BEYOND DISENTANGLEMENT: ADAPTIVE PARTIAL SHARING FOR SPATIAL REASONING CONTROL IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How should the feature subspaces used to control spatial reasoning in vision–language models be organized across relations: fully shared, relation-selective, or partially shared? We investigate this question through controlled spatial counterfactuals that connect changes in visual geometry to internal causal sensitivity and learned low-rank control. Our TriAxis-CF dataset comprises 300 factual/counterfactual image pairs, each reversing one spatial relation while preserving the other two, with geometry-derived labels and role-reversed questions. These paired contrasts provide controlled donors for activation interventions. Across three models evaluated under a shared protocol and an independent Qwen3 pilot, horizontal reasoning exhibits a repeated distribution of causal sensitivity across early visual regions, intermediate relation-text positions, and later final-input positions. Increasing visual state differences coexist with diminishing output effects, establishing a distinction between representation change and its causal influence on spatial judgments. A comparison of fully shared, relation-selective, and shared–private control in Qwen3 identifies adaptive partial sharing as an effective structure for balancing corrective intervention and answer preservation. Our SP-LoReFT combines a shared branch with relation-specific branches that adapt into partially overlapping subspaces. On held-out internal scenes, it attains the highest accuracy point estimate among the evaluated configurations, improving from 88.33% to 90.56% over shared rank-8 control. Break decreases from 6.18% to 3.89%, while Repair increases from 65.05% to 66.99%. These findings support adaptive partial sharing for spatial reasoning control, with its principal observed benefit arising from improved preservation of already-correct judgments.

1 INTRODUCTION

Depth, height, and horizontal reasoning concern different relationships among the same objects. Should the feature subspaces used to control these judgments in vision–language models (VLMs) be shared, relation-selective, or partially shared? We study low-dimensional hidden-state interventions that can alter spatial answers. Shared directions may retain common computations while interfering with other judgments; selective directions may avoid interference but exclude useful common information. Our question is which structure best balances correcting errors and preserving answers that are already correct.

Answering this question requires connecting control directions to internal responses to actual spatial changes. A question-answering (QA) margin gradient measures local decision sensitivity, whereas a factual/counterfactual state difference measures the response to a changed scene. Neither a large state difference nor a favorable score movement alone establishes useful spatial control. We therefore investigate counterfactual activation replacement before comparing learned control structures.

For this purpose, we construct **TriAxis-CF**: 300 rendered pairs that reverse one spatial relation while preserving the other two. Geometry-derived labels and reversed object-role questions identify which answers should change (Figure 1). These pairs provide both controlled QA and a scene-intervention

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.

reference for hidden-state replacement. The reference is not a pure semantic vector: occlusion and other rendering consequences may also change.

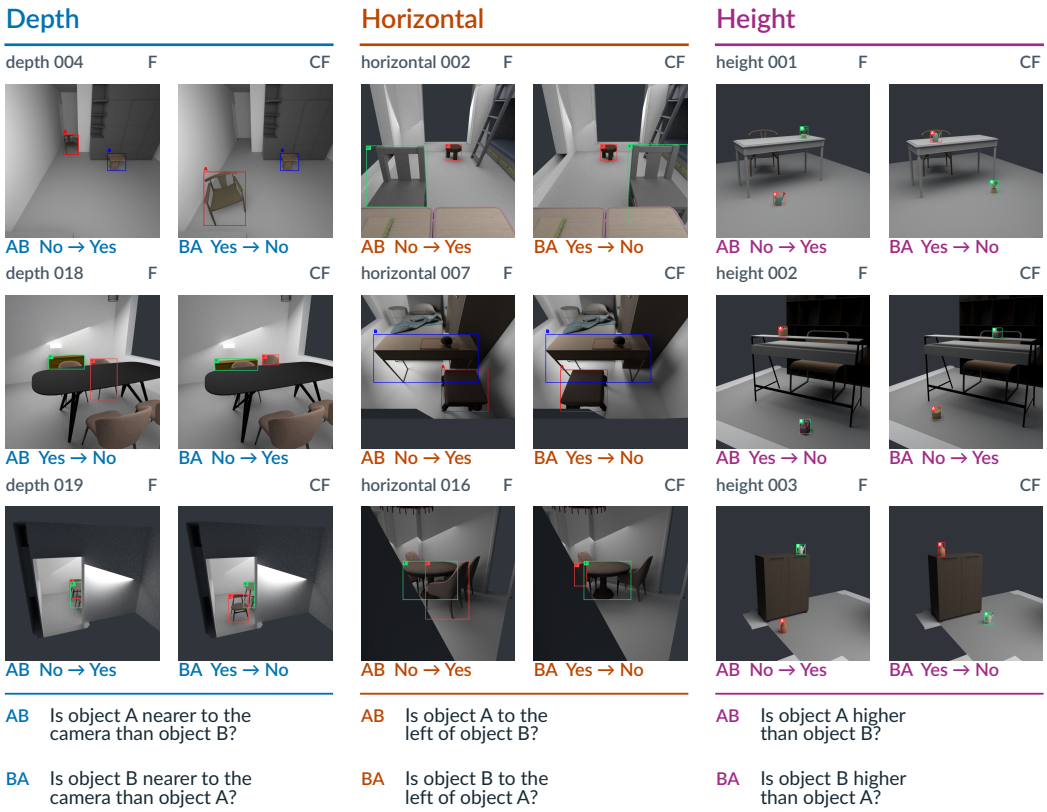


Figure 1: **Controlled spatial facts and linked questions.** Nine factual/counterfactual pairs span three intervention axes. Each column provides full AB/BA question stems; each pair reports all four target answers. Changing the image or the queried role reverses the truth, while changing both restores it. Answers are geometric ground truth, not model predictions.

Across three models evaluated on shared English inputs and an independent Chinese Qwen3 pilot, horizontal judgments exhibit repeated position-associated causal sensitivity. Visual state differences increase with layer depth while their output effects decline; donor-prediction transfer also differs from correct-relation transfer. These findings distinguish representation change from its downstream influence without identifying a complete causal pathway.

We return to control design through a common frozen-Qwen3 interface. LoReFT-MC is fully shared, RS-LoReFT emphasizes relation selectivity, and SP-LoReFT combines shared and relation-specific branches. RS-Frozen increases first-order selectivity by 187–$426\times$ but lowers QA utility. SP-Tune instead reaches 90.56% internal accuracy versus MC's 88.33%, with ten of twelve additional correct answers attributable to fewer destructive errors. The overall accuracy interval spans zero; the reduction in Break is clearer. Tuned shared/private branches overlap, supporting adaptive partial sharing as a structural prior rather than an identified orthogonal semantic decomposition.

Our contributions are: (1) a systematic comparison of spatial control sharing and the SP-LoReFT design (Sections 4–5); (2) cross-model counterfactual causal analysis separating state differences, prediction transfer, and truth transfer (Section 3); and (3) TriAxis-CF, connecting single-axis change, dual-axis preservation, and role-reversed QA to these investigations (Section 2).

2 TRIAXIS-CF: CONTROLLED SPATIAL COUNTERFACTUALS

TriAxis-CF makes each scene its own relational control: reversing a target ordering should change the corresponding answer while leaving the other two relations' answers stable. Factual (F) and counterfactual (CF) states form 300 pairs (100 per axis) and 600 original 768×768 images. House-disjoint training/validation/test partitions contain 210/45/45 pairs and 79/16/21 house identifiers.

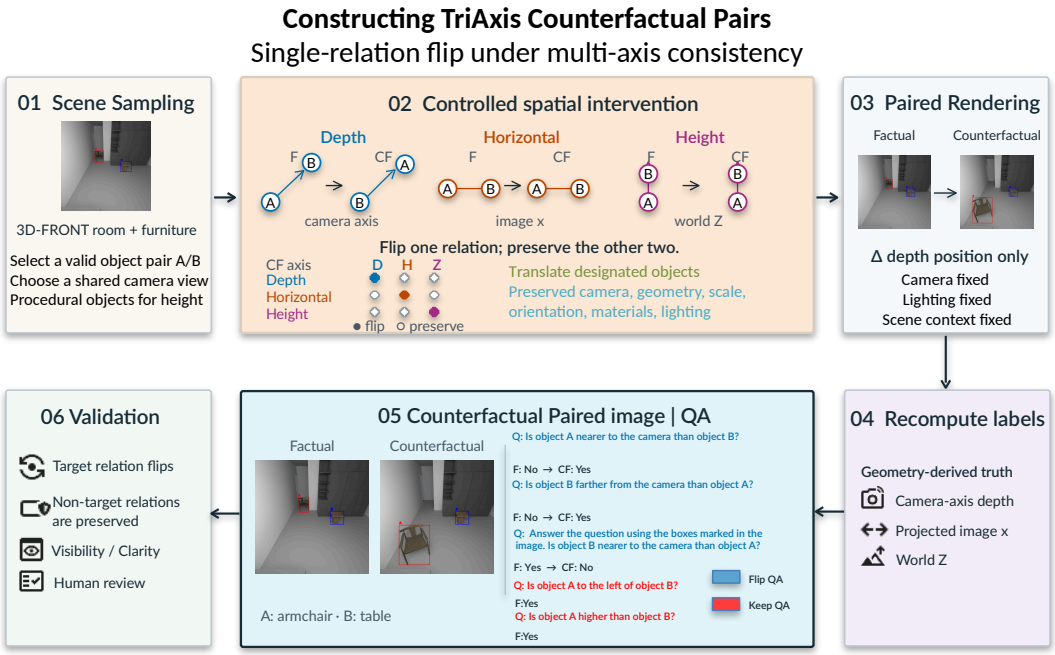


Figure 2: **TriAxis-CF construction and paired question answering.** The workflow selects scene objects, applies a controlled spatial intervention, and renders factual and counterfactual states under shared camera and lighting settings. Relations are recomputed from camera-axis depth, projected image coordinates, and world height before QA construction and validation. The paired-image panel illustrates the annotation interface.

Construction and truth. We render furnished MIDI-processed 3D-FRONT scenes Fu et al. (2021) with Blender Cycles. Depth and horizontal pairs use furniture; height pairs use procedural portable objects. Designated object translations preserve identity, mesh, scale, orientation, materials, camera, lighting, and non-intervened scene state. Both objects may move to supported placements. We recompute bounding-box-center relations in both states: depth uses camera-axis depth, horizontal uses projected image x, and height uses world Z. Height is not image-plane vertical order. All accepted pairs flip the target and preserve both non-target binary relations.

Paired QA and qualification. Each pair supplies two images $\times$ three relations $\times$ two role orders (AB/BA): 12 QA conditions per template. For truth $y \in \{-1, +1\}$, role reversal negates y; changing the image negates target truth and preserves non-target truth. Region-of-interest (ROI) boxes identify persistent objects A/B. Three binary templates and unmarked-image interfaces are available; the control study uses one template across all relations. Geometry labels, conservative clarity flags, and human acceptance are separate records: geometric validity need not imply perceptual clarity. All 300 pairs are accepted, but ambiguity flags and height-subset lateral regularities remain. Table 2 and Appendix A report complete view counts, qualification, provenance, and partition audits. Causal experiments retain their own English and Chinese snapshots rather than being pooled as identical inputs.

3 COUNTERFACTUAL CAUSAL ANALYSIS OF SPATIAL REASONING

We replace a recipient's decoder-output states with donor states at one layer and location group at a time. InternVL3.5-8B (Wang et al., 2025), Idefics3-8B-Llama3 (Laurençon et al., 2024), and DeepSeek-VL-7B-Chat (Lu et al., 2024) share an English snapshot; Qwen3-VL-8B-Instruct (Bai et al., 2025) supplies an independent corrected Chinese pilot. Each uses 300 pairs and six sites across 36/32/30/36 layers, respectively: 964,800 directed edges in total.

Sites and effects. **V-A/V-B** are the unions of object A/B's ROI token positions across both images; their donors change the image and keep the query fixed. **T-1/T-2** are the first/second object mentions, **T-R** is the relation expression, and **T-End** is the final input position before generation; their donors reverse query roles and keep the image fixed. Thus, location and donor type vary together; these are not demonstrated attention edges. ROI unions can contain background or other objects. Let m_r, m_d, m_p be recipient, donor, and patched yes-minus-no candidate log-score margins and y_d the donor truth. We measure

$$D_{\text{truth}} = y_d(m_p - m_r), \qquad D_{\text{pred}} = \text{sign}(m_d - m_r)(m_p - m_r). \tag{1}$$

Both use log-score units; raw magnitudes do not rank different models. ΔH is the change in donor-truth candidate accuracy, reported in percentage points (pp) in text. Strict transfer conditions on initially disagreeing predictions; empty denominators are NA. Definitions, coverage, and estimator details are in Appendices B and C.

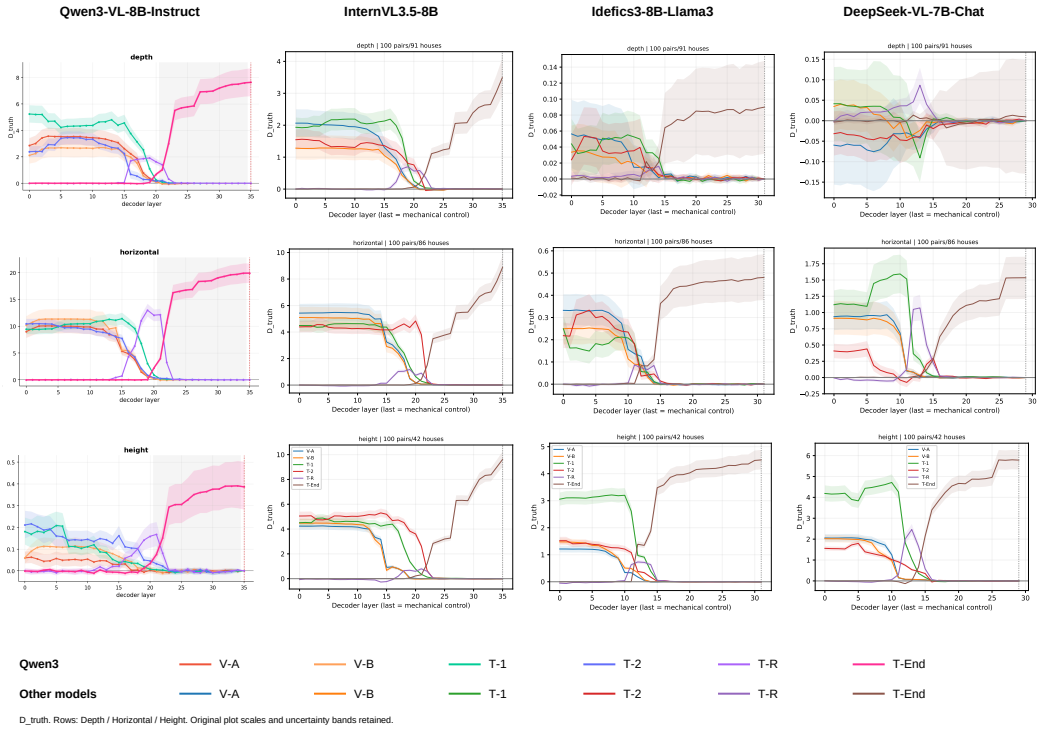


Figure 3: **Original experimental D_{truth} curves across all three relations.** Rows: depth, horizontal, height. Columns: Qwen3-VL-8B-Instruct, InternVL3.5-8B, Idefics3-8B-Llama3, DeepSeek-VL-7B-Chat. Panels retain source curves, scales, palettes, and 95% house-bootstrap bands (10,000 draws). Qwen3's original estimator is house-equal; the other models use pair-equal means. Numerical tables and the harmonized appendix curves use pair-equal estimates throughout. Qwen3 shading marks L21–34; dotted lines mark each model's final layer. All other metric plots are in Appendix D.

Layer-wise sensitivity. Horizontal reasoning shows early visual effects, intermediate relation-text peaks, and later final-input effects (Figure 3). T-R truth peaks occur at L19/L19/L11/L13 for Qwen3/InternVL3.5/Idefics3/DeepSeek-VL. The three English-protocol models also show the

pattern on height; depth does not support common positive truth transfer. Windows overlap and are descriptive. Final-layer non-T-End zeros and T-End copying are mechanical controls, excluded from window summaries (Table 4).

State change versus answer influence. Visual update norm and root-mean-square (RMS) increase across normalized layer-depth thirds while output effects generally decline. For horizontal InternVL3.5, early-to-late norm rises $686.4 \rightarrow 10140.0$ while D_{truth} falls $5.253 \rightarrow 0.008$. This distinguishes state difference from downstream causal influence; it does not show information loss.

Prediction versus truth transfer. DeepSeek depth visual patches have early $D_{\mathrm{pred}} = 0.319$ but $D_{\mathrm{truth}} = -0.019$. Horizontal T-2 has $D_{\mathrm{truth}} = 0.290$ but $\Delta H = -5.275$ pp (Table 5). Favorable score movement therefore need not improve discrete truth transfer. Complete curves for all three metrics, tasks, and models appear in Appendix D. These candidate-level findings do not establish four-model free-generation replication or a unique causal circuit.

4 Shared, Selective, and Shared–Private Control

Counterfactual motivation. For a fixed model, layer, prompt, and matched readout, $\Delta h_t^{\mathrm{cf}} = h(I_t^{\mathrm{cf}}) - h(I)$ describes the response to an external relation change; a QA gradient instead measures decision sensitivity. We hypothesize that useful common directions and relation-specific adjustments can coexist. The causal scans motivate testing this structure, not assuming that counterfactual geometry has already been decomposed. The following controllers use training-only gradient statistics; direct counterfactual-to-subspace alignment is not a reported result.

Common interface and fully shared control. All controllers use frozen Qwen3-VL-8B-Instruct and read native ROI states at zero-based block 3 with the neutral prompt *Describe the image.* Actions are added to the actual QA prompt's native ROI positions during prefill only. The queried relation selects a branch without using the answer. For $h \in \mathbb{R}^{4096}$, the LoReFT edit (Wu et al., 2024) is

$$\delta(h; R, W, b) = R^\top (Wh + b - Rh), \qquad RR^\top = I_r. \tag{2}$$

LoReFT-MC denotes our fully shared rank-8 baseline: one R, W, b across all three relations. MC is the experiment's baseline identifier.

Relation-selective control (RS). RS uses a rank-8 basis R_t for each relation with shared W, b. Training-only matrices A_t and B_t summarize target QA gradient energy and correctness-weighted non-target energy. Leading generalized eigenvectors of $A_t u = \lambda(B_t + \gamma I)u$, followed by Euclidean QR, initialize R_t. The selectivity ratio (SR) is $\mathrm{tr}(R_t A_t R_t^\top)/[\mathrm{tr}(R_t B_t R_t^\top) + 10^{-8}]$, a dimensionless first-order sensitivity proxy, not preservation accuracy. RS-Frozen fixes the bases; RS-Tune adapts them with cross-entropy (CE) plus a selectivity regularizer. Full matrix definitions are in Appendix E.

Shared–private control (SP). SP retains a common rank-4 branch and a rank-4 private branch for each relation:

$$\Delta h_t = \delta(h; R_s, W_s, b_s) + \delta(h; R_t^p, W_t^p, b_t^p). \tag{3}$$

The shared basis uses the top eigenvectors of the average trace-normalized A_t; private bases use projected relation-specific gradient geometry. Initially orthogonal shared/private branches are trained either with fixed bases (SP-Frozen) or with adaptable bases and the soft penalty $10^{-3} \frac{1}{3} \sum_t \|R_s (R_t^p)^\top\|_F^2$ (SP-Tune). Each basis is orthonormalized separately, permitting cross-branch overlap. Both start from the same untrained initialization. SP-Tune and shared LoReFT-r16 have exactly 131,088 trainable parameters; MC has 65,544. Matching parameters does not match active rank or routing. Training, selection, budgets, and regularizers are detailed in Appendix F.

5 Control Utility and the Repair–Stability Trade-off

Protocol and metrics. Training/validation/test contain 210/45/45 scene-disjoint families; test contains 540 QA (180 per relation) in 21 scene clusters. Geometry uses training inputs only; checkpoints

use validation only. All methods use greedy generation with an eight-token limit; invalid outputs count as errors (zero internally). B0 denotes the frozen, unmodified backbone. *Micro* is accuracy over 540 QA. *Repair* is the fraction of 103 B0 errors corrected; *Break* is the fraction of 437 B0-correct answers made wrong. *TwoOrder* requires both AB/BA answers for one image/relation to be correct (270 units), not all four answers across a pair. All table values are percentages; differences and confidence intervals (CIs) are in pp. Paired 95% percentile CIs use 2,000 scene-cluster bootstrap draws (seed 20260925), without multiplicity adjustment. This is a single-seed, single-backbone comparison; historical test reuse limits confirmation. Appendix H consolidates scales and abbreviations.

Method	Micro	Depth	Height	Horiz.	Repair	Break	TwoOrder
B0	80.93	72.78	77.78	92.22	0.00	0.00	68.52
LoReFT-MC	88.33	88.89	79.44	96.67	65.05	6.18	85.56
RS-Frozen	84.81	80.00	78.89	95.56	33.01	2.97	75.19
RS-Tune	87.96	87.22	80.00	96.67	57.28	4.81	83.33
LoReFT-r16	88.15	87.78	80.56	96.11	64.08	6.18	84.07
SP-Frozen	85.56	84.44	76.67	95.56	48.54	5.72	78.89
SP-Tune	90.56	90.00	85.56	96.11	66.99	3.89	88.15

Table 1: **Internal generated-QA control results (%).** Higher is better except Break. Horiz. means horizontal; MC/RS/SP are fully shared/relation-selective/shared–private control. Frozen fixes bases; Tune adapts them. Micro/Repair/Break/TwoOrder denominators are 540/103/437/270; each relation has 180 QA. B0's zero Repair/Break entries are identity comparisons. Shared r16 and SP-Tune each have 131,088 trainable parameters. Complete optimization details are in Appendix F.

Selectivity is not utility. RS-Frozen increases SR over MC by $193.5\times$, $187.4\times$, and $425.8\times$ for depth, height, and horizontal, yet loses 3.52 accuracy points (CI $[-5.81, -1.25]$). Its smaller Break accompanies fewer repairs. RS-Tune restores accuracy to 87.96%, while retaining much higher SR than MC. Cross-relation sensitivity is therefore not automatically harmful; useful common structure is a possible interpretation, not identified semantics.

Partial sharing and stability. SP-Tune attains the highest internal point estimate, $489/540$ versus MC's $477/540$, a $+2.22$-point difference (CI $[-0.64, +4.74]$). The count decomposition is $437 + 67 - 27 = 477$ for MC and $437 + 69 - 17 = 489$ for SP. Ten of twelve additional correct answers arise from fewer destructive errors. Break decreases by 2.29 pp (CI $[-4.07, -0.62]$); Repair increases by 1.94 pp (CI $[-8.93, +11.71]$). Height gains 6.11 pp (CI $[-0.47, +12.50]$), depth gains 1.11 pp, and horizontal loses 0.56 pp. The evidence is clearer for reduced damage than for an overall accuracy gain.

Capacity and adaptation. Parameter-matched shared r16 reaches 88.15%; SP's $+2.41$-point contrast has CI $[-0.96, +5.67]$. This controls total capacity, not rank, routing, or optimization. SP-Tune exceeds SP-Frozen by 5.00 pp (CI $[2.83, 7.21]$). Shared/private overlap $O_t = \|R_s(R_t^p)^\top\|_F$ rises from approximately zero to 0.720/0.388/0.167 across depth/height/horizontal; O_t is an unnormalized, dimensionless norm, not a percentage. Adaptation thus permits partial overlap rather than strict orthogonality. Frozen/Tune also differ in the soft penalty, so this is not an isolated trainability ablation. Action norms and external results are retained in Appendix F; the internal findings do not establish cross-domain superiority or measured counterfactual-subspace alignment.

6 Related Work

Spatial counterfactual evaluation. CLEVR provides rendered compositional diagnosis (Johnson et al., 2017); VSR and SpatialRGPT cover spatial judgments (Liu et al., 2023; Cheng et al., 2024). What'sUp tests controlled spatial contrasts (Kamath et al., 2023), and Winoground crosses role-sensitive images and captions (Thrush et al., 2022). TriAxis-CF jointly controls target change, non-target preservation, and query roles, supplying both QA and activation donors rather than claiming paired evaluation itself as novel.

Causal analysis and representation control. Interchange interventions and distributed alignment test hidden-state replacement (Geiger et al., 2022; 2024); causal tracing and patching examine output sensitivity (Meng et al., 2022; Zhang & Nanda, 2024). We apply these tools to controlled spatial contrasts. ReFT learns low-rank representation interventions (Wu et al., 2024); our contribution is the comparison of relation sharing and SP parameterization, not a new low-rank editing primitive.

7 Limitations and Conclusion

The causal evidence combines three shared-input models with an independent pilot; candidate-level effects are not free-generation replication. ROI patches lack matched background, token-count, and norm controls. The control study uses one seed and backbone; sample CIs do not capture training or selection uncertainty. SP's overall accuracy contrasts span zero, and external SURDS generation degrades (Appendix G). Synthetic scenes, clarity differences, and layout regularities limit dataset coverage. Branch semantics and counterfactual geometric alignment are not identified by utility alone.

Within this scope, TriAxis-CF connects controlled spatial change to internal causal investigation and learned control. The experiments distinguish state difference from output influence and show that stronger relation selectivity does not guarantee better QA. Adaptive shared–private control provides the best internal point estimate, chiefly through preserving already-correct answers. Spatial control benefits from balancing shared directions and relation-specific adaptation rather than maximizing separation.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093. Curran Associates, Inc., 2024. doi: 10.52202/079017-4293. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f38cb4cf9a5eaa92b3cfa481832719c6-Paper-Conference.pdf.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10913–10922, 2021. doi: 10.1109/ICCV48922.2021.01075.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 7324–7338, 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

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

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00566. URL http://dx.doi.org/10.1162/tacl_a_00566.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. URL https://arxiv.org/abs/2403.05525.

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.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962. Curran Associates, Inc., 2024. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/75008a0fba53bf13b0bb3b7bff986e0e-Paper-Conference.pdf.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

A TRIAXIS-CF: CONSTRUCTION, INTERFACES, AND AUDIT

Snapshot and counting unit. All dataset statistics in this appendix describe the accepted September 25, 2026 snapshot. The independent construction unit is an image pair, not an individual question or a template variant. Table 2 consolidates the corpus and QA statistics. The compatibility view

Corpus and QA view	**Train**	**Val.**	**Test**	**Total**
Image pairs	210	45	45	300
per intervention axis	70	15	15	100
House / scene-group identifiers	79	16	21	116
`native`, per response format	840	180	180	1,200
`cross`, per response format	2,520	540	540	3,600
`cross_filtered`, per format	1,384	296	300	1,980
`native_roi`, three templates	2,520	540	540	3,600
`cross_all_roi`, three templates	7,560	1,620	1,620	10,800
`cross_roi`, filtered	1,704	384	384	2,472

Both-state qualification	**Queried relation**			**Sum**
Intervention (100 pairs each)	Depth	Horizontal	Height	
Depth	32	94	1	127
Horizontal	59	100	0	159
Height	10	99	100	209

Table 2: **TriAxis-CF statistics.** The corpus contains 600 original 768×768 images and their ROI-marked counterparts. Upper block: QA views overlap; each description-based view has both multiple-choice and sentence-answer formats. Lower block: pairs passing the queried relation's clarity screen in both images, out of 100 per cell. The 495 retained pair–relation units yield 1,980 role-crossed questions per format. These are coverage counts, not model accuracies; raw labels remain available for all pairs.

`cross_roi_main` repeats the counts of `cross_roi`; `height_original` contains 420/90/90 train/validation/test rows (600 total). The 600 current ROI images are annotated derivatives of the 600 original renders, not additional scenes. The package also retains four obsolete marked images that are not referenced by current QA. Its 33,504 stored QA rows include overlapping views, response formats, templates, and compatibility copies; this total is not a count of independent examples.

Geometry and intervention semantics. Object relations are defined using 3D bounding-box centers and the recorded camera: depth compares positive camera-axis depths, horizontal order compares projected image x coordinates, and height compares world Z coordinates. Scene-normalized distances used in quality checks are not physical meters. Object identity is fixed across the two images; AB/BA changes which object is queried relative to the other, without relabelling the objects in the scene. Depth and horizontal subsets use original scene furniture. The height subset uses procedural portable objects in furnished room contexts; its construction checks support and collision at the two endpoints, not the dynamics of moving between them. Both objects can move, including lateral displacement needed to reach supported positions, while the non-target orderings remain unchanged.

Question interfaces and qualification. Unmarked-image questions use invariant appearance descriptions, with one wording per relation and both role orders. Multiple-choice questions append two directional options to the same question stem used by the sentence-answer interface. ROI questions instead refer to visibly labelled A/B boxes and use three binary question templates. Their `native_roi` and `cross_all_roi` views cover all pairs. The filtered `cross_roi` and `cross_roi_main` views require both depth and horizontal qualification in both images for depth/horizontal interventions, and all three relations for height interventions. They therefore do not share the relation-wise selection rule of `cross_filtered`. Despite its name, `height_original` contains questions on all three relations; it uses unmarked images, bounding boxes in metadata, and directional word answers. It is retained for compatibility, not treated as a complete role-crossed unit.

For the full role-crossed binary views, each image contributes opposite truth values under AB and BA. This gives 1,800 yes and 1,800 no answers in `native_roi`, and 5,400 of each in `cross_all_roi`, without removing visual layout correlations. Factual target directions are nearer in $53/100$ depth pairs, left in $44/100$ horizontal pairs, and higher in $51/100$ height pairs. Scene groups contribute one to seven pairs (median two), so questions are dependent observations within scenes. In particular, object A is left of B in all 200 height images. Raw geometric truth and conservative perceptual qualification must therefore be reported separately. Qualification flags include small margins, overlapping visible depth spans, vertical interval overlap, and conflicts between world height and visible vertical order. Human acceptance closes the review workflow without changing these flags or making an ambiguous relation eligible for a filtered view.

Audit, provenance, and partition integrity. The package retains per-pair geometry, labels, confidence flags, stable referents, split membership, construction code, and repair records with image/scene hashes. Repairs synchronize images, boxes, relations, and affected QA rather than retaining answers derived from superseded geometry. The final acceptance record covers all 300 pairs; it records dataset-wide approval, not an independent multi-annotator agreement study. The latest construction audit reports no cross-partition overlap in pair, house, or scene-group identifiers, source scene files, identical RGB images, or near duplicates at its pHash Hamming-distance threshold of six. These checks do not establish the absence of all visual similarity or pretraining overlap. The original scene assets are subject to the upstream 3D-FRONT access conditions and are not bundled; reconstruction requires obtaining them separately.

Experimental version boundary. The control study uses fixed audited inputs with 21 test scene clusters. The causal English snapshot and corrected Chinese pilot are documented in Appendix B; their results are not pooled as a same-input four-model experiment. Snapshot identity is established through input records, not inferred from matching numbers of QA.

Scope of the contribution. TriAxis-CF supports within-pair tests of sensitivity and preservation with geometry-derived labels, explicit object roles, and inspectable ambiguity. It does not cover arbitrary spatial predicates, unseen hypothetical outcomes, physical dynamics, or unconstrained object localization. Its limited number of scene groups, synthetic rendering, procedural height objects, and fixed height-subset lateral layout limit generalization. These restrictions make it a focused diagnostic resource rather than a replacement for broad natural-image spatial benchmarks.

B Metrics and Scope of the Cross-Model Causal Analysis

Inputs and intervention units. The three shared-protocol models use the English EXP-010 snapshot; the Qwen3 pilot uses an earlier corrected Chinese snapshot. Image replacement and question-role reversal define directed donor–recipient edges in the four-cell unit. A location name identifies where a donor state is installed in the recipient, not a demonstrated attention edge. V-A/V-B use the union of the corresponding ROI token indices across both images. T-1/T-2 identify the first and second object mentions, T-R the relation expression, and T-End the final input position. Position windows can overlap. The final layer is excluded from the reported summaries because its null and donor-copy outcomes can follow mechanically from the intervention location.

Continuous effects. For candidate margin $m = \log P(\text{yes} \mid X) - \log P(\text{no} \mid X)$, using the experiment's fixed candidate strings and sequence-scoring convention, let $\Delta = m_p - m_r$. Equation 1 gives the truth- and prediction-directed effects. D_{pred} is undefined when $m_d = m_r$; this exclusion does not remove the edge from D_{truth}. Donor-score proximity is

$$E_{\text{match}} = |m_r - m_d| - |m_p - m_d|.$$

Positive D_{pred} can coexist with negative E_{match} when a patch overshoots the donor score. Neither quantity alone establishes semantic correctness. For an edited matrix with n token positions and hidden dimension d, report $U = \|H'_S - H_S\|_F$ and $U_{\text{RMS}} = U/\sqrt{nd}$.

Discrete outcomes and denominators. Donor match is $\mathbb{1}[\hat{a}_p = \hat{a}_d]$, while clean donor match uses $\hat{a}_r$ in place of $\hat{a}_p$. Prediction change is $\mathbb{1}[\hat{a}_p \neq \hat{a}_r]$. Strict transfer measures donor match only

on initially disagreeing clean donor–recipient predictions. Its denominator can be empty, in which case the value is NA. The donor-truth hit change is

$$\Delta H = \mathbb{E}[\mathbb{1}[\hat{a}_p = a_d^*] - \mathbb{1}[\hat{a}_r = a_d^*]] \,.$$

For valid binary candidates and opposite donor/recipient truths, $\Delta \mathrm{Acc}_r = -\Delta H$; this identity need not hold with invalid generated outputs. Candidate classifications and parsed generations are therefore reported separately.

Aggregation and corrected pilot estimates. Within a fixed task, layer, and position, aggregate eligible edges within pairs before taking the equal-weight mean across pairs, retaining any prompt averaging defined by the source experiment. Conditional rates are pair-macro rather than pooled edge-micro ratios; empty conditional denominators remain NA. Qwen3's original D_{truth} and strict-transfer summaries used house weighting, while its other ten metrics used pair weighting. The current summary uses the corrected pair-weighted estimates for all metrics. The old two confidence intervals are not reused. Figure 4 recomputes pointwise house-cluster intervals for the three primary metrics of all four models under the equal-pair estimator; descriptive windows are not converted into new significance tests. Visual summaries give V-A/V-B equal weight. Early/middle/late summaries divide normalized layer depth $\ell/(L-1)$ into thirds and exclude layer $L-1$. The 90% plateau and sustained-transfer thresholds are post hoc descriptions, not preregistered tests.

Limits of replication. The shared horizontal pattern and the additional height evidence concern candidate-level intervention sensitivity. Depth is a counterexample to a universal positive truth-transfer claim. The different pilot language and snapshot preclude a matched four-model comparison; raw norm or margin scales are not cross-model effect-size rankings. ROI-union patches are not object-exclusive, and the all-layer scan lacks matched background, token-count, and update-norm controls. Incomplete visual-ROI generation checks and the Idefics3 strict-generation invalid rate prevent a four-model free-generation claim. These limitations are not remedied by a large number of intervention edges.

C Causal Sites and Numerical Summaries

Code	Recipient positions replaced	Donor contrast
V-A	Union of object A's visual ROI token indices in F and CF	Image changes; query fixed
V-B	Union of object B's visual ROI token indices in F and CF	Image changes; query fixed
T-1	First object mention in the question (A in AB; B in BA)	Role order reverses; image fixed
T-2	Second object mention (B in AB; A in BA)	Role order reverses; image fixed
T-R	Tokens representing the queried spatial relation	Role order reverses; image fixed
T-End	Final input position before answer generation	Role order reverses; image fixed

Table 3: **Patch locations and donor contrasts.** V denotes visual positions and T denotes text positions. A/B are persistent object identities, whereas 1/2 are question-order roles. T-End is an input position, not a generated answer token. All four models use the visual versus query contrasts shown here; position and donor type therefore vary together. ROI unions may include background or other objects. The codes identify intervention sites, not edges of a discovered information-flow graph.

D Complete Cross-Model Causal Curves

Together with Figure 4, Figures 5 and 6 display every layer, relation, and patch site for D_{truth}, D_{pred}, and ΔH in all four models. All 12 recorded metrics are independently aggregated in the

(a) Horizontal: layer windows and peaks Model	Visual plateau	T-R truth peak	T-End onset
Qwen3-VL-8B-Instruct (Bai et al., 2025)	L1–12	L19	L23
InternVL3.5-8B (Wang et al., 2025)	L0–14	L19	L31
Idefics3-8B-Llama3 (Laurençon et al., 2024)	L0–7	L11	L21
DeepSeek-VL-7B-Chat (Lu et al., 2024)	L0–8	L13	L26

(b) Horizontal: early → late visual values Model	Norm	RMS	D_{truth}	Early ΔH (pp)
Qwen3-VL-8B-Instruct (Bai et al., 2025)	$740.4 \to 5261.9$	$0.856 \to 5.612$	$10.328 \to -0.004$	$+45.01$
InternVL3.5-8B (Wang et al., 2025)	$686.4 \to 10140.0$	$0.569 \to 7.682$	$5.253 \to 0.008$	$+47.57$
Idefics3-8B-Llama3 (Laurençon et al., 2024)	$80.3 \to 298.5$	$0.081 \to 0.296$	$0.269 \to \approx 0$	$+5.90$
DeepSeek-VL-7B-Chat (Lu et al., 2024)	$1514.5 \to 3294.7$	$1.839 \to 4.030$	$0.913 \to 0.002$	$+5.76$

Table 4: **Repeated horizontal sensitivity with different behavioral efficacy.** (a) Descriptive, post hoc layer summaries, excluding the final layer; T-End onset denotes strict transfer sustained at $\geq 90\%$ through the penultimate layer. (b) Early/late thirds of normalized layer depth $\ell/(L-1)$, again excluding the final layer; V-A and V-B receive equal weight. Values describe within-model changes, not rankings of raw magnitudes across architectures or languages. Qwen3 is an independent Chinese-snapshot pilot; the other three share the English protocol. Pair-weighted point estimates are shown without a new significance claim.

Model	Task / site	D_{truth}	D_{pred}	ΔH (pp)
DeepSeek-VL-7B-Chat (Lu et al., 2024)	Depth / visual	-0.019	0.319	-3.575
DeepSeek-VL-7B-Chat (Lu et al., 2024)	Horizontal / T-2	0.290	0.997	-5.275

Table 5: **Continuous and discrete transfer can diverge.** All three primary metrics, recomputed from raw DeepSeek-VL-7B-Chat edges. Early-layer summaries exclude the final layer and average the first third of normalized layer depth; visual averages V-A/V-B equally. Depth shows prediction-directed transfer without net truth-directed transfer. Horizontal T-2 shows a positive mean truth-directed margin change but a negative net discrete answer change. Full layer curves for both cases are provided in Figure 3 and Appendix D.

accompanying source tables (28,944 layer/site/metric entries); additional Qwen3 diagnostics below illustrate agreement, strict transfer, and state-change magnitude. Each task/model has 100 pairs before metric-specific conditional exclusions. Qwen3 uses its independent Chinese snapshot; the other three models share English inputs. Equal sample counts do not make these snapshots identical.

Resampling houses with replacement reweights their constituent pairs and recomputes the equal-pair estimator. We use 2,000 draws and seed 20260926, retaining all layers/sites of a sampled pair together. Bands are pointwise percentile intervals, not simultaneous bands or corrected layer-selection tests. The additional nine Qwen3 metrics are descriptive point estimates. All layers are displayed, including each model's final layer; interpretation of high-effect windows excludes the mechanically constrained final layer. All partitions are pooled for descriptive scans, rather than presented as a held-out control evaluation. Conditional empty sets remain NA.

E Detailed Control Formulation

E.1 A common read/write interface

The backbone is frozen Qwen3-VL-8B-Instruct. All compared controllers read native ROI states from the neutral prompt *Describe the image.* at zero-based decoder block 3. Their actions are added at the corresponding native ROI positions of the actual QA prompt during prefill only; other positions and decoding steps are not edited. The relation t is known from the question and selects

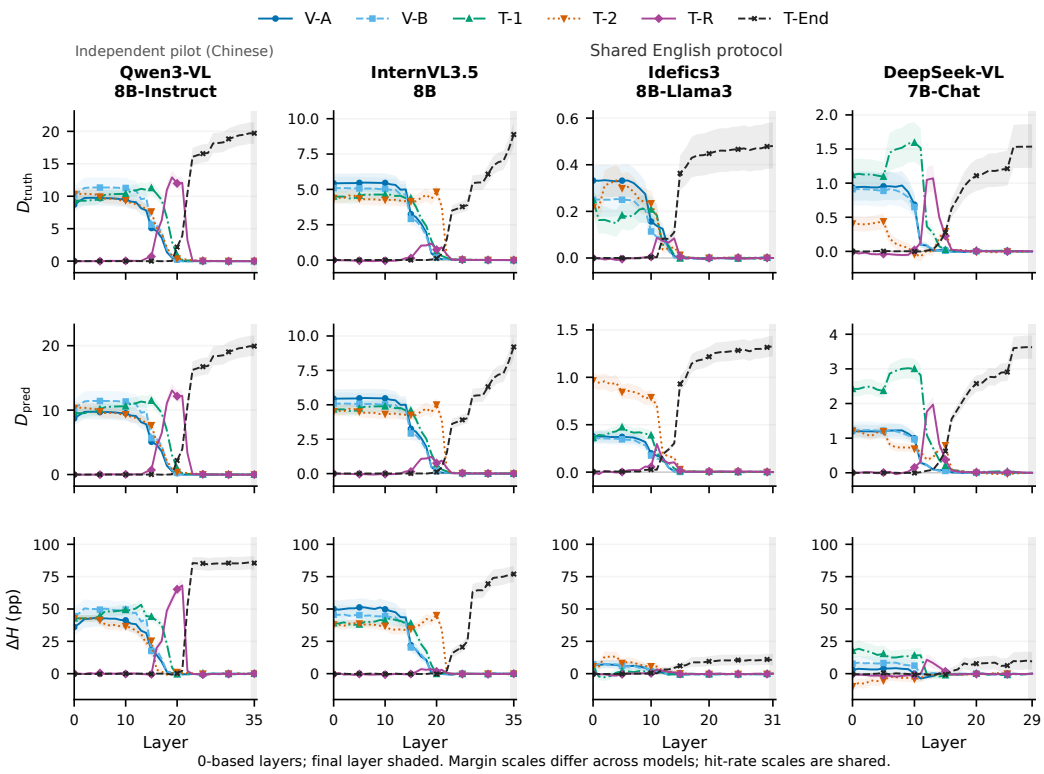


Figure 4: **Horizontal curves under the harmonized pair-weighted estimator.** Unlike the source-output composition in Figure 3, these curves are recomputed from raw edges, with 2,000-draw pointwise house-cluster intervals and hit-rate changes in percentage points. This preserves the corrected numerical reference alongside the original experimental plots.

any relation-specific branch without using the answer. This neutral-read, native-write interface is fixed across the comparison; the control-space structure varies.

For a column-vector hidden state $h \in \mathbb{R}^{4096}$, write

$$\delta(h; R, W, b) = R^\top (Wh + b - Rh), \qquad RR^\top = I_r. \tag{4}$$

The additive action belongs to the row span of R. The backbone stays fixed; we optimize only the controller. A relation-conditioned action can be useful without its basis being an object-exclusive or native causal representation.

E.2 Fully shared control: LoReFT-MC

MC uses one rank-8 basis and one W, b for all relations: $\mathcal{S}_d = \mathcal{S}_h = \mathcal{S}_l$ and $\Delta h = \delta(h; R, W, b)$. This parameterization has 65,544 trainable parameters. It preserves any useful common directions but does not explicitly separate task-specific effects. Whether sharing produces destructive errors is assessed by generated answers, not inferred from overlap alone.

E.3 Relation-selective control: RS-LoReFT

At zero control, compute the gradient $G_q \in \mathbb{R}^{N_q \times 4096}$ of the truth-directed candidate margin with respect to native ROI perturbations. For each relation, estimate training-only sensitivity matrices

$$A_t = \frac{1}{|Q_t|} \sum_{q \in Q_t} \frac{G_q^\top G_q}{N_q}, \qquad B_t = \sum_{j \neq t} \frac{1}{|Q_j|} \sum_{q \in Q_j} c_q^0 \frac{G_q^\top G_q}{N_q}, \tag{5}$$

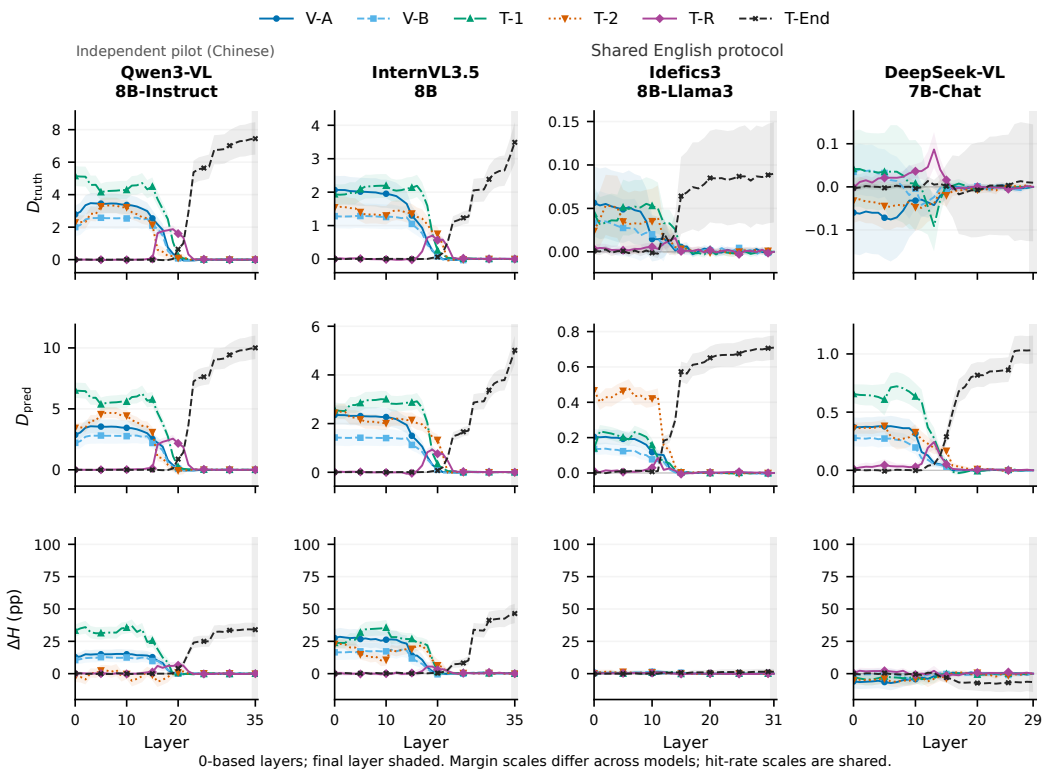


Figure 5: **Complete depth layer curves.** All four models, six sites, and three primary metrics. Columns and estimator follow Figure 4; raw margin scales are model-specific and hit-rate scales are shared. The varied signs and magnitudes show why the horizontal pattern is not a universal positive depth-transfer result.

where c_q^0 indicates that the frozen baseline originally generated the correct answer. Keeping the full $|Q_j|$ denominator avoids reweighting to only the correct subset. Token gradients are not averaged before forming $G_q^\top G_q$.

RS selects eight leading generalized eigenvectors of $A_t u = \lambda(B_t + \gamma I)u$, with $\gamma = 10^{-3}\,\mathrm{tr}(B_t)/4096$, then applies Euclidean QR to obtain the actual LoReFT basis R_t. The shared W, b remain common across relations. Its measured selectivity is

$$\mathrm{SR}_t = \frac{\mathrm{tr}(R_t A_t R_t^\top)}{\mathrm{tr}(R_t B_t R_t^\top) + 10^{-8}}. \tag{6}$$

This is a training-distribution first-order energy ratio, not test-set preservation accuracy. The generalized-eigenvector construction followed by QR is not claimed to globally optimize every Euclidean trace-ratio objective. RS-Frozen fixes the bases; RS-Tune adapts them with cross-entropy plus a selectivity regularizer (Appendix F).

E.4 Shared–private control: SP-LoReFT

RS tests whether emphasizing relation selectivity helps. Its utility shortfall is consistent with the hypothesis in Section 4: suppression of cross-relation-sensitive directions may remove useful common variation. SP instead allows common directions and relation-specific residuals to coexist. We parameterize

$$\Delta h_t = \delta(h; R_s, W_s, b_s) + \delta(h; R_t^p, W_t^p, b_t^p), \qquad \mathcal{S}_t = \mathcal{S}_s + \mathcal{S}_t^p. \tag{7}$$

The sum denotes available shared/private components, not an identified orthogonal semantic decomposition. Each branch has rank 4, so each QA uses a total nominal rank of 8. The shared branch has one set of parameters; the three private branches have independent parameters.

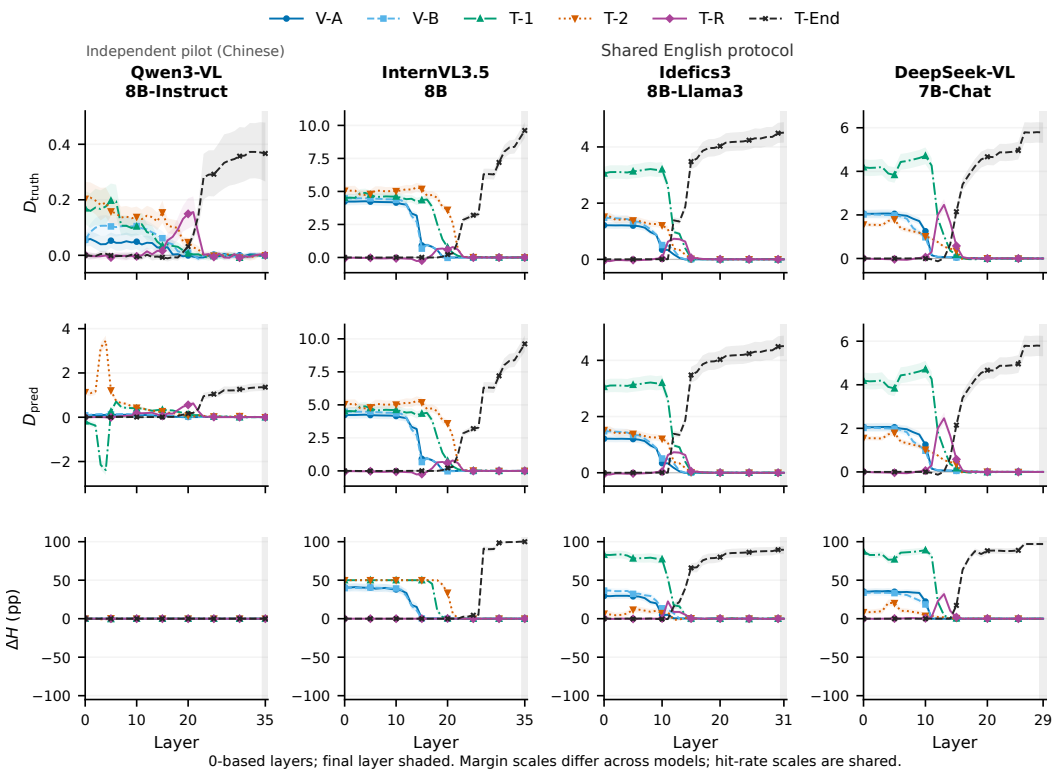


Figure 6: **Complete height layer curves.** Same model/site definitions and estimator as Figure 4. Signed effects are retained, including negative values. Qwen3 visual-region answer change is zero throughout, despite nonzero state updates. Qwen3 is an independent pilot, not a matched-input fourth replication.

Initialize the shared basis from the top four eigenvectors of $A_s = \frac{1}{3}\sum_t A_t/(\mathrm{tr}(A_t) + 10^{-8})$. With shared column basis U_s, let $P_\perp = I - U_s U_s^\top$. Private bases use projected target matrices $P_\perp A_t P_\perp$ against $P_\perp(\sum_{j\neq t} A_j)P_\perp$, followed by projection and QR. Unlike RS, this private initialization includes all other-relation QA, without correctness weighting. Shared/private bases are initially orthogonal.

SP-Frozen trains only branch W, b using cross-entropy. SP-Tune starts from the same untrained initialization, adapts all branch parameters, and adds $10^{-3}\,\frac{1}{3}\sum_t \|R_s(R_t^p)^\top\|_F^2$ to the loss. Each four-row basis is retracted separately; neither the combined eight-row basis nor different private branches are forced to remain mutually orthogonal. SP is consequently a prior for *adaptive partial sharing*. Frozen-versus-Tune changes both basis trainability and the soft penalty, so it does not isolate their individual causal contributions.

E.5 Capacity control and model selection

SP-Tune has 131,088 globally trainable parameters, versus 65,544 for MC; matching the per-QA nominal rank does not match capacity. We therefore also train an entirely shared LoReFT-r16 with exactly 131,088 parameters. This matches global parameter count, but not active rank or FLOPs: r16 uses rank 16 for every QA, whereas SP activates shared rank 4 plus private rank 4. Initialization, routing, constraints, and parameter reuse still differ.

The new models use AdamW, learning rate 10^{-3}, zero weight decay, seed 42, and answer-token/EOS cross-entropy. Checkpoints are selected only on validation, lexicographically by macro accuracy, micro accuracy, and then negative Break, with earlier checkpoints winning exact ties. Geometry is estimated from training QA only. MC is a reused run selected by macro/micro without Break; total

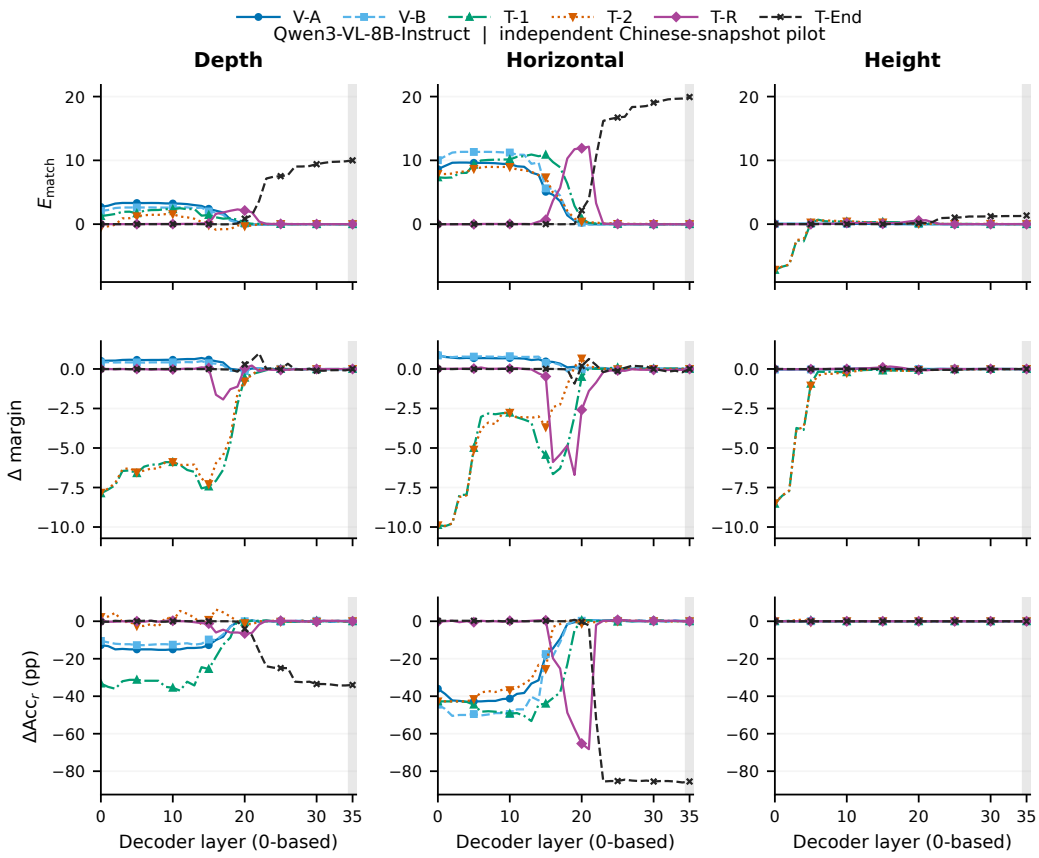


Figure 7: **Donor-score proximity and signed margin/accuracy changes.** E_{match} is the reduction in distance to the donor margin; $\Delta = m_p - m_r$ retains the original yes/no sign; recipient accuracy change is exactly the negative of donor-truth hit change for these opposite-truth binary edges. Same groups, layer indices, and pair weighting as Figure 4. Values are descriptive means without bands.

training exposure is not identical across every model. Appendix F records budgets and selection differences.

F Subspace-Control Protocol and Additional Evidence

Training-only geometry. EXP-016 estimates each A_t from 840 actual training QA gradients, for 2,520 QA in total, using per-image token normalization and FP64 matrix operations. RS uses correctness-weighted non-target covariances; SP reuses A_t but constructs its private competing covariance from all non-target QA. Cholesky whitening solves the regularized generalized eigenproblem. Projection, sign fixing, and Euclidean QR produce the bases actually used by the forward pass. The test set is excluded from these statistics. Input audits verify zero scene and image-hash overlap across train, validation, and test.

Optimization and budgets. RS-Frozen trains only shared W, b (32,776 parameters); RS-Tune trains three rank-8 bases and shared W, b (131,080). MC trains 65,544 parameters. SP-Frozen trains 65,552 branch W, b parameters; SP-Tune and LoReFT-r16 each train 131,088. SP stores four rank-4 branches but activates only the shared branch and one relation-private branch per QA. Parameter count, active rank, routing, and parameter reuse are distinct comparison axes.

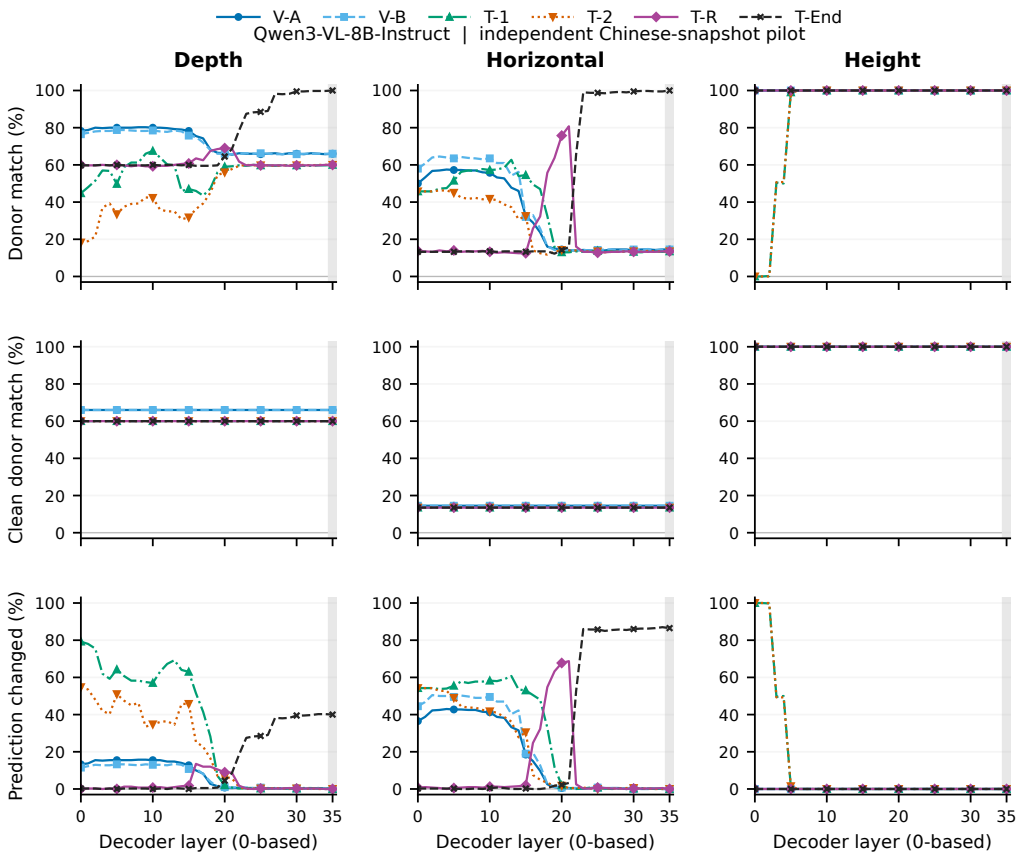


Figure 8: **Agreement versus actual answer change.** Donor match, clean donor match, and prediction-change percentages distinguish agreement induced by patching from agreement already present beforehand. Qwen3 height visual-region patches have donor match of 100% and no prediction change; early object-mention patches can instead break clean agreement. These effects are retained rather than collapsed across locations. All six locations are retained even when curves coincide.

All new runs use seed 42, AdamW with betas $(0.9, 0.999)$, epsilon 10^{-8}, learning rate 10^{-3}, and zero weight decay. Each epoch exposes 210 families, one image per family with image direction alternating by epoch, and six QA per image. Six families accumulate into an update, giving 35 updates per epoch. Validation uses all 540 QA each epoch. The minimum is five epochs, the maximum 100; RS/SP paired runs stop jointly when both have three epochs without improvement in their validation selection key. Each retains its own best checkpoint. Validation CE plateaus separately halve the learning rate, with minimum 10^{-5}. Early stopping is a validation plateau criterion, not a claim of global convergence.

RS-Tune adds $0.01\mathcal{L}_{\text{geo}}$ to CE, where

$$\mathcal{L}_{\text{geo}} = -\frac{1}{3}\sum_{t} \log \frac{T_t + 10^{-8}}{I_t + 10^{-8}}, \quad T_t = \operatorname{tr}(R_t A_t R_t^\top), \quad I_t = \operatorname{tr}(R_t B_t R_t^\top).$$

The numerator epsilon differs slightly from the displayed SR definition. Each basis undergoes QR retraction after an update. SP-Tune instead adds the 0.001 overlap penalty in Section 4; SP-Frozen uses CE only. Both SP runs start from the same untrained geometry and W, b initialization, not from the trained MC or trained SP-Frozen checkpoint.

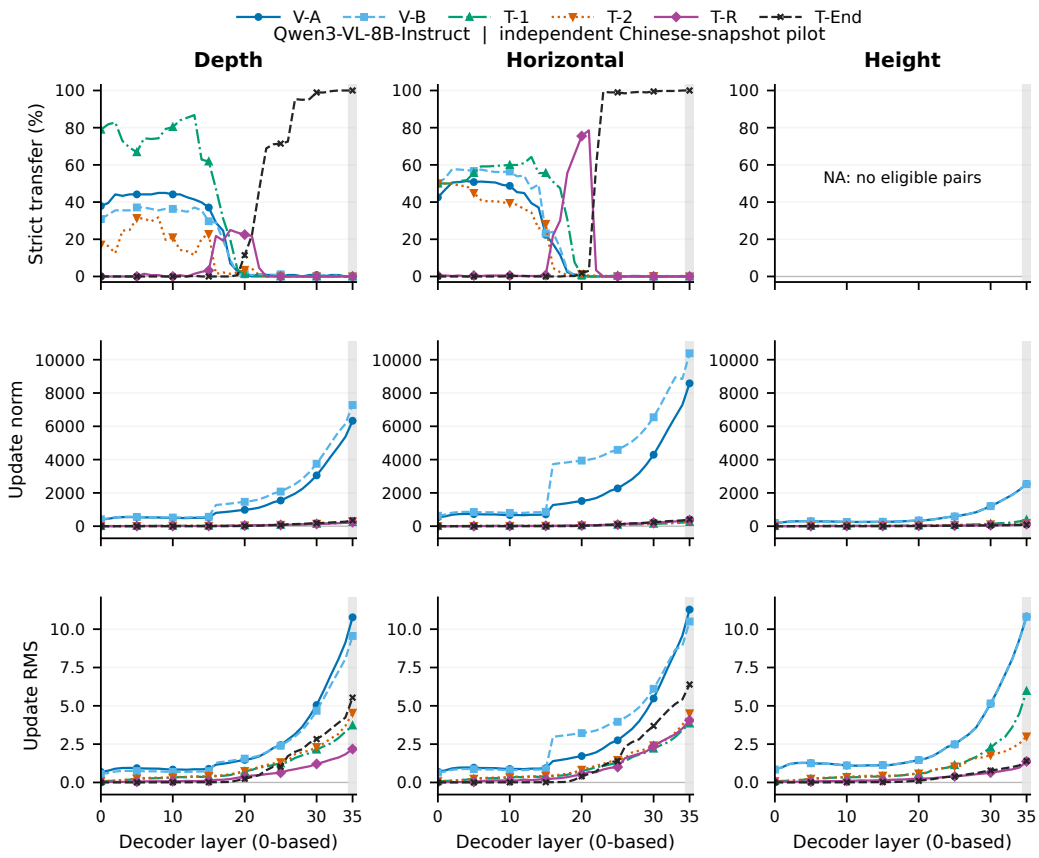


Figure 9: **Conditional transfer and the magnitude of state replacement.** Strict transfer averages within-pair rates conditional on initially disagreeing predictions; height is NA because no eligible pairs exist. It is not a zero success rate. Update norm and RMS measure the donor–recipient state difference at the patched positions. The norm/RMS panels use common linear scales across relations, without rescaling each curve to its own peak.

Method	Selected epoch	Stopping epoch	Total updates
MC (reused EXP-014)	10	13	455
RS-Frozen	7	12	420
RS-Tune	9	12	420
SP-Frozen	8	11	385
SP-Tune	6	11	385
LoReFT-r16	5	8	280

MC was selected by macro/micro accuracy, whereas the new runs add negative Break as a third lexicographic criterion. Thus, MC is a same-data reference, not a fully exposure- and selection-matched retraining. r16 and SP-Tune have identical family/direction exposure through r16's stopping epoch: 1,680 family and 10,080 QA exposures. The best epochs differ. Every checkpoint is frozen before final test evaluation; the comparison nevertheless follows iterative experiments on the same historical test set rather than an independent confirmatory replication.

Selectivity is a proxy, not preservation. The following SR values are calculated on training sensitivity matrices:

Relation	MC	RS-Frozen	RS-Tune
Depth	1.61606	312.720	39.3717
Height	0.31971	59.9052	10.9423
Horizontal	0.619705	263.849	71.6462

The RS initialization remains fixed in RS-Frozen. Tuning reduces extreme selectivity while recovering much of the downstream accuracy. Because the ratio is an energy proxy derived near zero control, it does not directly measure non-target answer preservation under the learned finite intervention.

Overlap and action scale. Shared/private overlap is the unnormalized Frobenius quantity $O_t = \|R_s(R_t^p)^\top\|_F$. It is near 2×10^{-9} at initialization and reaches 0.719957/0.387777/0.167295 in the selected SP-Tune checkpoint. Action norms are full-ROI Frobenius norms, computed once for each unique context/relation, not duplicated for AB/BA. There are 90 such internal contexts per relation. No across-seed uncertainty is attached to these means. Nonzero branch actions show that they are written, not that each branch is independently causally necessary; dedicated branch ablations are absent. The data support a useful partially overlapping parameterization, not an orthogonal decomposition of true spatial semantics.

G External Subspace Evaluation: A Boundary, Not a Main Claim

External tests never select checkpoints and do not support the main internal claim. We retain their adverse outcomes rather than treating internal point gains as evidence of cross-domain superiority. SpatialRGPT uses 638 QA in 200 clusters; SURDS uses 117 QA in 40 clusters, with only AB-order questions, so TwoOrder is undefined there.

	SpatialRGPT		SURDS	
Method	Correct / 638	Acc. (%)	Correct / 117	Acc. (%)
MC	588	92.16	104	88.89
SP-Frozen	582	91.22	77	65.81
SP-Tune	583	91.38	86	73.50
LoReFT-r16	592	92.79	105	89.74

On SpatialRGPT, SP-Tune minus MC is -0.78 points with interval $[-2.38, +0.80]$. On SURDS, the corresponding difference is -15.38 points with interval $[-20.54, -10.09]$. SURDS invalid outputs are 3/117 for MC, 32/117 for SP-Frozen, 20/117 for SP-Tune, and 1/117 for r16; all remain errors. Candidate-score accuracy is substantially higher for SP (105/117 for both variants), but is a different endpoint and does not replace generation accuracy. A post hoc 64-token diagnostic did not change the parsed answers under the same prefix parser. The internal sharing result therefore neither establishes broad generation robustness nor a general external advantage.

H Metric Scales, Abbreviations, and Experimental Conventions

F/CF denote factual/counterfactual states. A/B denote persistent object identities, whereas AB/BA specify query role order. QA means question answering; ROI means region of interest. V-A/V-B and T-1/T-2/T-R/T-End are defined in Table 3. B0 is the unmodified frozen backbone; MC is the fully shared baseline's experiment identifier; RS means relation-selective and SP means shared–private. LoReFT is the low-rank representation-finetuning parameterization used for every controller. r is nominal active rank and r16 is the shared rank-16 control. CE is cross-entropy; EOS is the end-of-sequence target; QR is orthogonal–triangular factorization; BF16/FP32/FP64 indicate numeric precision; SDPA is scaled dot-product attention; CI is confidence interval; NA is an undefined or empty conditional quantity, not zero.

Hardware and precision records. The SP experiment records one GPU per training worker on two NVIDIA GeForce RTX 4080 SUPER devices (32,760 MiB each as reported by the environment), with Python 3.12.3, PyTorch 2.8.0+cu128, Transformers 5.9.0, and CUDA build 12.8.

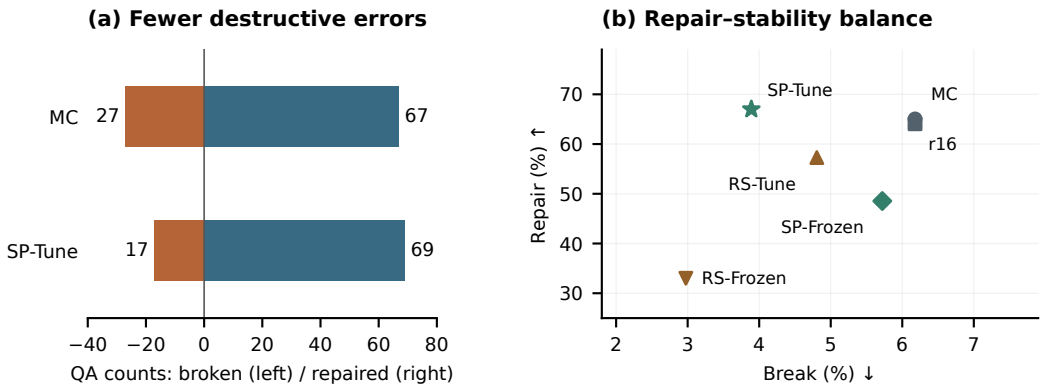


Figure 10: **Utility requires both repair and stability.** Left: actual counts underlying the MC and SP-Tune accuracy difference; net gains over B0 are 40 and 52 QA. Right: conditional Repair versus Break for the six trained structures, showing point estimates on the same internal test set; lower Break and higher Repair are preferable. The narrowest Break is not the highest-utility controller. Paired uncertainty is reported in the text rather than inferred from distances between points.

Symbol / name	Definition or interpretation	Scale / reporting unit
m_r, m_d, m_p	Yes-minus-no candidate log-score margin for clean recipient, clean donor, and patched recipient	Signed log-score units
D_{truth}	$y_d(m_p - m_r)$; positive means movement toward donor truth	Signed log-score units
D_{pred}	$\text{sign}(m_d - m_r)(m_p - m_r)$; excludes equal clean margins	Signed log-score units
E_{match}	$\lvert m_r - m_d\rvert - \lvert m_p - m_d\rvert$; reduction in donor-score distance	Signed log-score units
Δ	$m_p - m_r$, without truth orientation	Signed log-score units
ΔH	Change in fraction of answers matching donor truth	$[-1, 1]$; text in pp
ΔAcc_r	Change in recipient-truth accuracy; $-\Delta H$ on these opposite-truth binary edges	$[-1, 1]$; text in pp
Donor match	Patched prediction equals clean donor prediction	Fraction or %
Clean donor match	Clean recipient prediction equals clean donor prediction	Fraction or %
Prediction changed	Patched prediction differs from clean recipient	Fraction or %
Strict transfer	Donor match conditional on clean prediction disagreement, computed within pair	Fraction or %; empty = NA
Update norm / RMS	Frobenius norm U of the replaced state difference; RMS $= U/\sqrt{nd}$ for n tokens and width d	Model-dependent activation units
SR	Target/non-target gradient-energy ratio	Dimensionless; not accuracy
O_t	$\lVert R_s(R_t^p)^\top\rVert_F$; shared/private basis overlap	Dimensionless, $[0, 2]$ for rank 4
Micro / TwoOrder	QA accuracy / both role orders correct for one image and relation	%; denominators 540 / 270
Repair / Break	B0 errors corrected / B0-correct answers corrupted	%; denominators 103 / 437

Table 6: **Metric reference.** pp means percentage points: a fractional change of 0.01 equals 1 pp. Original Figure 3 preserves native score axes; original auxiliary exports may use fractional rates, whereas reconstructed hit-rate curves use pp. Raw margin and activation units are not normalized across model architectures. All percentage columns use a 0–100 scale, not a 0–1 scale.

The backbone uses BF16, controllers use FP32, geometry estimation uses FP64, and attention uses SDPA. These describe recorded runs, not claimed minimum hardware requirements. Selected/stopping epochs and total optimizer updates are reported in Appendix F; early stopping means training exposure differs across methods. The causal experiments use native model preprocessing, including the frozen 728-pixel Idefics3 setting; they are not a common-resolution performance ranking.

AI USE STATEMENT

AI tools assisted the authors in preparing this work in three areas:

- **Paper outline and writing.** AI assistants helped develop and revise the manuscript outline, organize the research narrative, and edit portions of the text for clarity and consistency.
- **Experimental code.** AI assistants helped write portions of the experimental and analysis code, including data-processing, metric-aggregation, and visualization utilities.
- **Literature retrieval.** AI assistants helped search for relevant literature and retrieve bibliographic information for author review.

AI assistance did not replace the authors' responsibility for research design, experimental decisions, interpretation, and the accuracy of the submission. The authors retain responsibility for checking the code, reported results, references, and final manuscript. No AI system is listed as an author.