EQUAL STATE FIT, DIFFERENT FINITE RESPONSES IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-model steering uses maps between hidden states to reuse activation interventions in another model. We ask whether fitting paired states determines the effect of a mapped intervention on held-out target inputs. In an eight-dimensional residual subspace, four paired examples constrain at most six dimensions, leaving maps with the same calibration fit and target update norm that can produce different finite responses. We measure this difference by the response diameter, the range of target responses across these maps. Across three target vision-language models, 123 of 249 controlled groups have scalar diameters above a conservative numerical threshold. The separation recurs on image-disjoint Visual Question Answering v2 directions, including 19/24 Qwen-to-IDEFICS2 and 16/24 LLaVA-to-IDEFICS2 groups, and on a fixed Multimodal Visual Patterns subset spanning eight of nine visual patterns. All evaluable natural responses remain stable when the intervention step is halved, with further evidence from fixed angle grids and a second residual site. These results distinguish the role of state calibration from target-response evaluation: calibration fixes the map on the observed span, while updates outside that span require direct comparison at matched target norms.

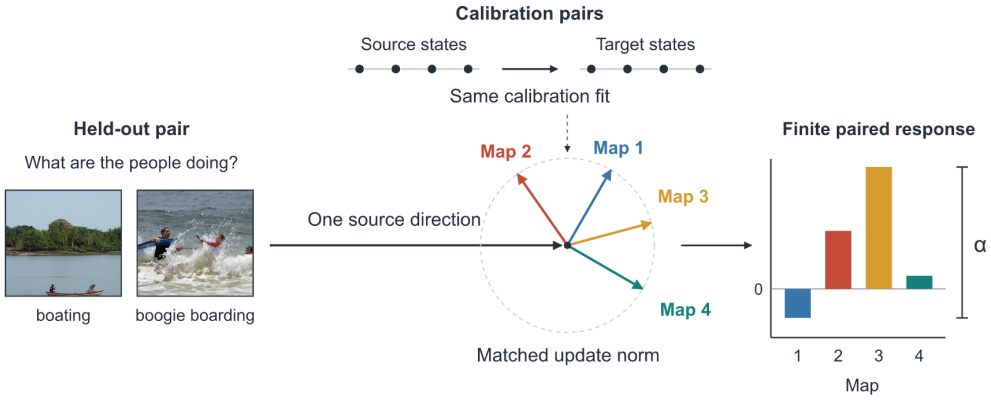

Figure 1: Equal state fit can leave different finite responses. Calibration pairs constrain the maps on the observed span (top). A separate held-out image pair defines one source direction. The circle marks equal-norm target updates under four equal-fit maps. Their finite paired responses on separate evaluation images span α, the response diameter. Colors identify maps; directions and bars are schematic.

1 INTRODUCTION

Activation interventions can change behavior in language models and vision-language models (VLMs) without updating their weights (Turner et al., 2023; Rimsky et al., 2024; Su et al., 2025; Sivakumar et al., 2025). Cross-model steering seeks to reuse these interventions by fitting maps between models (Huang et al., 2025; Oozeer et al., 2025; Chen et al., 2025; Syed, 2026). The practical question is whether a map fitted to paired hidden states determines what the mapped update does on

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.

held-out target inputs. State calibration measures correspondence, whereas this question concerns behavior.

Finite calibration identifies a map on the span supplied by its examples. Two maps can fit every calibration state equally well and assign the same norm to a new update, yet differ on the part of the source space absent from calibration. The target model may respond differently to those updates. Representational agreement, functional stitching, and a successful selected transfer each provide useful evidence (Raghu et al., 2017; Kornblith et al., 2019; Bansal et al., 2021; Huh et al., 2024). None determines whether all maps compatible with a finite state fit produce the same held-out target response.

Figure 1 illustrates the comparison. Calibration pairs constrain the maps on their observed span. A separate held-out image pair keeps the question fixed, changes the correct answer, and defines one source direction. Equal-fit maps send this direction to distinct target updates; the circle indicates their matched norm. We measure the range of finite paired responses on image-disjoint evaluation contexts. A positive response diameter means that the same calibration evidence permits more than one target response.

The identification question becomes exact in a finite intervention subspace. For an eight-dimensional residual subspace and K paired calibration examples, separate centering of base states and paired differences gives the rank bound $\operatorname{rank} X_m(K) \leq \min\{8, 2(K-1)\}$. Four pairs constrain at most six dimensions. We construct four isometric completions that share the fitted action and operator norm and differ only on the two-dimensional complement. A paired target-versus-foil score measures their finite responses. Because state fit and update norm are matched, the response range isolates information absent from calibration.

We test this boundary from controlled queries to natural images. Across three targets and 249 controlled groups, each response diameter is nonzero and reproduced exactly in a second run. A conservative sensitivity analysis resolves 129 vector and 123 scalar diameters. We then replace canonical complement queries with image-disjoint paired-image directions from Visual Question Answering v2 (VQAv2) (Goyal et al., 2017). This panel spans four checkpoints from the Qwen2.5-VL, LLaVA, and IDEFICS2 families (Bai et al., 2025; Liu et al., 2023; Laurençon et al., 2024). Response diameter is resolved in 19/24 Qwen-to-IDEFICS2 groups and 16/24 LLaVA-to-IDEFICS2 groups. With the construction fixed, a Multimodal Visual Patterns (MMVP) extension resolves 13/45 source-target-pattern cells across three edges and eight of nine patterns (Tong et al., 2024). Both source families show their strongest recurrence in the same target model.

The results distinguish what state calibration identifies from what target responses establish. The rank result states what paired states determine, matched completions expose the remaining freedom, and response diameter tests whether that freedom matters to target behavior. Controlled and natural evaluations cover three model families, five directed edges, two intervention steps, and a second residual site. This yields a direct evaluation rule: report the span fixed by calibration, compare maps at matched target norms, and measure held-out target responses wherever the proposed update leaves that span.

2 Related Work

Representation comparison establishes whether models organize observed inputs in related ways. Canonical correlation, projection-weighted correlation, and centered kernel alignment quantify this agreement (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019), and surveys distinguish representational from functional similarity (Klabunde et al., 2025). Model stitching goes further by feeding mapped states into another network (Bansal et al., 2021), while linear identifiability characterizes when representations agree up to an invertible map (Roeder et al., 2021). Shared-representation hypotheses likewise ask whether independently trained models converge toward common structure (Huh et al., 2024; Hu et al., 2026). These results describe observed states or one selected relation. We ask whether the full set of maps compatible with finite calibration identifies one held-out intervention response.

Activation steering supplies the behavioral side of this question. Activation engineering, contrastive activation addition, representation engineering, and function vectors show that residual directions can change language-model behavior without parameter updates (Turner et al., 2023; Rimsky et al.,

2024; Zou et al., 2023; Todd et al., 2024). Intervention-based evaluation treats behavioral control as evidence for an internal explanation (Bhalla et al., 2024). In VLMs, Activation Steering Decoding and SteerVLM measure hallucination or instruction-following behavior after hidden-state interventions (Su et al., 2025; Sivakumar et al., 2025). We adopt the same behavioral standard for the freedom left by cross-model state fitting.

Cross-model steering evaluates whether an intervention discovered in one network can be used in another. Existing approaches fit linear maps, nonlinear transfer functions, or stitching layers and then test a selected mapped direction (Huang et al., 2025; Oozeer et al., 2025; Chen et al., 2025; Syed, 2026). Cross-model transfer also extends to safety directions for visual generation (Poppi et al., 2026). Related studies examine sparse-feature steering in medical VLMs (Nooralahzadeh et al., 2026), causal probes of visual representations (Deng et al., 2026), and feature-direction heterogeneity across modalities (Lee et al., 2026). Successful transfer establishes that a selected mapping can work. We ask whether alternative maps with the same state fit give the same held-out target response. Causal abstraction formalizes the related requirement of intervention correspondence between models (Geiger et al., 2025).

Standard linear alignment tools make this comparison exact. The minimum-norm solution follows the pseudoinverse, orthogonal controls use the Procrustes solution, and numerical rank follows conventional singular-value analysis (Schönemann, 1966; Golub & Van Loan, 2013). We use them to hold state fit and update norm fixed while measuring the target response after intervention.

3 What Paired Calibration Identifies

3.1 Paired states and finite response

We use paired-image examples to define calibration states and target responses. Let s and t denote source and target models at fixed residual sites. Each example e contains a question, a reference image, a paired image, and two concise answers. If $h^0_{m,e}$ and $h^1_{m,e}$ are fixed token-level states in model m, an orthonormal basis $B_m \in \mathbb{R}^{d_m \times r}$ gives intervention coordinates

$$b_{m,e} = B_m^\top h^0_{m,e}, \qquad d_{m,e} = B_m^\top (h^1_{m,e} - h^0_{m,e}), \tag{1}$$

with $r = 8$ throughout the experiments. A linear state map $T : \mathbb{R}^8 \to \mathbb{R}^8$ is fitted from source to target coordinates.

Target response is measured after the residual update is applied. Let $\ell_m(x, y; u)$ be the teacher-forced log likelihood of answer y after adding update u. For target answer y^+ and foil y^-, the paired score is

$$\begin{aligned} Q_{m,e}(u) = &[\ell_m(x^1, y^+; u) - \ell_m(x^1, y^-; u)] \\ &- [\ell_m(x^0, y^+; u) - \ell_m(x^0, y^-; u)]. \end{aligned} \tag{2}$$

For a unit hidden-space direction v and residual amplitude a, the natural-direction evaluation uses the odd finite response

$$R^a_{m,e}(v) = \frac{Q_{m,e}(av) - Q_{m,e}(-av)}{2}. \tag{3}$$

A positive value means that changing $-av$ to av increases the paired target-versus-foil contrast. The paired score removes common image and answer shifts; the odd response measures the change under sign reversal without assuming local linearity.

3.2 Rank of the calibration span

Paired calibration supplies two centered sets of constraints. For K pairs, collect the base coordinates and paired differences in matrices $H_m(K), D_m(K) \in \mathbb{R}^{8 \times K}$, and center each matrix across examples:

$$X_m(K) = \left[H_m(K) - \bar{b}_m \mathbf{1}^\top,\ D_m(K) - \bar{d}_m \mathbf{1}^\top\right]. \tag{4}$$

The fitted map satisfies $TX_s(K) \approx X_t(K)$. The following bound states the information budget of this design.

Proposition 1 (Paired calibration rank)**.** *For the matrix in Equation 4,*

$$\operatorname{rank} X_m(K) \le \min\{8, 2(K-1)\}. \tag{5}$$

Hence $K = 4$ leaves a complement of dimension at least two, while $K = 5$ is the first budget that can generically reach rank eight.

The proof is direct: each centered $8 \times K$ block has rank at most $K-1$, and concatenation adds at most the two ranks. The bound separates the action observed by a finite state fit from the action that can remain unspecified. Appendix D gives the proof and the assumptions for the equal-fit construction.

3.3 Equal-fit completions and response diameter

Let $S = \operatorname{range} X_s(4)$ and let P_S be its projector. If T_1 and T_2 fit the same calibration states exactly, then

$$(T_1 - T_2)P_S = 0. \tag{6}$$

The two maps must agree on S, while calibration imposes no equality on $S^{\perp}$. We compare this remaining freedom under matched budgets. Let U_s and U_t be orthonormal two-dimensional bases for the source complement and an orthogonal target complement. Let $T_0 = T_0 P_S$ be the fitted action on S, and set $\kappa = \|T_0\|_2$. For the planar rotation matrix $O(\theta)$, define

$$T_j = T_0 + \kappa U_t O(\theta_j) U_s^{\top}, \qquad \theta_j \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}. \tag{7}$$

The orthogonal domain and range blocks give every completion the same state fit and operator norm κ. Target directions are normalized before intervention in the natural-direction evaluation, which also fixes their query-level update norm. Appendix D states the construction and its assumptions in full.

Lifting normalized mapped coordinates through B_t gives a finite set of unit target directions $\mathcal{U}_{t,e}$. Define its response diameter

$$\alpha^a_{t,e} = \max_{u,v \in \mathcal{U}_{t,e}} \left| R^a_{t,e}(u) - R^a_{t,e}(v) \right|. \tag{8}$$

If $\alpha^a_{t,e} > 0$, the observed state constraints admit distinct finite responses within this set. Its scale also has a direct prediction meaning.

Corollary 1 (Finite-response certificate)**.** *For any finite response set in Equation 8, the best map-independent scalar prediction has worst-case absolute error*

$$\inf_{c \in \mathbb{R}} \max_{u \in \mathcal{U}_{t,e}} \left| R^a_{t,e}(u) - c \right| = \frac{\alpha^a_{t,e}}{2}. \tag{9}$$

The midpoint of the minimum and maximum responses attains the bound, and the triangle inequality proves that no scalar can improve it. This guarantee applies to the tested maps, inputs, site, score, and amplitude.

4 Evaluation Design

4.1 Controlled and natural-image evaluations

We evaluate the identification claim in two stages. The controlled stage fixes observed and complement queries and uses one Qwen2.5-VL-3B source with three targets: Qwen2.5-VL-7B, LLaVA-1.5-7B, and IDEFICS2-8B. Four calibration pairs produce rank-six state designs. The map set contains the minimum-norm solution and the four completions in Equation 7. Each target has 83 groups: 3 construction controls, 32 controlled image pairs, and 48 public VQAv2 pairs. We fix maps, queries, checkpoints, scores, and groups before measuring responses.

Each candidate map produces one finite score change on each of the two complement basis queries. We collect these changes as a two-coordinate response vector. The vector diameter is the largest Euclidean distance between map responses, while the scalar diameter is the largest absolute difference on either coordinate. We report the observed query separately. Its diameter should be zero because all maps agree on the calibrated span.

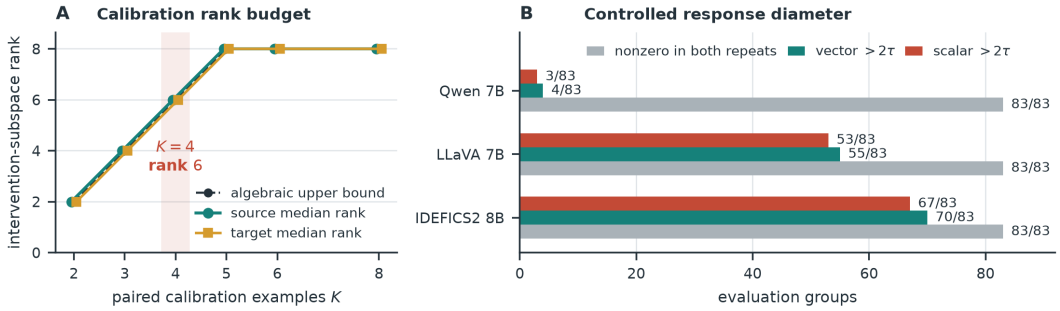

Figure 2: Calibration reaches its rank budget, and unconstrained directions change target response. (A) Source and target curves overlap because both attain the algebraic bound: $K = 4$ reaches rank six and $K = 5$ can reach rank eight. (B) For every target, each of its 83 groups has a nonzero diameter reproduced exactly in a second deterministic run. Vector and scalar bars report a conservative post hoc 2τ sensitivity analysis.

The natural stage tests the same information limit on task-linked visual directions. At each checkpoint and site, the leading right singular vectors of 24 unit-normalized paired-image state differences form an eight-dimensional intervention subspace. The pairs cover object or action, color, scene text, and count questions. Disjoint sets provide nested calibration, anchor candidates, source-only acquisition, test donors, and evaluation contexts. Appendix G gives the complete 120-pair allocation. Donors and contexts share neither rows nor underlying images, and response scores do not enter the subspace, maps, anchors, or donor-context assignments.

4.2 Models and intervention sites

The natural panel contains four checkpoints from three model families. Qwen2.5-VL-3B and LLaVA-1.5-7B serve as sources. Qwen2.5-VL-7B, LLaVA-1.5-7B, and IDEFICS2-8B serve as targets, giving five directed edges. The primary residual site lies at 55% of model depth, and a fixed eight-pair subset repeats two representative edges at 75% depth. We add normalized directions with amplitude $a = 0.004$ at the selected answer-token residual state. This fixes the update budget within each target. Absolute response scales are not pooled across target families.

4.3 Comparisons at matched information budgets

Each comparison is indexed by the state information it receives. At $K = 4$, we evaluate the minimum-norm map, four equal-budget completions, state ridge, scaled Procrustes, and a wrong-correspondence control. A fixed $K = 5$ subset evaluates the three standard map families after one additional pair. The $K = 8$ minimum-norm map provides a higher-budget reference. The target-native direction, measured from the target's own donor states, serves as an evaluation oracle. Appendix H reports additional correspondence-selection policies and their information budgets.

4.4 Numerical resolution and stability

To separate target response from numerical variation, one ordered four-row forward pass scores the two images and two answers in Equation 2. We evaluate each sign twice in 32-bit floating point. Each example e shares its edge's numerical threshold,

$$\tau_e = \max\{10\delta_e,\ 100\epsilon_{32}\max(1, L_e)\}, \tag{10}$$

where δ_e is that edge's largest exact-repeat discrepancy and L_e its largest absolute row score. A diameter is resolved when $\alpha > 2\tau_e$. Target-native activity uses the analogous test. The two-step check requires activity at both steps, matching signs, and at most 25% relative derivative difference. Coverage associations use edge-wise Spearman correlation and a source-blocked permutation test. We report finite-panel counts and permutation values, with complete criteria for additional correspondence policies in Appendix I.

5 RESULTS

5.1 CALIBRATION REACHES THE RANK BOUNDARY

The measured state designs attain the algebraic rank bound at every nested budget. Across all five directed edges and all fixed subsets, $K = 2, 3, 4$ give ranks $2, 4, 6$ in both source and target coordinates. Every $K = 5$ subset reaches rank eight, as do the $K = 6$ and $K = 8$ designs. Figure 2A therefore marks a sharp boundary: $K = 4$ leaves a two-dimensional complement, while $K = 5$ can identify a full-rank linear map in this subspace. The overdetermined $K = 8$ fit is a higher-information reference.

5.2 EQUAL STATE FIT LEAVES DIFFERENT TARGET RESPONSES

The calibrated span behaves exactly as the theory predicts. The observed query has zero response diameter in all 249 target groups because every map delivers the same update where calibration fixes its action. Outside that span, each group has a nonzero two-coordinate diameter reproduced exactly in a second deterministic run. All 249 groups remain separated when the comparison is restricted to the four isometric completions, which match both state fit and complement-update norm.

Table 1: Equal-fit maps produce distinct controlled responses. The exact column reports nonzero diameters that agree across deterministic repeats. The 2τ columns report a conservative post hoc numerical sensitivity analysis.

Target	Groups	Exact $\alpha > 0$	Vector $> 2\tau$	Scalar $> 2\tau$
Qwen2.5-VL-7B	83	83	4	3
LLaVA-1.5-7B	83	83	55	53
IDEFICS2-8B	83	83	70	67
All	249	249	129	123

The stricter numerical analysis retains a clear target-dependent pattern. Table 1 reports 129/249 vector diameters and 123/249 scalar diameters above 2τ. The scalar statistic resolves 67/83 IDEFICS2 groups and 53/83 LLaVA groups, compared with 3/83 Qwen groups. Each resolved scalar diameter gives an $\alpha/2$ lower bound on the worst-case error of using one response to represent the equal-fit map set. For these 123 groups, map choice changes the scalar response by more than twice the conservative numerical threshold even after update magnitude is matched.

5.3 THE SEPARATION RECURS ON NATURAL VISUAL DIRECTIONS

Natural paired-image directions reveal the same target-dependent separation. Table 2 and Figure 3A report all five edges. Qwen-to-IDEFICS2 has resolved completion diameters in 19/24 groups, and LLaVA-to-IDEFICS2 in 16/24, with resolved groups in all four content categories. At this fixed update norm, both source families produce resolved response differences most often in IDEFICS2.

Table 2: Natural response separation concentrates on IDEFICS2 at the primary site. The two columns ask whether a target-native donor direction is active and whether equal-fit completions disagree. Every row contains 24 fixed donor-context groups.

Source $\rightarrow$ target	Target-native active	Diameter $> 2\tau$
Qwen 3B $\rightarrow$ Qwen 7B	0/24	0/24
Qwen 3B $\rightarrow$ LLaVA 7B	0/24	0/24
Qwen 3B $\rightarrow$ IDEFICS2 8B	13/24	19/24
LLaVA 7B $\rightarrow$ Qwen 7B	0/24	0/24
LLaVA 7B $\rightarrow$ IDEFICS2 8B	13/24	16/24

Halving the intervention step preserves every evaluable natural response. All 24 Qwen-to-IDEFICS2 and 36 LLaVA-to-IDEFICS2 method-context rows retain their sign, and their relative derivative difference is at most 3.13%. At the second residual site, LLaVA-to-IDEFICS2 retains a resolved

diameter in 5/8 groups. The separation is therefore locally stable and persists beyond the primary site on the fixed subset.

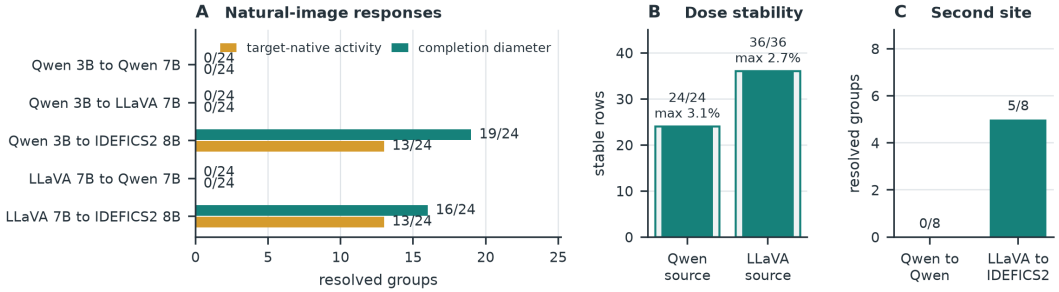

Figure 3: Equal-fit maps separate most often in IDEFICS2, and the response persists under two local checks. (A) Target-native activity and completion diameter on five edges at a matched within-target norm. Zero counts indicate no resolved groups at this dose. (B) Every evaluable row is stable when the step is halved, with maximum relative derivative difference below 3.2%. (C) The second site retains 5/8 LLaVA-to-IDEFICS2 groups.

The separation also recurs on MMVP without fitting on MMVP responses (Tong et al., 2024). Comparing the minimum-norm map and four rotations resolves 13/45 source-target-pattern cells across three edges and eight of nine visual patterns. Ten of 18 IDEFICS2 cells are resolved, compared with 3/27 cells for the other targets. A one-sided Fisher exact test supports this target concentration ($p = 0.0019$), as does an exact pattern sign-flip test ($p = 0.0117$). Section M reports every cell and the descriptive baselines.

The MMVP separation persists under fixed changes to angle and dose. On the previously responsive Qwen-to-IDEFICS2 edge, a follow-up resolves 5/9 patterns with the original four rotations and 7/9 after a 45° shift; at half dose, each grid resolves 4/9. The effect therefore occurs under both grids and doses. Section N gives the full comparison, including two reverse edges into Qwen-3B.

5.4 Source-dependent coverage association

The part of a source query outside the calibrated span is associated with response diameter on LLaVA-sourced edges. We measure this residual coverage as $\|(I - P_S)v\|_2/\|v\|_2$ for donor direction v. Across their 48 groups, the mean edge-wise Spearman coefficient is 0.458, with a one-sided stratified permutation value of $p = 0.0027$. The 72 Qwen-sourced groups show no detectable association ($\rho = -0.074$, $p = 0.1934$; Figure 4).

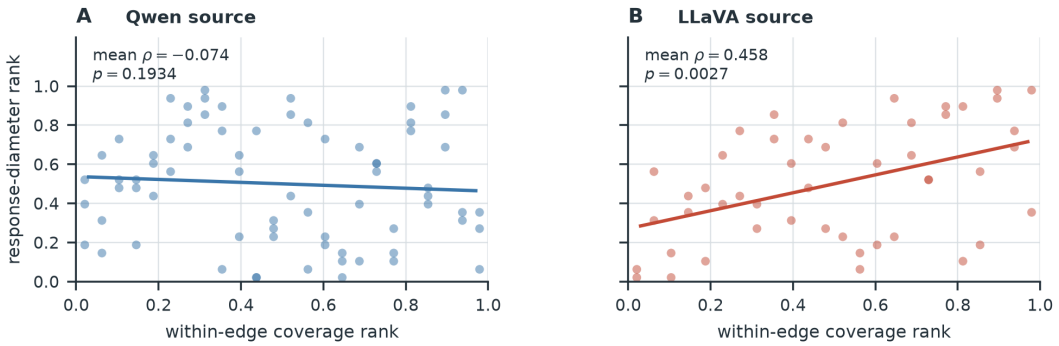

Figure 4: Coverage tracks response diameter for one source family. Residual coverage correlates with completion diameter on LLaVA-sourced edges, while Qwen shows no detectable association. Points are ranked within each edge before pooling edges with the same source. Lines are descriptive, and reported p values come from the stratified permutation test.

6 DISCUSSION

6.1 STATE ALIGNMENT IDENTIFIES A SPAN

State alignment identifies how a map acts on observed directions. Similarity measures and shared-representation results show that models can organize observed inputs in related ways (Raghu et al., 2017; Kornblith et al., 2019; Huh et al., 2024; Hu et al., 2026), and linear identifiability explains when one representation is recoverable from another (Roeder et al., 2021). Our results locate the intervention boundary of that evidence. A finite fit fixes linear action on its observed span. The response of a held-out mapped update remains a target-model quantity.

The rank curve shows how additional calibration constrains the map. At $K = 4$, the two-dimensional complement follows from the calibration design, and response diameter measures whether the target model is sensitive to that freedom. Every tested $K = 5$ subset reaches full rank and closes this particular ambiguity for an eight-dimensional linear map. Functional transfer still requires source and target responses because neither appears in the state-fit equations.

This boundary also changes how a successful selected transfer should be interpreted. When equal-fit maps disagree, the observed target response depends on how the map acts outside the calibration span. State correspondence alone cannot attribute that response to shared representation. Response diameter exposes this dependence and shows whether more calibration or direct target evaluation is needed.

6.2 CALIBRATION FOR THE INTENDED INTERVENTIONS

Mapped interventions need calibration that reflects their intended queries. Cross-model steering studies evaluate the transfer of selected directions (Huang et al., 2025; Oozeer et al., 2025; Chen et al., 2025; Syed, 2026), while direct intervention evaluations test whether an update changes the intended behavior (Bhalla et al., 2024; Su et al., 2025; Sivakumar et al., 2025). Our results connect these forms of evidence: effective rank and query coverage describe what the states identify, matched target norms control intervention size, and held-out responses establish the behavior being claimed.

Query coverage gives a direct design rule for calibration. Each intended query decomposes into its projection on the observed span and its complement. A new state pair is informative when its centered base or difference direction adds rank in a complement used by those queries. Among designs with similar state coverage, the better calibration is the one that contracts response diameter on held-out task queries. Calibration can thus target the intended interventions instead of optimizing global state similarity alone.

Calibration can stop for the evaluated queries when their mapping is identified or the tested completions become response-indistinguishable at the chosen dose. The first condition requires the queries to lie in a well-conditioned observed span; target responses test the second within the sampled map set. Causal abstraction likewise requires intervention correspondence in addition to observational correspondence (Geiger et al., 2025), which explains the continuing value of functional stitching and direct steering evaluations (Bansal et al., 2021; Klabunde et al., 2025).

Target update norm must remain fixed during this comparison. A fitted map can change direction and magnitude, and a larger residual can produce more output movement without improving state correspondence. Our controlled completions share operator and complement-query norms, while natural directions are normalized after mapping. Response diameter then measures variation due to direction at a matched intervention budget.

6.3 DEPENDENCE ON THE TARGET MODEL

Target dependence is the clearest empirical pattern. VQAv2 separates both source families into IDEFICS2. MMVP resolves 10/18 IDEFICS2 cells and a further 3/9 Qwen-to-LLaVA cells. Together, these certificates span eight visual patterns. The response induced by map freedom therefore depends strongly on the target, consistent with architecture- and modality-dependent visual features (Nooralahzadeh et al., 2026; Deng et al., 2026; Lee et al., 2026). State correspondence alone cannot reveal this response scale.

Source representation also changes which queries expose the complement. Residual coverage tracks response diameter on LLaVA-sourced edges and has no detectable relation on Qwen-sourced edges. Rank locates the unconstrained subspace, coverage measures how strongly a query uses it, and matched completions show whether the target responds to that freedom.

6.4 Evaluating mapped interventions

State and response measurements answer different scientific questions. Calibration rank locates the source directions fixed by the state fit. Matched-budget response diameter tests whether the remaining map freedom changes a held-out target response. Source and target effects and local stability complete the transfer claim in Table 3.

Table 3: Reliable mapped-intervention evaluation separates six questions.

Measurement	Question and evidence
State calibration	Identified source directions: effective rank and state-fit residual
Map sensitivity	Query agreement across state-compatible maps: matched-budget response diameter
Intervention size	Equal target budget: actual residual update norm
Source behavior	Effect available to transfer: held-out source response and resolution rule
Target behavior	Mapped effect: held-out target response on the task score
Local stability	Persistence: step and site comparisons where evaluable

A reliable evaluation follows these measurements in order. Effective rank first asks whether calibration fixes the proposed query. If map freedom remains, response diameter tests whether that freedom changes behavior. In our panel, every nested $K = 5$ design reaches full rank, both resolved IDEFICS2 edges retain their response in every evaluable step comparison, and LLaVA-to-IDEFICS2 persists on the second-site subset. Identification, response separation, and local stability can therefore be read together instead of collapsed into one score.

State alignment remains useful when its identified object is stated precisely. A well-conditioned full-rank calibration removes the linear completion freedom studied here. Rank-deficient calibration can still support a specific intervention when its proposed update is evaluated directly on held-out target responses. State measurements determine which part of the map is fixed; held-out responses establish what the mapped intervention does.

7 Limitations

Our conclusions concern finite sets of linear maps in an eight-dimensional residual subspace, a paired answer score, two residual sites, and matched within-target update budgets. These choices make the identification boundary exact and the natural evaluation tractable. Extension to nonlinear maps, larger subspaces, full-benchmark evaluation, and response-active source interventions remains open. The next question is whether query-targeted calibration can make held-out functional transfer predictable.

8 Conclusion

Cross-model steering can reuse a discovery reliably only when the mapped update has a determined target response. Equal-fit, equal-budget completions produce different responses in all 249 controlled groups, and the separation recurs on image-disjoint VQAv2 and MMVP directions. Resolved VQAv2 responses remain stable when the intervention step is halved, and the LLaVA-to-IDEFICS2 separation persists on a fixed second-site subset. These findings yield a practical criterion: use effective rank to determine whether calibration fixes the intended query, then use response diameter to test whether the remaining map choices change its target response. Cross-model steering should pair state evidence with held-out target behavior.

AI USE STATEMENT

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

REPRODUCIBILITY STATEMENT

Sections 3 through 5 define the state and response objects, calibration budgets, comparisons, and numerical rules. The appendix gives proofs, model and layer specifications, data partitions, image-disjointness checks, information budgets, complete results, and numerical checks. The anonymous code supplement provides a single-cell reproduction program with fixed input definitions and pinned model and public-data revisions. It writes input manifests, computed states and maps, raw scores, and numerical result tables.

The complete reproduction code and experimental results are provided in the Supplementary Material.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, pp. 225–236, 2021.

Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, and Himabindu Lakkaraju. Towards unifying interpretability and control: Evaluation via intervention. *arXiv preprint arXiv:2411.04430*, 2024.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching. In *Advances in Neural Information Processing Systems 38*, pp. 54114–54146, 2025. doi: 10.52202/085713-1621.

Zehao Deng, Tianjie Ju, Zheng Wu, Liangbo He, Jun Lan, Huijia Zhu, Weiqiang Wang, and Zhuosheng Zhang. Causal probing for internal visual representations in multimodal large language models. *arXiv preprint arXiv:2605.05593*, 2026.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Zhimin Hu, Lanhao Niu, and Sashank Varma. Language models represent and transform concepts with shared geometry. *arXiv preprint arXiv:2607.04525*, 2026.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704, 2025. doi: 10.18653/v1/2025.acl-long.185. URL https://aclanthology.org/2025.acl-long.185/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20617–20642, 2024.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3728458.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems 37*, pp. 87874–87907, 2024. doi: 10.52202/079017-2789.

Chungpa Lee, Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Same concept, different directions: Cross-modal feature heterogeneity in sparse autoencoders. *arXiv preprint arXiv:2606.29888*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, pp. 34892–34916, 2023. doi: 10.52202/075280-1516.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Farhad Nooralahzadeh, Benjamin Gundersen, Nicolas Deperrois, Hidetoshi Matsuo, Mizuho Nishio, Thomas Frauenfelder, Ahmed Allam, Christian Blüthgen, Michael Moor, and Michael Krauthammer. Universal boosts, specific suppressors: Sparse autoencoder steering of medical vision-language models. *arXiv preprint arXiv:2605.24977*, 2026.

Narmeen Fatimah Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amir Abdullah. Activation space interventions can be transferred between large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47212–47286, 2025. URL https://proceedings.mlr.press/v267/oozeer25a.html.

Tobia Poppi, Silvia Cappelletti, Sara Sarto, Florian Schiffers, Garin Kessler, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Do models share safety representations? cross-model steering for safe visual generation. *arXiv preprint arXiv:2606.05290*, 2026.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9030–9039, 2021.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665, 2025. doi: 10.18653/v1/2025.findings-emnlp.1285. URL https://aclanthology.org/2025.findings-emnlp.1285/.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, 2025. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Abdul Rafay Syed. Actionable activation directions for detecting and mitigating emergent misalignment across language model families. *arXiv preprint arXiv:2606.20225*, 2026.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Appendix Guide

This appendix separates mathematical scope, experimental construction, complete results, and reproducibility. Table 4 gives a reading guide. Sections B through D define the claim and prove the rank and response-diameter results. Sections E through I specify the measured score, intervention subspace, data partitions, maps, and inference rules. Sections J through Q report every target, panel, category, method, step, and site result, including the fixed Multimodal Visual Patterns (MMVP) external evaluation and the null analyses. Sections P through S document numerical checks and reproduction.

Table 4: Appendix reading guide and its relation to the main claims.

Question	Location
What is identified by K paired states?	Sections B and D; Table 12
How is target response measured?	Section E
How are intervention subspaces and maps constructed?	Sections F and H
Are donor and context images independent?	Section G
Does the result recur on another public benchmark?	Section M
Does it persist across angle grids and doses?	Section N
What information does each method receive?	Section H; Table 11
What are the complete positive and null results?	Sections J through Q
How were numerical errors and repeats handled?	Section P
What exactly can be reproduced?	Section S

B Notation and Claim Scope

The paper distinguishes state observations from intervention responses. This distinction controls every formal and empirical claim. For model m, $h^0_{m,e}$ and $h^1_{m,e}$ denote the residual states for the reference and paired images of example e. The matrix B_m is an orthonormal intervention basis in the model's hidden space. The coordinates $b_{m,e} = B_m^\top h^0_{m,e}$ and $d_{m,e} = B_m^\top (h^1_{m,e} - h^0_{m,e})$ are used for state fitting. Matrices $H_m(K)$ and $D_m(K)$ collect these coordinates over K calibration pairs. A state map T acts on the resulting eight-dimensional coordinates. The score $Q_{m,e}$ and finite response $R^a_{m,e}$ are evaluated after the mapped update is inserted into the target model.

Table 5: Notation used in the main paper and appendix.

Symbol	Meaning
m, s, t	A model, source model, and target model
e	A paired visual question-answer example
$B_m \in \mathbb{R}^{d_m \times 8}$	Orthonormal residual intervention basis for model m
$b_{m,e}, d_{m,e} \in \mathbb{R}^8$	Base-state coordinate and paired state difference
$H_m(K), D_m(K)$	Base-coordinate and paired-difference matrices for K examples
$X_m(K)$	Separately centered base and difference design for K pairs
$S = \operatorname{range} X_s(4)$	Source subspace observed by the primary calibration
T	Linear source-to-target state map in intervention coordinates
$Q_{m,e}(u)$	Paired target-versus-foil teacher-forced score
$R^a_{m,e}(u)$	Odd finite response at residual amplitude a
$\mathcal{U}_{t,e}$	Finite set of normalized target directions induced by state-compatible maps
$\alpha^a_{t,e}$	Response diameter over $\mathcal{U}_{t,e}$
τ_e	Numerical threshold shared by example e's edge

A response diameter above $2\tau_e$ distinguishes target responses among the tested equal-fit maps at the specified site and update norm. The exact-repeat controlled result uses the weaker condition $\alpha > 0$ and is reported separately from the conservative post hoc 2τ analysis.

Table 6: Roles of the controlled and natural evidence layers. The natural layer contains the primary VQAv2 panel and the fixed MMVP extension. Counts are not pooled across response definitions or panels.

Aspect	Controlled core	Image-disjoint natural evaluation
State evidence	Rank-six paired calibration design	Rank-six nested design in a task-linked intervention subspace
Query directions	Observed and source-complement queries	Held-out paired-image donor directions
Target response	One-sided finite score change $F(u)$	Central odd finite response $R^a(u)$
Primary outcome	Exact-repeat response diameter $\alpha > 0$	Resolved diameter $\alpha > 2\tau$
Additional check	Post hoc 2τ sensitivity	Target and source activity rules
Main role	Controlled identifiability construction	Natural, image-disjoint evaluation

C Evidence Layers and Outcome Terms

The evaluation has one controlled layer and two natural panels. The controlled layer isolates the map class with canonical observed and complement queries. The image-disjoint Visual Question Answering v2 (VQAv2) evaluation replaces the complement basis queries with task-linked paired-image directions and uses a central odd response. A fixed MMVP extension applies the same construction to a second public benchmark without fitting on MMVP responses. Table 6 records the differences between controlled and natural evidence so that their counts are interpreted separately.

We distinguish exact and numerically resolved outcomes. An *exact positive* has nonzero diameter under exact repeated evaluation, while a *resolved diameter* exceeds the numerical scale 2τ. A *low-activity inconclusive* cell has neither a resolved completion diameter nor an identifiable target-native response at the fixed update norm. Compound analyses are summarized only after all pre-specified conditions are evaluated, and their individual estimates remain visible. This terminology separates numerical zero, unresolved response, and an unmet compound criterion.

D Proofs and Mathematical Details

D.1 Rank of the paired calibration design

For K paired examples, let

$$C_b = H_m(K) - \bar{b}_m \mathbf{1}^\top, \qquad C_d = D_m(K) - \bar{d}_m \mathbf{1}^\top. \tag{11}$$

Both matrices have K columns whose sum is zero. Therefore each has a right null vector $\mathbf{1}$, and

$$\operatorname{rank} C_b \le K - 1, \qquad \operatorname{rank} C_d \le K - 1. \tag{12}$$

Since $X_m(K) = [C_b, C_d]$ has eight rows,

$$\operatorname{rank} X_m(K) \le \min\{8, \operatorname{rank} C_b + \operatorname{rank} C_d\} \tag{13}$$

$$\le \min\{8, 2(K-1)\}. \tag{14}$$

This proves Proposition 1. The bound does not require the base and difference blocks to be independent. Equality is an empirical property, which is checked for every fixed subset in Table 12.

D.2 State-compatible map class

Let $X_s, X_t \in \mathbb{R}^{8\times n}$ be the primary source and target designs, each of rank six. Assume an exact fitted action exists on $S = \operatorname{range} X_s$. Let P_S be the orthogonal projector onto S and let T_0 be the minimum-norm exact solution. Then $T_0 = T_0 P_S$ and $T_0 X_s = X_t$. If A is any linear operator with $AP_S = 0$, then

$$(T_0 + A)X_s = T_0 X_s + AP_S X_s = X_t. \tag{15}$$

Thus calibration cannot distinguish T_0 from $T_0 + A$ using only X_s and X_t.

The experiment restricts this general family to a matched finite set. Let $U_s, U_t \in \mathbb{R}^{8\times 2}$ contain orthonormal bases for the source complement and the orthogonal complement of the fitted target range. For $O \in O(2)$, define

$$T_O = T_0 + \kappa U_t O U_s^\top, \qquad \kappa = \|T_0\|_2. \tag{16}$$

Proposition 2 (Equal-fit, equal-budget completions)**.** *For every $O \in O(2)$, the map T_O satisfies*

$$T_O X_s = X_t, \qquad \|T_O\|_2 = \kappa. \tag{17}$$

It agrees with T_0 on S. For every unit $q \in S^\perp$, it also satisfies $\|T_O q\|_2 = \kappa$.

The completion term annihilates X_s because $U_s^\top X_s = 0$, which proves equal state fit. The two summands act on orthogonal domain subspaces and have orthogonal ranges. Hence the singular values of T_O are the singular values of $T_0|_S$ together with κ, κ, and

$$\|T_O\|_2 = \kappa \tag{18}$$

for every O. For a unit complement query, orthonormality of U_s, U_t, and O gives $\|T_O q\|_2 = \kappa$. We use the four quadrature rotations as the isometric comparison set. The maps have equal calibration fit, equal operator norm, and equal mapped-query norm on every unit complement direction.

D.3 Observed and unobserved queries

For any two state-compatible maps T_1, T_2 and any $q \in S$,

$$(T_1 - T_2)q = (T_1 - T_2)P_S q = 0. \tag{19}$$

A deterministic target evaluator therefore receives the same update and returns the same response. The zero observed-query diameter is an algebraic implementation check. For $q \in S^\perp$, calibration does not impose $(T_1 - T_2)q = 0$. This absence of a constraint does not prove that the nonlinear target response differs. A positive response diameter must be measured.

D.4 Decision radius lower bound

Let the finite response set be $\mathcal{R} = \{R(u) : u \in \mathcal{U}\} \subset \mathbb{R}$, with diameter $\alpha = \max_{r_1, r_2 \in \mathcal{R}} |r_1 - r_2|$. For any scalar summary c, choose a pair r_1, r_2 attaining the diameter. The triangle inequality gives

$$\alpha = |r_1 - r_2| \le |r_1 - c| + |r_2 - c|. \tag{20}$$

At least one term is therefore at least $\alpha/2$. A single response assigned to the entire state-compatible set has worst-case error at least $\alpha/2$. The same argument holds for a vector response under any norm.

D.5 What changes at full rank

If $X_s \in \mathbb{R}^{8 \times n}$ has row rank eight and exact consistency holds, then the linear action is unique:

$$T = X_t X_s^\top (X_s X_s^\top)^{-1}. \tag{21}$$

The completion class in Equation 16 then disappears because $S^\perp = \{0\}$. This fact explains the $K = 5$ transition in our nested design. It does not establish source-to-target response equality. That functional statement contains the additional functions Q_s and Q_t, which are absent from the state equations.

E Paired Score and Response Estimands

E.1 Teacher-forced paired score

Each logical score contains four teacher-forced log-likelihood rows in a fixed order: reference image with target answer, reference image with foil answer, paired image with target answer, and paired image with foil answer. If these values are $\ell_{0+}, \ell_{0-}, \ell_{1+}, \ell_{1-}$, respectively, then

$$Q(u) = (\ell_{1+}(u) - \ell_{1-}(u)) - (\ell_{0+}(u) - \ell_{0-}(u)). \tag{22}$$

All answer tokens are scored by teacher forcing and summed. This paired difference controls for an answer preference shared across images and for an image effect shared across candidate answers. It measures how the intervention changes the answer contrast between the two images.

The natural evaluation uses the odd finite response in Equation 3. It also records the even component

$$C^a(u) = \frac{Q(au) + Q(-au)}{2} - Q(0), \tag{23}$$

which detects finite-step asymmetry but is not used to select maps. The primary residual amplitude is $a = 0.004$. It is the product of a unit coordinate direction, basis scale 0.08, and intervention step 0.05. The stability subset also evaluates $a = 0.002$.

E.2 Core finite effect

The controlled core uses the one-sided finite effect

$$F_{t,e}(u) = Q_{t,e}(au) - Q_{t,e}(0). \tag{24}$$

It evaluates one observed query and a two-vector source-complement basis. Coordinate arguments below are lifted through B_t before intervention. For map T_j, let $z_j = (F_{t,e}(T_j q_1), F_{t,e}(T_j q_2))$ collect the effects on the two complement queries. The vector diameter is $\max_{j,k} \|z_j - z_k\|_2$, and the scalar diameter is $\max_{j,k,c} |z_{j,c} - z_{k,c}|$. The observed-query diameter is computed separately. Because the controlled and natural evaluations use different estimands and update scales, their counts are not pooled into a common rate.

E.3 Numerical resolution

The natural threshold in Equation 10 combines exact-repeat discrepancy with a conservative floating-point scale. The largest repeat discrepancy is zero on every edge, so the second term determines τ_e. A resolved completion diameter requires $\alpha > 2\tau_e$. Target-native activity and source activity use their respective model thresholds. These are measurement-resolution rules, not smallest effect sizes of practical interest.

The controlled core uses deterministic positivity as its primary outcome. Applying the natural evaluation's 2τ rule to the core is a post hoc sensitivity analysis. We report it so that exact nonzero values near numerical scale are not read as equally strong evidence. The primary outcome and the sensitivity analysis remain separate.

F Intervention Subspace Construction and Model Sites

F.1 Task-linked residual subspace

The natural intervention subspace is constructed separately for every model and site. For each of 24 discovery pairs, we read the answer-token residual states for the reference and paired images, form the difference, and normalize it to unit norm. Stacking these 24 differences gives a $24 \times d_m$ matrix. The leading eight right singular vectors form B_m. No response score is used in this construction.

The retained eight-dimensional energy ranges from 47.4% to 64.1% across model-site combinations. The smallest ratio σ_8/σ_1 is 0.485, and the smallest relative gap between the eighth and ninth singular values is 0.00454. Both are far above the fixed 10^{-7} numerical-rank floor. Orthonormality and duplicate state captures are checked before map fitting.

The primary site is 55% of model depth. The confirmatory subset uses 75% depth. A site fraction is converted to a valid language-model residual layer before any response is read. Each target direction is normalized in its own target intervention subspace before adding the fixed residual norm. This removes mapped-vector amplification from comparisons within an edge.

F.2 Projection checks

Every calibration, anchor, and donor direction must have nonzero full hidden difference and a projected-to-full norm ratio above 10^{-7}. Every mapped direction must also satisfy

$$\frac{\|Tv\|_2}{\|T\|_2 \|v\|_2} > 10^{-7}. \tag{25}$$

Every evaluated direction exceeds these numerical floors, so projection and normalization are defined throughout the evaluation.

Table 7: Model checkpoints, residual sites, and intervention-subspace diagnostics. Fractions index depth from input to output. Revision identifiers pin public checkpoint contents.

Checkpoint revision	Site	Layer	Hidden dim.	Energy	σ_8/σ_1	Relative gap
Qwen/Qwen2.5-VL-3B-Instruct, 66285546	0.55	19	2048	0.641	0.485	0.00454
Qwen/Qwen2.5-VL-3B-Instruct, 66285546	0.75	26	2048	0.588	0.611	0.01066
Qwen/Qwen2.5-VL-7B-Instruct, cc594898	0.55	15	3584	0.569	0.659	0.01193
Qwen/Qwen2.5-VL-7B-Instruct, cc594898	0.75	20	3584	0.558	0.651	0.01254
llava-hf/llava-1.5-7b-hf, b234b804	0.55	17	4096	0.570	0.647	0.04038
llava-hf/llava-1.5-7b-hf, b234b804	0.75	23	4096	0.572	0.602	0.03451
HuggingFaceM4/idefics2-8b, 2c42686c	0.55	17	4096	0.499	0.736	0.02152
HuggingFaceM4/idefics2-8b, 2c42686c	0.75	23	4096	0.474	0.760	0.01294

G Data and Image-Disjoint Evaluation

G.1 Paired public examples

The natural evaluation uses the paired-image construction of VQAv2 (Goyal et al., 2017). Each pair has the same question and two images with different concise answers. We use four task categories: object or action, color, scene text, and count. Questions are normalized only by stripping outer whitespace and collapsing repeated internal whitespace. Case, punctuation, word order, answers, row identifiers, and image identifiers remain unchanged.

Table 8: Data roles in the natural evaluation. Every role is balanced across four categories. Response data are read only for donors and evaluation contexts after all state-only construction is fixed.

Role	Pairs	Pairs/category	Use
Discovery	24	6	Build each model's intervention basis
Calibration	8	2	Nested $K = 2, 3, 4, 5, 6, 8$ state maps
Anchor candidates	16	4	Candidate extra correspondences
Source-only acquisition	24	6	Select coverage and cosine anchors
Test donors	24	6	Provide held-out natural source directions
Evaluation contexts	24	6	Measure target response
Total	120	30	240 VQAv2 rows

The donor and context partitions contain 96 distinct images in total. Their row identifiers and image identifiers are mutually disjoint. Both are also disjoint from 384 rows and 368 images used in the separate behavior-only validation set. The acquisition set is disjoint from the test donors, so coverage selection cannot read the direction whose response is later evaluated. Each donor is assigned one same-category evaluation context, which makes the donor-context pair the statistical unit.

G.2 Information flow before response evaluation

The experimental order keeps state construction separate from response evaluation. Every data role is assigned before a model response is measured, and each stage receives only the inputs needed for its specified output. Table 9 records this order.

The donor-context split and the information order address different dependencies. Image-disjoint roles prevent an evaluation image from appearing in donor construction. Response separation prevents a measured target outcome from changing the intervention subspace, map, selected anchor, direction, or context. Both conditions are checked from identifiers stored with the raw records.

G.3 Controlled core panels

The controlled evaluation contains three construction controls, 32 controlled image pairs, and 48 public paired visual question answering (VQA) examples. The controlled panel is balanced over color, count, text, and shape. The public panel is balanced over object or action, color, text, and count. All 83 groups are evaluated for each of the three targets. This panel isolates the finite map

Table 9: Information flow in the natural evaluation. Response scores become available only after bases, maps, anchor choices, donor directions, and evaluation contexts are fixed.

Stage	Permitted inputs	Fixed output	Response access
Role assignment	Public rows, questions, answers, and image identifiers	Six disjoint data roles	None
Intervention-basis construction	Discovery residual states	Basis B_m and projection diagnostics	None
Calibration	Calibration residual coordinates	$X_s(K)$, $X_t(K)$, and candidate maps	None
Anchor selection	Candidate and acquisition states	One anchor per policy and category	None
Query assignment	Test-donor states and context identifiers	Directions and donor-context units	None
Response evaluation	Fixed maps, directions, contexts, and update norms	Atomic score records	Source and target
Reconstruction	State and response records	Metrics, tests, tables, and figures	Read only

construction, while the image-disjoint natural panel provides held-out task-linked directions. The map and query set are fixed independently of both response evaluations.

H Map Families and Information Budgets

H.1 Zero-additional-correspondence maps

The primary $K = 4$ design uses one calibration pair from each category. The minimum-norm map is $T_{\min} = X_t X_s^{+}$. Four completion maps add the equal-norm operators in Equation 16. State ridge minimizes squared state residual plus $\lambda\|T\|_F^2$; its regularization parameter is selected by leave-one-pair-out angular state loss over a fixed global path. Scaled Procrustes fits an orthogonal map followed by one scalar. The wrong-correspondence control permutes category-matched columns before fitting.

The four completion maps satisfy a common set of algebraic and experimental invariants. Table 10 lists the checks used to isolate map freedom from state fit, map scale, and response-based selection.

Table 10: Invariants of the four isometric completions. The equalities hold for every fixed $K = 4$ calibration subset.

Check	Value for every completion T_j	Purpose
Observed action	$T_j P_S = T_0 P_S$	Matches every observed calibration direction
Calibration residual	Identical across j	Holds state fit fixed
Operator norm	$\|T_j\|_2 = \kappa$	Holds the map budget fixed
Complement action	$\kappa U_t O_j U_s^{\top}$	Changes only the unobserved block
Unit complement query	$\|T_j v\|_2 = \kappa$	Holds the core query norm fixed
Rotation set	$0^\circ, 90^\circ, 180^\circ, 270^\circ$	Fixes the finite comparison set
Response access	None during construction	Prevents outcome-based map choice

The minimum-norm map is reported as a familiar reference, but it is not needed for the central separation. The isometric-only analysis removes it and recomputes the diameter over the four completion maps. Positivity in all 249 core groups survives this removal, so the result does not depend on comparing a zero complement with a nonzero complement.

Table 11: Information budgets of all method families. The target-native direction is an evaluation oracle and is never treated as a deployable method.

Tier	State information	Methods
$K = 4$	Four paired examples	Minimum norm, four isometric completions, state ridge, scaled Procrustes, wrong correspondence
$K = 5$ nested	Five paired examples	Minimum norm, state ridge, scaled Procrustes
One-direction	Four pairs and one difference direction	Coverage, donor cosine, D-optimal, four fixed random policies, matched wrong target
$K = 8$ reference	Eight paired examples	Minimum-norm overdetermined reference
Oracle	Donor-specific target direction	Target-native donor direction, evaluation only

H.2 Additional-state-information maps

The fixed $K = 5$ tier adds the next nested calibration pair and refits minimum norm, state ridge, and scaled Procrustes. Every $K = 5$ source and target design has rank eight. The $K = 8$ reference uses all calibration pairs. It also has full source, target, and map rank, but its overdetermined fit is not exact, as Table 21 shows.

One-direction augmentation adds a normalized paired-state difference from one of 16 anchor candidates to the four-pair design. Each method selects a source anchor and then observes its corresponding target difference. This adds one constraint direction and raises the source rank from six to seven; a fifth complete pair supplies both a base state and a difference. The coverage policy uses the independent source-only acquisition set. If P_S projects onto the $K = 4$ source span and A_c contains acquisition directions in category c, it chooses

$$a_c^* = \arg\min_{a \in \mathcal{A}_c} \frac{1}{|A_c|} \sum_{v \in A_c} \|(I - P_{[S,a]})v\|_2. \tag{26}$$

The donor-cosine policy uses the same acquisition directions and candidate pool but maximizes mean absolute cosine. D-optimal selection maximizes the augmented state-design determinant. Four random policies use fixed candidate priorities. The matched wrong-target control uses the coverage-selected source anchor paired with an incorrect target correspondence.

These tiers answer different questions. Comparisons within a tier can isolate selection or map choice at a matched information budget. Comparisons across tiers show what changes with more state information, but they are not a single method ranking.

I Statistical Units and Decision Rules

The natural panel has 24 donor-context units per edge, six in each category, with the same assignments shared across edges. Coverage association is the mean edge-wise Spearman correlation between source residual coverage and completion response diameter. Each permutation shuffles coverage within categories and applies the same shuffle to every edge with that source. This preserves dependence from shared donors and contexts. The reported p values concern this fixed panel.

Response defect measures disagreement with the target-native donor response at the same update norm. For a method j and the 24 contexts $\mathcal{E}$ of one edge, its activity-normalized L_1 defect is

$$D_j = \frac{\sum_{e \in \mathcal{E}} |R^a_{t,e}(u_{j,e}) - R^a_{t,e}(u_{\text{native},e})|}{\sum_{e \in \mathcal{E}} |R^a_{t,e}(u_{\text{native},e})|}. \tag{27}$$

The denominator is positive on every edge. Table 22 reports this ratio of sums, including low-activity contexts. The activity requirement below determines whether a defect comparison supports response repair.

Table 12: Ranks across all fixed calibration subsets. Each entry is the common source and target rank; all five edges have the same result.

K	Subsets per edge	Rank bound	Observed rank
2	28	2	2
3	56	4	4
4	70	6	6
5	56	8	8
6	28	8	8
8	1	8	8

Table 13: Controlled response diameter by target and panel. The exact column uses repeated deterministic evaluation. Vector and scalar counts apply the conservative post hoc 2τ sensitivity.

Target	Panel	Groups	Deterministic	Vector $> 2\tau$	Scalar $> 2\tau$
Qwen2.5-VL-7B	Construction	3	3	0	0
Qwen2.5-VL-7B	Controlled	32	32	1	1
Qwen2.5-VL-7B	Public VQA	48	48	3	2
LLaVA-1.5-7B	Construction	3	3	3	3
LLaVA-1.5-7B	Controlled	32	32	22	21
LLaVA-1.5-7B	Public VQA	48	48	30	29
IDEFICS2-8B	Construction	3	3	2	2
IDEFICS2-8B	Controlled	32	32	31	30
IDEFICS2-8B	Public VQA	48	48	37	35

The one-correspondence analysis reports 21 methods and tests repair against 13 specified opponents: the eight $K = 4$ maps, four random-anchor policies, and matched wrong-target control. For each edge and opponent, the coverage policy must have a negative activity-normalized defect difference, a one-sided paired sign-flip value no larger than 0.05, a category-stratified resampling upper bound below zero, and no positive category mean. All 24 target-native responses must also be identifiable. Donor cosine and D-optimal use matched source information and are reported as policy sensitivities.

Source coverage acquisition uses a parallel criterion. Coverage selection must improve the held-out source residual over the base span and each of four fixed random policies under the same point, sign-flip, resampling, and category conditions. Source-to-target sign portability requires identifiable source and target effects before their signs can be compared. Low-activity cases remain undefined.

J Complete Quantitative Results

The following tables report all calibration subsets, every controlled target-panel cell, all natural edges and categories, both sites, every method, and every pre-specified analysis. They are generated directly from the result records and independent reconstruction reports.

Compact tables use Q3 for Qwen2.5-VL-3B, Q7 for Qwen2.5-VL-7B, L7 for LLaVA-1.5-7B, I8 for IDEFICS2-8B, and VQA for visual question answering.

K Controlled Results: Additional Interpretation

K.1 Deterministic construction and numerical sensitivity

The controlled evaluation contains $3 \times 83 = 249$ target groups. Every group has a nonzero full-vector diameter reproduced exactly in a second deterministic run and a zero observed-query diameter. The same 249 groups remain positive after deleting the minimum-norm map and retaining only the four isometric completions. At the aggregate level, all nine target-panel diameters are unchanged by this deletion. The maximum spread in isometric complement query norm is 2.22×10^{-16}, so update-norm differences do not explain the result.

Table 14: Core response scales. Diameters are scalar finite-response diameters; the hook norm is the actual residual norm after the intervention-basis scale. Cross-target magnitudes are not pooled.

Target	Min. diameter	Median	Max.	Hook norm
Qwen2.5-VL-7B	8.67e-05	5.47e-04	0.0029	0.0169
LLaVA-1.5-7B	2.86e-05	0.0017	0.0093	0.0608
IDEFICS2-8B	3.81e-04	0.0039	0.0195	0.0067

Table 15: Natural panel by content category. Each cell reports target-native identifiable groups and resolved completion-diameter groups out of six.

Edge	Measure	Object	Color	Text	Count
Q3→Q7	Native activity	0/6	0/6	0/6	0/6
	Completion diameter	0/6	0/6	0/6	0/6
Q3→L7	Native activity	0/6	0/6	0/6	0/6
	Completion diameter	0/6	0/6	0/6	0/6
Q3→I8	Native activity	2/6	4/6	4/6	3/6
	Completion diameter	5/6	4/6	6/6	4/6
L7→Q7	Native activity	0/6	0/6	0/6	0/6
	Completion diameter	0/6	0/6	0/6	0/6
L7→I8	Native activity	2/6	4/6	4/6	3/6
	Completion diameter	5/6	4/6	5/6	2/6

The conservative sensitivity analysis identifies separations that are large relative to numerical scale. Qwen-7B retains 4/83 vector and 3/83 scalar cases above 2τ. LLaVA retains 55/83 and 53/83, while IDEFICS2 retains 70/83 and 67/83. Table 13 gives the panel breakdown. The exact result establishes the finite construction, and the resolved counts quantify its numerical strength.

K.2 Response scale and source activity

Median scalar diameter ranges from 5.47×10^{-4} for Qwen-7B to 3.92×10^{-3} for IDEFICS2. The actual hook norms differ across targets, as Table 14 reports. These values prevent a common cross-target effect interpretation. Within each target, however, candidate maps receive matched construction budgets and the isometric candidates have identical query norm.

The source effect is below the 2τ sensitivity rule in all 83 source groups. The controlled evidence therefore establishes target-response underdetermination for source-state-compatible maps. Functional transfer of a response-active source intervention remains a separate question.

L Natural Panel: Per-Edge Findings

The Qwen 3B to Qwen 7B edge is low-activity inconclusive at the fixed residual norm. All 24 target-native effects and all 24 completion diameters are below 2τ. The median diameter is 1.74×10^{-4}, compared with $2\tau = 1.09 \times 10^{-3}$. All 80 step-check rows are below the activity rule.

The Qwen 3B to LLaVA 7B edge is also low-activity inconclusive. All 24 target-native effects and completion diameters are below the edge rule. The largest diameter is 8.53×10^{-4}, below $2\tau = 9.07 \times 10^{-4}$.

The Qwen 3B to IDEFICS2 8B edge has resolved completion diameter in 19/24 groups: 5/6 object, 4/6 color, 6/6 text, and 4/6 count. Target-native activity is identifiable in 13/24 groups under its separate rule. The median diameter is 2.17×10^{-3} and the maximum is 5.32×10^{-3}. Every one of the 24 evaluable two-step rows is stable; the maximum relative derivative difference is 3.13%.

The LLaVA 7B to Qwen 7B edge also falls below numerical resolution at the fixed scale. All target-native and completion responses are below the fixed-budget rule. The largest diameter is 4.37×10^{-4}, compared with $2\tau = 1.09 \times 10^{-3}$.

Table 16: Natural-panel numerical scale and observed diameter. The threshold is edge specific. Diameter summaries cover 24 donor-context groups.

Edge	2τ	Min. α	Median α	Max. α	Source active
Q3→Q7	0.0011	4.45e-05	1.74e-04	5.27e-04	0/24
Q3→L7	9.07e-04	2.72e-05	2.51e-04	8.53e-04	0/24
Q3→I8	0.0011	4.23e-04	0.0022	0.0053	0/24
L7→Q7	0.0011	5.21e-05	1.42e-04	4.37e-04	0/24
L7→I8	0.0011	2.32e-04	0.0013	0.0062	0/24

Table 17: Dose stability under step halving. Rows below the fixed activity threshold are reported as not evaluable (n.e.), not as unstable.

Edge	Stable/evaluable	Low activity	Mean rel. diff.	Max rel. diff.
Q3→Q7	0/0	80	n.e.	n.e.
Q3→L7	0/0	80	n.e.	n.e.
Q3→I8	24/24	56	0.0085	0.0313
L7→Q7	0/0	80	n.e.	n.e.
L7→I8	36/36	44	0.0093	0.0270

The LLaVA 7B to IDEFICS2 8B edge has resolved completion diameter in 16/24 groups: 5/6 object, 4/6 color, 5/6 text, and 2/6 count. Target-native activity is identifiable in 13/24 groups. All 36 evaluable two-step rows retain their sign, with maximum relative derivative difference 2.70%. At the second site, 5/8 completion groups and 4/8 target-native groups are resolved. The second-site analysis measures persistence on a fixed subset.

The resolved-diameter count can exceed the target-native activity count because they test different objects. The former asks whether equal-fit completions disagree with each other. The latter asks whether the target's own donor direction has a large odd response. Neither event contains the other.

Table 18: Second-site checks on two fixed eight-group subsets. Completion diameter and target-native activity use their separate resolution rules.

Edge	Completion resolved	Target-native identifiable
Q3→Q7	0/8	0/8
L7→I8	5/8	4/8

Table 19: Source-block coverage association. The coefficient is the mean of edge-wise Spearman correlations after preserving edge blocks.

Source	Mean ρ	Permutation p	Edge coefficients
Qwen 3B	-0.074	0.1934	-0.116, -0.069, -0.038
LLaVA 7B	0.458	0.0027	0.231, 0.685

Table 20: Outcomes of the additional natural-panel analyses. Each row summarizes the complete pre-specified analysis.

Analysis	Observed result
One-direction rank and state fit	Rank 7; residual $< 10^{-7}$
Source coverage acquisition	Mean coverage decreases
Target response repair	Mixed by source
Coverage-conditioned repair	No consistent improvement
Source-to-target sign portability	Unresolved

Table 21: Full-rank $K = 8$ reference. Rank is eight for source, target, and map on every edge, but the overdetermined state fit is not exact.

Edge	Source rank	Target rank	Map rank	Relative fit error
Q3→Q7	8	8	8	0.390
Q3→L7	8	8	8	0.484
Q3→I8	8	8	8	0.386
L7→Q7	8	8	8	0.491
L7→I8	8	8	8	0.414

Table 22: Activity-normalized L_1 response defect for every specified method and edge. Lower is better within an edge. Information-budget tiers are reported separately.

Method	Q3→Q7	Q3→L7	Q3→I8	L7→Q7	L7→I8
Minimum norm	0.841	1.059	1.202	1.279	0.755
Completion 0°	1.109	1.174	1.214	1.015	0.795
Completion 90°	0.790	1.213	1.250	1.409	0.817
Completion 180°	1.035	0.922	0.906	1.622	0.658
Completion 270°	1.450	0.802	0.898	1.394	0.696
State ridge	0.841	1.089	1.202	1.197	0.755
Scaled Procrustes	1.063	1.461	1.513	1.393	0.755
Wrong correspondence	1.369	0.998	1.469	1.397	1.631
$K = 5$ minimum	1.416	1.018	0.953	1.330	0.931
$K = 5$ ridge	1.416	1.018	1.272	1.319	0.931
$K = 5$ Procrustes	0.981	1.163	1.414	1.228	0.788
Coverage anchor	0.832	0.853	0.747	1.395	0.956
Donor-cosine anchor	1.151	1.100	1.131	1.234	0.981
D-optimal anchor	0.889	0.897	1.086	1.278	1.002
Random anchor 0	1.040	1.100	0.799	1.138	1.040
Random anchor 1	1.054	0.930	1.146	1.378	0.855
Random anchor 2	0.948	0.959	0.615	1.330	0.906
Random anchor 3	0.807	0.857	0.726	1.307	0.740
Matched wrong target	0.831	1.060	0.826	1.385	1.212
$K = 8$ reference	0.553	0.871	1.243	1.340	0.803
Target-native oracle	0	0	0	0	0

Table 23: Descriptive one-correspondence comparison. Negative values favor the coverage-selected anchor over the mean of four fixed random policies. The complete repair criterion is not met.

Edge	Guided minus random-mean normalized defect
Q3→Q7	-0.130
Q3→L7	-0.109
Q3→I8	-0.075
L7→Q7	0.107
L7→I8	0.071

M MMVP External Paired Evaluation

MMVP provides a second public distribution of paired images across nine basic visual patterns (Tong et al., 2024). We select two image-disjoint pairs from each pattern: one supplies a donor direction and one supplies the held-out evaluation context. This design yields 18 pairs and nine source-target-pattern cells for each of the five directed edges. The VQAv2 intervention bases, state maps, primary layer, residual amplitude $a = 0.004$, five-map decision set, paired score, and numerical decision rule remain fixed. MMVP responses do not select any of these quantities.

The external panel resolves 13/45 edge-pattern cells. The certificates span three edges, two target architectures, and eight of nine visual patterns. Resolution remains concentrated on IDEFICS2, with 10/18 cells compared with 3/27 for the other targets. A one-sided Fisher exact test gives $p = 0.0019$. Treating each visual pattern as a paired block, an exact sign-flip test of the IDEFICS2-minus-other resolution difference gives $p = 0.0117$. The Qwen-to-LLaVA edge contributes three additional resolved cells, while both Qwen-7B target edges remain unresolved at the fixed dose. Tables 24 and 25 retain every edge and pattern.

Table 24: Fixed MMVP external evaluation by edge. Each edge contains nine visual-pattern cells. The threshold τ follows the numerical rule applied to that edge's repeated scores.

Edge	Native active	Diameter $> 2\tau$	τ
Q3→Q7	0/9	0/9	2.67×10^{-4}
Q3→L7	0/9	3/9	5.98×10^{-5}
Q3→I8	6/9	5/9	2.44×10^{-4}
L7→Q7	0/9	0/9	2.67×10^{-4}
L7→I8	6/9	5/9	2.44×10^{-4}

The baseline scores in Table 26 describe this fixed 18-pair subset. They characterize the visual questions seen by each checkpoint. The external result measures response variation among equal-fit maps on this subset.

N Fixed Rotation, Dose, and Reverse-Edge Evaluation

This follow-up asks whether the MMVP response separation persists when the rotation grid and intervention dose change. It also adds two large-to-small edges: Qwen2.5-VL-7B to Qwen2.5-VL-3B and IDEFICS2-8B to Qwen2.5-VL-3B. Qwen2.5-VL-3B to IDEFICS2-8B serves as a reference because it was responsive in the original MMVP panel. The follow-up reuses the same nine donor-context assignments, calibration observations, and intervention bases. It is a sensitivity evaluation on those contexts; the original external-panel estimates remain unchanged.

N.1 Fixed construction and comparisons

Reverse maps are fitted from the reversed state correspondence. We exchange the source and target calibration matrices and recompute the minimum-norm map and their orthogonal complements. Each reversed design has rank six. Inverting a fitted forward map would impose its particular complement action, so it would not represent the family compatible with the reversed calibration data. All completion maps have relative calibration residual below 4.2×10^{-14} and share the fitted operator norm.

The rotation comparison uses two disjoint, equally spaced four-angle grids. The original grid is $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, the shifted grid is $\{45^\circ, 135^\circ, 225^\circ, 315^\circ\}$, and their union contains eight rotations. We evaluate each direction at hidden update norms 0.002 and 0.004. These grids and doses are fixed before measuring their responses. Diameter is computed over rotations only. Minimum-norm and target-native directions are evaluated separately.

The numerical comparison uses one threshold per edge across all methods and both doses. Equation 10 gives $\tau = 2.124 \times 10^{-4}$ for each reverse edge and 2.441×10^{-4} for the forward reference. Every score is newly evaluated under the same target-phase layout, with two repeats per sign. Repeats agree exactly; the floating-point scale term determines the thresholds. The comparisons

Table 25: MMVP response-diameter matrix. Cells report $\alpha/(2\tau)$, and bold values exceed the numerical resolution rule. Each row uses one fixed donor pair and one image-disjoint evaluation pair from the named visual pattern.

Visual pattern	Q3→Q7	Q3→L7	Q3→I8	L7→Q7	L7→I8
Orientation and direction	0.054	0.693	**1.550**	0.034	0.935
Presence of specific features	0.112	0.422	**2.434**	0.081	**1.987**
State and condition	0.050	**1.386**	0.857	0.052	**1.083**
Quantity and count	0.034	**1.068**	0.300	0.107	0.687
Positional and relational context	0.160	**1.530**	**3.884**	0.295	**6.002**
Color and appearance	0.177	0.725	0.967	0.117	0.922
Structural and physical characteristics	0.037	0.661	0.474	0.104	**1.493**
Text	0.409	0.996	**4.982**	0.146	**3.163**
Viewpoint and perspective	0.053	0.789	**4.576**	0.053	0.951

Table 26: Descriptive choice-log-likelihood accuracy on the fixed MMVP subset. Pair accuracy requires both images in a pair to be answered correctly.

Checkpoint	Image accuracy	Pair accuracy
Qwen2.5-VL-3B	25/36	7/18
LLaVA-1.5-7B	22/36	5/18
Qwen2.5-VL-7B	29/36	11/18
IDEFICS2-8B	27/36	9/18

comprise 2,160 response evaluations and 18 baselines. The statistical units are 27 edge-pattern cells sharing nine evaluation contexts, not the repeated score evaluations.

N.2 Why these angles preserve the comparison

The completion circle preserves mapped-query norm before normalization. Write $T_\theta = T_0 + \kappa U_t O(\theta) U_s^\top$ with the notation of Equation 16. Orthogonality of the fitted and complementary target ranges gives, for any source query v,

$$\|T_\theta v\|_2^2 = \|T_0 v\|_2^2 + \kappa^2 \|U_s^\top v\|_2^2. \tag{28}$$

The right-hand side is independent of θ. Thus normalization uses the same divisor along the circle, and angle changes direction without changing state fit or update size. The evaluated matrices are rotations in $SO(2)$; the general $O(2)$ construction also permits reflections, which these grids do not sample.

Uniform angles have a quantitative interpretation when response is affine on this circle. Suppose the scalar response of the normalized mapped query has the form

$$f(\theta) = A + B\cos\theta + C\sin\theta. \tag{29}$$

For an even number N of uniformly spaced angles, let D_N be their response diameter and D_{circle} the diameter over the full circle. Then

$$\cos(\pi/N) D_{\mathrm{circle}} \le D_N \le D_{\mathrm{circle}}. \tag{30}$$

To see this, write $f(\theta) = A + r\cos(\theta - \phi)$, where $r = \sqrt{B^2 + C^2}$. A grid point lies within π/N of the maximum, and its antipodal point is also in the grid. Their response difference is at least $2r\cos(\pi/N)$, while the continuous diameter is $2r$. The bound is immediate when $r = 0$.

Four and eight angles therefore retain at least 70.7% and 92.4% of the continuous diameter under this affine-response condition. The empirical certificates use measured finite responses and do not require that condition. For arbitrary nonlinear responses, adding angles can only increase the observed diameter because the original set is contained in the union; that increase alone is not evidence of improved robustness. The two separate four-angle grids provide the equal-size sensitivity comparison.

Table 27: Fixed rotation and dose sensitivity on nine MMVP contexts per edge. Counts are resolved rotation diameters; native is target-native activity. Every grid and dose uses one common numerical threshold within its edge.

Source → target	Norm	Original four	Shifted four	All eight	Native
Qwen 7B → Qwen 3B	0.002	0/9	0/9	0/9	0/9
Qwen 7B → Qwen 3B	0.004	0/9	0/9	0/9	0/9
IDEFICS2 → Qwen 3B	0.002	0/9	0/9	0/9	0/9
IDEFICS2 → Qwen 3B	0.004	0/9	0/9	0/9	0/9
Qwen 3B → IDEFICS2	0.002	4/9	4/9	5/9	3/9
Qwen 3B → IDEFICS2	0.004	5/9	7/9	7/9	6/9

Table 28: Qwen-to-IDEFICS2 response diameter relative to numerical resolution, $\alpha/(2\tau)$. Values above one are resolved. Columns compare the original four, shifted four, and all eight rotations at each update norm.

Visual pattern	Norm 0.002			Norm 0.004		
	Original	Shifted	Eight	Original	Shifted	Eight
Orientation	0.85	1.11	1.11	1.55	2.16	2.16
Specific features	1.16	0.88	1.16	2.43	1.97	2.43
State and condition	0.36	0.56	0.56	0.86	1.09	1.09
Count	0.04	0.08	0.08	0.07	0.03	0.10
Spatial relations	1.91	1.55	1.91	3.88	2.95	3.88
Color and appearance	0.60	0.66	0.66	0.97	1.21	1.21
Physical structure	0.24	0.23	0.26	0.47	0.58	0.58
Text	2.47	2.72	2.72	4.98	5.38	5.38
Viewpoint	2.35	2.24	2.35	4.58	4.40	4.58

N.3 Complete results and interpretation

The forward reference retains response separation on both grids and at both doses. At norm 0.004, the original and shifted grids resolve 5/9 and 7/9 patterns. At norm 0.002, both resolve 4/9. The eight-angle union resolves 7/9 and 5/9, respectively. Tables 27 and 28 report the complete grid comparison and per-pattern response scales. These results show that the observed separation is present beyond the original four angles and single dose.

The reverse edges remain below numerical resolution together with their target-native responses. Both grids and both doses resolve 0/9 patterns for each edge. Table 29 reports their response ranges. Because target-native activity is also 0/9, these measurements establish low response at the tested norm, not uniqueness of the reversed state map's response. They do not add a resolved large-to-small instance.

Individual responses also remain locally stable where both doses are measurable. On Qwen-to-IDEFICS2, all 13 evaluable rotation responses and all four evaluable reference responses retain their sign. Their maximum relative derivative differences are 6.68% and 2.37%, respectively, below the original 25% criterion. Table 30 includes the full denominators. The resolved comparisons support local dose stability without assuming that every response on the circle is linear.

O Coverage and One-Correspondence Results

Coverage association is source dependent. The Qwen-sourced edges have Spearman coefficients -0.116, -0.069, and -0.038, giving mean -0.074 and $p = 0.1934$. The LLaVA-sourced edges have coefficients 0.231 and 0.685, giving mean 0.458 and $p = 0.0027$. The association is resolved for the LLaVA-sourced edges in this finite panel. This source-level diagnostic pools two edges, including one low-activity target, and measures association within the panel.

The source-only acquisition objective reduces coverage relative to the base span. Mean guided-minus-base coverage is -0.195 for Qwen and -0.130 for LLaVA, with all category means nonpositive. It does not improve on every fixed random policy. Several guided and random policies choose

Table 29: Reverse-edge response scale across all nine contexts. Diameter and native response are divided by the same 2τ. Both remain below one in this panel.

Source → target	Norm	Min. diameter	Max. diameter	Max. native
Qwen 7B → Qwen 3B	0.002	0.039	0.119	0.125
Qwen 7B → Qwen 3B	0.004	0.047	0.135	0.212
IDEFICS2 → Qwen 3B	0.002	0.037	0.086	0.125
IDEFICS2 → Qwen 3B	0.004	0.049	0.132	0.212

Table 30: Two-dose stability of individual method-context responses. Evaluable responses exceed 2τ at both doses. Stable responses retain their sign and have relative derivative difference at most 25%. References comprise the minimum-norm and target-native directions; n/a denotes no evaluable response.

Source → target	Methods	Evaluable	Stable	Max. difference
Qwen 7B → Qwen 3B	Rotations	0/72	0	n/a
Qwen 7B → Qwen 3B	References	0/18	0	n/a
IDEFICS2 → Qwen 3B	Rotations	0/72	0	n/a
IDEFICS2 → Qwen 3B	References	0/18	0	n/a
Qwen 3B → IDEFICS2	Rotations	13/72	13	6.68%
Qwen 3B → IDEFICS2	References	4/18	4	2.37%

the same anchor, producing legitimate ties. We retain these ties instead of replacing them with the next random choice.

Response repair is mixed. The coverage anchor has lower activity-normalized defect than the mean random policy on the three Qwen-source edges, with differences -0.130, -0.109, and -0.075. The signs reverse on both LLaVA-source edges, with differences 0.107 and 0.071. The pre-specified joint criterion requires identifiable activity and improvement over every opponent on every edge, and it is not met. Table 22 reports all methods and comparisons.

A new state pair narrows the linear map class and can improve source-space coverage. These changes need not move the target's finite score response toward its target-native response. Target response calibration supplies information that a state correspondence does not contain.

P Numerical Checks and Robustness

P.1 Exact repeats and reconstruction

Every sign of every natural response is evaluated twice. All primary repeats are exactly equal in the stored score representation. The controlled evaluation contains 332 response units and 10,458 atomic four-row measurements, with zero repeat differences and reconstructed-effect discrepancies. The natural evaluation contains 224 response units and 14,528 atomic measurements, and all primary duplicate checks also agree exactly. Summary metrics are independently reconstructed from raw records before figure and table generation.

The MMVP extension contains 72 state objects, 72 baseline objects, and 1,080 response objects. Independent reconstruction recovers every paired effect, edge threshold, response diameter, target-native diagnostic, and exact test from the raw records. All 45 edge-pattern decisions and the 13 resolved certificates agree with the reported summary.

P.2 Dose stability

The stability subset repeats ten headline methods at residual norms 0.002 and 0.004 on eight fixed contexts per edge. A row is evaluable only when both finite effects exceed the activity rule. It then requires matching signs and relative derivative difference

$$\frac{|R^{0.004}/0.004 - R^{0.002}/0.002|}{\max(|R^{0.004}|/0.004, |R^{0.002}|/0.002)} \leq 0.25. \tag{31}$$

Table 31: Interpretation supported by each evidence layer, including the complete auxiliary outcomes.

Evidence	Interpretation
Calibration rank	$K = 4$ paired calibration leaves a two-dimensional complement in the eight-dimensional design.
Controlled response	Equal-fit, equal-budget completions have nonzero deterministic target-response diameter in all 249 controlled groups.
Natural response	Image-disjoint directions have resolved completion diameter for two source families into IDEFICS2 at the fixed budget.
MMVP recurrence	Fixed equal-fit maps yield 13/45 resolved edge-pattern cells across three edges and eight visual patterns.
Step stability	Two-step responses are stable on every evaluable IDEFICS2 row.
Coverage association	Coverage is associated with diameter on LLaVA-sourced edges in the fixed panel.
Added correspondence	One additional state correspondence does not reliably repair target response across the five edges.
Source dependence	Coverage is not a source-independent predictor in this panel.
Sign portability	Source response sign remains unidentified across the five edges.
Other update norms	Completion uniqueness for Qwen-7B and LLaVA targets remains unmeasured at prospective update norms.
Study scope	The claim is conditional on the evaluated map class, model panel, response score, intervention sites, and matched within-target update budgets.

The two IDEFICS2 edges have 24/24 and 36/36 stable rows. The remaining 340 rows are low-activity and are not reclassified as unstable.

P.3 Fixed norm and target scale

Normalizing every target direction before intervention prevents one map from gaining response merely through a larger mapped norm. This is the fairness condition for the within-target response diameter. The same absolute residual norm occupies different fractions of each model's natural paired-state scale. Consequently, absolute diameters and low-activity rates are not compared as model-family rankings. A model-relative dose would define a different finite-response estimand and is outside this fixed panel.

P.4 Second site

The second site evaluates the minimum-norm map, four completions, and target-native oracle on two fixed edges. Qwen 3B to Qwen 7B has 0/8 resolved completion groups and 0/8 identifiable target-native groups. LLaVA to IDEFICS2 has 5/8 and 4/8. The result establishes persistence on the fixed LLaVA-to-IDEFICS2 subset.

Q Additional Results and Scope

Table 20 reports every additional natural-panel analysis. Augmented calibration reaches the expected rank and state fit. Source coverage acquisition and added-correspondence policies do not consistently reduce response defect, while source effects remain below the identification threshold on every natural edge and in the conservative controlled sensitivity analysis. The available responses therefore do not determine source-to-target sign portability.

These findings complement earlier work. Shared geometry and representational similarity describe observed state structure (Kornblith et al., 2019; Huh et al., 2024; Hu et al., 2026). Cross-model steering studies provide evidence for the specific maps and responses they evaluate (Huang et al., 2025; Oozeer et al., 2025; Chen et al., 2025; Syed, 2026), while direct intervention evaluations measure the responses of selected updates (Bhalla et al., 2024; Su et al., 2025; Sivakumar et al., 2025). Our contribution is the matched comparison across alternative maps with the same observed state fit.

R Implications for Intervention Evaluation

Table 3 in the main paper separates calibration rank, map sensitivity, target update norm, source activity, target response, and robustness. The separation matters because each object rules out a different explanation. High rank addresses unobserved linear directions. Matched target norm removes amplification. Source and target responses establish the two behavioral quantities that a transfer claim relates.

This evidence hierarchy gives state alignment a constructive role. A well-conditioned full-rank design can remove the linear completion ambiguity studied here. A rank-deficient design can still support an intervention claim when the proposed map is evaluated directly on target responses. Response diameter identifies where this additional behavioral evidence is needed.

Numerical resolution determines how each cell can be interpreted. When the target score falls below the activity rule, map agreement at that dose does not establish response uniqueness. With unresolved source activity, the measured certificate concerns target-response underdetermination. Each outcome identifies the behavioral quantity established by the experiment.

S Reproducibility Details

S.1 Fixed public components

The reproduction program pins the four public checkpoint revisions listed in Table 7, the VQAv2 data revision used to materialize paired examples, and the public MMVP revision used by the external panel. Its fixed input definitions specify the 120-pair VQAv2 roles, the MMVP 18-pair selection, and exclusions. During execution it writes these manifests together with row-to-image records, raw question and answer fields, and image identifiers. These records allow the row-level and image-level separation claims to be checked directly from the public data.

The reported execution uses PyTorch 2.3.0a0, CUDA 12.3, NumPy 1.24.4, and Transformers 4.49.0. The cell pins Accelerate 1.2.1, Hugging Face Hub 0.28.1, and Pillow 11.3.0. Controlled text images use Pillow's default 10-pixel font, independent of installed system fonts. The Supplementary Material contains the plain-text program, a README, and the reported run's computed states and maps, raw responses, per-edge reports, and numerical tables. The program recomputes these quantities from public inputs without reading the supplied results.

S.2 Executable experiment paths

The supplement provides a plain-text program that runs in one Google Colab cell. It downloads the public data and pinned model weights, generates controlled images, captures residual states, constructs intervention bases, fits the maps, and evaluates all four experimental stages. The MMVP stage uses VQAv2 bases and maps computed earlier in the same run. The rotation-and-dose stage then uses the newly computed donor states and measures its own responses. Fixed sample assignments, prompts, seeds, angles, and doses specify the experiment; no previous result archive or precomputed state is required.

Each response record identifies the model, revision, layer, input pair, direction, sign, and repeat. Interrupted executions reuse completed records from the same run when these fields match. Map fitting and analysis use fixed numerical settings. A rank-deficient Procrustes fit has multiple optimal extensions on the unobserved complement, so its selected extension can depend on the linear-algebra backend. The output retains fitted maps, raw scores, and the execution environment for direct comparison.

S.3 Independent result reconstruction

For the reported run, separate CPU analyses rebuilt the main response statistics from raw records. They checked response cardinalities, exact repeats, state-fit residuals, query norms, response diameters, thresholds, per-category counts, Spearman coefficients, Fisher and sign-flip values, and coverage permutation values. The supplied experiment program computes these statistics and the

Table 32: Traceability from paper claims to stored evidence. The experiment program and separate analysis scripts implement the checks listed here.

Claim	Primary record	Reconstruction check
Rank budget	Calibration states and singular values	Recompute every subset rank and compare with Equation 5
Equal state fit	Candidate maps and calibration matrices	Rebuild residuals and observed-query diameter
Matched map budget	Completion matrices and query vectors	Recompute operator and mapped-query norms
Core diameter	Atomic target score rows and fixed map identities	Rebuild vector and scalar diameters for all 249 groups
Natural diameter	Donor, context, map, sign, and repeat records	Rebuild all edge-category counts
MMVP recurrence	Fixed bases and maps, paired states, raw scores, signs, and repeats	Rebuild all 45 cells and both exact tests
Step stability	Paired records at $a = 0.002$ and $a = 0.004$	Recompute activity, sign, and derivative conditions
Coverage association	Source residuals and group diameters	Recompute edge ranks, source means, and blocked permutation values
Repair outcome	Per-method target responses and oracle responses	Recompute every opponent, category, and joint criterion
Image separation	Public row and image identifiers	Check all specified cross-role intersections

pre-specified comparisons and joint repair criteria from its own records. The paper's figures and quantitative tables use the reconstructed results and experiment reports.

S.4 Claim-to-artifact traceability

Each empirical claim is linked to its primary records and analysis procedure. Table 32 gives the path from stored evidence to the corresponding manuscript result. Record cardinalities are checked before summary computation.

The experiment program checks expected record sets before aggregation. A missing sign, repeat, map, donor, context, site, or edge prevents summary computation instead of reducing a denominator. Separate reconstruction checks the raw-response cardinalities used by the main analyses.