DIAGNOSING TEMPORAL INSTANCE-STATE BINDING AND ITS STAGE-DEPENDENT REPAIRABILITY IN VISION-LANGUAGE MODELS: A CONTROLLED STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal visual understanding requires more than detecting that a state change occurred. A model must also identify which persistent instance underwent that change. We introduce **TIB-Bench**, a controlled synthetic benchmark designed to isolate event recognition from instance-state binding. Its counterfactual pairs end in pixel-identical final frames, but differ in history and in the correct event owner. On Qwen3-VL-2B, perception and event occurrence are both correct on 210 of 480 episodes. Yet the owner is wrong in 109 of those cases (51.9%), revealing a failure that event-level accuracy does not capture. Layerwise probing shows that owner-related information remains accessible in downstream representations, while late-layer activation interventions perturb binding decisions substantially more than event judgments. We then examine how this failure can be repaired. Replacing matched ordinary examples with counterfactual-generated examples raises counterfactual-history consistency from 0.006 to 0.850 without an explicit pairing objective. Under matched eight-layer LoRA budgets, early and middle adaptation also substantially outperform late adaptation. Temporal-order perturbations further show that repaired models rely more strongly on the canonical history. Together, these results separate event recognition, owner-related representation, and history-sensitive owner attribution within a controlled temporal visual setting.

1 INTRODUCTION

Understanding a visual sequence requires more than recognizing the objects and events that appear in it. A model must also maintain which persistent object each event belongs to as the scene evolves. This distinction becomes especially important when several visually similar instances move, disappear, reappear, or change state. Recognizing that a cup moved, for example, does not necessarily imply knowing which cup moved. A model may identify the event correctly while assigning it to the wrong persistent instance, yet this failure can remain hidden when evaluation focuses primarily on event recognition.

We study this failure through *temporal instance-state binding* (TIB): whether a recognized state change is assigned to the persistent object that actually underwent it. The central question is deliberately narrower than general video understanding: when the event itself is recognized, can its owner still be wrong? To isolate this relation, we introduce TIB-Bench, a controlled synthetic benchmark built from counterfactual histories. Each counterfactual pair reaches a pixel-identical final frame but follows a different preceding history and therefore has a different correct event owner. The construction holds the final visual observation fixed while changing the history-dependent event-to-owner relation, requiring the model to recover ownership from temporal evidence.

This controlled setting reveals a substantial recognition–binding dissociation. On Qwen3-VL-2B, perception and event occurrence are both correct on 210 of 480 episodes, yet the model assigns the change to the wrong owner in 109 of those cases (51.9

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.

We next ask whether the behavior can be repaired. Adding ordinary supervision improves pointwise binding while leaving counterfactual-history consistency near zero, whereas replacing matched ordinary examples with counterfactual-generated examples raises this consistency from 0.006 to 0.850 without an explicit pairing objective. Repair also depends strongly on adaptation placement: under matched eight-layer LoRA budgets, early and middle adaptation substantially outperform late adaptation even though owner labels are most strongly decodable downstream. Together, these results separate event recognition, owner-related representation, and history-sensitive owner attribution as distinct components of temporal visual understanding.

2 Related Work

Object persistence, state change, and binding. Maintaining object identity across changing visual observations is central to both object tracking and state-change understanding. Prior work studies objects through occlusion and containment, state-changing object localization, and open-world object-state transitions (Soucek et al., 2022; Van Hoorick et al., 2023; Yu et al., 2023; Xue et al., 2024). A related distinction between the availability of visual information and the correctness of its association has long been studied in human vision: illusory-conjunction experiments show that visual elements can be individually available while incorrectly bound (Treisman & Schmidt, 1982), while object-file theories emphasize maintaining persistent identity as visual states evolve (Kahneman et al., 1992). TIB builds on this distinction but asks specifically whether a VLM can recognize a temporal event while assigning it to the wrong persistent instance.

Temporal and compositional video reasoning. Existing benchmarks evaluate causal and compositional reasoning over videos (Yi et al., 2020; Grunde-McLaughlin et al., 2021; Xiao et al., 2021), broad temporal understanding in multimodal models (Li et al., 2024; Chen et al., 2024; Fu et al., 2025; Hong et al., 2025), and explicit agent–action association across multiple events (Saravanan et al., 2025). These benchmarks establish substantial weaknesses in temporal and multi-event reasoning, but typically evaluate settings in which event evidence, object identity, and event ownership vary together. TIB-Bench is complementary: its counterfactual construction holds the final visual observation fixed while changing the history-dependent owner, allowing event recognition and event-to-instance attribution to be evaluated separately.

Counterfactual supervision and internal representations. Our repair experiments are related to work showing that counterfactually augmented or bias-conflicting examples can reduce reliance on spurious correlations (Kaushik et al., 2020; 2021; Nam et al., 2020; Liu et al., 2021). We use counterfactual-generated coverage for a different purpose: exposing the model to histories in which similar final observations correspond to different event owners, without requiring an explicit pairwise objective. Our representation analysis is likewise diagnostic rather than a localization claim: linear probing measures the accessibility of owner-related information (Hewitt & Manning, 2019; Voita & Titov, 2020), while activation interventions follow prior work that exchanges internal states to test their influence on downstream predictions (Meng et al., 2022; Wang et al., 2023).

3 Controlled Isolation of Temporal Instance-State Binding

3.1 Counterfactual Construction

Temporal instance-state binding is difficult to study when event recognition and event ownership change together. We therefore construct a controlled benchmark that isolates the relation between a state change and the persistent object that underwent it. TIB-Bench contains 480 storyboard episodes, including 240 independent episodes and 120 counterfactual pairs. Each episode contains three timesteps and three persistent objects, and covers one of four event families: MOVE, ATTRIBUTE_CHANGE, STATE_TOGGLE, and OCCLUDE_REAPPEAR.

We formalize each episode as a visual history $x = (I_1, \ldots, I_T)$ containing a set of persistent instances $\mathcal{O} = \{o_1, \ldots, o_K\}$ and a target event e. Temporal instance-state binding asks for the owner $y \in \mathcal{O}$ that underwent e. A counterfactual pair (x, x') preserves the final visual observation, $I_T = I'_T$, while changing the preceding history such that the correct owners differ, $y \neq y'$.

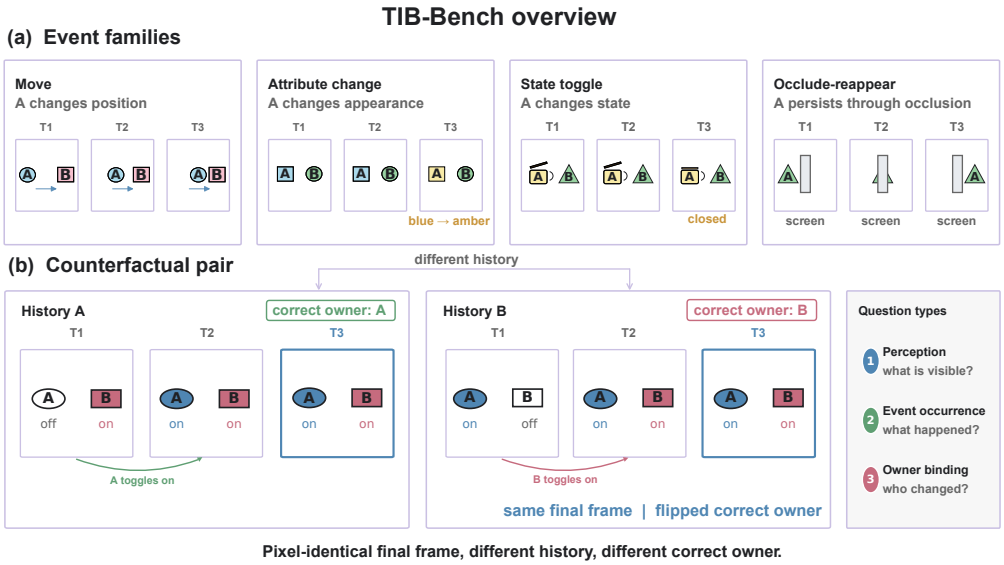


Figure 1: **TIB-Bench isolates temporal instance-state binding through controlled histories.** (a) Four event families require tracking a persistent instance across three time steps. (b) Each counterfactual pair follows a different history but converges to a pixel-identical final frame; the correct event owner therefore changes while final visual evidence is held fixed. (c) Separate perception, event-occurrence, and owner-binding questions allow event recognition to be distinguished from event-to-instance attribution.

This construction removes the final frame as a sufficient cue: any rule that ignores the preceding history must produce the same evidence for both members even though the correct answer changes.

All 120 counterfactual pairs pass this audit. The construction also separates counterfactual control from visual difficulty. Objects retain persistent identities across the three panels, while event type, object count, identity similarity, and answer position are explicitly controlled during generation. Counterfactual partners share a pixel-identical final visual configuration but differ in the preceding visual history and therefore in the correct event owner. Answer options are independently randomized by the evaluation interface. The construction therefore controls the final visual observation while requiring the model to recover owner identity from temporal evidence.

3.2 Tasks and Evaluation

Each episode is paired with questions that separately target perception, event occurrence, and owner binding. The perception question checks whether the model reads the visible scene correctly. The event question asks whether the target state change occurred and serves as an event marginal. The binding question asks which object underwent that change. This separation allows us to distinguish failures of event recognition from failures of owner attribution.

We report three main metrics. Raw binding accuracy is the accuracy of the binding question over all episodes. Event marginal accuracy measures whether the model correctly identifies the target event. Conditional binding accuracy evaluates the binding question only on episodes for which both perception and the event marginal are correct. We also report counterfactual-history consistency (CFC), defined over the 120 counterfactual pairs as the fraction of pairs for which both members receive the correct owner after the history, and therefore the correct owner, is changed. CFC is a stringent pair-level measure of whether owner attribution remains correct across counterfactual histories. CFC is therefore used as a stringent pair-level success criterion rather than as a capability dimension independent of pointwise binding accuracy.

Let p_i, e_i, and b_i denote correctness of perception, event recognition, and owner binding for episode i. Conditional binding is defined as

$$\mathrm{CondBind} = \frac{\sum_i \mathbb{1}[p_i = 1, e_i = 1, b_i = 1]}{\sum_i \mathbb{1}[p_i = 1, e_i = 1]}, \tag{1}$$

so recognition failures cannot inflate the measured binding error. For the M counterfactual pairs, we define counterfactual-history consistency as

$$\mathrm{CFC} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}[b_j = 1 \wedge b'_j = 1], \tag{2}$$

where b_j and b'_j denote binding correctness for the two members of pair j. CFC is intentionally stricter than pointwise accuracy: a model receives credit only when it assigns the correct owner under both histories. A model that performs well by exploiting a history-insensitive shortcut can therefore obtain reasonable single-episode accuracy while still achieving low CFC.

The benchmark also reduces several obvious static and positional shortcuts. Gold answer positions are balanced across the four options, and leftmost or rightmost heuristics remain at chance. In an auxiliary generation-harness control, a final-frame-only input preserves some event information but reduces conditional binding accuracy to 0.081. This provides additional evidence that owner attribution in TIB-Bench depends on temporal history rather than on the final visual state alone. Additional shortcut checks and auxiliary controls are reported in the appendix.

4 When Event Recognition Is Not Enough

4.1 A Binding Gap Remains After Event Recognition

The central failure appears even after the measured perceptual and event-level controls are correct. On Qwen3-VL-2B, both perception and event occurrence are correct on 210 of 480 episodes. Yet the model assigns the change to the wrong owner in 109 of those 210 cases, corresponding to an error rate of 51.9% (95% CI: 44.8–58.6%). Equivalently, 109 of all 480 episodes (22.7%) exhibit the joint pattern of correct perception, correct event recognition, and incorrect owner attribution. The effect appears across all four event families and is dominated by identity-swap errors rather than random guesses, suggesting that the model often recognizes the event while attaching it to the wrong persistent instance.

Conditioning on correct perception and event recognition is important because an incorrect owner answer is otherwise ambiguous. If the model failed to detect the event or misread the scene, the subsequent owner error provides little evidence about binding itself. The conditional analysis removes these cases and asks a stricter question: after the model has already demonstrated access to the relevant perceptual and event-level evidence, does it associate that evidence with the correct persistent instance? Passing these controls does not certify complete frame-level identity correspondence or trajectory recovery; the analysis isolates owner errors after the measured perceptual and event-level checks rather than uniquely localising them to a dedicated binding mechanism. The 51.9% conditional owner error shows that these capabilities can fail separately within the same episode.

The error structure is also inconsistent with a purely random failure mode. Conditional errors occur across all four event families and are dominated by assignments to another plausible instance in the same scene rather than arbitrary responses. We refer to these cases as identity-swap errors. Their prevalence motivates treating the failure as an association problem: the model frequently identifies the relevant change but transfers it to the wrong persistent object. Event-wise breakdowns and the full error taxonomy are provided in the appendix.

The same separation between event recognition and owner attribution is not limited to the smallest model. Qwen3-VL-4B and Qwen3-VL-8B reduce the conditional owner error, but do not eliminate it. InternVL3.5-4B provides a stronger example: its conditional binding accuracy reaches 0.858, while its counterfactual-history consistency is only 0.458. Thus, strong pointwise binding does not guarantee stable owner attribution across counterfactual histories. The same recognition–binding dissociation appears in an automatic real-video diagnostic from MOT17, MOT20, and KITTI (Appendix A.7).

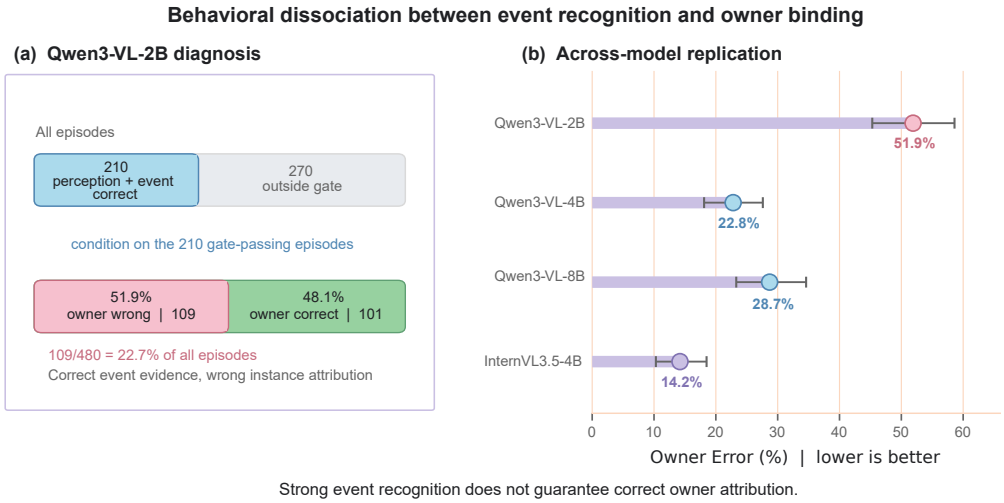


Figure 2: **Event recognition does not guarantee correct owner binding.** (a) Qwen3-VL-2B correctly answers both perception and event occurrence on 210 of 480 episodes, yet assigns the event to the wrong owner in 109 of those cases (51.9%). (b) Conditional owner errors remain across stronger Qwen3-VL variants and InternVL3.5-4B. Error bars show 95% episode-bootstrap confidence intervals.

Table 1: Cross-model diagnosis on TIB-Bench. Conditional binding is evaluated only when perception and event recognition are both correct. Owner Error is $P(\text{wrong owner} \mid \text{perception and event correct})$. CFC measures pair-level correctness across counterfactual histories.

Model	Conditional Binding	Owner Error ↓	CFC ↑
Qwen3-VL-2B	0.481	0.519	0.025
Qwen3-VL-4B	0.772	0.228	0.517
Qwen3-VL-8B (4-bit)	0.713	0.287	0.342
InternVL3.5-4B	0.858	0.142	0.458

4.2 Owner Labels Remain Decodable in Downstream Representations

A natural explanation for behavioral misbinding is that the model has simply lost the information needed to identify the correct owner. We test this possibility by examining the residual-stream representation at the final input position across all 28 layers. Pair-split linear probes show that owner identity remains decodable, with binding accuracy peaking at 0.694 at layer 24, compared with a behavioral binding accuracy of 0.505 under the same probing harness. Event information is easier to recover, reaching a peak probe accuracy of 0.958. These results show that gold-owner labels are linearly decodable from downstream representations on this split, without establishing that the model functionally uses this signal when producing its answer.

Layerwise probing. At each of the 28 language-model layers, we extract the residual-stream state at the final input position and train an L_2-regularized linear classifier to predict the correct event owner. Counterfactual partners are kept within the same data split so that a probe cannot train on one history and test on its paired counterpart. We apply the same procedure to the event marginal as a reference signal. The resulting curves should not be read as evidence that a particular layer uniquely represents ownership; rather, they measure where owner-related information becomes linearly recoverable from the state used by the downstream answer process (Hewitt & Manning, 2019; Voita & Titov, 2020).

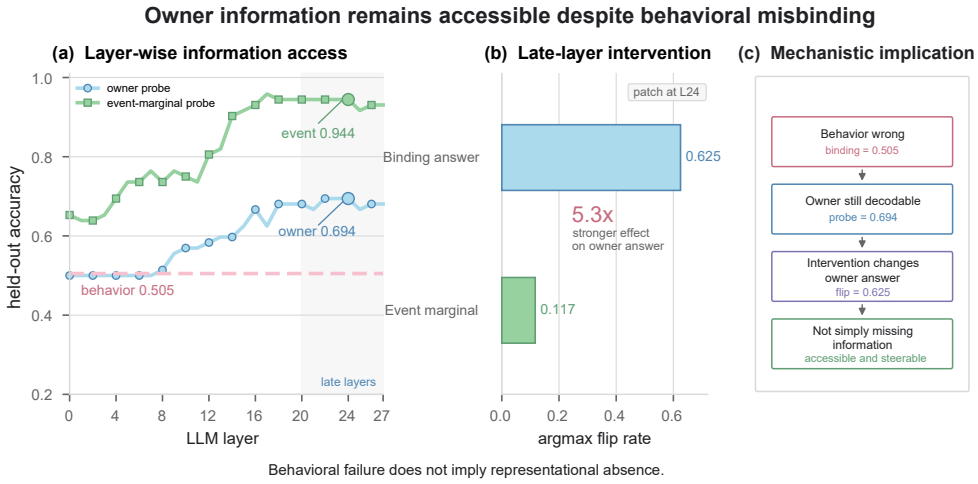


Figure 3: **Owner-related information remains accessible despite behavioral misbinding.** (a) Layerwise linear probes recover owner identity increasingly well in downstream representations, peaking at 0.694, while event-marginal information reaches 0.958. (b) Counterfactual activation patching at a late layer changes the binding answer in 62.5% of interventions but the event-marginal answer in only 11.7%, showing a substantially larger intervention effect on binding decisions than on the event marginal. (c) Together, the results show that predictive owner-related signal is linearly recoverable and that late answer states are intervention-sensitive, while leaving their causal organisation unresolved.

A targeted 16-dimensional ownership subspace produces partial effects, while the complementary residual state retains most of the intervention effect and norm-matched random directions remain non-zero controls (Appendix A.3, Table A.11). We therefore interpret the combined evidence as showing that gold-owner labels are decodable and that late answer-position states are intervention-sensitive, not that a single low-dimensional or owner-specific causal representation explains the behavioural error.

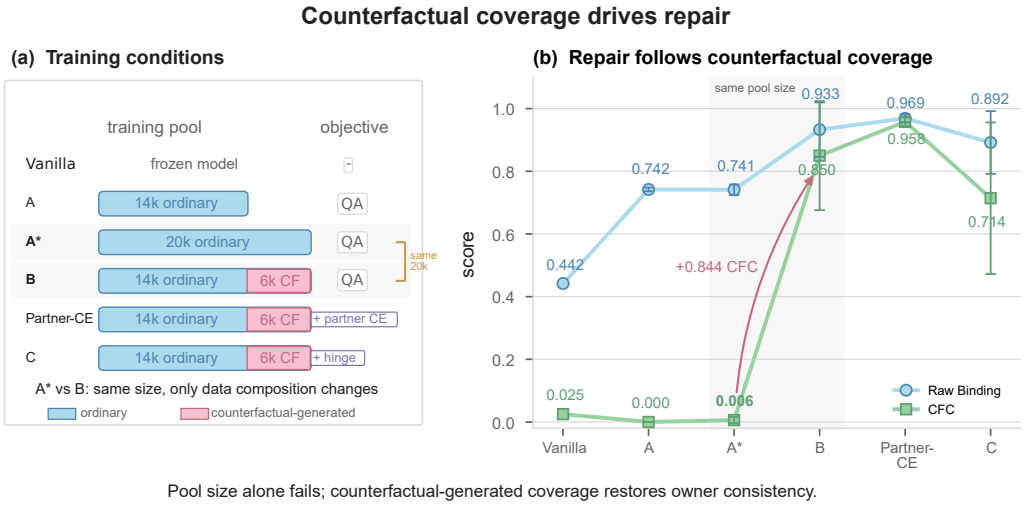
Counterfactual intervention. To move beyond passive decodability, we use a causal activation intervention, following prior work that restores or exchanges internal states to test their influence on downstream predictions (Meng et al., 2022; Wang et al., 2023). Specifically, we replace the recipient residual state at a selected layer with the corresponding state from its counterfactual partner and measure whether the predicted answer changes. Late-layer patching produces a 0.625 binding flip rate and a 0.117 event-marginal flip rate. These raw intervention effects are substantially different, although the two tasks also differ in output cardinality and should not be compared as a calibrated causal-effect ratio. Moreover, same-answer and random donors also produce substantial binding perturbations in the donor-control analysis. We therefore interpret full-state patching as evidence that late answer-position states are behaviourally influential and context-sensitive, rather than as evidence of owner-specific causal mediation.

5 Repairing Temporal Instance-State Binding

The previous analysis suggests that behavioral misbinding is not simply caused by the absence of owner-related information. We next ask what training signal can convert this accessible information into reliable, history-sensitive owner attribution.

5.1 Counterfactual-Generated Coverage Supports Repair

We first isolate the effect of the training data while keeping the base model, early-layer LoRA placement, optimizer, training schedule, step budget, evaluation protocol, and random seeds fixed. Arm A contains 14,000 ordinary single episodes. Arm A* adds 6,000 newly generated ordinary episodes,



Pool size alone fails; counterfactual-generated coverage restores owner consistency.

Figure 4: **Counterfactual-generated coverage supports repair beyond pool size alone.** (a) A* and B use equally sized training pools and the same QA objective, but differ in whether the additional 6,000 examples are ordinary or counterfactual-generated. (b) Size-matched ordinary supervision leaves CFC near zero (0.006), whereas counterfactual-generated supervision raises it to 0.850. Error bars show mean ± SD over three seeds for adapted conditions.

matching the 20,000-example size of the full training pool. Arm B instead replaces these 6,000 additional examples with counterfactual-generated episodes, but still trains every episode independently, without a pairwise loss or explicit access to its counterpart. The resulting pools are closely matched on event type, object count, identity similarity, and answer position. This size-matched comparison tests whether counterfactual-generated coverage provides information beyond simply adding more training examples while controlling these observed factors.

Increasing the number of ordinary examples is not sufficient to repair counterfactual owner attribution. Arm A* reaches a raw binding accuracy of 0.741 but a CFC of only 0.006 ± 0.010, essentially unchanged from the unadapted model. In contrast, replacing the additional ordinary examples with counterfactual-generated episodes raises raw binding to 0.933 and CFC to 0.850 ± 0.174. Crucially, this gain does not require explicit paired training: Arm B treats each episode independently, and only 13–22 complete counterfactual pairs are co-sampled during a 1,500-step run. The result therefore points to counterfactual coverage in the training distribution, rather than repeated exposure to paired examples. These results provide evidence that counterfactual-generated coverage, beyond pool size and the matched observed factors, is a major contributor to the observed repair.

Explicit relational objectives are also not necessary for the effect. Adding ordinary cross-entropy on the counterfactual partner yields a CFC of 0.958 ± 0.014, while the hinge-based objective reaches 0.714 ± 0.242. With three seeds, we do not observe an additional benefit from the hinge over counterfactual-generated supervision itself. These controls reinforce the same conclusion: the strongest change comes from exposing the model to histories in which visually similar final states correspond to different event owners, rather than from the particular relational loss used during training.

5.2 LoRA Repair Depends on Adapter Placement

We next study whether repair effectiveness varies with adapter placement on the primary Qwen3-VL-2B model under the tested eight-layer LoRA intervention family. We place identical rank-16 LoRA adapters over eight consecutive early (layers 0–7), middle (8–15), or late (20–27) layers while keeping the training pool, optimization schedule, evaluation protocol, and trainable module type fixed. We repeat this comparison under both the hinge-free counterfactual condition (Arm B) and the hinge-augmented condition (Arm C).

Table 2: Placement-dependent LoRA repair on Qwen3-VL-2B under matched eight-layer adaptation budgets. Results are mean ± SD over three seeds. Both supervision conditions show substantially stronger repair with early or middle adaptation than with late adaptation.

Objective	Metric	Early (0–7)	Middle (8–15)	Late (20–27)
B: Counterfactual-generated	Raw Binding	0.933 ± 0.087	0.911 ± 0.026	0.464 ± 0.048
B: Counterfactual-generated	CFC	0.850 ± 0.174	0.736 ± 0.121	0.069 ± 0.070
C: + relational hinge	Raw Binding	0.892 ± 0.100	0.895 ± 0.037	0.493 ± 0.030
C: + relational hinge	CFC	0.714 ± 0.242	0.750 ± 0.139	0.117 ± 0.113

The stage effect is large and consistent across both objectives. Under Arm B, early, middle, and late adaptation reach raw binding accuracies of 0.933, 0.911, and 0.464, with corresponding CFC scores of 0.850, 0.736, and 0.069. Arm C shows the same pattern: raw binding is 0.892, 0.895, and 0.493, while CFC is 0.714, 0.750, and 0.117. Early and middle adaptation therefore repair the behavior far more effectively than late adaptation under the same eight-layer LoRA budget. We do not infer a reliable ordering between early and middle layers, nor do we interpret these results as locating the origin of the failure.

We test two alternative explanations for the stage effect. First, the advantage of early or middle placement is not reproduced by simply distributing a similar trainable-parameter budget across the entire network: the parameter-matched all-layer adapter reaches substantially lower mean performance and exhibits much larger optimization variance. Second, the result is not specific to one learning rate, rank, or training duration. An independent validation sweep preserves the early-versus-late separation across stable configurations, and the validation-selected settings reach 0.978 ± 0.012 for early adaptation compared with 0.517 ± 0.046 for late adaptation on the frozen benchmark. These controls support a difference in repairability under the tested intervention family rather than an isolated hyperparameter accident.

The contrast with Section 4.2 is descriptive: gold-owner labels are most linearly decodable downstream, whereas the tested LoRA intervention is substantially more effective when placed earlier. We use *decodability–repairability mismatch* as a label for this empirical contrast under the present probe and adaptation interface, not as a mechanistic localisation claim.

5.3 Repair Increases Sensitivity to Temporal History

Higher pointwise accuracy alone does not show that a model has learned to use temporal history. This distinction is visible in the ordinary-data controls. Arms A and A* improve conditional binding, yet the predicted owner changes across counterfactual partners in only about 0.48–0.56 of pairs, remaining close to what would be expected from history-insensitive behavior. In contrast, successful counterfactual-trained conditions change the predicted owner in 0.79–0.99 of pairs. The improvement is therefore not merely a more accurate static decision rule; it is accompanied by a stronger dependence of owner attribution on the preceding history.

We further test this dependence by perturbing the temporal order at evaluation time. Reversing or cyclically rotating the panels leaves the unadapted model largely unchanged: raw binding is 0.442 in the canonical order, 0.438 under reversal, and 0.498 under rotation. The repaired early-layer models behave differently. Their binding accuracy drops by 19–33 percentage points when the same panels are presented in a non-canonical order. Thus, successful adaptation is accompanied by greater sensitivity to the canonical panel order. Because these permutations can also alter or invalidate the episode semantics, we treat this result as supporting evidence of temporal-order sensitivity rather than as a standalone demonstration of correct temporal reasoning.

We also test whether the repaired behavior depends on the exact synthetic configurations observed during adaptation. The structural OOD evaluation changes one factor at a time, including an unseen object count, event–similarity combinations that never co-occur during training, and an unseen object shape. Counterfactual-trained models retain CFC values between 0.69 and 0.91 across these splits, whereas the unadapted model remains at or below 0.03. These splits therefore test structural generalization within the controlled generation distribution and argue against a narrow explanation based on memorizing the exact object or event templates present in the training pool.

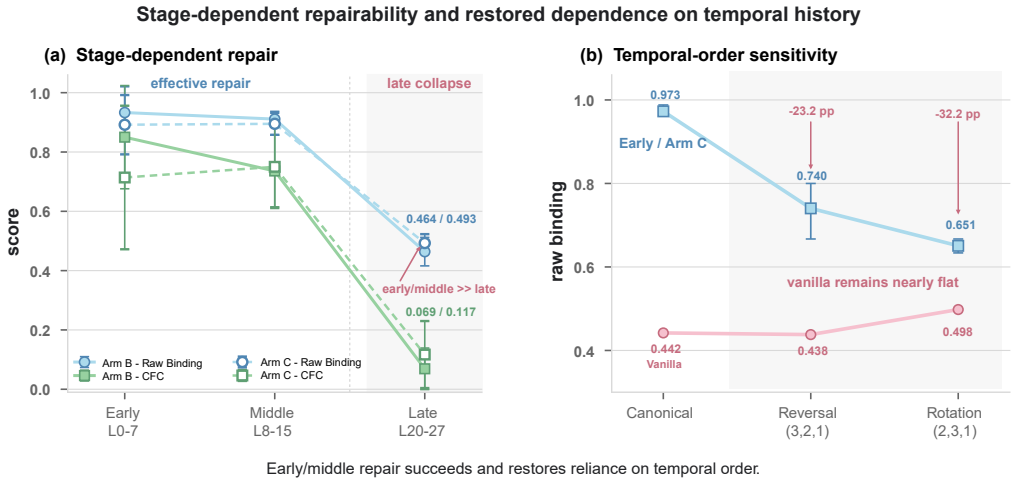


Figure 5: **LoRA repair varies with adapter placement and successful adaptation increases sensitivity to temporal order.** (a) Under matched eight-layer LoRA budgets, early and middle adaptation substantially outperform late adaptation for both raw binding and counterfactual-history consistency under Arms B and C. Error bars denote mean ± SD over three seeds. (b) Temporal-order perturbations leave the vanilla model largely unchanged but substantially reduce binding accuracy in a successfully repaired early-layer model, indicating that repair is accompanied by stronger dependence on the canonical temporal history.

6 Discussion

Our results reveal a *decodability–repairability mismatch*: gold-owner labels are most linearly decodable in downstream representations, whereas the same eight-layer LoRA intervention is substantially more effective at early or middle layers of the primary Qwen3-VL-2B model. Late-layer interventions nevertheless show substantial answer-state sensitivity. This contrast is descriptive and does not identify a dedicated ownership representation or the causal stage at which the error originates; whether the same repair profile holds across VLM architectures or adaptation interfaces remains open.

The supervision experiments further distinguish pointwise learning from history-conditioned ownership. Additional ordinary examples improve binding on individual episodes while leaving counterfactual owner behavior largely unchanged, whereas counterfactual-generated examples improve correctness across controlled history pairs even without an explicit pairing objective. Together with the size-matched A* control, this indicates that counterfactual-generated coverage contributes beyond pool size and the matched observed factors, without excluding other distributional differences between the generation procedures. This pattern is consistent with prior work on counterfactually augmented or bias-conflicting data (Kaushik et al., 2020; 2021; Nam et al., 2020; Liu et al., 2021). More broadly, independently sampled episodes may overestimate whether a model has learned the underlying temporal relation. Our claims therefore concern controlled temporal instance-state binding rather than domain-general video understanding.

7 Conclusion

Using controlled counterfactual histories, we show that event recognition and owner binding can dissociate substantially in vision-language models even when owner-related information remains linearly decodable. Counterfactual-generated supervision substantially outperforms size-matched ordinary supervision, while the tested LoRA intervention on Qwen3-VL-2B is markedly more effective at early or middle than late placements. Together, these results highlight the importance of maintaining and correctly using event–instance relations, rather than merely recognizing what happened.

AI USE STATEMENT

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

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

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

ETHICS STATEMENT

This work uses a synthetically generated controlled benchmark and publicly available tracking datasets for an auxiliary real-video diagnostic. It does not involve human subjects, private data, or interventions involving people. The work concerns the evaluation and interpretation of vision-language model behavior and makes no claims of clinical or neurological equivalence.

REPRODUCIBILITY STATEMENT

Appendix A.6 documents the model versions, inference and training protocols, LoRA configurations, random seeds, deterministic settings, evaluation harnesses, and statistical procedures used in this work. Appendix A.1 provides benchmark construction and integrity checks; Appendix A.4 reports per-seed repair results, training-pool composition, exposure audits, and placement controls; and Appendix A.5 defines CFC and the history-sensitivity analyses. We will additionally provide an anonymous repository containing the dataset-generation, training, evaluation, and auditing code used for the reported experiments.

REFERENCES

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary Chase Lipton. Explaining the efficacy of counterfactually augmented data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=HHiiQKWsOcV.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/

068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

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 *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

property	value
episodes	480 = 240 single-history episodes + 240 counterfactual-pair members (120 pairs)
timesteps / objects	3 (T1–T3) / 3 persistent objects per episode
panel resolution	384×384 px; frozen set drawn from the `low`-similarity stratum
event families	MOVE 127; OCCLUDE_REAPPEAR 120; ATTRIBUTE_CHANGE 117; STATE_TOGGLE 116
pair audit (inline, per member record)	240/240 records with pixel-identical final frame; 240/240 records with all paired answers flipped; 120/120 pairs valid
pair gold answers	gold binding *text* differs in 120/120 pairs; gold *letters* differ in 91/120, because option letters are shuffled per question
dataset audit	PASS; 0 annotation errors; 0 prompt–option text overlaps
manifest	`data/generated/phase1_full/manifest.json`; the SHA-256 of `episodes.jsonl` recomputed for this appendix matches the manifest entry

Table A.1: TIB-Bench: frozen configuration and integrity audit. Pair validation is stored inline with every pair member, so the pixel-identity and answer-flip claims can be checked without regenerating data.

APPENDIX A SUPPLEMENTARY RESULTS

This appendix documents the evidence behind the three claims of the main text. Section A.1 covers the benchmark, its prompts and its static shortcut controls; Section A.2 the behavioural diagnosis under both measurement harnesses; Section A.3 the representation and intervention evidence; Section A.4 the repair arms and their controls; Section A.5 the formalisation of counterfactual-history consistency (CFC), the history-sensitivity evidence and the in-setting structural out-of-distribution splits ;Section A.6 reproducibility and Section A.7 the auxiliary real-video diagnostic replication. All appendix values were audited against the archived experiment outputs and committed prediction artifacts listed in Section A.6. Quantities for which complete episode-level intermediate rows are available were additionally recomputed from those rows.

A.1 TIB-BENCH: CONSTRUCTION, INTEGRITY, PROMPTS, AND STATIC SHORTCUT CONTROLS

Episodes are produced by the committed generator: three persistent objects are rendered over three timesteps (T1–T3) at 384 px, one event from a controlled family occurs mid-sequence, and the frozen set mixes single-history episodes with counterfactual pairs whose two members share a pixel-identical final frame but flip the gold owner. Table A.1 lists the configuration and the integrity audit, Table A.2 the verbatim prompt and question design, and Table A.3 the static controls that bound how much of the binding question can be answered without using the history. A worked pair, with the shared final frame and the two gold owners, is main-text Figure 1. The representation-level evidence, the layerwise probes, counterfactual activation patching, and the 16-dimensional ownership subspace with its residual and random-direction controls is reported in Section A.3.

Final-frame-only control, and where the metric definitions live. The static controls include the final-frame-only input probe: the model is shown only the final panel, with the history removed, and is asked the same binding question. Its conditional binding is 0.081 (Table A.3), and because it was run through the generation pipeline rather than the canonical letter-logit one, its full row is reported with the harness baselines of Section A.2 (Table A.5) rather than mixed into the canonical tables. Raw binding, event marginal and conditional binding are defined in the main text; Section A.5 gives the pair-level CFC formalisation and associated history-sensitivity analyses.

A.2 BEHAVIOURAL DIAGNOSIS: BOTH HARNESSES, CROSS-MODEL SCOPE, PER-EVENT BREAKDOWN, TAXONOMY

The canonical harness is the letter-logit argmax (Section A.1). The earlier generation harness (free-text answers, parsed) is retained as a robustness check only: it is never the headline denominator, and its rows are never mixed with letter-logit rows inside one table. Table A.4 gives the joint diagnostic

item	content
prompt template	"The image shows three panels from left to right: the leftmost panel is time T1, the middle panel is T2, and the rightmost panel is the final time T3. Reply with only the option letter (A, B, C or D)." followed by `Question:`, `Options:` and `Answer:`
questions per episode	2 perception (object visibility at T1 and T3), 2 marginal (object count in the final scene; history-required event existence), 1 binding ("Which object *event*?"), plus an integrity question on 360/480 episodes
binding options	the three object names plus "None of the objects"; option letters are shuffled per question
scoring	argmax over the four option-letter logits at the final *input* token; no free-text parsing, no parse failures
letter balance (binding, $n = 480$)	A 122 / B 120 / C 119 / D 119 ($\chi^2 = 0.05$ on four cells; entropy 1.9999 bits)

Table A.2: Prompt, question design and scoring rule of the canonical letter-logit harness (`src/tiber/models/base.py`, `src/tiber/data/questions.py`). All headline numbers and every adaptation table in this paper use this harness.

control	score	coverage
best constant letter (chance 0.25)	0.254	100%
leftmost object at T1	0.246	100%
rightmost object at T1	0.244	100%
open-state change	1.000	19%
colour change	1.000	18%
any field change	0.830	56%
region change	0.511	20%
final-frame-only input (auxiliary harness)	0.081	$n = 74$

Table A.3: Static shortcut controls on the frozen benchmark. Heuristic rows are computed from each episode's ground-truth state metadata, not from pixels, so they bound how much of the binding question is answerable by simple position or state-change rules; the only rules above chance read the changed attribute itself, which already requires solving the change-detection step that the event marginal measures. The score of the final-frame-only row is conditional binding, measured under the auxiliary generation pipeline (Table A.5), not the letter-logit harness.

under both harnesses, Table A.5 the auxiliary prompting and readout baselines, Table A.6 the cross-model scope check, Table A.7 the per-event breakdown, and Table A.8 the error taxonomy.

Two conventions for the rest of the appendix. First, all pointwise intervals are percentile bootstrap intervals over episodes (5,000 resamples), and all CFC intervals are percentile bootstrap intervals over counterfactual pairs (5,000 resamples), because the pair is the unit on which that statistic is defined. Second, two different binding accuracies appear in this appendix and must not be conflated: 0.481 is the frozen 480-episode conditional binding of the canonical harness (main-text Table 1), whereas 0.505 is the behavioural binding accuracy measured on the 240-episode probing subset used by the representation analyses of Section A.3. They are the same harness on different episode sets; neither may be quoted as if it were the other.

A.3 Representation and intervention

The representation and intervention analyses use the 240 episodes that form the 120 counterfactual pairs of the frozen set (the pair split never straddles train and test). On that subset the vanilla model's conditional binding is 0.505; the corresponding number on the full 480-episode frozen set is 0.481 (Table 1 of the main text), because the subset conditions on the paired episodes. Table A.9 reports the probing protocol and layer curve, Table A.10 the activation-patching protocol and results, and Table A.11 the ownership-subspace intervention with its residual and random-direction controls. Figures A.1 and A.2 plot the probing and patching quantities.

harness	both correct (of 480)	P(owner wrong | both)	joint rate (of 480)
letter-logit argmax (canonical)	210 = 0.4375	109/210 = 0.519 [0.448, 0.586]	109 = 0.227 [0.190, 0.265]
generation, free text parsed (alternate)	222 = 0.4625	110/222 = 0.496 [0.428, 0.559]	110 = 0.229 [0.194, 0.267]

Table A.4: The headline diagnostic under both harnesses (vanilla Qwen3-VL-2B, frozen 480 episodes). "Both correct" means the two perception questions and the history-required event-existence question are all correct; the third column conditions on that set and the fourth is the unconditional joint co-occurrence. Intervals are 95%episode-level bootstrap intervals (5,000 resamples); the two harnesses agree on the joint rate to within 0.2 pp.

method (generation harness)	event marginal	cond. binding	BDG (pp)
vanilla full-history	0.669	0.505	+16.4
final-frame-only	0.435	0.081	+35.4
oracle textual ledger	1.000	0.983	+1.7
trajectory CoT	0.419	0.390	+2.9

Table A.5: Prompting and readout baselines under the auxiliary generation harness (solvability/history diagnostic). Every row, including the vanilla reference row, is measured under the same generation harness, so the BDG column is internally consistent; the table is not comparable row-by-row with the letter-logit tables. The canonical letter-logit vanilla reference is the Qwen3-VL-2B row of Table 1 of the main text (conditional binding 0.481), and the validation-selected early-layer configuration (0.978 ± 0.012) is reported in Section 5.2.

What these analyses do and do not establish. Linear decodability is not evidence that the model "knows" the owner, nor that it uses that information when it answers: the probe is an external readout trained on pair-disjoint splits, and on the same subset the model's own conditional binding is 0.505. Activation patching shows that the late-layer answer state is behaviourally steerable by a counterfactual donor; it does not identify a complete mechanism, because same-answer and random donors perturb the answer as well (Table A.10) and the targeted 16-dimensional subspace accounts for only part of the flip, with the complementary residual carrying the majority of the effect and norm-matched random directions moving the answer too (Table A.11); the subspace is therefore selectively but only partially causally steerable, and we make no claim that it accounts for the whole behavioural effect. The stage pattern of Section A.4 characterises where adaptation is effective - an adaptation-interface property - and is not offered as a causal localisation of the failure.

A.4 Repair: arms, pools, per-seed results, stage controls

All repair results use the deterministic protocol (all RNGs seeded; two runs of the same configuration produce bit-identical checkpoints), the frozen 480-episode benchmark for evaluation, and an independent 240-episode validation set for any selection decision. Table A.12 defines the arms, Table A.13 gives the per-seed results, Table A.14 the pool composition and the matching audit, Table A.15 the exposure audit, and Table A.16 the adaptation configuration.

A.5 History sensitivity, the CFC formalisation, and in-setting structural splits

This section defines CFC exactly, relates its observed values to the independence expectation implied by the pointwise marginals, and reports the perturbation families that test whether repaired owner attribution depends on the counterfactual history rather than on a static cue: object-level answer flips on the frozen pairs, temporal-order permutations, and in-setting structural splits of the generation distribution.

Definition and reference quantities. Each of the 120 counterfactual pairs in the frozen benchmark consists of two episodes that share a pixel-identical final frame but differ in visual history and in the gold owner identity; 91 of the 120 pairs also differ in the gold option letter because option order is

model	perception	event marginal	raw binding	cond. binding	P(wrong | both) [95% CI]	joint rate	CFC [95% CI]
Qwen3-VL-2B	0.718	0.617	0.442	0.481	0.519 [0.448, 0.586]	0.227	0.025 [0.000, 0.058]
Qwen3-VL-4B	0.855	0.777	0.740	0.772	0.228 [0.181, 0.276]	0.133	0.517 [0.425, 0.608]
Qwen3-VL-8B (4-bit)	0.864	0.606	0.594	0.713	0.287 [0.233, 0.346]	0.148	0.342 [0.258, 0.425]
InternVL3.5-4B	0.823	0.694	0.683	0.858	0.142 [0.103, 0.185]	0.085	0.458 [0.367, 0.542]
SmolVLM2-2.2B (scope check)	0.444	0.504	–	0.828	–	–	–

Table A.6: Cross-model scope check under the letter-logit harness (frozen 480 episodes, identical prompt, scoring rule, evaluation set and bootstrap for the Qwen rows and InternVL3.5-4B). The 8B model runs in NF4 4-bit for the 11 GB device budget, so its row carries a quantisation confound, and InternVL3.5 uses its own dynamic-tiling preprocessing; treat both as scope probes rather than like-for-like comparisons. The SmolVLM2-2.2B row is an earlier preliminary scope check whose conditional-binding cell rests on $n = 29$ episodes; its CFC is not reported here. High pointwise accuracy does not guarantee stable counterfactual owner behaviour: Qwen3-VL-4B and InternVL3.5-4B are close in conditional binding (0.772 vs 0.858) yet differ in CFC (0.517 vs 0.458).

event (n)	2B event marginal	Qwen3-VL-2B	Qwen3-VL-4B	Qwen3-VL-8B (4-bit)	InternVL3.5-4B
MOVE (127)	0.480	0.386 / 0.286 / 0.030	0.543 / 0.543 / 0.212	0.457 / 0.432 / 0.091	0.606 / 0.631 / 0.303
ATTRIBUTE_CHANGE (117)	0.940	0.701 / 0.718 / 0.067	0.966 / 0.966 / 0.867	0.966 / 0.974 / 0.867	0.974 / 0.974 / 0.900
STATE_TOGGLE (116)	0.388	0.336 / 0.125 / 0.000	0.707 / 0.917 / 0.630	0.336 / 0.846 / 0.148	0.319 / n/a / 0.111
OCCLUDE_REAPPEAR (120)	0.667	0.350 / 0.193 / 0.000	0.758 / 0.650 / 0.400	0.625 / 0.455 / 0.267	0.833 / 0.873 / 0.500

Table A.7: Per-event breakdown under the letter-logit harness. Each cell is raw binding / conditional binding / CFC; the second column is the Qwen3-VL-2B event-existence accuracy for that family. The effect is present in every family and is weakest on MOVE, the family with the lowest event-existence accuracy. InternVL3.5-4B's STATE_TOGGLE cell is thin (its own perception and event accuracy are low there), so that cell is not a binding result. Single frozen set; no per-family significance claim is made.

independently randomized. With $a_i, b_i \in \{0, 1\}$ indicating whether the member A and member B episodes of pair i were answered correctly,

$$\mathrm{CFC} \;=\; \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} a_i b_i,$$

the pair-both-correct rate. Because the gold owner identity differs across all counterfactual pairs, a policy that predicts the same owner identity for both members cannot be correct on both. Option letters, however, are independently randomized across pair members and should not be interpreted as a pair-invariant identity representation. Independent uniform four-way answering scores $1/16 = 0.063$ on the pair-both-correct rate, and the unadapted model scores 0.025. Independence expectation, re-pairing null and bootstrap are defined in Table A.20.

Observed CFC versus the independence expectation. For the stable high-performing configurations - early-placement adaptation, the partner-CE control, the validation-selected early configuration and the two higher-accuracy seeds of the full-pool QA arm - the observed CFC matches $\hat{p}_A\hat{p}_B$ to within 0.002, comfortably inside the 0.008 tolerance. Deviations above 0.01 occur only where the arm or seed itself is weak or unstable: the unadapted baseline (+0.010), full-pool QA at early layers seed 111 (+0.045, CFC 0.650 against 0.933/0.967 for its sibling seeds), hinge early seed 222 (+0.051, CFC 0.467), hinge late seeds (−0.053 and +0.033, CFC 0.233/0.108), validation-selected late seed 333 (+0.028, CFC 0.117) and middle-placement seed 111 in both arms (+0.012 / +0.019). Across all 39 arm-seed combinations the largest deviation is 0.053. The re-pairing null reproduces $\hat{p}_A\hat{p}_B$ to within 1×10^{-4} in every arm. We therefore interpret CFC as a stringent pair-both-correct success metric whose magnitude is largely determined by the pointwise marginals in stable high-performing configurations, rather than as a capability dimension orthogonal to pointwise accuracy. Its value here is that success requires the same model to answer both controlled histories correctly.

Changing the answer versus changing it correctly. The unadapted model changes its predicted binding text on 0.483 of pairs, but those changes carry almost no signal about correctness (CFC 0.025; CFC given a flip 0.052), and the history-free controls behave identically (A 0.000, A* 0.000-0.030). The adapted arms flip far more often (early placement 0.958-0.983, full-pool QA 0.792-0.983, partner-CE 0.967-0.983) and their flips are correct: CFC given a predicted-text flip is 0.890/1.000/0.915 for the three early-placement seeds and 0.958-1.000 for partner-CE. Late-placement arms and the validation-selected late configuration stay intermediate - flip rates rise only

harness	binding-error taxonomy
letter-logit, canonical	268 errors = 265 assignments to another instance + 3 "None of the objects"; among the 171 errors on episodes whose event question was correct, 171/171 are instance assignments
generation, alternate	274 errors = 145 identity swaps + 129 perception-gated + 0 substitutions

Table A.8: Error taxonomy. An *instance assignment* names a different object than the gold owner; a *substitution* answers "None of the objects"; a *perception-gated* error is one where the gold object was not identified in the perception questions. The canonical counts were re-derived from the committed letter-logit prediction rows and the benchmark option texts for this appendix; the generation-harness counts come from the earlier free-text run and are labelled as such.

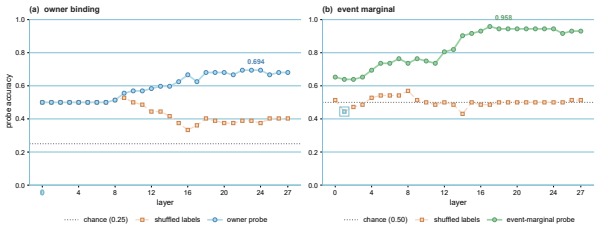


Figure A.1: Layer-wise linear-probe accuracy for owner binding and for the event marginal, with shuffled-label controls. The ownership probe peaks at 0.694 (layer 24) while the model's own conditional binding on this subset is 0.505. This establishes linear decodability on this split, not that the model functionally uses the decoded signal when answering. Decodability is not evidence that the model "knows" the answer.

modestly (0.425-0.525 and 0.433-0.600) and a minority of those flips are both-correct (0.000-0.270 and 0.058-0.458). The object-level measure therefore separates changing the answer from changing it correctly, and it is the correct part, not the change itself, that the counterfactual pool restores.

Temporal-order perturbations. Permuting the panel content (reverse: 3,2,1; rotate: 2,3,1; labels kept in their canonical positions) leaves the unadapted model essentially unchanged - 0.442 raw binding becomes 0.438 under reversal and 0.498 under rotation - while the repaired early-layer arms lose 19-33 pp and their CFC falls from 0.875/0.958/0.892 to 0.558/0.733/0.683 (reversal) and 0.542/0.525/0.508 (rotation). The contrast is consistent with greater sensitivity to temporal order after adaptation, while not by itself identifying which temporal relation the model has learned.

In-setting structural splits. Adapters trained on the synthetic pool transfer to object counts never trained (2-3 to 4), to event $\times$ similarity cells never co-trained, and to an unseen shape: CFC 0.69-0.91 for the adapted arms versus at most 0.03 for vanilla, with no reliable difference between the hinge arm and full-pool QA on these single-seed splits. Because these splits are carved out of the same synthetic generator, they test structural generalisation within the generation distribution rather than transfer to a different domain.

A.6 Reproducibility

Models and hardware. Primary model: Qwen3-VL-2B-Instruct (FP16; sdpa attention) from the local Hugging Face cache; the 8B comparison rows use NF4 4-bit quantisation. All runs used one NVIDIA RTX 2080 Ti (11 GB).

Inference protocol. All headline numbers use the frozen letter-logit harness: the model receives the fixed prompt ('The image shows three panels from left to right: T1, T2, T3. Reply with only the option letter.') and the answer is the argmax over the four option-letter logits at the final position, so no text parsing is involved and there are no parse failures. Evaluation uses the frozen 480-episode benchmark only, which is never used for training.

layer	binding probe	binding control	marginal probe	marginal control
0	0.500	0.500	0.653	0.514
8	0.514	0.514	0.736	0.569
12	0.583	0.444	0.806	0.500
16	0.667	0.333	0.931	0.486
20	0.681	0.375	0.944	0.500
24	**0.694**	0.375	**0.944**	0.500
27	0.681	0.403	0.931	0.514

Table A.9: Probing protocol and selected layers (L2 logistic regression on the answer-position residual state; chance 0.25 for the 4-way binding target and 0.50 for the binary marginal target). Split 70/30 *by counterfactual pair* ($n_{\text{train}} = 168$, $n_{\text{test}} = 72$; the two members of a pair never straddle the split), all 28 transformer layers are probed, and a shuffled-label control is reported next to every cell. The full 28-layer curve is plotted in Figure A.1.

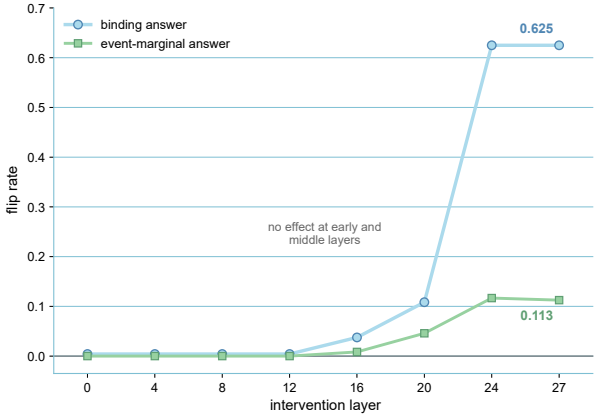


Figure A.2: Activation-patching flip rates under the paired counterfactual donor. Late answer-position replacement yields raw binding and event-marginal flip rates of 0.625 and 0.117, respectively; generic donor controls below show that the binding effect is not owner-specific.

Supervision arms. LoRA rank 16, alpha 32, applied to q/k/v/o of 8 consecutive LLM layers (early 0-7, middle 8-15, late 20-27); AdamW, lr 3×10^{-4}, gradient clipping 1.0; 1,500 steps, batch size 1, 256-px training panels; loss = CE(binding) + 0.5 CE(marginal) + 0.5 KL(marginal-preserve). Arm A drops all 6,000 pair members (14,000 episodes remain). Arm A* adds 6,000 newly generated independent singles from a separate generator seed, giving a 20,000-episode pool with no counterfactual members, matched to the full pool on event type, object count, similarity and answer letter. Arm B keeps the full 20,000-episode pool and answers every episode independently (no pair lookup, no hinge); the partner-CE control additionally scores the partner's binding answer with ordinary cross-entropy. Arm C adds the counterfactual hinge (margin 0.5) on pair members. The training pool is disjoint from the benchmark by construction, and a manifest audit matches the arms on event type, object count, similarity and answer letter.

Determinism. The initial exploratory protocol did not seed the global RNG, so LoRA initialisation varied between runs and same-seed re-runs differed by up to 0.18 CFC. The headline arms were therefore repeated under a fully seeded protocol: python, numpy and torch (CPU and CUDA) seeds set from the run seed, with `cudnn.deterministic`, `cudnn.benchmark=False`, `torch.use_deterministic_algorithms(True, warn_only=True)` and `CUBLAS_WORKSPACE_CONFIG=:4096:8`. Under this protocol two runs of the same configuration produce bit-identical checkpoints (matching SHA256) and identical loss traces, and each run writes a `run_manifest.json` recording the seed, architecture, step budget and the SHA256 of the training pool and config. The one exception is the parameter-matched all-layer control, which predates the seeded protocol and whose run directory carries no `run_manifest.json` (only checkpoints and evaluation logs); it is reported only as supporting

protocol element	setting / result
intervention	replace the answer-position residual state of member A with member B's (and vice versa) at layers $\{0, 4, 8, 12, 16, 20, 24, 27\}$; measure the next-token letter-logit argmax
main result (120 pairs)	binding flip 0.625 vs marginal flip 0.117 at layers 24 and 27; layers 0–20 give 0.00–0.11 (binding) and 0.00–0.05 (marginal)
direction asymmetry	in the B $\leftarrow$ A direction the answer moves to A's correct actor in 60% of cases, against 10% toward "None of the objects" in the opposite direction
20-pair donor controls (layer 24 / layer 27)	counterfactual donor 0.50 / 0.45 (binding) vs 0.15 / 0.05 (marginal); same-answer donor 0.65 / 0.60; random donor 0.75 / 0.60; non-answer-token injection 0.00 at every layer

Table A.10: Activation patching. The intervention is position-specific and late-layer-specific. Under the counterfactual donor the raw binding flip rate exceeds the event-marginal flip rate, but the two tasks differ in output cardinality and the ratio is not interpreted as a calibrated causal-effect comparison. Same-answer and random donors also perturb binding strongly, so full-state patching is interpreted as answer-state sensitivity rather than owner-specific causal mediation.

intervention	layer 20	layer 24	layer 27
full donor state	0.889	0.855	0.856
ownership projection (U_b, 16-d)	0.105	0.261	0.295
complementary residual	0.717	0.572	0.606
norm-matched random direction	0.166	0.178	0.200
orthogonal random direction	0.150	0.233	0.194
marginal flip (all variants)	0.000	0.000	0.000
ownership share of the full effect	12%	31%	34%
residual share of the full effect	81%	67%	71%

Table A.11: Ownership-subspace intervention (5-seed mean, $n_{\text{test}} = 36$ held-out pairs per seed). U_b is the top-16 PCA basis of normalised training-pair difference vectors. The subspace effect is partial and is not consistently larger than the random-direction controls at every layer; the event-marginal answer remains unchanged across all variants. The residual carries the majority of the effect, and random directions flip a non-trivial fraction, so the 16-dimensional subspace is not a complete account of the behavioural flip.

evidence (Table A.17), and every headline number in this paper comes from the deterministic protocol.

Headline rerun. Nine runs (early and late placement; singles-only vs full-pool QA; seeds 111/222/333) evaluated on the frozen 480 episodes give singles-only QA 0.742 raw binding and CFC 0.000 at all three seeds, full-pool QA at early layers 0.933 raw binding and CFC 0.850, and full-pool QA at late layers 0.464 and CFC 0.069. Both headline effects reproduce: the coverage effect is $+0.85$ CFC and the stage effect is $+0.47$ raw binding, matching the initial protocol. Seed-to-seed spread remains real (full-pool early CFC 0.650 / 0.933 / 0.967 across the three seeds) and is reported alongside every mean in this appendix rather than hidden behind it.

Metrics. Raw binding = unconditional binding accuracy; event marginal = history-required event-existence accuracy; conditional binding = binding accuracy on perception- and marginal-correct episodes; BDG = (event marginal $-$ conditional binding) in pp with a 5,000-iteration episode bootstrap 95% CI; CFC reported separately for binding and marginal pairs and never pooled; parse failures count as incorrect.

Test suite and consistency gates. The repository test suite (`tests/`, 9 files) passes 61/61 tests with no model weights loaded. The stale-value gate (`scripts/audit_stale_values.py`) checks 11 active artifacts for retired pre-letter-logit values (0.429 raw binding, 222/480 joint denominator, 110/222, 0.496, 49.6%) and passes. The benchmark-integrity, pool-matching,

arm	training pool	counterfactual pair members	extra forward	objective
A: singles-only QA	14,000 singles	0	no	CE(binding) + 0.5 CE(marginal) + 0.5 KL(marginal-preserve)
A*: equal-size, no counterfactual episodes	14,000 singles + 6,000 newly generated independent singles (separate generator seed)	0	no	same as A
B: full-pool QA	20,000 = 14,000 singles + 6,000 pair members	6,000	no	same as A; every episode answered independently, no pair lookup, no hinge
partner-CE control	same as B	6,000	yes	same as A + CE on the partner's binding answer (non-relational)
C: pair-augmented hinge	same as B	6,000	yes	same as A + relational hinge (margin 0.5) on pair members

Table A.12: Supervision arms. Everything else is identical across arms: LoRA rank 16, alpha 32 on q/k/v/o of 8 consecutive LLM layers, AdamW with learning rate 3×10^{-4} and gradient clipping 1.0, 1,500 steps, batch size 1, 256-px training panels, seeds 111/222/333. Training logs confirm that Arm B's counterfactual loss is 0.000 at every logged step, and the A* extras are separately generated independent singles rather than same-scene interventions, so the A*/B contrast changes the presence of counterfactual-generated pair members while closely matching the observed event-type, object-count, similarity and answer-position marginals reported below. Arm B has no explicit pairwise objective.

arm (placement)	raw binding, seeds 111/222/333	CFC, seeds 111/222/333	raw binding (mean $\pm$ SD)	CFC (mean $\pm$ SD)
A: singles-only (early)	0.735 / 0.744 / 0.746	0.000 / 0.000 / 0.000	0.742 ± 0.006	0.000 ± 0.000
A*: equal-size, no CF (early)	0.721 / 0.750 / 0.752	0.000 / 0.000 / 0.017	0.741 ± 0.017	0.006 ± 0.010
B: full-pool QA (early)	0.833 / 0.981 / 0.985	0.650 / 0.933 / 0.967	0.933 ± 0.087	0.850 ± 0.174
B: full-pool QA (middle)	0.931 / 0.881 / 0.921	0.858 / 0.617 / 0.733	0.911 ± 0.026	0.736 ± 0.121
B: full-pool QA (late)	0.519 / 0.444 / 0.429	0.150 / 0.025 / 0.033	0.464 ± 0.048	0.069 ± 0.070
partner-CE (early)	0.975 / 0.977 / 0.954	0.967 / 0.967 / 0.942	0.969 ± 0.013	0.958 ± 0.014
C: +hinge (early)	0.912 / 0.783 / 0.981	0.725 / 0.467 / 0.950	0.892 ± 0.100	0.714 ± 0.242
C: +hinge (middle)	0.937 / 0.875 / 0.873	0.850 / 0.592 / 0.808	0.895 ± 0.037	0.750 ± 0.139
C: +hinge (late)	0.508 / 0.458 / 0.512	0.233 / 0.008 / 0.108	0.493 ± 0.030	0.117 ± 0.113
vanilla (no adapter)	0.442	0.025	–	–

Table A.13: Per-seed repair results under the fully seeded deterministic protocol (frozen 480 episodes, 120 CFC pairs). The seed-to-seed spread is real: Arm B early CFC is 0.650/0.933/0.967 and Arm C early CFC is 0.725/0.467/0.950, so single-seed CFC differences below roughly 0.15 are not individually resolvable. Paired per seed, B $-$ A* is +0.192 raw binding (SD 0.069, 3/3 seeds positive) and +0.844 CFC (SD 0.169, 3/3 positive); the partner-CE control exceeds the hinge by +0.076 raw binding (2/3) and +0.244 CFC (2/3, test underpowered), which is why we report a higher mean and lower variance for partner-CE and claim no detectable additional hinge benefit rather than equivalence.

error-taxonomy and CFC cells of this appendix were additionally re-derived read-only from the committed prediction files by `scripts/audit_appendix_evidence.py`, which writes `results/appendix_evidence_audit.json`; the mechanism tables are transcriptions of their committed report JSONs. No new model forward passes were run for any of it.

Seed-policy disclosure. The project standard for training runs is six seeds with the best three reported (ADR-001), and every affected experiment also records all six values. The LoRA adaptation grid in this paper is an exception: it was run with three fixed deterministic seeds (111/222/333) and reports all three rather than any best-of-k readout, so no seed selection enters a headline number; seed spread is instead disclosed wherever it is large (full-pool early CFC 0.650/0.933/0.967).

pool	episodes	pair members
Arm B (full pool)	20,000	6,000 (3,000 pairs)
Arm A* (equal size, no counterfactual episodes)	20,000	0
matching audit (A* vs B)	max. absolute share difference: event type 0.65 pp; object count 0.84 pp; similarity 0.08 pp; gold binding letter 0.21 pp	
pool integrity	all 6,000 pair members in Arm B carry a pixel-identical final frame and flipped paired answers; the pool is disjoint from the benchmark by construction	

Table A.14: Training-pool composition and the matching audit behind the equal-size control. Both pools contain 20,000 episodes; matching is assessed on event type, object count, similarity and the gold binding letter. A* replaces the 6,000 counterfactual pair members with 6,000 newly generated independent singles. The comparison therefore contrasts independently generated ordinary episodes with counterfactual-generated episodes at equal pool size while closely matching the observed marginals listed above; unmeasured higher-order distributional differences are not excluded.

arm B, seed	episodes seen	pair members seen	complete counterpart pairs co-sampled	shared-final-frame groups
111	1,500	451	22	27
222	1,500	406	13	25
333	1,500	450	18	33

Table A.15: Exposure audit for Arm B (each episode is seen once per run). Simultaneous exposure to both members of a pair is rare: only 13–22 of the 3,000 available counterpart pairs are completely observed during a run (0.4–0.7%), while 406–451 of the 6,000 available pair members are sampled individually (6.8–7.5%). The recovery therefore tracks the presence of counterfactual-generated episodes in the pool rather than simultaneous pair exposure; this is why the claim is about pool composition, not about an implicit pairwise objective.

Availability. The benchmark, the training pools, the evaluation rows and the analysis scripts are local artifacts of this project; the paths and hashes in Table A.25 identify the exact revisions behind the synthetic-benchmark results in this appendix.

configuration element	setting
adaptation	LoRA, rank 16, alpha 32, on q/k/v/o projections of 8 consecutive LLM layers
placements	early 0–7, middle 8–15, late 20–27
trainable parameters	2,097,152 per adapter (identical for every placement and for the parameter-matched all-layer control)
optimisation	AdamW, lr 3×10^{-4}, gradient clipping 1.0, 1,500 steps, batch size 1, 256-px panels
seeds	111 / 222 / 333 (deterministic protocol)
evaluation	frozen 480-episode benchmark; 120 CFC pairs; independent 240-episode validation set never used for the frozen numbers

Table A.16: Adaptation configuration shared by every arm. The stage contrast at this configuration (Arm B and Arm C at early/middle/late placement with per-seed values) is reported in Table A.13; both arms place early and middle substantially above late. We make no claim ranking early above middle.

condition	raw binding, seeds 111/222/333	CFC, seeds 111/222/333	raw binding (mean $\pm$ SD)	CFC (mean $\pm$ SD)
early placement (8 layers, rank 16)	0.963 / 0.988 / 0.967	0.875 / 0.958 / 0.892	0.972 ± 0.013	0.908 ± 0.044
all-layer placement (28 layers, ranks 5/4)	0.779 / 0.254 / 0.865	0.608 / 0.075 / 0.633	0.633 ± 0.331	0.439 ± 0.315

Table A.17: Parameter-matched all-layer control. Both conditions carry exactly 2,097,152 trainable parameters, the same objective, optimiser, step budget and seeds; the all-layer condition distributes rank 5/4 across all 28 layers instead of rank 16 across 8 layers. Distributed all-layer adaptation is both worse on average and far less stable: seed 222 reaches 0.927 binding at step 500 and collapses to 0.250 by step 1500 (the deterministic early-placement arm never shows this behaviour). Matching the total parameter count therefore does not reproduce the early-placement advantage. Because the large spread is dominated by optimisation instability, and because the two conditions differ in per-layer capacity as well as placement, we treat this as supporting evidence rather than a clean capacity control. The early row here is the same arm as Arm C early under the earlier non-deterministic protocol; its deterministic rerun (Table A.13) gives 0.892 ± 0.100 raw binding and CFC 0.714 ± 0.242, which is why headline numbers use the deterministic protocol.

setting (independent validation set, 240 episodes)	early	late	early $-$ late
baseline: lr 3×10^{-4}, rank 16, 1,500 steps	0.917	0.550	+0.367
lr 1×10^{-4}	0.967	0.546	+0.421
lr 6×10^{-4}	0.250 (collapse)	0.487	-0.237
rank 8	0.908	0.579	+0.329
rank 32	0.900	0.588	+0.312
750 steps	0.804	0.479	+0.325
3,000 steps	0.996	0.475	+0.521

Table A.18: One-factor-at-a-time placement robustness on the independent validation set (240 episodes, disjoint from the frozen benchmark). Early placement beats late in every setting except the lr 6×10^{-4} cell, where early training collapses (the late arm is unaffected because late placement never learns this behaviour under any setting).

validation-selected configuration	raw binding	conditional binding	CFC
early-best (3,000 steps)	0.978 ± 0.012	0.980	0.933 ± 0.058
late-best (rank 32)	0.517 ± 0.046	0.595	0.139 ± 0.126

Table A.19: Each position's best configuration selected on the validation set only, then run for three deterministic seeds on the frozen benchmark. The gap under each position's own validation-selected optimum is 0.461 raw binding (3/3 seeds), and late placement remains close to the unadapted baseline (0.517 vs 0.442), so the stage effect is not an artefact of one hyperparameter choice.

element	definition	value
pair inclusion	the two members share a pixel-identical final frame and differ in the binding gold text (91/120 pairs also differ in the gold letter)	120 pairs (240 member episodes)
CFC (binding)	P(member A correct $\wedge$ member B correct)	0.025 (vanilla)
independent uniform guessing	arithmetic chance for the pair-both-correct rate under four-way independent answering	0.063
independence expectation	$\hat{p}_A\hat{p}_B$, the pair-both-correct rate implied by the two pointwise member accuracies	per arm, Table A.21
re-pairing null	member A resampled against member B of another pair (5,000 permutations), preserving the marginals and destroying the within-pair relation	reproduces $\hat{p}_A\hat{p}_B$ within 1×10^{-4}
confidence interval	5,000-iteration bootstrap resampling pairs (not episodes)	reported with every cell

Table A.20: The formal apparatus behind the CFC column. All quantities are computed on the same 480 frozen episodes / 120 pairs; the statistic is never pooled across the binding and marginal pair sets.

arm (position)	CFC, seeds 111/222/333	independence expectation $\hat{p}_A\hat{p}_B$	difference
vanilla (unadapted)	0.025	0.015	+0.010
early placement	0.875 / 0.958 / 0.892	0.876 / 0.958 / 0.893	−0.001 / 0.000 / −0.002
middle placement	0.792 / 0.933 / 0.633	0.780 / 0.934 / 0.636	+0.012 / 0.000 / −0.003
late placement	0.000 / 0.025 / 0.142	0.000 / 0.022 / 0.110	0.000 / +0.003 / +0.032
A: singles-only QA	0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000
A*: equal-size no-CF pool	0.000 / 0.000 / 0.017	0.000 / 0.000 / 0.017	0.000 / 0.000 / 0.000
B: full-pool QA, early	0.650 / 0.933 / 0.967	0.605 / 0.934 / 0.967	+0.045 / −0.001 / 0.000
B: full-pool QA, middle	0.858 / 0.617 / 0.733	0.840 / 0.609 / 0.729	+0.019 / +0.007 / +0.004
C: +hinge, early	0.725 / 0.467 / 0.950	0.731 / 0.416 / 0.942	−0.006 / +0.051 / +0.008
C: +hinge, middle	0.850 / 0.592 / 0.808	0.855 / 0.572 / 0.810	−0.005 / +0.020 / −0.002
C: +hinge, late	0.233 / 0.008 / 0.108	0.286 / 0.006 / 0.076	−0.053 / +0.003 / +0.033
partner-CE, early	0.967 / 0.967 / 0.942	0.967 / 0.967 / 0.942	0.000 / 0.000 / 0.000
validation-selected early	0.958 / 0.867 / 0.975	0.959 / 0.867 / 0.975	0.000 / 0.000 / 0.000
validation-selected late	0.275 / 0.025 / 0.117	0.265 / 0.016 / 0.089	+0.010 / +0.009 / +0.028

Table A.21: Observed CFC against the independence expectation implied by the pointwise member accuracies, per arm and seed. If CFC measured a capability dimension orthogonal to pointwise accuracy it would systematically exceed $\hat{p}_A\hat{p}_B$; the stable high-performing configurations match it to within 0.002, and every deviation above 0.01 sits in the unadapted baseline or in an arm/seed whose own CFC is far below the level its sibling seeds reach.

arm (position)	P(predicted binding text flips)	CFC	CFC given text flip
vanilla (unadapted)	0.483	0.025	0.052
A: singles-only QA	0.483-0.508	0.000	0.000
A*: equal-size no-CF pool	0.475-0.558	0.000-0.017	0.000-0.030
early placement	0.958-0.983	0.875-0.958	0.890-1.000
middle placement	0.775-0.967	0.633-0.933	0.817-0.966
late placement	0.425-0.525	0.000-0.142	0.000-0.270
B: full-pool QA, early	0.792-0.983	0.650-0.967	0.821-0.983
B: full-pool QA, middle	0.900-0.975	0.617-0.858	0.685-0.896
C: +hinge, early	0.750-0.967	0.467-0.950	0.622-0.983
C: +hinge, middle	0.775-0.983	0.592-0.850	0.763-0.924
C: +hinge, late	0.483-0.650	0.008-0.233	0.017-0.424
partner-CE, early	0.967-0.983	0.942-0.967	0.958-1.000
validation-selected early	0.917-0.992	0.867-0.975	0.945-0.992
validation-selected late	0.433-0.600	0.025-0.275	0.058-0.458

Table A.22: Object-level history sensitivity on the 120 frozen pairs (ranges over the three seeds of each arm). The predicted binding text is the object state the model actually names; a history-insensitive model may still flip its answer between the two members by noise, which is why the informative quantity is the both-correct rate conditional on a flip. The unadapted model flips its text on about half of the pairs but is both-correct on only 0.052 of them; the history-free controls flip at the same rate and reach 0.000-0.030.

condition	vanilla binding	early-layer arm 1/2/3 binding	arm CFC 1/2/3
unshuffled (reference)	0.442	0.963 / 0.988 / 0.967	0.875 / 0.958 / 0.892
reverse content (3,2,1)	0.438	0.667 / 0.800 / 0.754	0.558 / 0.733 / 0.683
rotate content (2,3,1)	0.498	0.652 / 0.660 / 0.640	0.542 / 0.525 / 0.508

Table A.23: Temporal-order perturbation of the frozen benchmark (480 episodes per cell). Panel content is permuted while the option labels keep their canonical positions. The unadapted model is essentially unchanged by either permutation, whereas the repaired early-layer arms lose 19–33 pp of binding accuracy and 0.16–0.45 CFC. This provides evidence that successful adaptation increases sensitivity to canonical panel order; because the permutations can alter episode semantics, the control is not treated as a standalone demonstration of correct temporal reasoning.

split (seed 111)	vanilla binding / CFC	full-pool QA binding / CFC	hinge arm binding / CFC
object count 4 (2-3 never trained)	0.379 / 0.000	0.875 / 0.692	0.875 / 0.700
compositional holdout	0.342 / 0.017	0.969 / 0.900	0.969 / 0.908
unseen shape	0.402 / 0.025	0.931 / 0.767	0.896 / 0.750

Table A.24: In-setting structural splits (single seed, exploratory). Each split removes part of the training distribution and re-evaluates the repaired arms against vanilla on the same episodes; the repaired arms retain most of their consistency, so the effect is not tied to the exact co-occurrences seen in training. These are structural generalisation checks inside the synthetic generation in-setting distribution, not out-of-domain transfer.

artifact	path	recorded reference
frozen benchmark (480 episodes / 120 pairs)	`data/generated/phase1_full/episodes.jsonl`	sha256 `15f1fbb5...`
full training pool (20,000 episodes / 3,000 pairs)	`data/generated/tba_train/episodes.jsonl`	sha256 `f4710d27...`
equal-size control pool (20,000 episodes, no pairs)	`data/generated/tba_train_astar/episodes.jsonl`	sha256 `ddc3f227...`
training configuration	`configs/experiments/tba_train.yaml`	sha256 `49a7ca5b...`
per-run record	`outputs/**/run_manifest.json`	seed, layers, steps, pool/config hashes, determinism flags
frozen per-episode predictions	`outputs/reports/mechanism/frozen_rows/`, `outputs/reports/repro_eval/`	per-episode rows + per-run JSON
CFC re-derivation	`results/appendix_evidence_audit.json`, `results/cfc_formalization.json`	CPU-only, read-only from the prediction files

Table A.25: Artifact inventory for the synthetic-benchmark results reported in this paper. Hash prefixes are the values recorded in the corresponding manifest or run record; the frozen benchmark hash was recomputed from the file for this appendix and matches the value recorded when the benchmark was frozen.

A.7 Real-Video Diagnostic Replication

To assess whether the recognition–binding dissociation is specific to synthetic scenes, we additionally constructed an automatic diagnostic from MOT17, MOT20, and KITTI tracking footage. The resulting set contains 384 episodes from 21 source sequences and was evaluated with the same letter-logit readout used for TIB-Bench. Among episodes that passed the measured perception and event controls, conditional owner binding remained low for Qwen3-VL-2B (0.333; 28/84), Qwen3-VL-4B (0.410; 16/39), and InternVL3.5-4B (0.258; 16/62), corresponding to conditional owner-error rates of 0.667, 0.590, and 0.742, respectively. Video-clustered 95% confidence intervals for owner error were [0.558, 0.770], [0.480, 0.824], and [0.588, 0.907]. We treat this analysis as an external replication of the behavioral recognition–binding dissociation on real-video imagery, rather than as evidence of cross-domain repair. The real-video diagnostic does not provide the pixel-identical counterfactual histories required by TIB-Bench CFC, and its storyboards differ from TIB-Bench in visual-token budget; we therefore reserve CFC and all repairability claims for the controlled synthetic benchmark.