over those layers of the fraction after patching at each layer; "along / orthogonal": the donor's final-token change projected on the other model's same-item change, and its orthogonal rest.

intervention (fraction of the margin gap)	seed 1101	seed 1104
E → Base, all image features at the LM input	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
background image tokens only	0.85 [0.83, 0.87]	0.88 [0.86, 0.90]
object-stroke tokens only	0.20 [0.17, 0.23]	0.19 [0.17, 0.22]
target-object tokens only	0.08 [0.06, 0.10]	0.08 [0.06, 0.09]
size-matched random background tokens	0.22 [0.20, 0.25]	0.24 [0.22, 0.27]
norm-matched random perturbation	0.11 [0.10, 0.11]	0.10 [0.10, 0.11]
E → Base, text tokens, layers 12–23	0.99 [0.98, 0.99]	0.99 [0.98, 0.99]
E → Base, final token, layers 12–23	0.89 [0.88, 0.90]	0.88 [0.88, 0.89]
along F's delta / orthogonal to it	1.04 / −0.05	1.03 / −0.07
F → Base, image tokens, layers 12–23	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
F → Base, text tokens, layers 12–23	0.98 [0.97, 0.99]	1.00 [1.00, 1.01]
F → Base, final token, layers 12–23	0.84 [0.84, 0.85]	0.87 [0.86, 0.87]
along E's delta / orthogonal to it	0.62 / 0.51	0.63 / 0.53
Base → E, background tokens at the LM input (necessity)	0.80 [0.77, 0.83]	0.81 [0.78, 0.83]
Base → E, object-stroke tokens at the LM input (necessity)	0.15 [0.13, 0.17]	0.12 [0.10, 0.14]

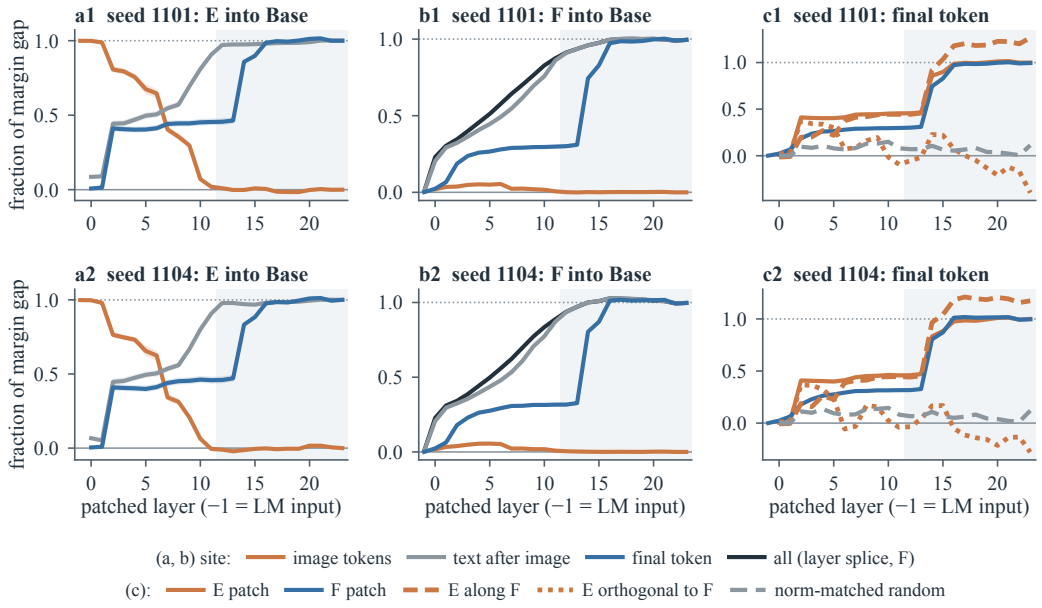

Figure 20: G02 activation patching by layer (rows: seeds 1101 and 1104; primary items). (a) E's states put into Base at the image tokens, at the text tokens after the image, and at the final token; (b) F's states put into Base (all positions = the layer splice), at the image tokens, the text tokens, and the final token; (c) the final token: E's and F's patches, E's change projected on F's same-item change and its orthogonal rest, and a norm-matched random perturbation. Bands: 95% scene-family intervals; shaded: layers 12–23; layer −1 is the language-model input.

Result: decodable is not used. E's target-object tokens decode the target shape type better than Base's (probe balanced accuracy 78/80% vs. 66%; Appendix M), yet put into Base they recover only 0.08/0.08 of E's repair. Base's option tokens decode the target option (81%) although Base answers 24% of the rows correctly.

Reading. On G02 the location of the parameter change is not the location of the computation that repairs. E adapts only the vision encoder, but its repair is carried by the background positions of the visual token field and computed by the frozen language model, whose text tokens read E's image keys and values in the lower layers and whose final-token query then carries the change (the late query side of Appendix N); F changes the text computation itself and makes no causal use of image-token states. The two different causal routes meet in one downstream decision state at the final token, where E's change is a sub-direction of F's. Together with the decodability results, this gives the RQ3 picture for G02: Base already carries task-relevant information that its computation does not use, and interventions in different modules repair the same failure by changing that downstream computation along different routes. Status for the candidate families of Section 4.3: an upstream change that reroutes frozen downstream computation (B5): supported (G02, both seeds); several functional leverage points meeting on one late decision state (B1/B5): supported (G02, both seeds); information present but under-used (A1): supported for G02 at the level of decodability (a probe shows that the information is present, not that Base could use it); a target-localized visual repair and a key-only V-side mechanism: not supported. *Caveats.* The last layer of the vision encoder mixes all patches, so a background position can carry information about the objects: "background" describes positions, not content, and whether the repair uses information distributed over the visual field (A2) is unresolved. The projection test uses a direction defined by F's own change and is a subspace intervention, whose interpretation needs care (Makelov et al., 2024; Wu et al., 2024); here the projection carries all of E's effect and the orthogonal rest none. The identity runs (resumed, suffix, and batched forwards without a patch) stay within 0.500/0.625 nats of the clean margin.

O.2 G02: DO THE TWO REPAIRS NEED THE SAME ATTENTION HEADS?

Question. E and F both raise the target-over-distractor query–key contrast at the answer position in late layers (Appendix N). Do they need the same attention heads there?

Design. For each seed and model, the query heads of layers 12–23 are ranked by the query-side effect of Appendix N (attention scores only, no margin or prediction), the top five (F5, E5) and top ten (F10, E10) form the ablation sets, and five random sets matched on layer and disjoint from both top-ten sets are the controls; a head is ablated by zeroing its slice of the attention output at every position (zero ablation; Wang et al. (2023) use mean ablation), and the outcome is the fraction of each model's margin repair over Base removed, on the 319 official rows. The decision rule was fixed with the head sets before any ablation: a set "removes most" of a repair if it removes at least half with its paired contrast to the random mean above zero, and the verdict is head-level convergence (most of E's and F's repair, random sets clear), different heads (the owner's repair only), or distributed (neither).

Result. In attention scores the two repairs use largely the same heads: E's and F's top-five sets share 4 heads in seed 1101 and 4 in seed 1104, their top-ten sets 7 and 7 (Table 47). Ablating them removes little of either repair (Table 48): the verdict is distributed for every set in both seeds, every top-five set removes at most a fifth of the margin repair of E, F, and VL, the top-ten sets also remove less than half, and no set moves the accuracy repair by more than a few points. F's top five removes a small share of F's repair that is above its random sets (contrast 0.04 [0.04, 0.05] and 0.08 [0.07, 0.09]); E's own top five removes almost none of E's (-0.00 [-0.01, 0.00] and -0.00 [-0.01, 0.00]). F's top ten removes more of E's repair than each of its random sets and more than of F's own (0.19/0.21 vs. 0.14/0.14); in Table 48 the largest shares of E's repair are removed by the sets that contain the one layer-14 head of F's top ten (Table 47), the layer at which both repairs reach the final token (Appendix O.1). That head was not ablated alone, so attributing the share to it is a hypothesis.

Reading. At head level the convergence of the two repairs is visible in the attention scores but is not a small set of heads that both repairs need: by the pre-set rule both repairs are distributed. This is consistent with the shared decision state of Appendix O.1 being carried by many heads, or being rebuilt by other heads when some are removed; zero ablation measures necessity given the other heads, so it cannot separate the two. Status: a small set of heads as the meeting point of the two routes (B1/B5 at head level): not supported (both seeds); how the shared state is carried (many heads, or compensation by other heads): unresolved.

Table 47: Query heads selected for ablation on G02 (layer, head), by seed. F5/F10 and E5/E10: the top five/ten heads by the query-side effect of L (F) and of V (E).

set	seed 1101	seed 1104
F5	L16H1, L17H8, L15H12, L15H11, L17H3	L14H13, L16H1, L15H12, L17H8, L15H11
E5	L15H11, L15H12, L16H1, L21H12, L17H8	L15H12, L15H11, L16H1, L17H8, L21H12
F10	L16H1, L17H8, L15H12, L15H11, L17H3, L21H12, L14H13, L20H7, L21H6, L18H11	L14H13, L16H1, L15H12, L17H8, L15H11, L17H3, L21H12, L20H7, L15H13, L21H6
E10	L15H11, L15H12, L16H1, L21H12, L17H8, L20H7, L20H8, L21H6, L21H9, L21H2	L15H12, L15H11, L16H1, L17H8, L21H12, L20H7, L20H8, L21H6, L21H9, L20H11

Table 48: G02 head ablation: fraction of the margin repair over Base removed by zeroing each head set (E = V, F = L, G = VL; all official rows), with 95% scene-family intervals; random: mean (maximum) over the five layer-matched random sets; verdict: the pre-set decision rule.

seed	set	E removed	F removed	G removed	random, E	random, F	verdict
1101	F5	0.00 [−0.01, 0.01]	0.04 [0.04, 0.05]	0.05	−0.05 (0.09)	−0.00 (0.11)	distributed
	E5	−0.00 [−0.01, 0.00]	0.04 [0.03, 0.04]	0.05	0.03 (0.07)	0.01 (0.03)	distributed
	F10	0.19 [0.18, 0.21]	0.14 [0.14, 0.15]	0.15	0.04 (0.13)	0.04 (0.08)	distributed
	E10	0.03 [0.02, 0.04]	0.06 [0.05, 0.06]	0.07	−0.03 (0.11)	0.02 (0.12)	distributed
1104	F5	0.16 [0.15, 0.18]	0.11 [0.10, 0.12]	0.11	−0.03 (0.17)	0.03 (0.17)	distributed
	E5	−0.00 [−0.01, 0.00]	0.02 [0.01, 0.02]	0.02	−0.03 (0.13)	0.06 (0.21)	distributed
	F10	0.21 [0.19, 0.22]	0.14 [0.13, 0.15]	0.13	−0.01 (0.05)	0.01 (0.06)	distributed
	E10	−0.01 [−0.02, −0.00]	0.02 [0.02, 0.03]	0.03	0.01 (0.12)	0.03 (0.14)	distributed

O.3 Is the G02 convergence task-specific? Final-token states on G02, B02, and B07

Question. On G02 the V-only and the L-only repair meet in one final-token state (Appendix O.1). Do the two single-module repairs of B02 and B07 converge in the same way?

Design. From the probe captures of Appendix M (official rows, residual stream at the final prompt token after every layer, official scoring regime), we compute per item the changes $\Delta_E = h_E - h_{\text{Base}}$ and $\Delta_F = h_F - h_{\text{Base}}$ and summarize them over layers 12–23 on the primary items of each task (Base wrong, E and F right in that seed's official predictions): the same-item cosine C, its excess X over the cosine between different items of the task (the part of the alignment specific to the item), and the share of $\|\Delta_F\|^2$ along Δ_E of the same item. The labels (convergent: lower bound of C at least one half and of X above zero; shared direction only: the same with X not above zero; partial: lower bound of C between zero and one half), the task contrasts, and a logit-lens readout (each model's own final norm and output head applied to its intermediate final-token states) were fixed before any statistic was computed.

Result. On G02 the two changes converge item by item; on B02 they converge only partially (Table 49). The same-item cosine is higher on G02 than on B02 by 0.34 [0.33, 0.35] and 0.29 [0.28, 0.29] and than on B07 by 0.25 [0.20, 0.28] and 0.24 [0.23, 0.25]; every G02 primary item (lowest 0.69/0.69) lies above every B02 primary item (highest 0.56/0.60); the item-specific excess is 0.174/0.168 on G02 and 0.042/0.051 on B02, and 0.64/0.62 of F's squared change lies along E's same-item change on G02 against 0.22/0.26 on B02. The shared G02 direction is specific to the task: the cosine of E's and F's mean changes is 0.84/0.83 within G02 but at most 0.37/0.39 between E's mean change on one task and F's on another. B07's primary set is small (3/20 items), so its primary-set values are descriptive and B07 is compared on all rows (G02 minus B07: 0.24/0.20). In the logit lens, the G02 answer emerges at layer 15/15 in E, F, and VL alike on all rows, and Base's G02 lens stays at chance at every layer (peak balanced accuracy 0.26), so no lens-readable answer is computed and later overwritten. On B02, Base's own lens reads the relation better at layer 15 (balanced accuracy 0.67 [0.65, 0.69]) than at its output (0.55; difference 0.12 [0.10, 0.15]); Base is shared by the two seeds, so this is one readout of one model.

Reading. The G02 narrow waist, two single-module repairs meeting in one final-token state, is specific to G02 among the tasks with a causal test (the pre-set rule labels B07's all-row changes convergent; Table 49): on B02 the V-only and L-only repairs reach their answers through largely different final-token states. This is representation similarity at one position, not a causal test; the causal counterpart for B02 is Appendix O.4. The late readout of B02's relation in Base, stronger

than at its output, is a lens observation compatible with a late overwrite, not evidence that late layers cause the failure. Status: convergence on one decision state (B1/B5) is supported for G02 and partial for B02 (representation level); B02's late-overwrite pattern: partial (readout only).

Table 49: Final-token convergence of the V-only (E) and L-only (F) repairs, layers 12–23. C: same-item cosine of Δ_E and Δ_F; cross: the cosine between different items of the task; X: item-specific excess (C minus cross); share: fraction of $\|\Delta_F\|^2$ along the same-item Δ_E (in parentheses: along a different item's Δ_E); label: the pre-set rule. Primary items (Base wrong, E and F right), except the B07 all-rows line. Brackets: 95% scene-family intervals.

seed	task	items	n	C	cross	X	share	label
1101	G02	primary	207	0.80 [0.79, 0.80]	0.62	0.174 [0.169, 0.178]	0.64 (0.41)	convergent
	B02	primary	108	0.46 [0.45, 0.47]	0.42	0.042 [0.028, 0.058]	0.22 (0.19)	partial
	B07	primary	3	0.55 [0.52, 0.60]	0.53	0.016 [−0.011, 0.066]	0.30 (0.29)	shared-direction only
	B07	all rows	720	0.54 [0.54, 0.55]	0.49	0.055 [0.052, 0.058]	0.30 (0.24)	convergent
1104	G02	primary	212	0.78 [0.78, 0.79]	0.62	0.168 [0.163, 0.172]	0.62 (0.40)	convergent
	B02	primary	109	0.50 [0.49, 0.50]	0.45	0.051 [0.038, 0.070]	0.26 (0.21)	partial
	B07	primary	20	0.55 [0.54, 0.56]	0.52	0.031 [0.020, 0.040]	0.30 (0.27)	convergent
	B07	all rows	720	0.57 [0.57, 0.58]	0.51	0.062 [0.059, 0.065]	0.33 (0.26)	convergent

O.4 B02: DO THE TWO REPAIRS TAKE DIFFERENT ROUTES, AND DO THEY MEET?

Question. On B02 (are two circles touching?), E and F also repair the same items. Do their repairs act along different routes and meet in one downstream state, as on G02?

Design. The G02 design of Appendix O.1 (models, scoring regime, metric, patch semantics) is ported to all official B02 rows with a reduced grid fixed before any B02 patched outcome: image sites at the language-model input are the *contact* region (tokens whose source cell meets a disk of half the radius at the midpoint of the two centres, the probes' contact site), the disks outside it, the background, and size-matched random background tokens; text, question, and final-token sites at every layer; F's layer splice and its reverse; the final-token projection test; norm-matched random controls. The decision rule, also fixed in advance, calls the convergence "like G02" only if both final-token states carry at least half of the repair in layers 12–23, cross half within two layers of each other, and E's final-token change projected on F's same-item change carries most of E's effect with a negligible orthogonal rest (the criteria G02 meets). The decision rule was fixed after a three-row plumbing smoke and before any set-level B02 result; the other claim thresholds were drafted after a 576-row partial analysis of both seeds, seven claims after the complete results, and the identity-run tolerance (0.2 nats) after that partial, so unlike on G02 seed 1104 is not an out-of-sample check for them.

Result. On the primary items (108/109 rows), E's repair enters through the image features (all of them recover 1.02 [1.01, 1.04] and 1.02 [1.01, 1.03] in Base) and is carried by the text positions in layers 12–23 (0.92 [0.91, 0.92] and 0.91 [0.90, 0.92]); F's image-token states carry none of F's repair (at most 0.03/0.04 in magnitude) and its layer splice carries it (0.95/0.94), so the upstream routes differ as on G02 (Table 50). Unlike on G02, E's image locus is centred on the contact region but, by the pre-set rule, mixed: E's contact-region features alone recover 0.48 [0.41, 0.55] and 0.40 [0.34, 0.44] of the gap, more than as many random background tokens (contrast +0.46 [+0.39, +0.53] and +0.38 [+0.33, +0.43]) and more than the disk tokens outside the contact region (+0.46 [+0.40, +0.53] and +0.38 [+0.32, +0.42]); its background features recover −0.38/−0.25 (G02: 0.85/0.88); Base's contact-region features put into E remove 1.40 [1.19, 1.65] and 1.28 [1.08, 1.44] of E's advantage; and a norm-matched random perturbation of the contact features recovers 0.00/0.00. Downstream, both repairs reach the final token (0.70/0.72 for E and 0.77/0.76 for F in layers 12–23; crossing half at layers 14/14 and 14/14), but, unlike on G02, E's final-token change is not contained in F's: projected on F's same-item change it recovers 0.36 [0.31, 0.42] and 0.29 [0.25, 0.33] of E's effect (G02: 1.04/1.03), its orthogonal rest 0.15/0.20. Norm-matched random final-token perturbations recover −0.03/−0.03 (E) and −0.01/−0.01 (F). Pre-set verdicts: E's image locus mixed (seed 1101) and mixed (seed 1104); convergence like G02: no/no.

Reading. On B02, too, the V-only and the L-only repair take different upstream routes, but by the pre-set rule they do not meet in one final-token state as on G02: E's repair is read mainly from the contact region of the image, and its final-token change is not contained in F's. With Appendix O.3,

the G02 picture of different routes converging on one decision state is therefore specific to G02; on B02 the two repairs remain different computations down to the answer position. Status: different upstream causal routes for one failure (B1): supported on G02 and B02 (both seeds); convergence on one late decision state (B1/B5): G02 only.

Table 50: B02 causal contrast (E = V, F = L; primary items, 108/109 rows in seed 1101/seed 1104): fraction of the margin gap recovered or removed, as in Table 46, with 95% scene-family intervals; G02 rows: the same statistic from Appendix O.1.

B02 intervention (fraction of the margin gap)	seed 1101	seed 1104
E → Base: all image features (LM input)	1.02 [1.01, 1.04]	1.02 [1.01, 1.03]
contact-region tokens only	0.48 [0.41, 0.55]	0.40 [0.34, 0.44]
as many random background tokens	0.03 [0.01, 0.04]	0.02 [0.01, 0.02]
disk tokens outside the contact region	0.02 [0.00, 0.04]	0.02 [0.01, 0.03]
background tokens	−0.38 [−0.61, −0.18]	−0.25 [−0.41, −0.05]
Base → E: contact-region tokens (necessity)	1.40 [1.19, 1.65]	1.28 [1.08, 1.44]
E → Base: text tokens, layers 12–23	0.92 [0.91, 0.92]	0.91 [0.90, 0.92]
E → Base: final token, layers 12–23	0.70 [0.69, 0.70]	0.72 [0.71, 0.73]
along F's delta / orthogonal (B02)	0.36 / 0.15	0.29 / 0.20
along F's delta / orthogonal (G02)	1.04 / −0.05	1.03 / −0.07
F → Base: image tokens, max over layers	0.03	0.04
F → Base: layer splice, layers 12–23	0.95 [0.92, 0.97]	0.94 [0.92, 0.97]
F → Base: final token, layers 12–23	0.77 [0.75, 0.79]	0.76 [0.74, 0.79]
along E's delta / orthogonal (B02)	0.13 / 0.37	0.22 / 0.23
along E's delta / orthogonal (G02)	0.62 / 0.51	0.63 / 0.53
final-token crossing layer E / F (B02)	14 / 14	14 / 14
final-token crossing layer E / F (G02)	14 / 14	14 / 14
0.5 reached by E's text tokens / F's splice (B02)	11 / 12	11 / 12
0.5 reached by E's text tokens / F's splice (G02)	6 / 5	6 / 6

P FUNCTION AND PARAMETER SPACE OF THE REPAIRS, AND THE PAIR-EMERGENT REGIME

This appendix collects the RQ3 analyses that ask whether the availability of several repairs is a property of function and parameter space (candidate family D), and how pair-emergent repair differs from multiple-single repair. *Notation.* Runs carry the letters of Section 3 (A=P, B=PL, C=VPL, D=VP, E=V, F=L, G=VL). X.M is module M as trained in run X, placed alone in Base context (M^S of Section 3.3; e.g. G.V = V^{VL}, F.L = L^L); X.M+Y.N is a recombination with Base elsewhere; a prime marks a module of the other seed's run (G′.L). *Item classes* of a Base error: *multiple-single* (MS) if at least two single sets repair it, *single-sole* if exactly one does, *pair-emergent* (globally new) if no single set and some pair does, *constituent-new only* if a pair repairs it and its own two single sets do not but the third single set does, *triple-emergent*, and *unrepaired* (Section 3.3). Unless stated otherwise the classes use the re-scored models of Section 4.2 (Appendix F), so counts can differ slightly from the RQ1 repair space. Values are seed 1101/seed 1104 with 95% scene-family intervals (2,000 replicates) within each seed; seeds are never pooled. Every qualitative sentence below holds in each seed unless it gives per-seed values and says that the seeds differ. All analyses reuse the official rows and were designed after the official results were seen (Section 3.4); they are exploratory.

P.1 MARGIN STRUCTURE OF THE REPAIR CLASSES

Question. When a pair or the triple repairs a Base error that no single set repairs, is the repair carried by one jointly trained module, by single-module effects that are each below the decision boundary and only cross it together, or by an interaction between the modules? And do the modules behind a multiple-single repair act redundantly?

Design. From the stored candidate log likelihoods of the re-scored models (no new scoring), the margin of model M on item i is $m_M(i) = \mathrm{LL}_M(\text{gold}) - \max_{c \neq \text{gold}} \mathrm{LL}_M(c)$ in nats ($m > 0$: correct). For a pair $X = ab$ with its modules alone in Base context, the additive prediction is

$m_{\text{add}} = m(a|X) + m(b|X) - m(\text{Base})$ and the interaction $I = m(X) - m_{\text{add}}$ (for VPL the three transplants enter and the reversions give the pairwise and three-way terms). A repair of X is *carried* if one module of X alone repairs it; otherwise it is *additive* if $m_{\text{add}} > 0$ and $|I| \leq \tau$, *superadditive* if $I > \tau$, and *other* otherwise, with $\tau = 0.5$ nats. For VL, recombination with the single-trained partners (G.V+F.L, E.V+G.L) asks whether the synergy needs the co-trained partner: a superadditive repair is *partner-specific* if its interaction with each independently trained partner stays within τ.

Result. Of the 6383 Base errors, 2893/2783 are multiple-single, 2075/2116 single-sole, 325/377 pair-emergent, 72/121 triple-emergent, and 1018/986 unrepaired (Table 51, Figure 21). *Pair-emergent repairs are rarely additive:* pooled over the three pairs, 1 [0, 1] and 1 [0, 3]% of them are sub-threshold additive, whereas 48/42% are carried by one module and 46/52% are superadditive. *Which of the two depends on the pair.* In PL, B.L alone makes 88 [77, 98] and 80 [72, 88]% of the pair-emergent repairs; in VP, D.V alone makes 57 [45, 69] and 44 [33, 56]%. In VL, 100/100% of the repairs without a carrier are superadditive, with a mean interaction of +9.8 [+7.0, +12.2] and +11.1 [+8.6, +13.4] nats against +0.55 [−0.08, +1.16] and +2.07 [+1.67, +2.48] on VL's multiple-single instances. The VL synergy has a consistent sign structure: on 92 [84, 96] and 97 [94, 99]% of its superadditive repairs G.L alone moves the margin away from gold (mean −8.4/−8.4 nats) while G.V alone moves it slightly toward gold (mean +0.8/+0.8). It is mostly not partner-specific: it survives an independently trained partner on 97 [94, 100] and 96 [91, 99]% of these repairs and is partner-specific on 3 [0, 6] and 4 [1, 9]%. The two single-trained modules together (E.V+F.L) repair 44 [28, 61] and 52 [43, 62]% of VL's pair-emergent items, although neither alone does and their additive prediction crosses the boundary on only 9/9%. Adding the log-likelihood vectors instead of margins gives the same category for 99/99% of the VL repairs without a carrier. *Triple-emergent repairs are essentially V×L:* C's V and L with Base P repair 99 [95, 100] and 98 [95, 100]% of them, the three-way term is +0.02 [−0.06, +0.09] and +0.01 [−0.07, +0.09] nats against a V–L term within C of +9.6 [+6.6, +12.0] and +7.4 [+3.8, +11.5], and the mean stand-alone margin change of P stays within ±0.21/0.18 nats in every set and class. *Multiple-single repairs mostly keep a stand-alone carrier:* at least one jointly trained module still repairs the item alone on 96/96% (PL), 98/97% (VP), 90/88% (VL), and 83/80% (VPL) of the instances. Their interaction is negative in PL (−1.09 [−1.24, −0.94] and −0.71 [−0.82, −0.60] nats) but positive in VP (+0.73 [+0.62, +0.83] and +0.83 [+0.69, +0.98]). *Unrepaired items are not simply the farthest from the boundary:* their Base margin is only slightly below that of multiple-single items (difference −0.33 [−0.53, −0.12] and −0.36 [−0.56, −0.13] nats) and not detectably different from that of pair-emergent items (−0.22 [−0.63, +0.18] and −0.14 [−0.45, +0.19]); the pair sets move 80 [76, 83] and 76 [72, 80]% of the unrepaired instances away from gold (pairs pooled), and the stand-alone effects would add up to a repair on 2 [1, 4] and 3 [2, 5]%. The pair-emergent instances concentrate on B04 (43/37% of the pooled pair-emergent instances), the case of Appendix P.3.

Reading. In margin space, pair-emergence is not the sum of single-module moves that each fall short of the boundary. Either joint training yields a module that carries the repair alone (most PL and many VP repairs): learning the repair needed the pair, but the learned solution belongs to one module, so emergence is not ownership. Or the two trained modules interact superadditively (VL), and the triple-emergent items are this V×L interaction realized in C. For the analysed VL pair-emergent repairs the synergy mostly does not require run-private code: it transfers to independently trained partners, and the single-trained modules alone compose on a substantial share. This pattern fits several leverage points arranged in series, a changed V output that becomes usable through a changed L readout (candidates B1/B3); the information families (A) are not tested by margins. PL, VP, and VPL have no recombination of a jointly trained module with a single-trained partner, so co-adaptation and multi-stage computation are not separable for them. A module that hurts alone in Base context (G.L) is also what an interface mismatch would produce (transplant incompatibility, family F), so the sign structure is descriptive. Redundancy is set-specific, and a negative interaction on multiple-single items is also what margin saturation alone produces. Unrepaired items are not repaired under the current adaptation procedure (data, budget, optimizer); the margins show that the adaptation moves most of them away from gold, not that they cannot be repaired. Margins include the runner-up switch and bf16 scoring quantizes log likelihoods; alternative strategies and shortcuts are not ruled out.

Table 51: How pair- and triple-emergent repairs are realized in margin space (Appendix P.1): number of repaired instances n and the share (%) that is carried by one module of the set alone (carr.), sub-threshold additive (add.), superadditive (super.), or other; $\tau = 0.5$ nats. *Constituent-new only*: the pair's own single sets fail but the third single set repairs the item.

	seed 1101					seed 1104				
repairs	n	carr.	add.	super.	other	n	carr.	add.	super.	other
PL, pair-emergent	50	88	2	8	2	81	80	1	16	2
VP, pair-emergent	80	59	1	15	25	81	46	5	28	21
VL, pair-emergent	241	37	0	63	0	293	31	0	69	0
three pairs, pair-emergent	371	48	1	46	6	455	42	1	52	4
three pairs, constituent-new only	287	70	4	21	5	389	65	3	28	4
VPL, triple-emergent	72	50	0	50	0	121	55	2	43	0

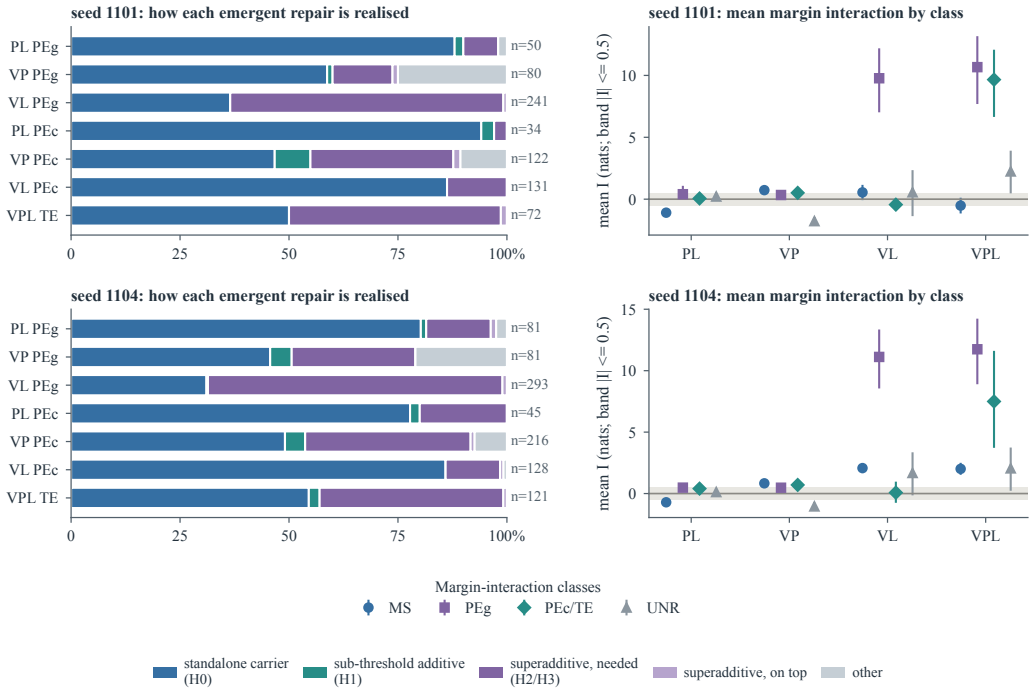

Figure 21: Margin structure by seed (Appendix P.1). Left: share of each emergent repair class carried by one module alone (legend H0), sub-threshold additive (H1), superadditive with an additive prediction at or below zero (needed) or above it (on top), or other; the legend's labels H0–H3 are those of the margin analysis, not the hypotheses of Appendix P.3. Right: mean margin interaction I (nats) by set and item class with 95% scene-family intervals; band: $|I| \leq \tau$.

P.2 First-order blockwise test at Base

Question. Does the local, first-order geometry of Base predict which modules can repair an item, and so separate multiple-single, pair-emergent, and unrepaired items (candidate D2, blockwise tangent spaces that reach the same output change)?

Design. On a stratified sample of 575 items (multiple-single, pair-emergent, and unrepaired, drawn per seed in proportion to each class's task mix) we computed the gradient g_X of the Base margin with respect to each module $X \in \{V, P, L\}$ in the scorer's numerical regime (the recomputed Base margin differs from the stored one by at most 0.250 nats). Everything is compared in output (margin) space: $g_X \cdot \Delta\theta_X$ is the first-order margin change of module X's actual single-set update (E for V, A for P, F for L; same seed). A module counts as *capable* on item i if an update of the realized size and typical alignment, $\eta_X \|g_X\| \|\Delta\theta_X\|$, exceeds the Base deficit $-m_{\text{Base}}(i)$, where η_X is the median

realized $|\cos(g_X, \Delta\theta_X)|$ over the sample (V 7.9×10^{-3}/8.1×10^{-3}, P 3.9×10^{-2}/4.0×10^{-2}, L 3.6×10^{-3}/3.4×10^{-3}). Without the alignment factor every module would count as capable on essentially every item.

Result (Table 52). L counts as capable on every sampled item of every class in both seeds (100/100% of multiple-single, 100/100% of pair-emergent, 100/100% of unrepaired items), so no sampled item has zero capable modules (0/0% of unrepaired items). V and P count as capable less often on pair-emergent and unrepaired items than on multiple-single items (point estimates; V: 63/61%, 48/43%, 35/40%). The first-order prediction ranks the observed single-set margin changes moderately for V and P but not for L on multiple-single items (Spearman -0.23/-0.04). The first-order sum of E's V and F's L predicts a crossing on 32/30% of unrepaired items, whereas the recombination E.V+F.L repairs 8/7% of them.

Reading. This is a partial, largely negative result. The first-order test does not separate the classes cleanly: L is locally capable in every class, so the Base Jacobian does not predict which repair training realizes, and unrepaired items are not items without first-order leverage. We make no claim about the neural tangent kernel or the full Jacobian: the test is first-order, per module, at Base, on a sample, and its capability threshold depends on the alignment normalization we chose.

Table 52: First-order blockwise test at Base (Appendix P.2), per seed and item class: sampled items n; share (%) of items on which V, P, or L counts as capable, and on which at least two modules do; Spearman correlation between the first-order and the observed margin change of each module's single-set update; share (%) on which the first-order sum of E.V and F.L predicts a crossing (pred.) and on which the recombination E.V+F.L repairs the item (obs.).

			capable (%)				Spearman			E.V+F.L (%)	
seed	class	n	V	P	L	≥ 2	V	P	L	pred.	obs.
1101	multiple-single	198	63	53	100	63	0.47	0.67	−0.23	83	90
1101	pair-emergent	126	48	35	100	48	0.41	0.57	0.50	29	47
1101	unrepaired	195	35	27	100	35	0.41	0.66	0.29	32	8
1104	multiple-single	189	61	53	100	62	0.43	0.72	−0.04	83	88
1104	pair-emergent	136	43	28	100	43	0.35	0.55	0.37	37	51
1104	unrepaired	192	40	34	100	40	0.41	0.65	0.25	30	7

P.3 A PAIR-EMERGENT CASE: COUNTING (B04) WITH VL

Question. What does joint VL training build on items that no single set repairs? We distinguish four hypotheses: one jointly trained module carries the repair (H0); two independent deficits that independently trained modules fix and that compose (H1); a code private to the co-trained pair (H2); or a computation that needs both modules changed in the joint way but whose interface is shared across independent joint runs (H3). H2 and H3 predict the same failures with Base or single-trained partners; only co-trained partners from another run separate them. This operationalization is ours.

Design. The case was chosen by a rule fixed before the mechanism analysis: in every seed a (task, pair) needs at least 40 globally pair-emergent items (official predictions) from at least 10 scene families, holding at least 60% of the task's globally pair-emergent items, with the same top pair in both seeds. Only B04/VL qualifies (runner-up B07/VP, with 34 items in its weaker seed); its item set S has 141/135 items from 15/20 scene families. We scored module compositions on S with the scorer of Section 4.2 (the extra campaign re-scores G bitwise identically in both seeds): each module of G alone; partners trained without the other module (E.V+F.L, G.V+F.L, E.V+G.L, G.V+B.L, D.V+G.L); co-trained partners from another run (C's modules, and the other seed's G); and single-trained partners across seeds. A composition *works* on a stratum if it repairs at least 50% of it and *fails* if it repairs at most 20%. B04 asks for a count; the gold count of the items in S is mostly 6 or 7, and an L trained without V almost never answers an even count, so S is split by the parity of the gold count into 74/91 odd and 67/44 even items (scene families are gold-pure, so the strata are family-disjoint). The split was chosen after the first exploration and before the extra compositions were scored. The *even core* are the 37/17 even items that neither module of G repairs alone.

Result (Table 53, Figures 22 and 23). *Odd counts: two deficits that compose (H1 in both seeds).* The single-trained modules together (E.V+F.L) repair 82 [65, 96] and 73 [60, 92]% of the odd items, although E and F alone repair none (by selection) and each module of G alone repairs few (G.V 1/8%, G.L 0/0%); each module of G also works with the single-trained partner (G.V+F.L 100/98%, E.V+G.L 53/59%), and single-trained partners from the other seed compose like those of the same seed (E.V+F′.L 81/75%, E′.V+F.L 82/75%). *Even counts: a jointly trained partner is needed, but not the own one (H3 in both seeds).* Every split with a partner trained without the other module fails (E.V+F.L 0/0%, G.V+F.L 3/0%, E.V+G.L 0/0%, G.V+B.L 1/2%, D.V+G.L 0/0%), and a single-trained L is a worse partner for G.V than Base's L (-33 [-73, -5] and -25 [-52, -6] pp). A co-trained partner from the other seed's independent run works: G.V+G′.L repairs 97 [91, 100] and 77 [53, 94]% and G′.V+G.L 61 [45, 84] and 91 [79, 100]% of the even items (100/100% and 68/100% of the even core), and a partner co-trained in C works in at least one direction (G.V+C.L 82/43%, C.V+G.L 64/86%). The L trained without V (F, B) answers an even count on 1.7/1.5% and 0.7/0.4% of all official B04 rows, against 35.2/37.7% for G.L and 58.2/58.2% for Base. *Information without function:* G.V alone orders gold 6 below gold 7 better than E (difference in the area under the curve of $\mathrm{LL}(7) - \mathrm{LL}(6)$ over all official rows with gold 6 or 7: +0.25 [+0.08, +0.42] and +0.18 [+0.05, +0.31]), than Base (+0.27 [+0.09, +0.46] and +0.25 [+0.06, +0.45]), and even than G (+0.10 [+0.02, +0.20] and +0.06 [+0.01, +0.11]), yet with Base's L it repairs only 36/25% of the even items. Which module of G alone repairs more of the even items differs between the seeds (G.V 36/25%, G.L 15/41%). *The even items are reproducible across seeds:* the other seed's VL repairs 52/89% of them against 6/17% of its other even-count Base errors. *Formation:* at the stored checkpoints, G (all of VL) repairs 73/100% of the even core by mid-training, while neither module of G alone repairs more than 20% of it at any stored update and no single set (A, E, F) repairs any even-core item at any stored update (0/0%); the modules that carry the even items alone at intermediate updates differ between the seeds, so only the end point of the formation replicates. *Location in L* (weight-level block transplants between G.L and F.L, with G.V fixed): F.L with G.L's tied embedding and output head repairs 3/0% of the even items and G.L with F.L's keeps 100/95%, so the embedding carries none of the even repair; F.L with G.L's layers 12–23 restores 51/61% and with its layers 0–11 22/14%, and no single block of G.L restores the repair fully: the joint L change is spread over decoder layers, more of it in the upper half.

Reading. The B04 VL pair-emergent items contain two regimes. The odd-count items are a complementary distributed repair that recombining single-trained modules already reaches: globally new at the level of trainable sets, but not a new capability at the module level. The even-count items need both modules changed jointly, yet not by the same run: partners from the other seed's joint run work, while every partner trained without the other module fails and a single-trained L is a worse partner for G.V than Base's L. In our operationalization this is H3 (distributed and reproducible), or co-adaptation to a class of partners rather than to a private one; either way the synergy of the analysed VL pair-emergent repairs mostly does not require run-private code. This does not extend to PL, VP, or VPL, for which no such recombinations exist. Limits: one case, chosen by a rule, with strata defined after the first exploration; few scene families per count, so intervals are wide; module compositions show which modules must change together, not that two compositions compute the same way; there is no activation-level patch or probe on B04; alternative strategies and shortcuts are not ruled out.

Table 53: Module compositions on the B04 VL pair-emergent items (Appendix P.3): share (%) of the selected items repaired, by gold-count parity. *Even core*: even items that neither module of G repairs alone. X.M: module M of run X (A=P, B=PL, C=VPL, D=VP, E=V, F=L, G=VL); a prime marks the other seed's run; unlisted modules are Base.

	seed 1101			seed 1104		
composition	odd	even	even core	odd	even	even core
G.V	1	36	0	8	25	0
G.L	0	15	0	0	41	0
E.V+F.L	82	0	0	73	0	0
G.V+F.L	100	3	3	98	0	0
E.V+G.L	53	0	0	59	0	0
G.V+B.L	100	1	3	99	2	6
D.V+G.L	32	0	0	46	0	0
G.V+C.L	97	82	86	98	43	59
C.V+G.L	91	64	51	85	86	94
G.V+G′.L	100	97	100	92	77	100
G′.V+G.L	99	61	68	91	91	100
E.V+F′.L	81	0	0	75	0	0
E′.V+F.L	82	0	0	75	0	0
C.V+C.L	95	49	38	98	64	82
G	100	100	100	100	100	100

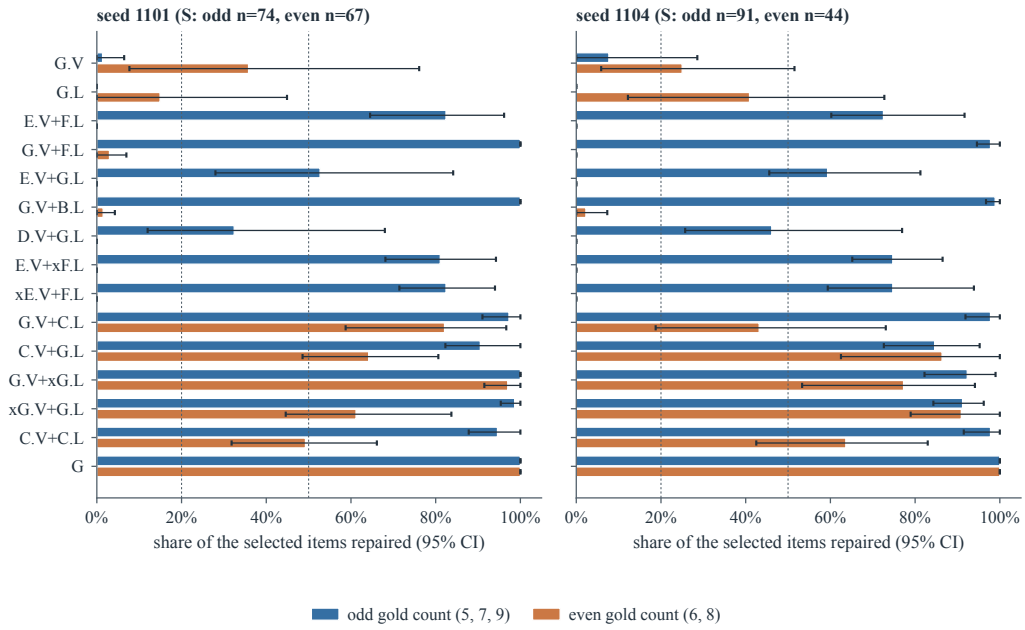

Figure 22: Module compositions on the B04 VL pair-emergent items by gold-count parity, with 95% scene-family intervals (Appendix P.3); an x prefix marks the other seed's run (a prime in Table 53). Dashed lines: the pre-set thresholds for a composition that fails (20%) or works (50%).

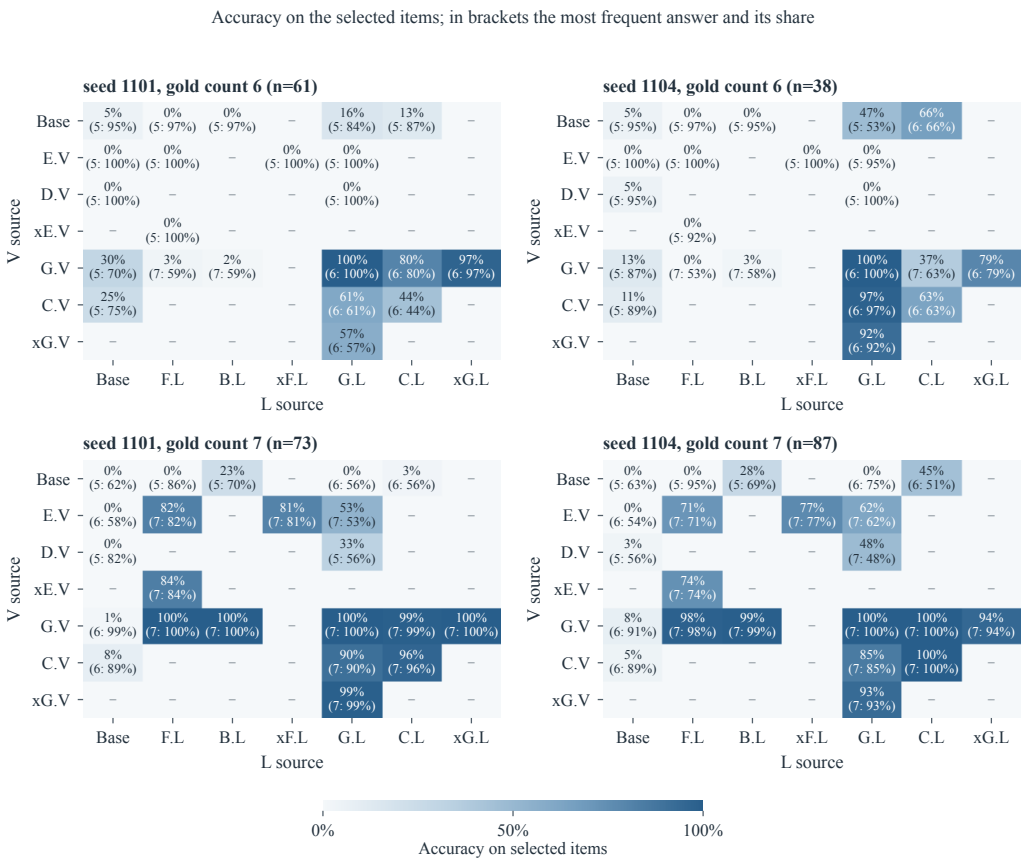

Figure 23: V source × L source on the B04 VL pair-emergent items with gold count 6 (top) and 7 (bottom), P from Base: accuracy and, in brackets, the most frequent answer and its share (Appendix P.3).

P.4 Single-task control on B04

Question. Is the VL pair-emergence on B04 a conflict between tasks in 13-task training, such that V or L trained on B04 alone would repair the items (candidate family E), rather than an interaction between V and L?

Design. In each seed, V with the recipe of E and L with the recipe of F were retrained on B04's training rows only, for as many B04 exposures as the 13-task runs give B04 (6,000 exposures in 189 updates), and scored on B04's official rows as sV (trained V, Base P and L), sL (Base V and P, trained L), and sV+sL. The rule was fixed before scoring: on the items S of Appendix P.3, the verdict is *conflict* if the better single-task module repairs at least 50%; *consistent with synergy* if it repairs at most 20% and a training guard holds (the single-task modules reproduce the B04 repairs of E and F); *inconclusive* if it repairs at most 20% without the guard; and *partial* otherwise.

Result (analysis status: complete; Table 54). On S, sV repairs 7% [1, 19] and sL 0% [0, 0] in seed 1101, and 5% [1, 13] and 7% [0, 16] in seed 1104; sV+sL repairs 7/20% (E.V+F.L of the 13-task runs: 43/49%). The guard: sV reproduces 50/67% of E's B04 repairs and sL 56/74% of F's. sL answers an even count on 1.6/0.0% of B04's official rows (Base 58.2/58.2%). Verdict: consistent with synergy in seed 1101 (odd items: consistent with synergy; even items: consistent with synergy) and consistent with synergy in seed 1104 (odd: consistent with synergy; even: consistent with synergy).

Reading. Seed 1101: neither module trained on B04 alone repairs these items although both pass the training guard, so their pair-emergence is not explained by multi-task training. Seed 1104: neither module trained on B04 alone repairs these items although both pass the training guard, so their pair-

emergence is not explained by multi-task training. The single-task L run fits B04's training rows almost perfectly (training loss near zero), so memorization of the training rows cannot be excluded; the official rows are held out but not counterfactual. The Adam step budget is not matched: the single-task runs take 189 updates, the 13-task runs 2,439, so a single-task failure may also be a budget false negative.

Table 54: Single-task control on B04 (Appendix P.4; status: complete): repair rate (%) of Base, sV, sL, and sV+sL on the VL pair-emergent items S and its gold-count strata, on the B04 repairs of E and F (training guard), and on B04's unrepaired items; last row: B04 accuracy (%).

seed	set	Base	sV	sL	sV+sL
1101	VL-only items S (%)	2	7	0	7
1101	odd gold	0	5	0	14
1101	even gold	4	9	0	0
1101	arm E's B04 repairs	5	50	39	48
1101	arm F's B04 repairs	1	51	56	73
1101	unrepaired	0	1	0	16
1101	B04 accuracy	29	31	29	39
1104	VL-only items S (%)	1	5	7	20
1104	odd gold	0	3	10	30
1104	even gold	5	9	0	0
1104	arm E's B04 repairs	8	67	77	81
1104	arm F's B04 repairs	1	51	74	82
1104	unrepaired	1	4	0	10
1104	B04 accuracy	29	33	35	42

P.5 Is multiplicity generic? Restricted, truncated, and random updates

Question. Would any subset of a module's parameters, any low-rank part of its update, or any perturbation of the same size repair the same failures? If so, multiplicity would be generic sensitivity of the failure; if not, it is a property of the trained updates (candidates D1, degeneracy of the trained solution, and A4, a dormant capability reached by a small update; and the artifact that items near the boundary flip under any perturbation).

Design. Each patch is built at load time from Base and one single-set update d (E for V, F for L, A for P; final update). *Restricted:* F's update on language-model blocks 0–11 only or 12–23 only, and E's on encoder layers 0–13 only or 14–26 only; everything else stays Base, including L's tied embedding and final norm (studied in Appendix P.6). *Truncated:* the best rank-k approximation of d for every matrix, $k \in \{1, 4, 16\}$, with vectors (norms, biases) kept; $k = 0$ keeps the vectors only. *Random:* per tensor an isotropic Gaussian direction scaled to that tensor's $\|d\|$, three draws per module. Rows are the official Base errors of G02 (all 244), B02 (200), B07 (200), and B04 (207; the last three capped by whole scene families). *Retention* is the share of the full single set's repairs that the patch also makes, and *release* the share of the errors that the full set does not repair that the patch repairs. Through the same path the full updates reproduce E, F, and A bitwise on G02. Rules fixed before scoring: a half *keeps most* if its retention exceeds 50%; a random draw *releases* multiple-single items if it repairs at least 10% of them and more of them than of the other items. Scoring status: complete.

Result (Table 55). *Within L, several parameter subsets carry the G02 repair.* F's update on blocks 0–11 keeps 100 [100, 100] and 100 [99, 100]% of F's G02 repairs and on blocks 12–23 67 [61, 74] and 66 [59, 72]%: both halves keep most, the upper half fewer. On B02 only the lower half keeps most (78/75% against 6/8%). *The trained direction matters, not the norm:* none of the three norm-matched random L directions repairs any of G02's Base errors (repair rate 0/0%, 0/0%, and 0/0%) or releases multiple-single items, and F's vector updates alone repair none (0/0%). *Low rank suffices, and multiple-single repairs are the most robust:* rank 4 keeps 87 [82, 91] and 87 [83, 91]% of F's G02 repairs and rank 1 27 [21, 33] and 23 [18, 28]%; at rank 4, F's multiple-single repairs survive more often than its single-sole ones (91/91% against 33/35%). Rank 4 also keeps most of F's repairs on B02 (70/60%) and pooled over the four tasks (70 [61, 79] and 67 [58, 77]%); on B07 and B04, where F repairs few items, no tested rank keeps most. *V is less degenerate on G02:* neither half of E's update keeps most of E's G02 repairs (layers 0–13 5 [2, 8] and 5 [2, 8]%, layers 14–26 33

[27, 39] and 27 [21, 34]%), rank 16 keeps 93 [89, 96] and 89 [85, 94]% and rank 1 11 [8, 16] and 12 [8, 17]%, and a norm-matched random V direction (first draw) repairs none (0/0%). On B02 each half of E's update keeps most of E's repairs and rank 1 already does (67/58%). By the pre-set rule, G02 reads as "generic degeneracy" in both seeds, through the L halves; the alternative that only the upper L half carries the repair does not hold. On B02 the rule classifies the random L directions as "releases MS items (3/3 draws)" in seed 1101 and "does not release MS items (0/3 draws)" in seed 1104; we do not interpret this difference. The P rows and the V rows at rank 4, at $k = 0$, and for random draws 2–3 are reported without interpretation.

Reading. Within the language model, several parameter subsets (either block half, a rank-4 truncation) carry the same G02 repair, while random directions of the same norm carry none. On G02 the multiplicity is therefore degeneracy of the trained solution, in where inside L the update sits and how few directions it needs, not generic sensitivity of the failure to any perturbation (D1 supported for L on G02; A4 partial, since low rank does not show that the capability was dormant in Base). F's G02 repair acts through the upper layers' query side (Appendix N), yet its update on blocks 0–11 alone keeps it: another case where the location of the parameter change differs from the location of the changed computation. Limits: rows are Base errors only, so breaks are not measured; B02, B07, and B04 are subsampled by scene family.

Table 55: Restricted, truncated, and random single-set updates (Appendix P.5): retention (%) of the full single set's repairs pooled over G02, B02, B07, and B04 (ret.) and on G02 (ret. G02); release (%) of the errors the full set does not repair; repair rate (%) on multiple-single items (MS rate). L: F's update, V: E's, P: A's.

	seed 1101				seed 1104			
patch	ret.	ret. G02	release	MS rate	ret.	ret. G02	release	MS rate
L blocks 0–11	85	100	9	89	81	100	3	89
L blocks 12–23	42	67	1	56	45	66	0	53
V layers 0–13	31	5	5	24	31	5	3	25
V layers 14–26	44	33	6	41	36	27	5	34
L vectors only	2	0	1	2	2	0	1	3
L rank 1	24	27	0	31	19	23	0	22
L rank 4	70	87	4	83	67	87	1	78
L rank 16	83	98	6	92	81	97	11	89
V vectors only	3	0	0	2	3	0	0	3
V rank 1	28	11	3	24	27	12	3	23
V rank 4	68	73	2	68	64	73	4	65
V rank 16	84	93	1	82	83	89	1	80
P vectors only	0	0	1	2	2	0	1	3
P rank 1	49	46	9	27	49	45	10	27
P rank 4	77	72	6	40	77	72	6	39
P rank 16	94	94	3	51	94	91	3	49
L random 1	2	0	1	3	1	0	0	2
L random 2	2	0	0	3	3	0	3	3
L random 3	4	0	2	6	2	0	1	2
V random 1	3	0	0	2	3	0	0	2
V random 2	3	0	0	3	3	0	0	2
V random 3	3	0	0	3	2	0	0	2
P random 1	1	0	1	3	2	0	1	2
P random 2	2	0	1	2	2	0	1	2
P random 3	1	0	1	3	3	0	1	3

P.6 The tied embedding: a readout handle only L has?

Question. In the language model the output head shares its matrix with the input embedding, so training L can move the answer-token readout directly, a handle that V and P lack. Does F's repair go through this handle or through its transformer blocks?

Design. The tie is broken at load time, and all eight models that take each of the input embedding, the output head, and the blocks (the decoder layers and final norm) from Base or from F are scored; the Base and F cells reproduce the original models bitwise on G02 and B05, with the clean and the

blank image. Rows are official Base errors; the analysis is complete (13 of 13 tasks in both seeds), and the pre-set summaries below are on the core tasks G01, G02, G03, B02, B05, B07 or on one of them. Metrics: repair rate, retention of F's repairs, the blank-image advantage over Base on the same rows, and factorial effects on the repair rate. The rule fixed before scoring supports the handle hypothesis only if the head alone repairs a prior-type (image-independent) subset *and* F without its head loses at least half of its blank-image advantage.

Result (Table 56). F's blocks alone, with Base's embedding and head, keep 100 [99, 100] and 100 [100, 100]% of F's repairs; the blocks' main effect on the repair rate is +75.3 [+72.7, +77.6] and +73.0 [+70.4, +75.6] pp, the head's +2.1 [+1.7, +2.5] and +1.2 [+0.9, +1.5] pp, and the input embedding's rounds to +0.2/+0.1 pp. F's head alone repairs 4 [4, 5] and 3 [2, 4]% of the core Base errors, most of them image-independent; on B02 it repairs 30 [27, 33] and 20 [16, 24]%, 95/100% of them with the same answer under the blank image, and raises blank-image accuracy on B02's Base errors by +42/+42 pp, whereas on G02 it repairs 0/0%. The head–blocks interaction is negative (-1.9 [-2.3, -1.5] and -1.2 [-1.5, -0.8] pp): the head's repairs are redundant with the blocks'. F with Base's head keeps its blank-image advantage over Base (+33/+33 pp against +32/+33 pp for F; on G02 +64/+61 pp), so the pre-set rule for the handle hypothesis fails in both seeds.

Reading. The tied embedding is a real but redundant L-only readout handle: on its own it shifts the answer prior (on B02 a prior-type subset of repairs), but F's repairs, including its image-independent blank-image behaviour, are carried by the transformer blocks. As a driver of L's repairs the handle is not supported; that it exists is partly supported. The final norm is counted with the blocks.

Table 56: Tied-embedding factorial on the core tasks G01, G02, G03, B02, B05, B07 (Appendix P.6): source (B: Base, F: L) of the input embedding, the output head, and the blocks; repair rate (%) of the Base errors, retention (%) of F's repairs, and blank-image advantage over Base (pp).

source			seed 1101			seed 1104		
emb	head	blocks	repair	ret. F	blank adv.	repair	ret. F	blank adv.
B	B	B	0	0	+0	0	0	+0
B	B	F	78	100	+33	75	100	+33
B	F	B	4	4	+5	3	2	+5
B	F	F	78	100	+32	75	100	+33
F	B	B	1	1	+0	1	1	+0
F	B	F	77	99	+33	75	100	+33
F	F	B	4	4	+5	3	2	+5
F	F	F	78	100	+32	75	100	+33

P.7 Summary: repair regimes and evidence status

Table 57 summarizes what these analyses say about each repair regime of RQ1, with a status per claim: *supported* (holds in both seeds under a pre-set rule or claim check), *partial* (holds in both seeds but leaves the named alternative open), or *unresolved*. Alternative strategies and shortcuts are not ruled out for any regime, since no counterfactual cue intervention was run.

Table 57: Repair regimes and the function/parameter-space evidence of Appendix P, with status.

regime	evidence	status
multiple-single	mostly a jointly trained module still repairs alone after joint training; within L the G02 repair survives either block half and rank 4, random directions of the same norm repair nothing (degeneracy of the trained solution, D1)	supported (G02; L)
multiple-single	margin interaction negative in PL, positive in VP: redundancy is set-specific; saturation confounds a negative interaction	partial
pair-emergent	sub-threshold additive repairs are rare; the rest are carried by one jointly trained module (PL, VP: emergence $\neq$ ownership) or superadditive (VL)	supported
pair-emergent, VL	synergy transfers to independently trained partners; on B04 odd counts compose from single-trained modules, even counts need a jointly trained but not run-private partner	supported (VL; one case study for B04)
pair-emergent, PL, VP, VPL	co-adaptation and multi-stage computation not separable (no recombination with single-trained partners)	unresolved
triple-emergent	C's V and L without its P repair almost all; three-way term negligible: a V$\times$L interaction	supported

Table 57: Repair regimes and the function/parameter-space evidence of Appendix P, with status. (continued)

regime	evidence	status
unrepaired	adaptation moves most items away from gold; Base margins close to other classes; L has first-order leverage everywhere; not observed under the current adaptation procedure	partial
local geometry (D2)	the first-order test at Base does not separate the classes	partial (largely negative)
tied embedding	a redundant readout handle; F's repairs live in its blocks	against as driver

Q REPAIR LOSS LANDSCAPE

The landscape of Section 4.3 scores $\theta(\alpha, \beta) = \theta_0 + \alpha\,\Delta V + \beta\,\Delta L$ at $\alpha, \beta \in \{0, 0.25, \ldots, 1.25\}$ (36 points per plane) on the official rows of G02, B02, and B07 in two planes, EF with ΔV of V and ΔL of L, and GG with both updates of VL; P stays at Base, and the corners reproduce the models of Section 4.2 item by item. Figures 24 and 25 give the accuracy change over Base and Tables 58 and 59 the corners and path summaries for seeds 1101 and 1104. Values below are seed 1101/seed 1104. On G02, 33/33 of the 36 points of the EF plane keep at least half the best gain. On B02 the V-only and L-only repairs are also subadditive (-6.5 [-9.2, -3.9] and -5.5 [-8.6, -2.6] pp) with little or no accuracy dip between them (0.0/+0.5 pp), whereas inside VL the two updates are superadditive: ΔL_{VL} alone changes accuracy by -0.2 [-1.6, +1.3] and +2.4 [+0.7, +4.4] pp, both together by +12.2/+16.1 pp (additivity +6.8 [+4.4, +9.2] and +3.1 [+0.6, +5.7] pp). So in seed 1101 the B02 repair of VL needs both of its updates, and in seed 1104 ΔL_{VL} alone makes a small part of it. We do not interpret B07: its Base accuracy is 2.2% and no grid point exceeds 33.2/32.6%.

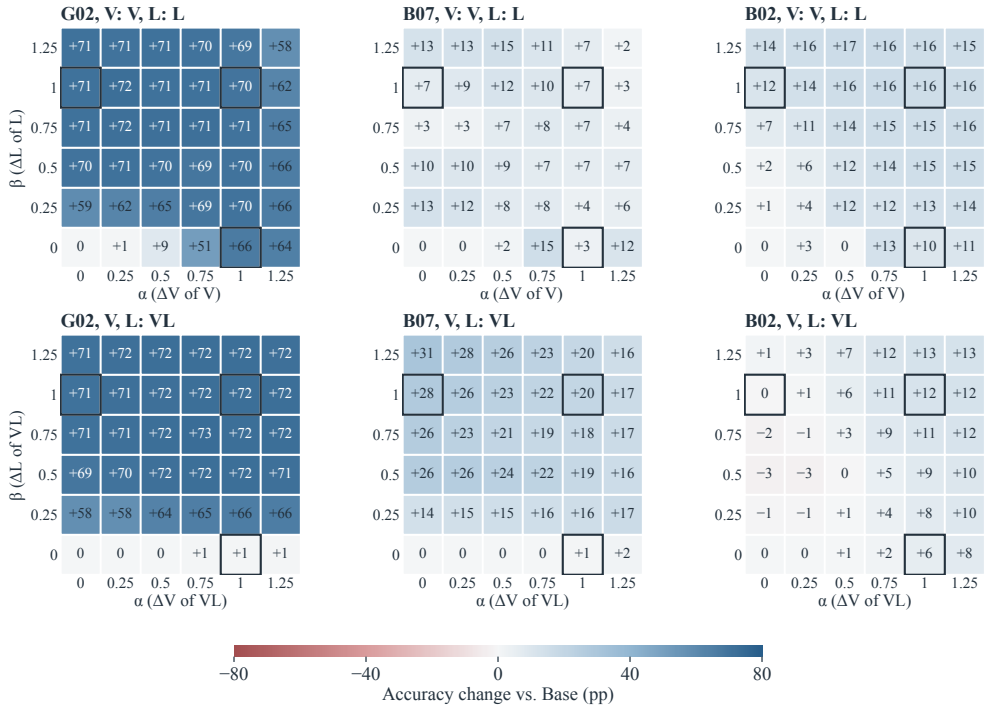

Figure 24: Accuracy change (pp) over Base along $\theta(\alpha, \beta) = \theta_0 + \alpha\,\Delta V + \beta\,\Delta L$ (seed 1101). Top: ΔV of V, ΔL of L; bottom: both of VL. Boxed: (1,0), (0,1), (1,1).

Table 58: Repair loss landscape $\theta(\alpha, \beta) = \theta_0 + \alpha\,\Delta V + \beta\,\Delta L$ (seed 1101). EF: ΔV of V and ΔL of L; GG: both of VL. Corners (0,0), (1,0), (0,1), (1,1): scene-first accuracy (%) and mean gold-answer NLL (nats). Dip: lower corner accuracy minus the minimum on $\alpha + \beta = 1$ (pp); additivity: gain(1,1) − gain(1,0) − gain(0,1) (pp); 95% scene-family intervals.

		Accuracy %				Gold NLL					
Grid	Task	(0,0)	(1,0)	(0,1)	(1,1)	(0,0)	(1,0)	(0,1)	(1,1)	Dip (pp)	Additivity (pp)
EF	G02	24.4	90.7	95.5	94.1	5.77	0.26	0.17	0.22	0.0 [0.0, 0.0]	−67.6 [−74.0, −61.3]
	B07	2.2	5.6	8.8	8.9	3.55	2.81	13.29	17.71	+0.3 [0.0, +2.5]	−3.2 [−6.4, +0.1]
	B02	78.4	88.8	90.2	94.1	0.53	0.45	1.99	8.44	0.0 [0.0, +0.7]	−6.5 [−9.2, −3.9]
GG	G02	24.4	25.2	95.2	96.6	5.77	5.24	0.20	0.13	0.0 [0.0, 0.0]	+0.5 [−1.7, +2.8]
	B07	2.2	2.9	29.7	22.4	3.55	3.08	7.95	8.87	0.0 [0.0, 0.0]	−8.1 [−11.8, −4.4]
	B02	78.4	84.1	78.2	90.6	0.53	0.40	2.48	2.11	+0.7 [+0.1, +1.7]	+6.8 [+4.4, +9.2]

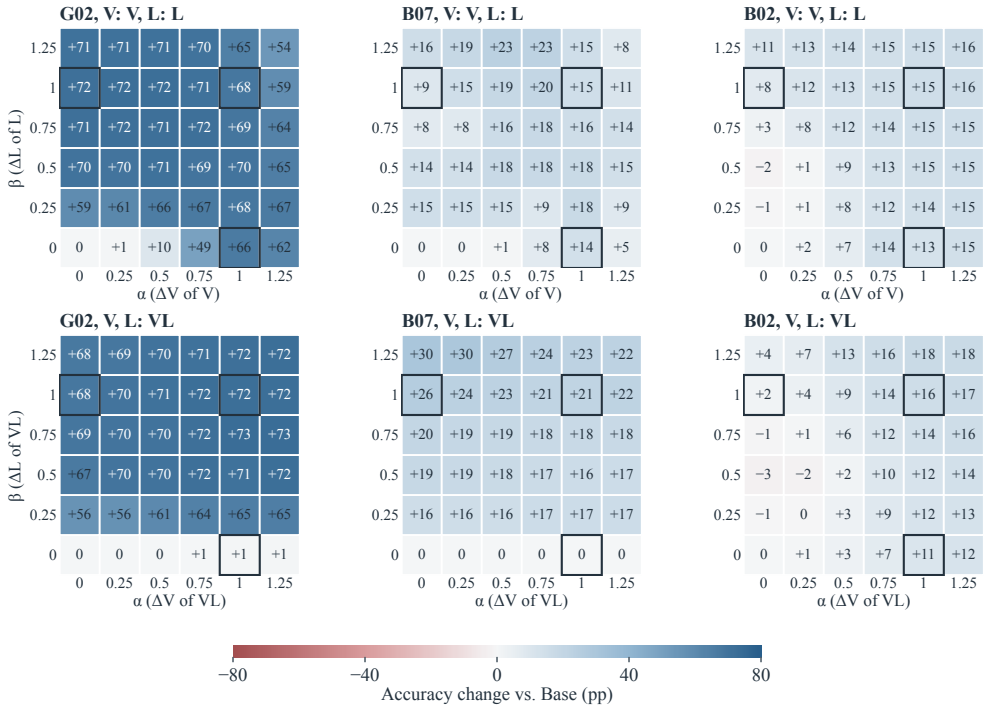

Figure 25: Accuracy change (pp) over Base along $\theta(\alpha, \beta) = \theta_0 + \alpha\,\Delta V + \beta\,\Delta L$ (seed 1104). Top: ΔV of V, ΔL of L; bottom: both of VL. Boxed: (1,0), (0,1), (1,1).

Table 59: Repair loss landscape $\theta(\alpha, \beta) = \theta_0 + \alpha\,\Delta V + \beta\,\Delta L$ (seed 1104). EF: ΔV of V and ΔL of L; GG: both of VL. Corners (0,0), (1,0), (0,1), (1,1): scene-first accuracy (%) and mean gold-answer NLL (nats). Dip: lower corner accuracy minus the minimum on $\alpha + \beta = 1$ (pp); additivity: gain(1,1) − gain(1,0) − gain(0,1) (pp); 95% scene-family intervals.

		Accuracy %				Gold NLL					
Grid	Task	(0,0)	(1,0)	(0,1)	(1,1)	(0,0)	(1,0)	(0,1)	(1,1)	Dip (pp)	Additivity (pp)
EF	G02	24.4	90.7	96.1	92.5	5.77	0.30	0.16	0.29	0.0 [0.0, +2.3]	−69.9 [−76.3, −63.0]
	B07	2.2	16.0	11.7	17.5	3.55	2.39	10.91	12.26	+1.2 [0.0, +4.9]	−7.9 [−13.2, −2.9]
	B02	78.4	91.2	86.5	93.8	0.53	0.29	2.93	7.42	+0.5 [0.0, +1.3]	−5.5 [−8.6, −2.6]
GG	G02	24.4	25.2	92.8	96.4	5.77	5.18	0.27	0.14	0.0 [0.0, 0.0]	+2.8 [+0.5, +5.5]
	B07	2.2	2.2	28.1	23.5	3.55	3.10	7.76	8.12	0.0 [0.0, 0.0]	−4.6 [−8.2, −1.0]
	B02	78.4	89.1	80.8	94.6	0.53	0.38	1.66	2.23	+1.4 [+0.6, +2.4]	+3.1 [+0.6, +5.7]