CAUSAL ANALYSIS OF PHYSICAL UNDERSTANDING IN VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding physical variables from visual representations does not establish how their encoded information contributes to task prediction. The challenge is to separate target-specific contributions from general intervention effects and the loss of information shared with related variables. We introduce CausalProbe, a framework for representation-level causal analysis. It progressively erases target-predictive directions while protecting selected readouts of related variables. Random and reverse-role erasure provide comparisons for interpreting the resulting task effects. Across six frozen encoders, erasing second-order geometry impairs grasping-affordance prediction on SuctionNet much more than erasing first-order geometry, supporting its stronger causal contribution. On Physion++, spatial erasure similarly affects future-contact prediction more than motion erasure. Removing either family without protection also reduces the other's predictability, indicating overlap between spatial and motion representations. Our framework moves beyond identifying what visual representations encode to examining how that information supports task prediction.

1 INTRODUCTION

The physical world is characterized by many variables, including local geometry, spatial position, and velocity, among others, that can directly influence the outcomes of specific tasks. Local surface geometry influences the contact a suction gripper can form with an object, while relative position and motion help determine whether two objects will make contact in the future (Cao et al., 2021; Tung et al., 2023). Visual foundation models offer representations for these tasks, but their features do not explicitly separate such physical variables. This raises a question beyond whether a model can predict them: **how does information about a physical variable contribute to prediction on a related task?**

Probing provides a starting point for examining these implicit representations. By fitting readouts to frozen features, Probe3D measures their accessibility for depth, surface orientation, and correspondence (El Banani et al., 2024); Video-based studies similarly examine the prediction of dynamic physical properties and the layerwise accessibility of motion variables (Zhan et al., 2026; Joseph et al., 2026). These analyses reveal what can be predicted from a representation and how these capabilities are distributed within the model. However, decoding a variable alone does not establish whether the encoded information contributes to prediction on a related task (Elazar et al., 2021).

Beyond probing, recent work has moved toward intervening on representations. GeoSubProbe decomposes geometric probe weights to isolate task-aligned subspaces and examines how retaining or removing these subspaces affects geometric prediction, revealing the concentration and compressibility of geometric information (Zhou et al., 2026). Video representation analyses also combine probing with targeted ablations to investigate distributed physical information (Joseph et al., 2026). Going further, we ask a distinct question: *does information predictive of a target variable make a causal contribution to task prediction, beyond information associated with related variables?*

Simply comparing the task prediction performance and the variable prediction only establishes **correlation** instead of **causal dependence**. To investigate the target information's contribution, we can erase target-predictive directions and compare the change in task performance with random erasure. However, the erased directions may also encode other related variables that contribute to task prediction. For example, erasing curvature-predictive directions may also disrupt surface-orientation,

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

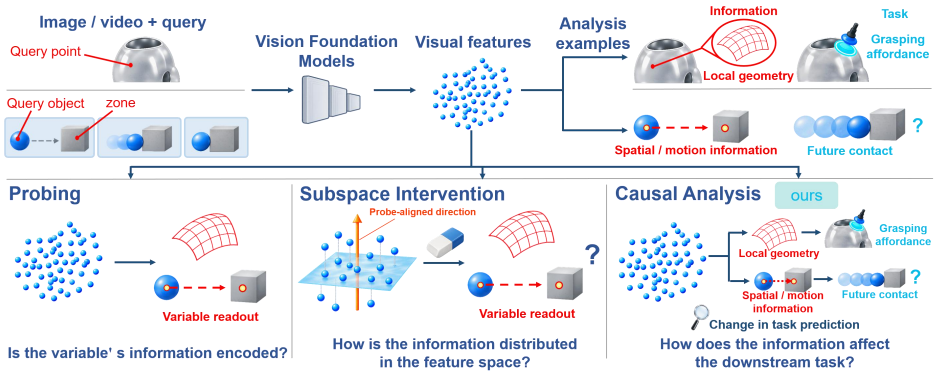

Figure 1: **Causal analysis of physical understanding in visual representations.** We extract query-aligned features from frozen vision foundation models to analyze physical variables and their contributions to task prediction. We study two tasks: grasping affordance prediction and future contact prediction. With grasping affordance as an example, **probing** tests whether local geometry is encoded; **subspace intervention** tests whether geometry information can be removed; and our **causal analysis** examines how local geometry changes grasping affordance prediction.

while erasing spatial directions may impair motion prediction. The resulting performance decline would combine these losses, **obscuring the target variable's specific contribution**.

To this end, we introduce CausalProbe, a framework for representation-level causal analysis through controlled subspace erasure. It progressively erases target-predictive directions while protecting selected readouts of related variables. After erasure, we refit probes and task predictors to assess changes in variable predictability and task performance. We compare these changes by erasing the same number of random directions and by reversing the roles of the target and protected variables. These comparisons help separate general erasure effects from changes associated with the target variable, strengthening evidence that target information contributes causally to task prediction beyond the protected readouts. We apply this framework to six frozen visual encoders in two complementary settings: static grasping affordance on SuctionNet and video-based future contact prediction on Physion++ (Fig. 1). Our contributions are threefold:

- **A framework for causal analysis of physical understanding in visual representations.** We propose a controlled intervention framework that selectively erases target relevant information while protecting predictions of related variables. This enables us to assess the representation-level causal contribution of target-variable information to task prediction.
- **Evidence for the causal contribution of local geometry.** Erasing second-order geometry while protecting first-order readouts consistently impairs grasping-affordance prediction more than random or reverse-role erasure, supporting its stronger representation-level causal contribution.
- **Extension to spatial and motion information in video.** We extend the framework to video-based future contact prediction, where spatial erasure has a larger effect than motion erasure. Erasing either without protection also impairs prediction of the other, indicating overlap between spatial and motion representations.

2 Related Work

Geometric and physical understanding in visual representations. Physical-understanding benchmarks evaluate plausibility and interaction reasoning through IntPhys, PHYRE, and CLEVRER (Riochet et al., 2018; Bakhtin et al., 2019; Yi et al., 2020), and future-event prediction through Physion and Physion++ (Bear et al., 2021; Tung et al., 2023). Dex-Net and SuctionNet connect local geometry to grasping and suction (Mahler et al., 2018; Cao et al., 2021). Complementing these task-level evaluations, probing examines information accessible from frozen representations (Alain & Bengio, 2016; Hewitt & Liang, 2019). Probe3D, Lexicon3D, and SeeSE3 examine ge-

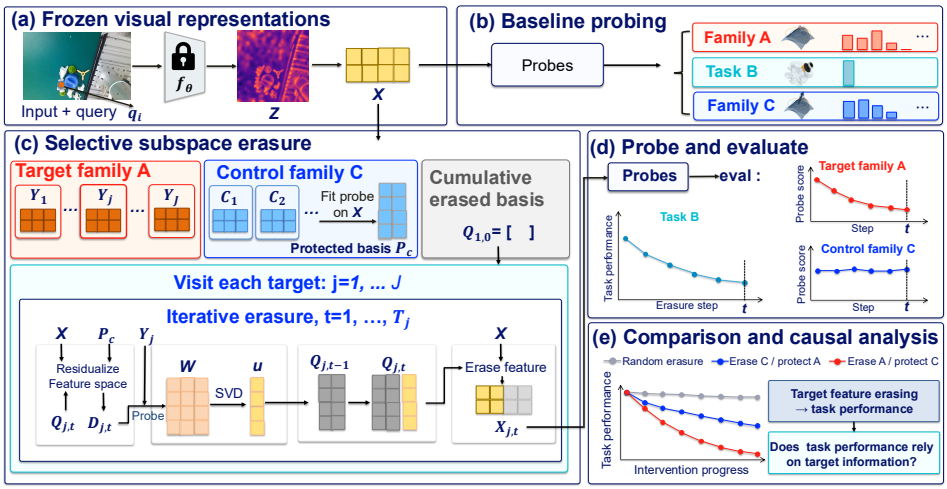

Figure 2: **CausalProbe framework.** (a) Extract query-aligned features from a frozen visual encoder. (b) Establish baseline target variable prediction, control variable prediction, and downstream task performance. (c) Visit target groups sequentially. For each family, iteratively fit a probe, extract an admissible direction, and update the cumulative erased basis while protecting control directions. Project this basis out of the original features and carry it across target groups. (d) Refit target/control probes and the task head on erased features at recorded checkpoints. (e) Compare task performance under forward, reverse, and random erasures, interpreted alongside target and control prediction.

ometric and spatial information in visual features (El Banani et al., 2024; Man et al., 2024; Chen et al., 2026). Video analyses further examine dynamic physical properties and the layerwise organization of motion information (Zhan et al., 2026; Joseph et al., 2026). However, decodability and representational structure alone do not establish the causal contribution of physical information to prediction on another task.

Subspace intervention and attribution. INLP iteratively projects out linearly decodable information (Ravfogel et al., 2020), and Amnesic Probing uses erasure to examine its role in prediction (Elazar et al., 2021); LEACE provides a closed-form linear erasure operator (Belrose et al., 2023). GeoSubProbe applies subspace interventions to characterize the concentration and accessibility of geometric information (Zhou et al., 2026). TCAV relates concept directions to model predictions (Kim et al., 2018), while analyses of activation patching show that intervention effects alone can give misleading attributions (Makelov et al., 2024). Our focus is the representation-level causal contribution of physical information to a distinct prediction task. CausalProbe uses variable protection and controlled erasure comparisons to strengthen this attribution.

3 Problem Formulation

Let $\mathcal{D} = \{(x_i, q_i, a_i, b_i, c_i)\}_{i=1}^{N}$ contain visual inputs x_i, queries q_i, target-variable values a_i, task labels b_i, and control-variable values c_i. We study the representation-level causal contribution of target family A to predicting B, while accounting for related control family C. As shown in Fig. 2, we probe frozen visual features, selectively erase target-predictive directions while protecting selected control readouts, and compare variable prediction and task performance across targeted, reverse-role, and random erasures.

4 Variable-Targeted Representation Interventions

4.1 Frozen Visual Representations

For input x_i and query q_i, a frozen encoder and query-aligned readout produce

$$z_i = \phi(f_\theta(x_i), q_i), \qquad X = T(Z) \in \mathbb{R}^{N \times d}. \tag{1}$$

Here Z stacks the sample features, and T applies training-fitted standardization and optional PCA. The resulting X contains N query-aligned rows and d feature coordinates (panel (a)). The encoder and transform remain fixed across splits and erasure steps.

4.2 BASELINE PROBING

The analysis begins by verifying whether the variables of interest can be decoded from frozen visual representations. The target family A comprises scalar- or vector-valued groups $Y_j{}_{j=1}^{J}$, each standardized using a transformation S_j fitted on the training data: $\widetilde{Y}_j = S_j(Y_j)$. Before erasure, we use linear probes to evaluate the prediction of target variables (A) and control variables (C) from X on held-out data, and a regression or classification head to evaluate downstream task performance (B; panel (b)). Test data are excluded from both fitting and model selection. Detailed target definitions and fitting protocols are provided in Sec. 5.1 and Sec. A.

4.3 SELECTIVE SUBSPACE ERASURE

A natural next step after probing is to erase directions predictive of A and measure the change in downstream task performance. However, these directions may also encode information about related variables C that contribute to predicting B. A performance drop could therefore reflect **the disruption of shared information, rather than a causal contribution specific to variable** A.

To better isolate **the causal contribution of variable** A **at the representation level**, we progressively erase A-predictive directions while protecting directions used by selected C decoders (Panel (c)). A task-performance drop with control prediction maintained provides stronger evidence for this contribution beyond the protected control information. To construct the protected subspace, we fit control decoders on X and define the protected basis as $P_C = \mathrm{orth}([W_{C,1}, \ldots, W_{C,L}])$. Here $W_{C,\ell}$ contains the feature-side weights of control decoder ℓ. The erased basis starts empty, $Q_{1,0} = \varnothing$.

Two nested loops. A target group may remain predictable after removing a single direction. We therefore refit its probe and erase additional directions iteratively, before moving to the next group. The outer loop visits target groups $j = 1, \ldots, J$ in a fixed order; the inner loop adds one erased direction per successful update, for $t = 0, \ldots, T_j - 1$.

1. Construct the discovery input. Direction discovery uses the component of X orthogonal to the protected and previously erased spans:

$$\Pi_{j,t} = I - P_C P_C^\top - Q_{j,t} Q_{j,t}^\top, \qquad D_{j,t} = X\Pi_{j,t}. \tag{2}$$

The resulting $D_{j,t}$ is used only for direction discovery.

2. Fit a probe and extract a target direction. We fit a probe to predict the current target from $D_{j,t}$, identifying information that remains readable outside the excluded spans:

$$(W_{j,t}, b_{j,t}) = \mathrm{FitProbe}_{\mathcal{O}_{j,t}}(D_{j,t}, \widetilde{Y}_j), \qquad W_{j,t} \in \mathbb{R}^{d \times q_j}, \tag{3}$$

where q_j is the target group's output dimension and $\mathcal{O}_{j,t}$ specifies the fitting protocol. We then extract one feature-space direction from the projected probe weights using SVD:

$$\Pi_{j,t} W_{j,t} = U_{j,t} \Sigma_{j,t} V_{j,t}^\top, \qquad u_{j,t} = U_{j,t}[:, 1]. \tag{4}$$

The leading nonzero singular direction captures the greatest sensitivity of the fitted decoder within the admissible span; it is not selected to maximize downstream task damage. Ridge and gradient-based fitting protocols are detailed in Sec. A.

3. Constrain and accumulate. Before erasure, we reorthogonalize the extracted direction against the protected and previously erased spans to correct numerical drift. We then normalize it and append it to the cumulative erased basis:

$$\widetilde{u}_{j,t} = \Pi_{j,t} u_{j,t}, \qquad Q_{j,t+1} = \left[Q_{j,t}, \frac{\widetilde{u}_{j,t}}{\|\widetilde{u}_{j,t}\|_2} \right]. \tag{5}$$

Numerically degenerate directions are rejected; each accepted update adds one new erased direction while preserving the basis constraints.

4. Erase from the original features. Using the updated basis, we remove all accumulated directions from the original features:

$$X_{j,t+1} = X(I - Q_{j,t+1}Q_{j,t+1}^{\top}). \tag{6}$$

Unlike the discovery input, $X_{j,t+1}$ retains the protected span. It has the same feature dimension as X and preserves the predictions of the protected fixed decoders. These erased features are passed to the probes and task head in Panel (d) to evaluate the intervention.

The inner loop continues until the non-test discovery monitor is nonpositive, no admissible direction can be extracted, or the local budget is reached. We then move to the next target while retaining the cumulative erased basis. After the final target, the method returns $Q_{\text{final}} = Q_{J,T_J}$ and $X_{\text{erased}} = X(I - Q_{\text{final}}Q_{\text{final}}^{\top})$, together with recorded checkpoints at each erasure step. Detailed budgets, fitting scopes, and reverse-path policies are provided in Sec. A.

4.4 Probe and evaluate during erasure

We next assess whether erasure reduces target predictability, affects control information, and changes downstream task performance (Panel (d)). After each rank-one erasure, we refit target/control probes and the task head on the updated features $X_{j,t}$ to assess how well the remaining representation predicts the target variables, control variables, and downstream task. Tracking these predictions across erasure steps reveals how target predictability, control predictability, and downstream task performance change as more directions are removed.

4.5 Comparison and representation-level causal analysis

We compare three interventions to assess whether task-performance changes reflect the removal of target-predictive information (Panel (e)). First, we erase directions predictive of A while protecting C. The resulting performance drop may arise from a general reduction in representation rank rather than the specific information removed. We therefore use **full-space random erasure at the same removed rank** as a reference for general rank-reduction effects.

A larger drop than under random erasure does not yet show whether the effect is specific to variable A or also occurs when removing directions predictive of another variable. We therefore introduce **reverse-role erasure**: erase C-predictive directions while protecting A. This comparison tests whether task performance responds differently to removing information about the two variable families. Each intervention is evaluated with a separately refitted task head. Together, these interventions enable **representation-level causal analysis** by comparing task responses to targeted, equal-rank random, and reverse-role erasure. Target and control prediction reveal which information is affected, helping interpret the task's dependence on the targeted information.

5 Experiments

5.1 Experimental setup

We evaluate physical-variable prediction and the effects of selective subspace erasure on downstream task performance using six frozen encoders: VGGT (Wang et al., 2025), VGGT-Omega (Wang et al., 2026), DINOv2 (Oquab et al., 2023), DINOv3 (Siméoni et al., 2025), LingBot-Vision (Fu et al., 2026), and V-JEPA 2 (Assran et al., 2025). The experiments cover two complementary aspects of physical understanding (Fig. 1): grasping-affordance prediction from static contact geometry and future-contact prediction from observed dynamics. We use SuctionNet (Cao et al., 2021) and Physion++ (Tung et al., 2023) for the respective tasks. Probe capacity is selected without test data. All intervention conditions use the same test examples, allowing direct comparison of their effects on prediction. Full experimental protocols are provided in Secs. B and C.

Local Geometry and Grasping affordance. We use SuctionNet (Cao et al., 2021) to examine how local geometry contributes to grasping affordance prediction. The target family A comprises second-order local geometry estimated from quadratic surface fits, the control family C comprises

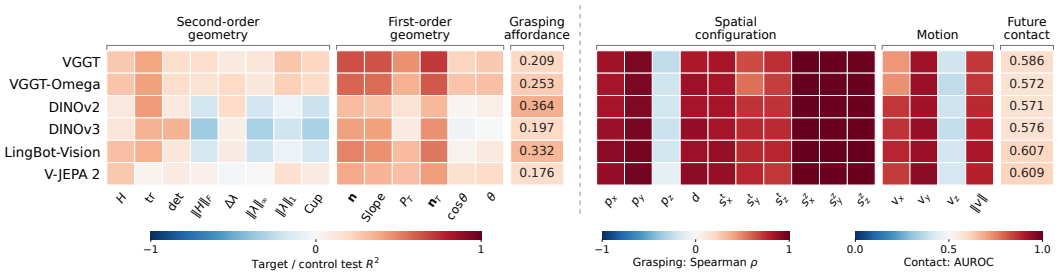

Figure 3: **Decodability before intervention.** Shared rows identify encoders; the dashed divider separates SuctionNet (left) from Physion++ (right). Target/control cells show test R^2; annotated task cells show grasping-affordance Spearman correlation and future-contact AUROC. Color scales center on zero for R^2/Spearman and 0.5 for AUROC. H, λ, and Cup denote the scale-aware quadratic matrix, its eigenvalues, and cup-rim deviation; s^t, s^z denote target/zone scales.

first-order orientation descriptors, and task prediction B is the seal score. Features combine a query-point readout with an independently encoded local crop. We erase second-order directions while protecting first-order directions, and exchange these roles for the reverse intervention.

Spatial/Motion information and Future contact prediction. We use Physion++ (Tung et al., 2023) to examine how spatial and motion information contribute to future target–zone contact prediction. The target family A comprises ten continuous spatial descriptors, including relative position, distance, and target/zone scales; the control family C comprises relative velocity and speed; and B indicates future contact. Encoders receive observed frames and a query-pair cue, while simulator states provide supervision only.

Test R^2 measures linear-probe prediction performance: $R^2 = 1 - \frac{\sum_i \|y_i - \hat{y}_i\|_2^2}{\sum_i \|y_i - \bar{y}_{\text{test}}\|_2^2}$, where y_i, $\hat{y}_i$, and $\bar{y}_{\text{test}}$ denote the ground truth, prediction, and test-set mean, respectively, with sums taken over test samples. A score of 1 indicates perfect prediction, 0 matches the constant test-mean predictor, and negative values are worse. Maximum test R^2 is the highest score across target groups (SuctionNet) or coordinates (Physion++). Grasping affordance performance is measured by test Spearman correlation between predicted and ground-truth seal scores. It measures how well predictions rank query points by seal quality: higher correlation indicates closer agreement with the ground-truth ordering (formal definition in Eq. (19)). Future-contact performance is measured by the AUROC of a refitted logistic readout (0.5: chance). Higher values indicate better prediction performance.

5.2 Decodability before intervention

Fig. 3 shows readable target and control information in all six encoders. On SuctionNet, geometry-probe performance varies across models. The result reveals that 3D vision foundation models and V-JEPA 2 can decode more geometry information than the three 2D vision transformers. On Physion++, spatial and motion prediction is generally strong, although data limitations prevent reliable probing of z-axis spatial and velocity information; vertical-velocity prediction remains consistently weak. Future-contact AUROC is modestly above chance.

5.3 Causal analysis of downstream predictions

We examine target and control prediction under protected erasure to establish what each intervention removes and preserves. We then assess changes in downstream task performance. Following our three-intervention design, random erasure provides a reference for general rank-reduction effects, while reverse-role erasure tests the response to removing information about the other variable family.

5.3.1 Target suppression and control preservation

Grasping affordance. Geometry erasure largely preserves prediction of the other family. In Fig. 4, second-order prediction falls sharply while first-order prediction remains stable (top). Reversing the roles produces the corresponding pattern (bottom). These trends support comparatively

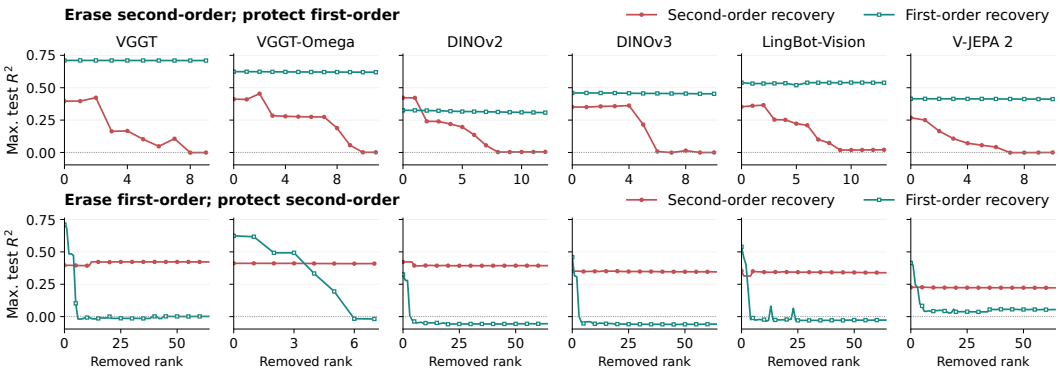

Figure 4: **Variable prediction during protected geometry erasure.** Columns identify encoders. Top: erase second-order geometry while protecting first-order directions; bottom: reverse roles. Refitted probes measure maximum test R^2 across eight second-order or six first-order groups. Lower values indicate lower probe performance; a stable maximum does not certify preservation of every group.

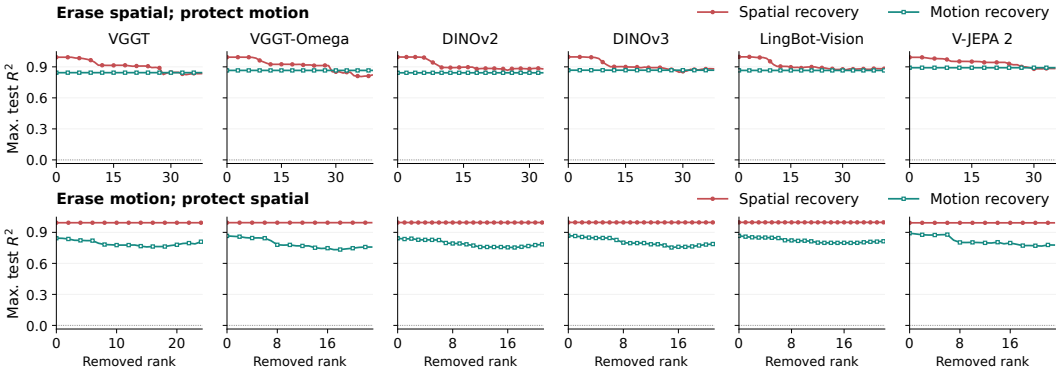

Figure 5: **Variable prediction during protected spatial/motion erasure.** Columns identify encoders. Top: erase spatial targets while protecting motion directions; bottom: reverse roles. Refitted probes report maximum test R^2 across ten spatial or four motion coordinates. Spatial targets exclude target type. High remaining probe scores indicate incomplete target suppression.

selective interventions, helping connect the task effects examined below to the geometry family targeted. Group-wise audits appear in Fig. 13.

Future contact prediction. Spatial and motion information remain predictable under protection. In Fig. 5, target prediction decreases only moderately in either direction, while control prediction remains stable. Unlike the geometry experiment, protecting the related family restricts target suppression. The downstream analysis therefore examines how this partial removal affects task performance.

5.3.2 Comparison and Causal Analysis

Grasping affordance. Second-order geometry makes a strong representation-level causal contribution to grasping affordance. Erasing second-order directions reduces performance in all six encoders despite preserving the selected first-order decoders (Fig. 6); DINOv2 and DINOv3 reach negative Spearman correlation. This is intuitive since local curvature is very relevant to grasping verification.

First-order erasure produces no comparable task impairment. When the roles are reversed, task effects are much smaller and performance sometimes even improves. The contrasting results support a contribution of second-order information beyond the protected first-order readouts, rather than a uniform effect of geometry-directed erasure. This asymmetry has a geometric interpretation. First-order geometry describes tangent-plane orientation; second-order geometry captures local bending.

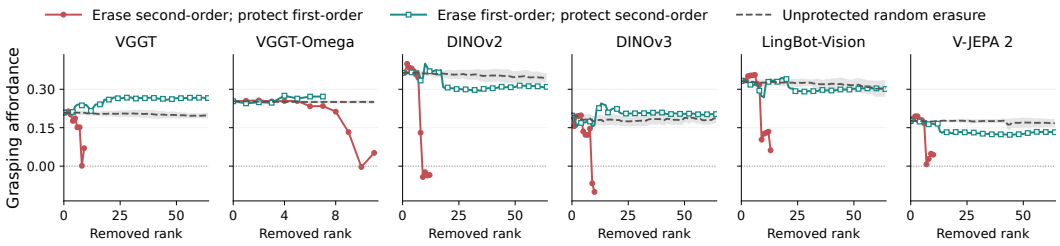

Figure 6: **Protected geometry erasure and grasping affordance.** Each encoder panel compares second-order erasure under first-order protection with the reverse path. Performance is refitted-head test Spearman correlation with seal scores. Gray: full-space random erasure matched by rank only (mean ± SD, eight projections). Protected matched-random controls appear in Fig. 41. Lines join evaluated ranks; path schedules and endpoints differ.

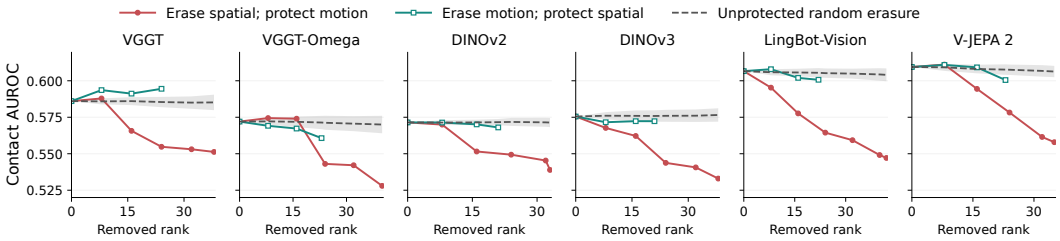

Figure 7: **Protected spatial/motion erasure and future contact.** Each encoder panel compares spatial erasure under motion protection with the reverse path using refitted-head AUROC (chance: 0.5). Gray: full-space rank-only random reference (mean ± SD, 50 projections). Paths have different terminal ranks. Protected matched-random controls appear in Fig. 42.

When the suction cup aligns with the local normal, bending determines surface deviations across the contact region, helping explain the stronger effect of second-order erasure on seal prediction.

Future contact prediction. Spatial information makes a consistent representation-level causal contribution to future contact prediction. Spatial erasure lowers AUROC while preserving the selected motion decoders (Fig. 7). The representation still supports contact prediction after erasure. As the preceding audits show (Fig. 5), spatial predictability decreases only modestly under motion protection. Even this limited suppression, however, produces a substantial decline in future-contact AUROC. In Sec. 5.3.3, we remove motion protection to assess how far spatial predictability can be reduced and how much future-contact performance declines.

Motion erasure has a smaller task effect under spatial protection. Reversing the intervention produces smaller performance changes. The late pre-cutoff frames and target–zone cue provide visual cues to relative position, distance, and object scales. These spatial cues may sustain contact prediction after velocity-predictive directions are removed. The preceding audits also show that motion remains decodable under spatial protection, motivating unprotected tests to examine the effects of further removal.

5.3.3 Cross-family coupling without protection

Variable prediction. Removing protection strengthens target suppression but also disrupts the other family. Compared with the protected trajectories, target prediction declines more strongly in Fig. 8(a). Spatial erasure also reduces motion prediction, and motion erasure reduces spatial prediction. This pattern is consistent with overlapping directions used to predict the two families, explaining why preserving one can limit removal of the other.

Future contact prediction. Spatial erasure remains more damaging without protection. In Fig. 8(b), spatial erasure causes a larger AUROC decline than motion erasure, while the random reference remains comparatively stable. The weaker motion effect thus persists without explicit spatial protection. As discussed above, the final observed target–zone configuration may provide

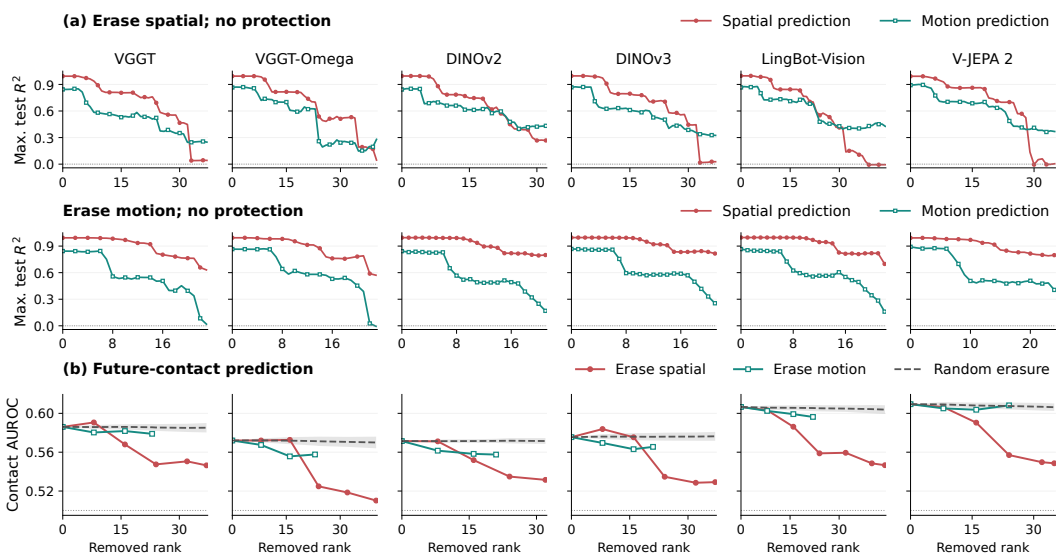

Figure 8: **Variable prediction and task effects under unprotected erasure.** Columns denote encoders; each erasure starts from original features. (a) Refitted-probe maximum test R^2 over ten spatial or four motion coordinates; colors denote predicted families. (b) Refitted-head future-contact AUROC; colors denote erased families. Gray: full-space rank-matched random erasure (mean $\pm$ SD, 50 projections); dotted line: chance (0.5). Curves connect evaluated ranks.

much of the signal needed for contact prediction, limiting reliance on the removed velocity information. Both unprotected erasures affect spatial and motion predictability; their task effects therefore include collateral changes in the other family.

5.4 Further Analysis of Task-Performance Restoration

We test whether predicted second-order geometry improves grasping-affordance prediction after erasure. A linear probe h_A predicts geometry values $\widehat{A} = h_A(X)$ from the original features X. We concatenate these predictions with the erased features X_{erased} and fit a new task predictor on $[X_{\text{erased}}, \widehat{A}]$. Training predictions use scene–frame-grouped out-of-fold splits. This addition improves performance across all six encoders (Fig. 9). Together with the controlled erasure results, these gains support the causal contribution of second-order geometric information to grasping-affordance prediction. The implementation details can be found in Sec. B.6.

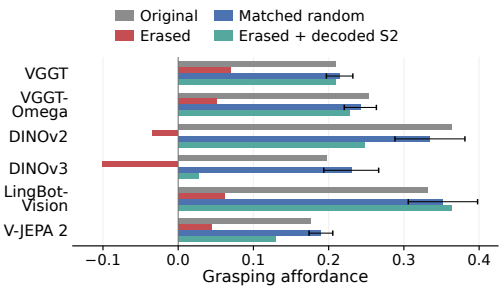

Figure 9: **Second-order restoration.** Test Spearman ρ; S2: decoded second-order geometry. Random: mean $\pm$ SD (8 projections).

6 Conclusion

We introduce CausalProbe, a framework to analyze visual representations' causal contribution to task prediction for physical understanding. Our method incorporates iterative erasure, protection of related readouts, and random and reverse-role comparisons. We identify which variable-predictive information supports task performance. The experiments reveal that second-order geometry erasure impairs grasping affordance than first-order geometry. Spatial erasure similarly impairs future-contact prediction, while spatial–motion information is more coupled in feature space. Reverse-role erasures produce smaller task effects in both settings. Our framework provides a systematic protocol for assessing how physical-variable information contributes causally to task prediction, providing a new analytical framework for visual representation learning.

AI USE STATEMENT

We used generative AI tools to assist with identifying and searching for relevant literature, generating code, deriving mathematical expressions, and editing the manuscript for clarity and readability. The authors manually reviewed the implementation of the core method and carefully checked the relevant core code and mathematical derivations for correctness and consistency with the described methodology. AI assistance supported these activities, while the authors retained responsibility for assessing and incorporating its outputs. We take full responsibility for the final content of this work, including its methods, mathematical expressions, experimental claims, references, and any text or artifacts produced with AI assistance.

ETHICS STATEMENT

This work analyzes frozen visual representations using the existing SuctionNet and Physion++ datasets, with attribution to their original creators. SuctionNet provides object-centric RGB-D scenes and suction annotations, while Physion++ provides simulated physical interactions and associated metadata. Our experiments concern geometric and physical prediction rather than human-subject evaluation or personal profiling. The study is an offline representation analysis and does not establish safe robotic deployment. In particular, seal-score prediction is not a guarantee of grasp success, and performance on simulated contact prediction need not generalize to real-world environments. The qualified SuctionNet subset and uneven Physion++ scenario coverage also restrict the scope of our conclusions. We report these limitations and distinguish interventions on learned representations from physical-world causal effects to reduce the risk of overstating model capabilities. Any downstream robotic use would require task-specific validation and appropriate safety assessment.

REPRODUCIBILITY STATEMENT

We support reproducibility through detailed documentation of the analysis pipeline and its evaluation. Section 4 and Appendix A specify the intervention formulation, protected subspaces, direction discovery, stopping policies, and probe and task-head evaluation, with pseudocode in Algorithm 1. Appendices B and C describe the data sources, sample selection, supervision construction, preprocessing, feature extraction, data splits, and fitting protocols. Tables 5 and 10 report partition sizes and grouping units; the accompanying text explains subset restrictions and scenario coverage. These appendices also document restoration procedures, per-model trajectories, additional results, and the interpretation of uncertainty estimates. Appendix F distinguishes shared random baselines from path-matched controls. Further implementation and diagnostic details appear in Appendices B.11, B.12, and D, allowing readers to trace the reported findings to their experimental conditions and assumptions.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Hanwen Cao, Hao-Shu Fang, Wenhai Liu, and Cewu Lu. Suctionnet-1billion: A large-scale benchmark for suction grasping. *IEEE Robotics and Automation Letters*, 6(4):8718–8725, 2021.

Caroline Chen, Sayna Ebrahimi, Fedor Kitashov, Ming-Hsuan Yang, Leonidas Guibas, Viorica Pătrăucean, and Maks Ovsjanikov. Seese3: Emergence of 3d space in vision features. *arXiv preprint arXiv:2607.14228*, 2026.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. In *CVPR*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception. *arXiv preprint arXiv:2607.05247*, 2026.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. *arXiv preprint arXiv:2602.07050*, 2026.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. *Advances in Neural Information Processing Systems*, 35, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David Gealy, and Ken Goldberg. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In *2018 IEEE International Conference on robotics and automation (ICRA)*, pp. 5620–5627. IEEE, 2018.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6: 1257–1267, 2022.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622. Association for Computational Linguistics, 2020.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D. Cotterell. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 2022a.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6034–6055. Association for Computational Linguistics, 2022b.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9413–9431. Association for Computational Linguistics, 2023.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

Guanqi Zhan, Xianzheng Ma, Weidi Xie, and Andrew Zisserman. Inferring dynamic physical properties from video foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5214–5224, 2026.

Weichen Zhou, Yawen Zou, Chunzhi Gu, Ran Dong, Haoran Xie, and Chao Zhang. Understanding geometric representations in self-supervised vision transformers via subspace intervention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

A Analysis Framework: Definitions and Implementation

Both studies implement the constrained discovery and fixed target sweep described in Secs. 3 to 4. They share the intervention rule. Target grouping, probe grids, and evaluation controls are specified for each dataset below.

A.1 Notation and the object of intervention

For sample i, (x_i, q_i) is the visual input and query, (a_i, b_i, c_i) provides target, task and control supervision, and g_i identifies a grouping unit. Grouping units define the split and resampling structure. Frozen feature construction gives $z_i = \phi(f_\theta(x_i), q_i)$ and $X = T(Z)$. The fitted transform T includes standardization and optional PCA. All rows, including held-out rows, use the same transform. PCA remains fixed throughout erasure.

Symbol	Meaning throughout the framework and case studies
x_i, q_i	Image/video and query: a surface point in SuctionNet, a target–zone pair in Physion++.
A, B, C	Target family, downstream response, and measured control family. They may be vector-valued and correlated.
$Y_j, \widetilde{Y}_j$	Raw target group and its training-standardized coordinates.
$X \in \mathbb{R}^{N\times d}$	Preprocessed feature matrix constructed from the visual inputs.
$P_C \in \mathbb{R}^{d\times r_C}$	Orthonormal columns spanning protected decoder directions; $P_C P_C^\top$ is the corresponding projector.
$Q_{j,t} \in \mathbb{R}^{d\times k_{j,t}}$	Orthonormal erased directions after t updates within target group j, carried cumulatively between groups.
$X_{j,t}, \mathcal{E}_{j,t}$	Erased features $X_{j,t} = \mathcal{E}_{j,t}(X) = X(I - Q_{j,t}Q_{j,t}^\top)$.
$W_{j,t}$	Feature-side decoder weights of shape $d \times q_j$ at local round t for target group j.
g_B, h_A, h_C	Downstream task head and target/control probes fitted to frozen features.
k_*, K, t_0	Selected/terminal global ranks and last visible physical time (Physion++ only).

Table 1: Notation ledger. The nested indices j, t denote target-group and local erase iterations; k is the cumulative global rank used for reporting. Neither is physical time. The local quadratic matrix is H; B always denotes the downstream response in the manuscript.

A.2 Target preparation is separate from protection

The target family is finite and explicitly specified. A family member can be a vector group, such as the three entries of a symmetric matrix, or a scalar descriptor. Target dimensionality, erased feature rank, and the number of underlying physical factors are distinct quantities.

Each target group is train-standardized:

$$\widetilde{Y}_j = S_j(Y_j). \tag{7}$$

Here S_j uses only the fitting partition's per-coordinate mean and scale. Scores are computed in the target units used by the declared evaluator. We keep the raw physical target definitions in forward, reverse, and joint interventions.

Two target formulations used for the appendix comparison. The main paper uses the raw formulation above. We compare this with a linear residual formulation. Given a declared control family C, fit a training-only map m_j and define

$$Y_j^{\text{raw}} = Y_j, \qquad Y_j^{\text{res}} = Y_j - \widehat{m}_j(C), \qquad \widetilde{Y}_j^s = S_j^s(Y_j^s), \quad s \in \{\text{raw}, \text{res}\}. \tag{8}$$

The residual path tests decodability and removal of the component left after linear regression on C. The raw path targets the complete measured A family. Both preserve the same selected C-decoder span. Changing the target can change native R^2, endpoint rank, and stopping time. We compare the two label families on the same endpoint features wherever both predictions are available, and report a smaller-rank check.

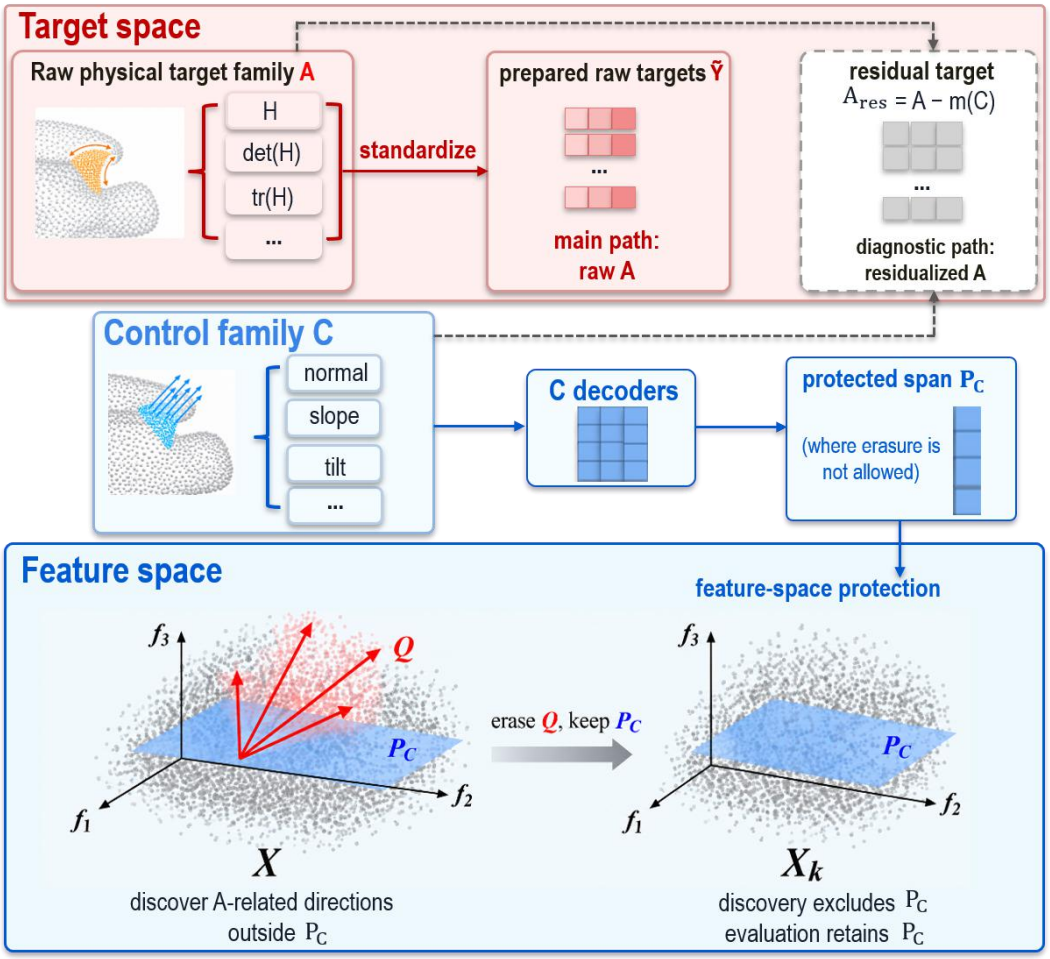

Figure 10: **Raw-target analysis framework.** Standardized physical target families guide constrained discovery. C-decoder protection acts on feature directions; evaluation retains the protected span. A fixed sweep determines the endpoint without test or downstream-task feedback. Fresh A/C probes, matched erasure controls, and target-channel restoration distinguish suppression, collateral loss, and task contribution.

Protection is defined in feature space. Fit control decoder weights $W_{C,\ell}$, concatenate their feature-side columns, and retain an orthonormal basis:

$$P_C = \mathrm{orth}([W_{C,1}, \ldots, W_{C,L}])\,. \tag{9}$$

SuctionNet uses the union of independently fitted first-order-group decoders. Physion++ uses the full span of one four-output motion decoder. The decoder weights determine the span's numerical rank. An unprotected intervention sets P_C empty. Decoder fitting scopes are specified for each study below.

The SuctionNet mirror reverses target and control roles: its protected path targets the raw first-order family and protects the second-order decoder union. Physion++ joint removal instead places all raw A and C scalar targets in the active family and uses no protection. This joint intervention evaluates their combined removal.

Choice	SuctionNet main path	Physion++ main protected path
Raw A	Eight second-order groups; ten coordinates from $r_c H$ and its descriptors	Ten continuous spatial coordinates; target type excluded
B	Seal score; ranking measured by Spearman	Future target–zone contact; ranking measured by AUROC
Control C	Six first-order geometry groups	Relative velocity (v_x, v_y, v_z) and speed
C coordinates	Fifteen coordinates across six groups	Four motion coordinates
Protected span	Union of six control-group decoder spans	Full four-output C-decoder span
Evaluation input	Same point-aligned X_k for A/C/B	Same history-derived X_k for A/C/B

Table 2: Mapping the abstract protocol to the main examples. Both studies target raw labels; protection is a feature-space constraint. Full data and fitting details follow in Secs. B and C.

A.3 Direction discovery and iterative removal

For discovery input $D_{j,t}$ at target group j and local round t, a declared probe-weight fitting operator supplies

$$(W_{j,t}, b_{j,t}) = \mathrm{FitProbe}_{\mathcal{O}_{j,t}}(D_{j,t}, \widetilde{Y}_j). \tag{10}$$

The operator $\mathcal{O}_{j,t}$ records the objective, optimizer, regularization or stopping rule, and fitting/selection split. In the main fixed-sweep runs it is instantiated by validation-selected ridge fitting; the gradient-based direction-discovery ablation in Sec. B.9 instantiates the same interface with a linear decoder trained by full-batch AdamW. Both solvers return feature-side weights to the same direction-extraction rule. With the $d \times q_j$ orientation above, the feature-side directions are the left singular vectors of $W_{j,t}$; under a transposed decoder convention they are equivalently the right singular vectors. For a scalar target, the direction is its normalized weight vector. The leading singular direction maximizes decoder sensitivity $\|W_{j,t}^\top u\|_2$ over unit u; without whitening, this should not be interpreted as maximizing explained target variance.

The operator also specifies the fitting and penalty-selection split. Some discovery monitors $m_{j,t}$ are evaluated after the selected decoder is refitted on the combined pool. These monitors guide direction discovery on the refitted pool.

Admissible directions. For orthogonal protected and erased bases at (j,t), define

$$\Pi_{j,t} = I - P_C P_C^\top - Q_{j,t} Q_{j,t}^\top. \tag{11}$$

The direction module explicitly decomposes the projected weights:

$$\Pi_{j,t} W_{j,t} = U_{j,t} \Sigma_{j,t} V_{j,t}^\top. \tag{12}$$

It retains the leading numerically nonzero left singular vector, reorthogonalizes it against P_C and $Q_{j,t}$, and normalizes it before appending it to $Q_{j,t}$. For a scalar target this is the normalized projected weight $\Pi_{j,t} W_{j,t}$, up to sign. The singular-value tolerance determines numerical rank; zero directions are rejected. Control bases are instead obtained from SVD of the control weights without exclusion, combining retained feature directions by QR. Decoder SVD identifies target-aligned directions in the coordinates defined by preprocessing PCA. In exact arithmetic, $Q_{j,t}^\top Q_{j,t} = I$ and $Q_{j,t}^\top P_C = 0$ imply

$$X_{j,t} W_C = X W_C \quad \text{whenever} \quad W_C \in \mathrm{span}(P_C). \tag{13}$$

This identity also preserves the intercept of a fixed decoder. It does not guarantee unchanged accuracy of another fitted C probe or preserve unmeasured control functions.

Numerical safeguards and returned state. The reorthogonalization in Eq. (5) repeats exclusion after discovery to maintain the stored orthogonality constraints in finite precision. Numerically zero directions are rejected before normalization. The cumulative projector reduces rank while preserving the feature dimension of X. After all target groups, the returned state is $Q_{\mathrm{final}} = Q_{J,T_J}$ and $X_{\mathrm{erased}} = X(I - Q_{\mathrm{final}} Q_{\mathrm{final}}^\top)$, together with the recorded checkpoints. Trajectory indices count cumulative removed rank.

Discovery features versus evaluation features. Both main studies fit discovery decoders on $D_{j,t} = X\Pi_{j,t}$, temporarily excluding protected directions. Evaluation uses $X_{j,t}$, retaining those directions. After each local target-group sweep, the cumulative basis is carried to the next j. Physion++ uses the same discovery routine as SuctionNet, with its dataset-specific inputs and probe grid.

Policy module	SuctionNet fixed sweep	Physion++ fixed sweep
Discovery input $D_{j,t}$	$X\Pi_{j,t}$	$X\Pi_{j,t}$
Eligible targets $\mathcal{J}_{j,t}$	Current group in a fixed eight-group order	Current scalar in fixed order: 10 A targets; then 4 C targets for joint removal
Target selection	Current group; no global maximum search	Current scalar; no global maximum search
Nonpositive monitor on $D_{j,t}$	Advance to next group	Advance to next scalar
Blocked direction	Advance to next group	Advance to next scalar
Budget	Six local rounds per group; one direction per round	Six per scalar: at most 60 protected / 84 joint updates
Report rank k_*	End of the recorded sweep	End of the recorded sweep ($k_* = K$)
Final audit	All eight groups on X_{k_*}, including earlier groups	Raw A, C and active targets on X_{k_*}

Table 3: Concrete policies used in Alg. 1. Both use zero as the active stopping threshold. SuctionNet additionally records a 0.005 all-family cleanliness tolerance; this is an endpoint diagnostic, not a different global stopping rule. The mirror ablation has its own maximum-group, 64-round policy.

Selection and evaluation. The algorithm's FITANDMONITOR returns decoder weights and a non-test monitor. SELECTTARGET and SELECTENDPOINT may read only those monitors, policy counters and numerical feasibility. The separate audit module fits fresh A/C probes on erased features and records B performance at declared checkpoints. Neither test labels/scores nor observed B damage choose a direction or rank.

A.4 VARIABLE-PREDICTION AUDITS AND TASK-EFFECT ESTIMATES

Every endpoint reports target variable prediction and control variable prediction to assess suppression and collateral effects. For each declared target group, R^2 compares squared prediction error with that group's constant-mean reference on the evaluated split. Zero-variance targets have undefined population-style R^2: Physion++ component reports mark them undefined, while the SuctionNet multivariate evaluator uses a numerical denominator floor. Such degenerate outputs cannot establish useful variable prediction. Group-aggregate R^2, a macro mean across groups, and the maximum across groups/components are different statistics; the plot specifies which is used. A maximum emphasizes remaining readable targets. Nonpositive R^2 indicates failure of the evaluated probe to beat its reference, not absence of all information.

The fixed-head metric uses the same fitted g_B on X and X_{k_*}. The retrained metric fits a new head on X_{k_*} with the declared regularization protocol. The fixed-head score measures dependence of the original predictor. Refitting measures the task performance recoverable under the specified training procedure.

For a higher-is-better held-out metric, let $\mathcal{M}_B^{\mathrm{ret}}(V)$ denote the performance of a task head fitted on representation V. The refitted-head loss under intervention $\mathcal{E}$ is

$$\Delta_B^{\mathrm{ret}}(\mathcal{E}) = \mathcal{M}_B^{\mathrm{ret}}(X) - \mathcal{M}_B^{\mathrm{ret}}(\mathcal{E}(X)). \tag{14}$$

For K_R path-specific matched-random interventions $\{\mathcal{R}_r\}_{r=1}^{K_R}$, excess loss is

$$\begin{aligned}\Delta_{B,\mathrm{excess}}^{\mathrm{ret}} &= \Delta_B^{\mathrm{ret}}(\mathcal{E}) - \frac{1}{K_R}\sum_{r=1}^{K_R}\Delta_B^{\mathrm{ret}}(\mathcal{R}_r) \\ &= \frac{1}{K_R}\sum_{r=1}^{K_R}\mathcal{M}_B^{\mathrm{ret}}(\mathcal{R}_r(X)) - \mathcal{M}_B^{\mathrm{ret}}(\mathcal{E}(X)).\end{aligned} \tag{15}$$

Algorithm 1 Iterative variable-targeted erasure with an explicit discovery policy

Require: Fixed features X; prepared target family $\{\widetilde{Y}_j\}$; protected basis P_C (possibly empty); fitting/monitor splits.
Require: Audit labels A/C/B and evaluation splits; policy π (Tab. 3); rank budget $K_{\max}$; threshold ϵ.
Ensure: Selected basis Q_{J,T_J}, terminal basis Q_K, monitor history and evaluation records.
$Q_{1,0} \leftarrow [\,]$; $k \leftarrow 0$; initialize outer target cursor j, local round t, and history.
Main-study policy: fixed outer sweep over j, six inner updates per target, $k_* = K$.
Fit original task head g_B; record rank-zero variable-prediction and task-performance baselines.
while the policy has unfinished targets **do**
    $X_{j,t} \leftarrow X(I - Q_{j,t}Q_{j,t}^\top)$; $\Pi_{j,t} \leftarrow I - P_C P_C^\top - Q_{j,t}Q_{j,t}^\top$.
    $(D_{j,t}, \mathcal{J}_{j,t}) \leftarrow \textsc{DiscoveryInputs}(\pi, X, X_{j,t}, \Pi_{j,t})$.
    $j_t \leftarrow \textsc{SelectTarget}(\pi, \mathcal{J}_{j,t})$; $(W_{j_t,t}, m_{j_t,t}) \leftarrow \textsc{FitAndMonitor}(D_{j,t}, \widetilde{Y}_{j_t})$.
    Append rank and monitor to history; record refitted A/C probe performance and fixed/retrained B performance on $X_{j,t}$.
    if $k \geq \min(K_{\max}, d - \operatorname{rank}(P_C))$ **then**
        break
    end if
    if $m_{j_t,t} \leq \epsilon$ **then**
        if π is fixed sweep **then**
            Set $Q_{j+1,0} \leftarrow Q_{j,t}$, advance j, reset $t \leftarrow 0$; **continue**.
        else
            break
        end if
    end if
    $u_{j,t} \leftarrow \textsc{AdmissibleDirection}(W_{j_t,t}, P_C, Q_{j,t})$.
    if u is numerically zero **then**
        if π is fixed sweep **then**
            Set $Q_{j+1,0} \leftarrow Q_{j,t}$, advance j, reset $t \leftarrow 0$; **continue**.
        else
            break ▷ No search over alternative blocked targets
        end if
    end if
    $Q_{j,t+1} \leftarrow [Q_{j,t}, u_{j,t}]$ with unit, reorthogonalized direction; $t \leftarrow t + 1$, $k \leftarrow k + 1$.
    At the local-round budget, set $Q_{j+1,0} \leftarrow Q_{j,t}$, advance j, and reset $t \leftarrow 0$.
end while
$K \leftarrow k$; $k_* \leftarrow \textsc{SelectEndpoint}(\pi, \text{monitor history}, K)$.
Audit every declared target/control on X_{k_*}; evaluate B and matched random erasures.
Save Q_{k_*}, Q_K, P_C, preprocessing, ranks, stop reason and fitting scope.

Measurement	Question	Companion evidence
A probe	Can a probe predict A?	Task responses to target-directed erasure.
Frozen-head erasure	Does the fixed head depend on removed directions?	Target and control prediction audits.
Retrained-head erasure	Can remaining features support B?	Remaining target predictability and matched random erasure.
Protected erasure	How much A predictability can be suppressed under protection?	Fresh control probes and the unprotected path.
Decoded/GT restoration	Can a target channel restore task performance?	Channel-matched shuffles and control-channel addback where available.

Table 4: Complementary measurements for representation-level attribution.

The second equality cancels the common original-feature baseline. Endpoint estimates use $\mathcal{E} = \mathcal{E}_{k_*}$. Positive excess loss means that targeted erasure causes a larger task-performance decrease than the average matched-random erasure. Here K_R counts random controls and is unrelated to terminal rank K.

The matched-random erasures used for excess loss match removed rank and the implemented training-covariance criterion; under protection they lie in the same admissible complement. SuctionNet uses eight projections and reports their standard deviation in trajectory plots. Physion++ uses 50 projections and reports paired test-group bootstrap intervals for selected contrasts. The bidirectional main task plots additionally show a shared unprotected, rank-only full-space random reference. Excess-loss estimates use the path-specific controls, as detailed in Sec. F.

A.5 Restoration as a separate supervised channel

For a declared prepared target block V (for example, concatenated standardized A descriptors), let $\widehat{V} = h_V(X)$ be decoded from the original features. An out-of-fold training construction gives

$$X_{\text{rescue}} = [X_{k_*}, \widehat{V}], \qquad X_{\text{GT}} = [X_{k_*}, V]. \tag{16}$$

A new B head is fitted for each condition. SuctionNet's main block is $V = A$, with GT normal-vector and shuffled GT descriptor controls. Motion-protected Physion++ uses $V = A$; joint Physion++ uses raw A, raw C or their concatenation. Its shuffled controls permute the same decoded/GT channel being tested. Cross-fitting units and fitting scopes differ by dataset and are specified in the following appendices.

Restoration measures the predictive benefit of an explicit target channel. GT addback supplies privileged supervision, while decoded addback can both reintroduce task-relevant information and change effective regularization. The identity $X_{k_*} + XQ_{k_*}Q_{k_*}^\top = X$ provides a separate implementation check. We assess restoration jointly with target suppression, control preservation, and matched erasure effects.

A.6 What the protocol establishes

The protocol estimates task-performance changes caused by interventions on a computational representation. Attribution to the target family is supported by three observations: reduced A predictability, preserved measured C readouts, and task loss beyond matched random erasure. Incomplete suppression limits conclusions from small task effects, while collateral loss limits target-specific attribution. Conclusions apply to the evaluated target families, probe classes, and feature distributions; effects of manipulating physical variables in the world are outside this intervention's scope.

B SUCTIONNET: DATA, READOUTS, AND INTERVENTIONS

This section documents the main six-encoder SuctionNet experiment, using frozen global–local features, raw second-order targets, and first-order-family protection. The corresponding Physion++ protocol appears in Sec. C.

B.1 EXPERIMENTAL PURPOSE

We test whether second-order surface information supports seal prediction while selected first-order readouts are preserved. Baseline probes measure decodability. Protected erasure, matched random projections, and decoded restoration then assess suppression, task loss, and recovery. Mirror and single-target ablations examine the direction of intervention and the choice of target representation.

B.2 DATA SOURCES

We use released SuctionNet RGB-D scenes, camera calibration and object poses, nontextured object meshes, suction candidates, and seal-score annotations (Cao et al., 2021). RGB is the encoder input. Depth, masks, poses, and meshes are used only to construct or qualify supervision. The query pixel specifies where to sample the feature map. The response B is SuctionNet's released seal score. It measures analytical seal quality.

The qualified subset contains 16,072 queries from 313 frames and 49 object identities. Splits are scene-disjoint. Object identities can recur across splits, and queries within a frame share the same image.

Split	Scenes	Frames	Query points
Training	10	140	8,345
Validation	3	48	2,935
Test	13	125	4,792

Table 5: SuctionNet analysis subset. Validation scenes are held out within the downloaded training scenes; test scenes are drawn from the downloaded official novel-object test partition. Object identities can recur across these scene partitions.

B.3 SAMPLE CONSTRUCTION AND DATA PREVIEW

From a candidate to a local target. We transform each object's mesh and candidate into the camera frame, project the candidate into RGB, and check visibility against the depth image and instance assignment. Candidates flagged for object collision are excluded. At a retained query, mesh vertices from the same object within a 16 mm neighborhood define a local surface patch. The supplied normal defines a tangent chart: camera x is projected into the tangent plane, with a deterministic fallback near degeneracy. A weighted least-squares fit estimates

$$h(u, v) = \beta_0 + \beta_1 u + \beta_2 v + \tfrac{1}{2} H_{11} u^2 + H_{12} uv + \tfrac{1}{2} H_{22} v^2. \tag{17}$$

The cup radius is $r_c = 10\,\text{mm}$. Weights are Gaussian in tangent-plane distance, with standard deviation 0.55 times the fit radius; rank-deficient fits are rejected. Targets use $r_c H$, so the matrix coefficients are scale-aware. Eigenvalues below refer to this fitted quadratic term; they are not exact principal curvatures under arbitrary chart error. The contact-region quadratic gain is $\text{clip}(1 - \text{MSE}_{\text{quadratic}}/\text{MSE}_{\text{plane}}, 0, 1)$ on cup support, with degenerate plane energy handled by the evaluator.

Qualification and sampling. Sample construction applies a 12 mm depth-visibility tolerance and checks local depth-derived geometry against the mesh target. Maximum depth/mesh normal discrepancy is $22.5°$; maximum relative second-order discrepancy is 0.75; depth and mesh fit RMSE divided by the fitting radius must not exceed 0.1. The scale-aware curvature-strength interval is $[0.246, 1.347]$. A curvature-sensitive screen also requires quadratic-fit improvement of at least 0.2

and plane-fit RMSE divided by cup radius in $[0.002, 0.35]$. These screens and the sampled candidate budget yield the counts in Tab. 5. The resulting subset emphasizes reliable, curvature-sensitive queries.

Target and control families. The eight A groups are the three packed entries of $r_c H$, trace, determinant, Frobenius norm, eigenvalue gap, maximum absolute eigenvalue, sum of absolute eigenvalues, and cup-rim deviation: ten scalar coordinates in total. Each raw coordinate is standardized using training data only. The same eight groups are used for discovery, endpoint auditing and the ten-coordinate restoration channel.

For C, write $w = -p/\|p\|$ for the viewing direction and (e_1, e_2) for a deterministic orthonormal basis perpendicular to w. With n the unit normal, define $a = (n^\top e_1, n^\top e_2)$ and $\gamma = \mathrm{clip}(n^\top w, 10^{-4}, 1)$. The six groups are

$$C = \left\{ n,\ -a/\gamma,\ \mathrm{pack}(I - nn^\top),\ a,\ \gamma,\ \arccos\gamma \right\}. \tag{18}$$

Their dimensions are $3, 2, 6, 2, 1, 1$, totaling 15. These six groups describe first-order geometry under the specified viewing context. Sec. B.10 explains the finite-family interpretation.

Seal Spearman correlation. Let s_i and $\hat{s}_i$ be the benchmark seal score and its prediction for test query i. We assign ranks $r_i = \mathrm{rank}(s_i)$ and $\hat{r}_i = \mathrm{rank}(\hat{s}_i)$ across all N_{test} test queries, using average ranks for tied scores. Seal Spearman is the Pearson correlation of these ranks:

$$\rho_{\text{seal}} = \frac{\sum_{i=1}^{N_{\text{test}}} (r_i - \bar{r})(\hat{r}_i - \bar{\hat{r}})}{\sqrt{\sum_{i=1}^{N_{\text{test}}} (r_i - \bar{r})^2} \sqrt{\sum_{i=1}^{N_{\text{test}}} (\hat{r}_i - \bar{\hat{r}})^2}}, \tag{19}$$

where $\bar{r}$ and $\bar{\hat{r}}$ are the mean ranks. Values of 1, 0, and -1 indicate identical ordering, zero rank correlation, and reversed ordering, respectively. The score measures seal-quality ranking across pooled test queries. Absolute calibration and robot grasp success require separate evaluation. The evaluator marks the correlation undefined if either score vector has standard deviation at most 10^{-12}.

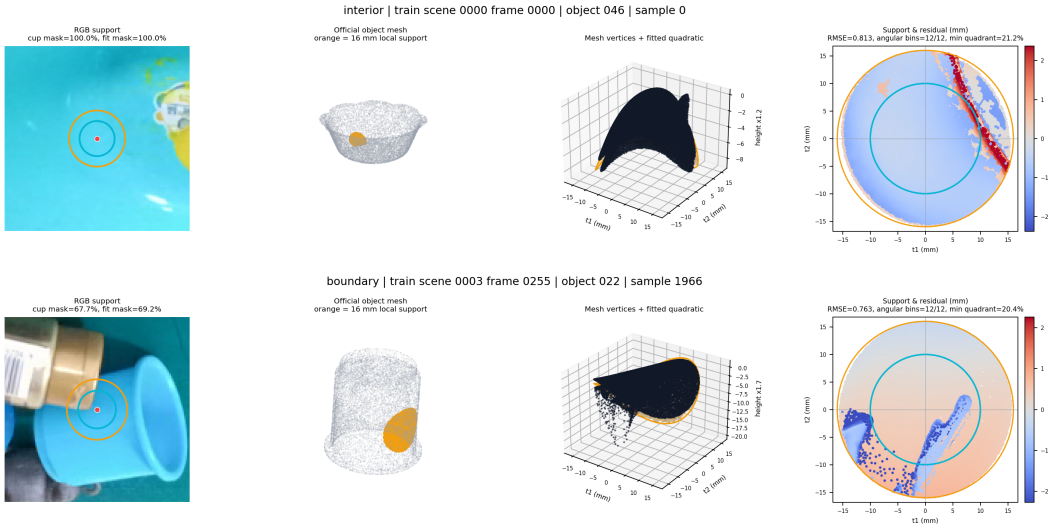

Figure 11: **Real SuctionNet samples and mesh-derived supervision.** Training examples: an interior query (sample 0) and a near-boundary query (sample 1966). Cyan marks the 10 mm cup scale; orange marks the 16 mm fit support. Surface height is exaggerated for visibility. Boundary/interior identify the illustrated query locations. Surface fits use only the query object's mesh.

B.4 Frozen feature extraction and probing

Each encoded input contains one RGB frame; duplicated inputs repeat that same view. We retain a dense feature map and bilinearly interpolate its four bracketing token centers at the query pixel. For

original image size $H \times W$ and feature-grid size $H_f \times W_f$, the continuous location is

$$(r, c) = \left((y + \tfrac{1}{2})H_f/H - \tfrac{1}{2},\ (x + \tfrac{1}{2})W_f/W - \tfrac{1}{2}\right), \tag{20}$$

clamped at the outermost token centers. This 1,024-dimensional global feature is concatenated with a 1,024-dimensional local feature from a query-centered, 128-pixel square RGB crop. Crops are edge-padded and resized to 224×224. Each crop is encoded separately, and its feature map is sampled bilinearly at the local query position. Global and local branches use separate attention. Local extraction batches two independent crops, duplicating a final odd crop; DINOv2 uses its cached local features. Coordinates, depth, masks, normals and mesh coefficients are not appended.

Encoder	Layer/component	Raw dimension	Intervention dimension
VGGT	11 / global	2048	512 (PCA)
VGGT-Omega	11 / global	2048	1024 (PCA)
DINOv2	23 / final	2048	512 (PCA)
DINOv3	23 / final	2048	1024 (PCA)
LingBot-Vision	23 / final	2048	256 (PCA)
V-JEPA 2	static repeated-frame spatial tokens	2048	512 (PCA)

Table 6: Extractor layer identifiers and fixed global–local capacities. All six apply training-fitted PCA after raw-channel standardization. The map remains fixed during erasure.

The global input wrapper resizes RGB to 518×518; model-specific preprocessing is retained by the extractor (including the configured 512 resolution for the VGGT-family global path). The local branch uses the crop resolution above. Ridge penalties are selected on validation from $\{10^{-5}, 10^{-4}, \ldots, 10^{3}\}$. Selected evaluation heads are refitted on training plus validation. Continuous-variable prediction is evaluated as test R^2, seal ranking as Spearman correlation, and normal orientation additionally as angular recall. The rank-zero heatmap uses separate target-group decoders. We retain negative R^2 and Spearman values; these scores measure prediction and ranking quality.

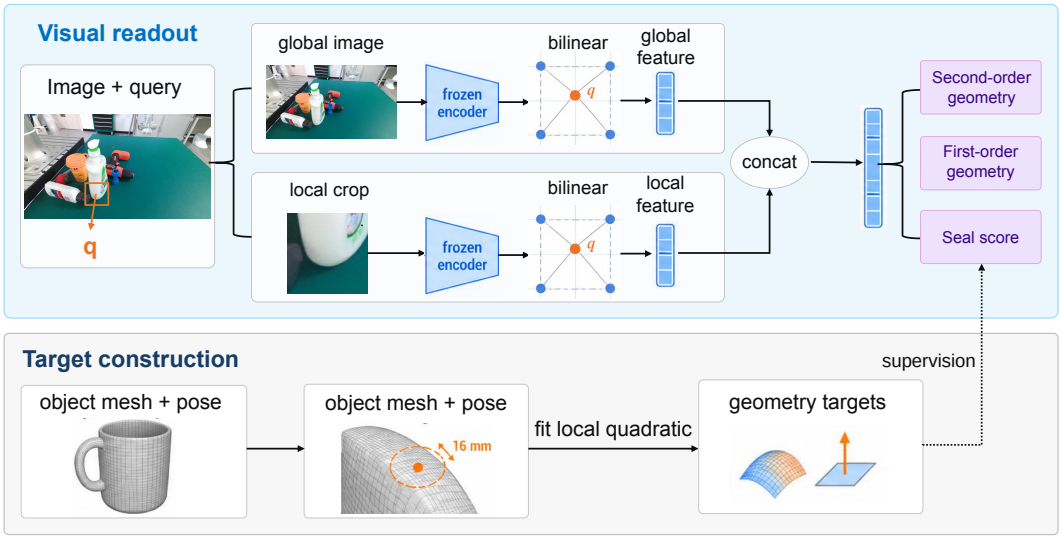

Figure 12: SuctionNet construction and frozen global–local readout. Supervision comes from the query object's mesh and the benchmark seal annotations; those labels are not concatenated to visual features.

B.5 Iterative erasure protocol

This instantiates the fixed-sweep policy in Alg. 1 and Tab. 3. Let X denote the fixed preprocessed features and P_C an orthonormal basis for the union of independently fitted first-order decoder spans. At rank k, evaluation features are $X_k = X(I - Q_k Q_k^\top)$, while discovery uses $D_k = X(I - P_C P_C^\top -$

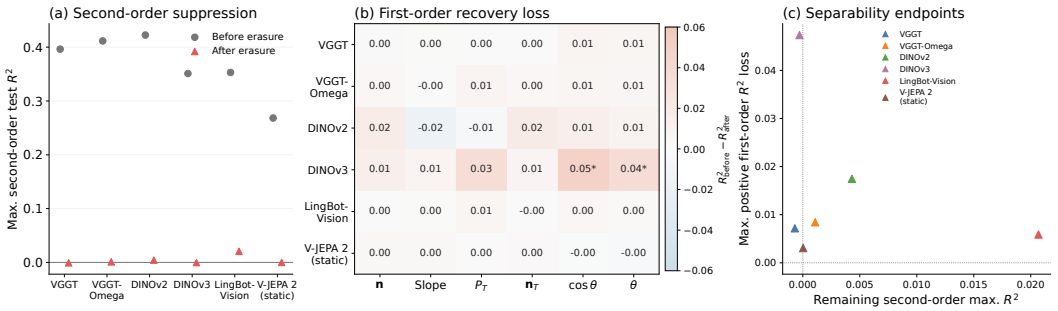

Figure 13: **Endpoint audits under first-order family protection.** (a) Maximum second-order test R^2 before and after erasure. (b) Fresh first-order probe loss, $R^2_{\text{before}} - R^2_{\text{after}}$; asterisks mark nonpositive baseline scores. (c) Remaining second-order maximum R^2 versus maximum first-order loss, clipped at zero. The group-wise audit measures changes in each control readout. Values are point estimates.

$Q_k Q_k^\top$). The declared probe-weight fitting operator for the active raw target supplies a feature-side direction; in the main run this operator is a refitted ridge fit, while Sec. B.9 replaces it with a full-batch AdamW fit. Numerical exclusion and reorthogonalization remove the direction's components in the protected and already removed spans before normalization. Thus $Q_k^\top P_C = 0$. This preserves the selected protected decoder outputs, as derived in Sec. A.

The cumulative erasure procedure visits matrix entries, trace, determinant, Frobenius norm, anisotropy, maximum absolute eigenvalue, absolute sum, and cup deviation in that order. It removes one direction per local round, with at most six rounds per group. The active group's nonpositive discovery-monitor R^2, or a zero admissible update, advances the schedule. Each group is visited once in this fixed order. Final auditing evaluates all eight groups. The six endpoints remove 9, 11, 12, 10, 13 and 10 directions, respectively, in Tab. 6's order. The endpoint audit checks suppression across the full family.

At each recorded rank we refit variable probes and task heads, evaluate the original fixed seal head, and compare with eight same-rank random erasures under the implemented training-covariance-spectrum matching criterion and the same protected span. Error bars show standard deviations across random projections. Fresh six-group C probes audit collateral change separately from fixed-decoder preservation. Some discovery monitors are computed after refitting on training plus validation. Test seal performance never selects the stopping point.

The fixed sweeps remove 9–13 directions. Maximum remaining raw-family test R^2 ranges from -0.001 to 0.021; maximum first-order-group loss ranges from 0.003 to 0.047.

B.6 Restoration protocol

Restoration uses the endpoint projection from erasure. From original features, a joint decoder estimates the ten raw A coordinates, using the same training-standardization convention as erasure. Five-fold out-of-fold predictions are grouped by $(\text{scene}, \text{frame})$ within training; validation and test predictions use a decoder fitted on training alone. Each appended channel is standardized with training statistics. The configurations are original visual features, erased features, erased plus decoded A descriptors, erased plus GT A descriptors, erased plus GT normal, and erased plus within-frame shuffled GT A descriptors (16 shuffles). GT normal addback supplies three coordinates from the normal-vector group. Each seal head selects its ridge penalty on validation and refits on training plus validation.

Decoded restoration tests the predictive value of a low-dimensional supervised channel; GT restoration supplies the corresponding measured descriptors. Within-frame shuffling preserves frame-level GT marginals and provides a control for GT addback. Internal folds are frame-grouped, and upstream preprocessing and discovery remain fixed.

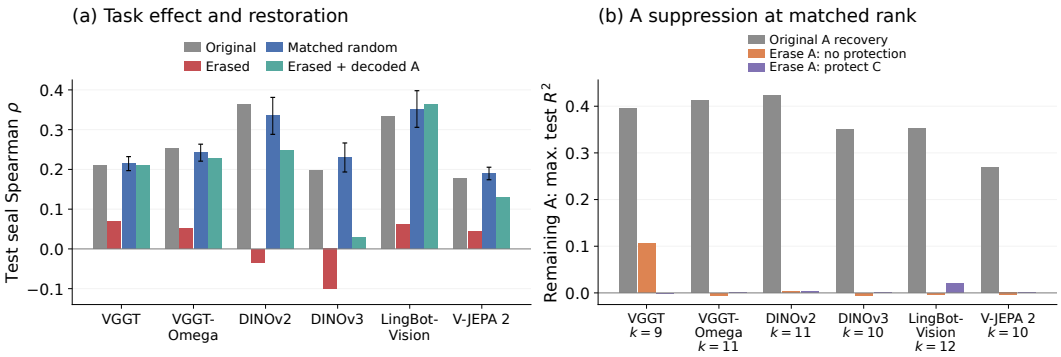

Figure 14: **Seal contribution, restoration, and rank-matched separability.** (a) Retrained seal Spearman for original features, protected A erasure, matched-random erasure, and erased features plus decoded A. Decoded training channels are frame-group cross-fitted; error bars show standard deviations across eight random projections. (b) Remaining maximum second-order test R^2 with and without C protection at each model's largest common recorded A-erasure rank k (labels). Gray bars show probe performance on original features; lower bars mean stronger suppression under the evaluated probes. Rank selection does not use test scores. Full restoration and control audits appear in Secs. B.6 and B.7.

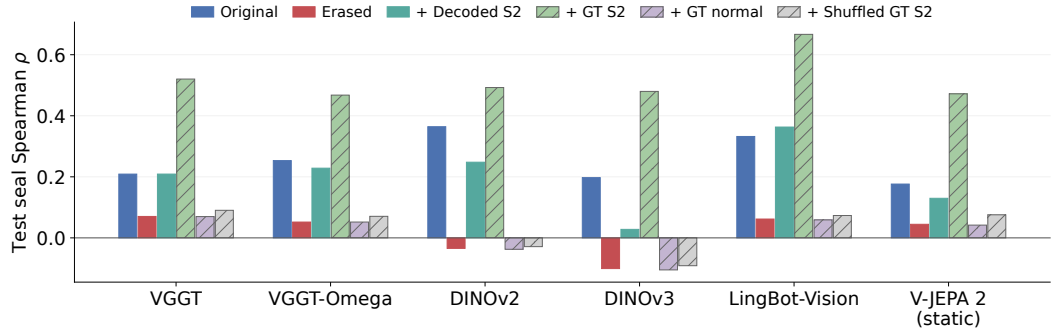

Figure 15: **Restoring second-order information after family-protected erasure.** Test seal Spearman for original and erased features, and erased features augmented with decoded second-order descriptors or ground-truth controls. All heads are retrained. Decoded training targets use frame-grouped out-of-fold predictions; validation and test estimates use a train-only decoder. Hatched bars use GT channels: raw S2, normal vector, or raw S2 shuffled within each frame and split. GT addback supplies privileged supervision. Values are point estimates.

Decoded-channel restoration. Appending decoded second-order descriptors improves seal prediction for all six encoders (Fig. 14a), with the degree of recovery varying across models. Training channels are constructed out of fold, and each seal head is retrained. This recovery supports the predictive value of an explicit second-order channel after erasure.

Full restoration comparison. Fig. 15 compares decoded channels with GT addback and shuffled controls. GT second-order descriptors provide a larger gain, whereas the three-dimensional GT normal channel adds little. Channel dimension and information content differ across these conditions. The compact summary in Fig. 14a combines original, erased, matched-random, and decoded-A results.

B.7 Control protection and collateral loss

To isolate the role of feature-space protection, we repeat the complete raw-family intervention with P_C empty. Both conditions use the same global–local features, raw A labels, splits, fitted preprocess-

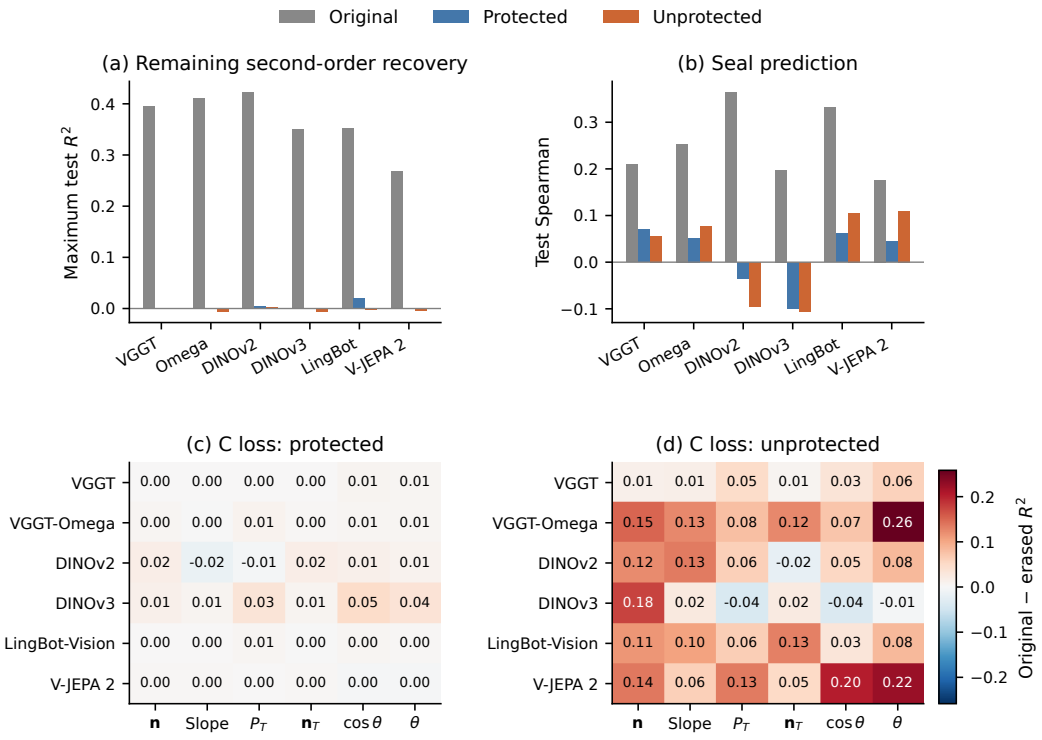

Figure 16: **Raw-family erasure with and without control protection.** Both paths target the same eight raw second-order groups. (a) Maximum second-order test R^2. (b) Retrained seal Spearman. (c–d) First-order-group test R^2 loss (original minus erased), with a shared color scale; negative values indicate improvement. Each path uses its own fixed-sweep endpoint, which can differ in removed rank. Values are point estimates.

ing, probe grids, target order and six-update local budget. Without protection, discovery excludes only previously removed directions. Fresh probes audit all six first-order groups in both conditions.

Without protection, the largest first-order-group test R^2 loss ranges from 0.059 to 0.258, compared with 0.003–0.047 under protection. It is larger without protection in 6 of six models. Unprotected endpoints remove 10–12 directions. Protection reduces the maximum measured first-order loss in all six encoders. Same-rank random controls accompany each path's endpoint seal effects.

Target suppression under protection. Fig. 14b compares the protected and unprotected A sweeps at the largest removed rank recorded by both paths for each model, selected without test scores. Five encoders have near-zero maximum second-order test R^2 in both conditions (at most 0.020); VGGT instead retains 0.105 without protection versus approximately zero with protection at rank 9. Strong A suppression remains possible under measured C protection. These finite-budget sweeps show no consistent suppression penalty across encoders.

Rank-matched suppression diagnostic. Fig. 14b uses the largest common recorded rank of the protected and unprotected A sweeps for each model: 9, 11, 11, 10, 12 and 10 in Tab. 6's order. These matched ranks can precede the protected endpoint used in panel (a). Both conditions audit the same eight raw A groups with fresh probes. Their common target order and local budget do not force identical target-switch times, because stopping monitors depend on the current representation. At these ranks, unprotected normal angular recall is lower than protected recall in five models, but slightly higher in VGGT; the full-family endpoint audit covers the remaining control groups.

B.8 PER-MODEL ERASURE TRAJECTORIES

Figs. 17 to 22 show all six raw-family cumulative trajectories. Target audits cover all eight second-order groups. The normal-recall panel measures one aspect of the control family; Fig. 4 adds maximum family-wise test R^2 from refitted probes at every recorded rank, and Fig. 13 reports group-wise endpoint losses. Task panels distinguish retrained and frozen heads and show rank-matched random controls.

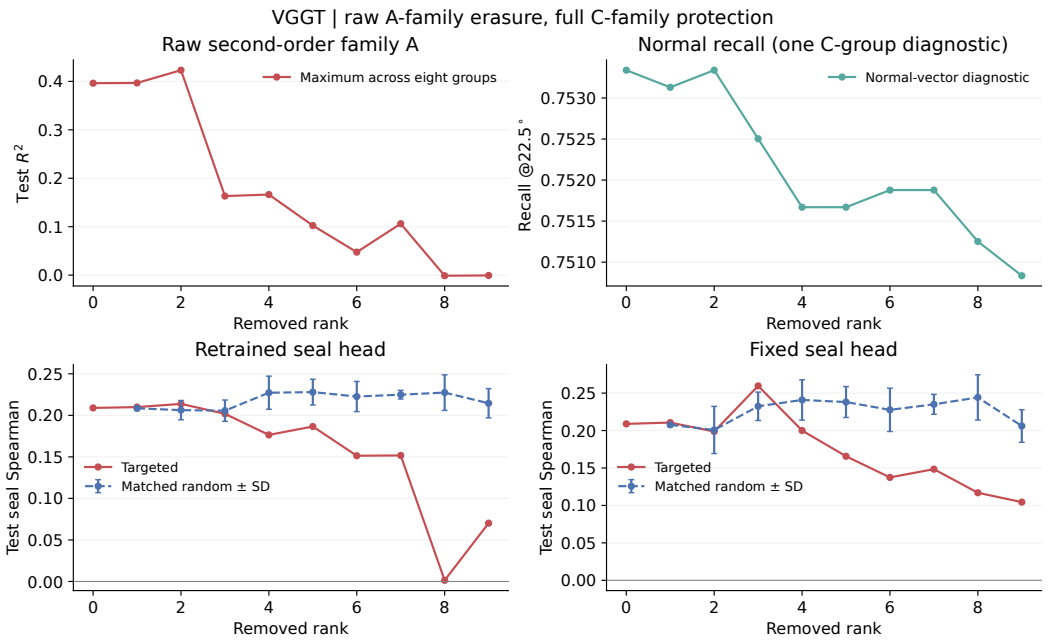

Figure 17: **VGGT: cumulative second-order erasure.** Trajectory under first-order-family protection. The target audit covers all eight second-order groups; the normal-recall panel evaluates normal orientation. Seal panels distinguish retrained and frozen heads, with excess loss computed for retrained heads. Error bars show standard deviations across eight random projections.

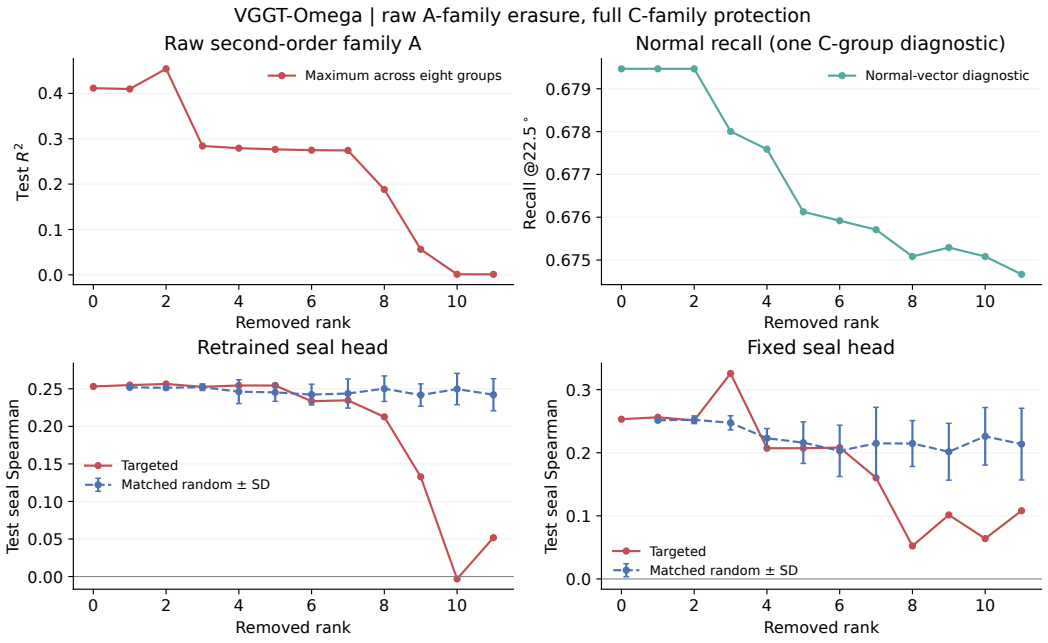

Figure 18: **VGGT-Omega: cumulative second-order erasure.** Trajectory under first-order-family protection. The target audit covers all eight second-order groups; the normal-recall panel evaluates normal orientation. Seal panels distinguish retrained and frozen heads, with excess loss computed for retrained heads. Error bars show standard deviations across eight random projections.

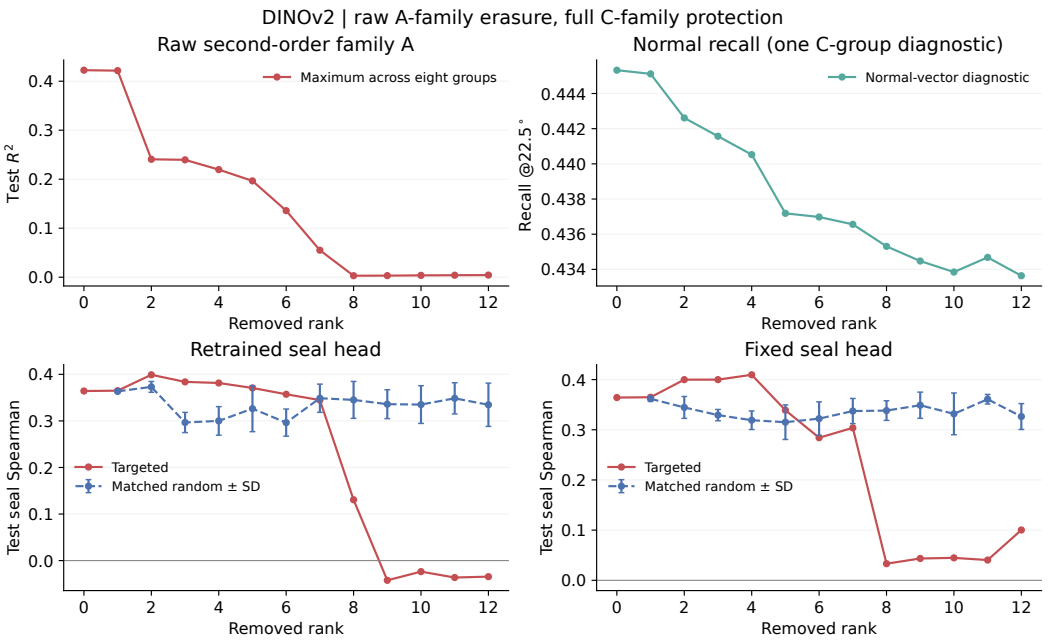

Figure 19: **DINOv2: cumulative second-order erasure.** Trajectory under first-order-family protection. The target audit covers all eight second-order groups; the normal-recall panel evaluates normal orientation. Seal panels distinguish retrained and frozen heads, with excess loss computed for retrained heads. Error bars show standard deviations across eight random projections.

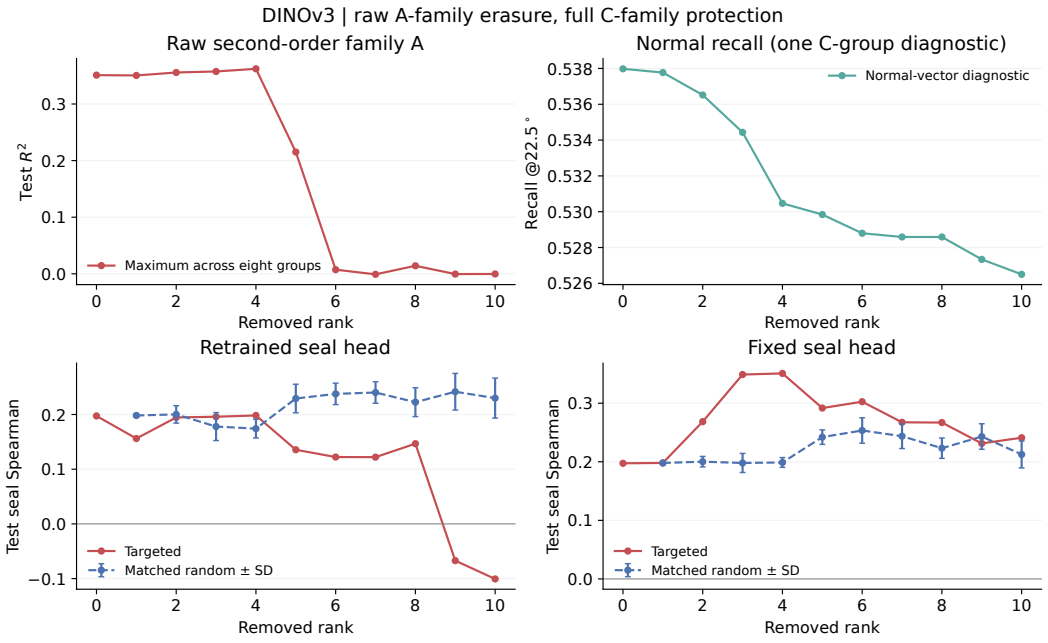

Figure 20: **DINOv3: cumulative second-order erasure.** Trajectory under first-order-family protection. The target audit covers all eight second-order groups; the normal-recall panel evaluates normal orientation. Seal panels distinguish retrained and frozen heads, with excess loss computed for retrained heads. Error bars show standard deviations across eight random projections.

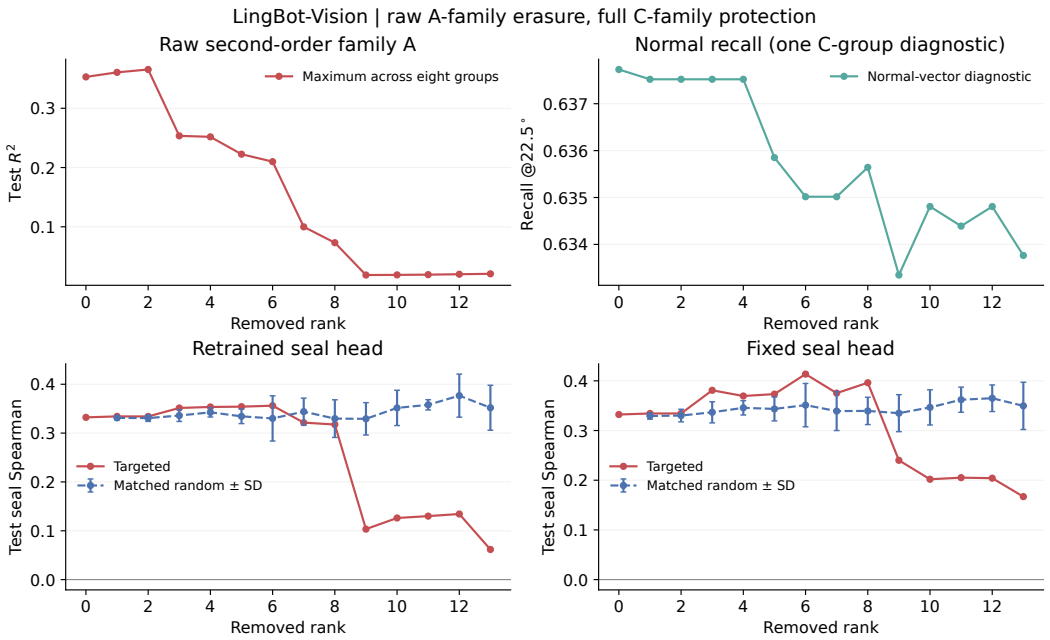

Figure 21: **LingBot-Vision: cumulative second-order erasure.** Trajectory under first-order-family protection. The target audit covers all eight second-order groups; the normal-recall panel evaluates normal orientation. Seal panels distinguish retrained and frozen heads, with excess loss computed for retrained heads. Error bars show standard deviations across eight random projections.

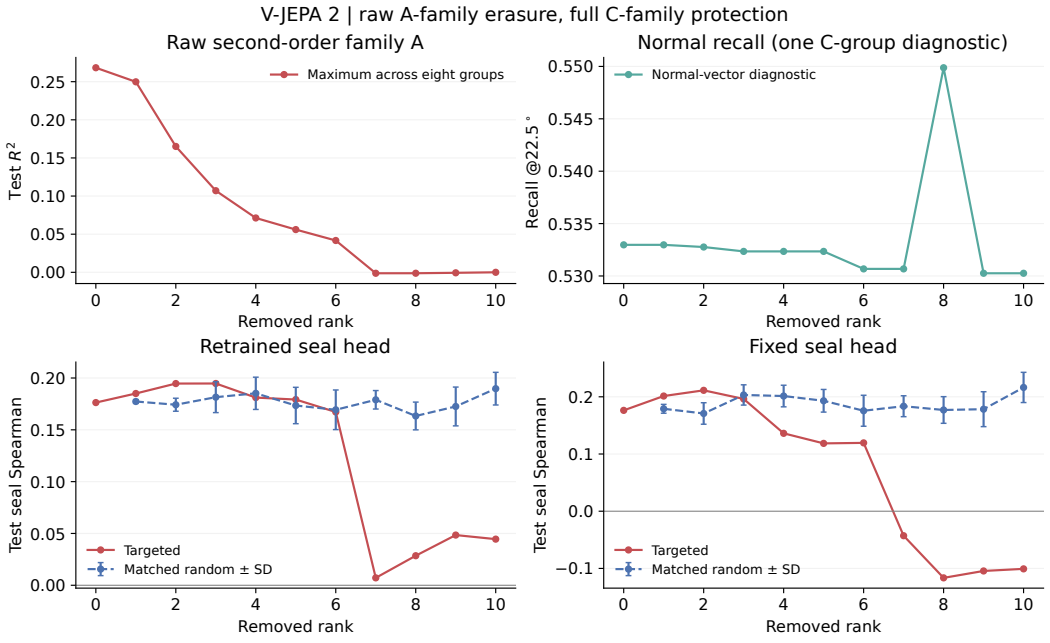

Figure 22: **V-JEPA 2: cumulative second-order erasure.** Static repeated-frame global–local readout under full first-order-family protection. The target audit covers all eight second-order groups; the normal-recall panel is a normal-vector diagnostic.

B.9 GRADIENT-BASED DIRECTION DISCOVERY

We repeat the cumulative intervention with the same global–local features, raw second-order family and full first-order-family protection, replacing the ridge discovery fit with a linear decoder trained by full-batch AdamW. Learning rate $\{0.003, 0.01\}$, weight decay $\{0, 10^{-4}\}$, and the stopping step are selected on validation, with at most 600 optimization steps per fit, evaluation every 10 steps, and patience 12. The decoder is refitted on training plus validation for the selected number of steps. Target order, rank-one updates, evaluation ridge probes, seal heads, and eight matched-random controls follow the main protocol. This compares discovery solver configurations, including their regularization and optimization budgets.

Fig. 23 shows all six encoders. The top row audits maximum test R^2 across the eight raw second-order groups; the bottom row reports retrained seal Spearman. Both solver paths use identical raw targets and protection. Their endpoints need not match in rank or remaining target predictability, so task losses must be read together with the audits.

AdamW endpoints remove 34–48 directions and leave maximum raw-family test R^2 of 0.030–0.297. Maximum discovery-monitor recovery remains 0.100–0.559.

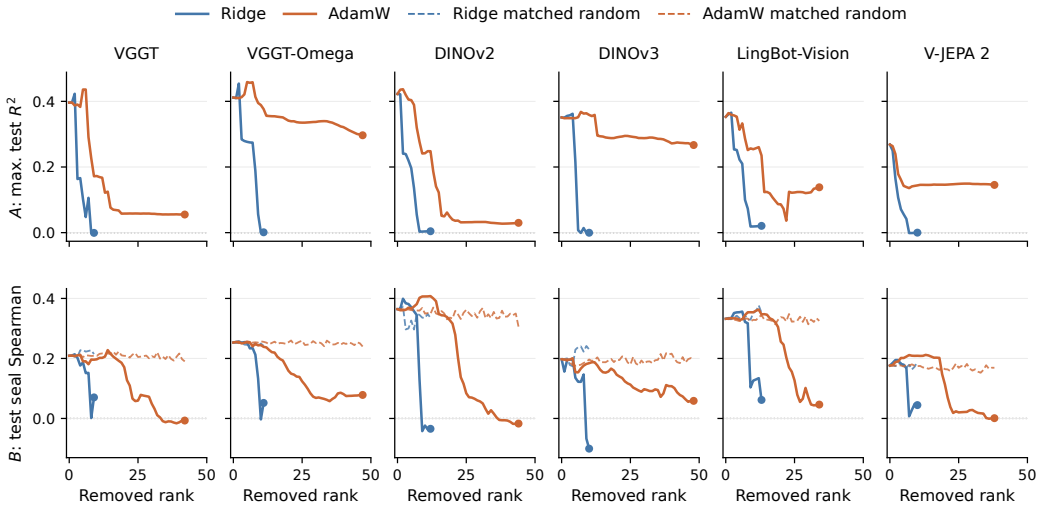

Figure 23: **Gradient-based direction discovery.** Columns show all six encoders. Top: maximum second-order test R^2 across eight groups. Bottom: test seal Spearman with retrained heads. Solid curves compare ridge and AdamW discovery; dashed curves show each path's mean over eight matched-random projections. All variable-prediction and task evaluations use ridge probes. Dots mark endpoints, which differ in rank and remaining target predictability. Lines join measured checkpoints.

Role	Examples in the evaluated families	Interpretation
Coordinate representation	Normal components; packed second-order matrix entries	Coordinates of the underlying object in a specified frame or chart.
Geometric descriptor	Tangent projector; trace, determinant, Frobenius norm, eigenvalue gap and absolute-eigenvalue summaries	Related measures of plane orientation, quadratic bending strength, or anisotropy.
Context-dependent measurement	View-relative slopes and normal components, incidence, tilt; cup-rim deviation	Geometry relative to a viewing direction or a fixed cup scale.

Table 7: Rationale for the finite geometric target families. Cup-rim deviation depends on cup scale; incidence and tilt depend on viewing direction.

B.10 Geometric target families: definition and scope

Definition and scope. A geometric target family is a fixed, finite collection of physically interpretable descriptors of a common underlying geometric object:

$$\mathcal{F}_\Theta = \{\psi_j(\Theta; \xi)\}_{j=1}^{J}, \tag{21}$$

where Θ is the geometric object and ξ denotes declared context such as a coordinate frame, viewing direction, or measurement scale. In SuctionNet, $\Theta_1 = \mathbf{n}$ describes the oriented local tangent plane, and $\Theta_2 = H$ describes the fitted quadratic surface term in the chosen chart. Family members are related measurements of the same geometric object.

Selection rationale. We distinguish coordinate representations, geometric descriptors, and context-dependent measurements (Tab. 7). Each target has an explicit formula and geometric interpretation, and the same listed families are evaluated across models. Discovery and auditing use this same descriptor set.

Why audit derived targets? A deterministic descriptor adds no information beyond its underlying geometric object and context. Probe accessibility can still differ. A linear readout may predict a nonlinear descriptor such as $\|H\|_F$ even when it poorly predicts H. Trace has a different role: it is a linear combination of matrix entries, so vanishing feature–entry cross-covariances imply vanishing feature–trace cross-covariance. Its probe checks consistency under finite samples, regularization, and imperfect removal. Held-out scores estimate predictive performance; the covariance statement concerns exact population quantities.

Raw families and feature protection. The family is defined on raw geometry before training-fitted target standardization. Both forward and reverse experiments retain these physical targets. Protection constrains feature directions; target-label relationships remain unchanged. Correlated or mathematically related descriptors may remain decodable through the protected span.

Family-level audit. Family-level auditing tests whether suppression extends across the listed representations of the same geometry. The single-target ablation motivates this broader audit by measuring what remains predictable after targeting one descriptor.

B.11 Geometry ablation details

The reverse-order evaluator selects the first-order group with the highest current discovery-monitor score and removes one feature direction per round, with a budget of 64 rounds per path. Eight covariance-matched random projections are evaluated at each rank. It stops when the maximum active validation monitor is nonpositive; the endpoint audit additionally records a 0.005 tolerance. Discovery monitors use refitted probes, as in the main protocol. No test score selects a stopping point.

Both reverse paths target the six raw first-order groups. The protected path preserves the union of the eight second-order-group decoder spans; fresh probes audit all eight groups. The forward experiment exchanges their roles. Both paths use the main global–local features and preprocessing.

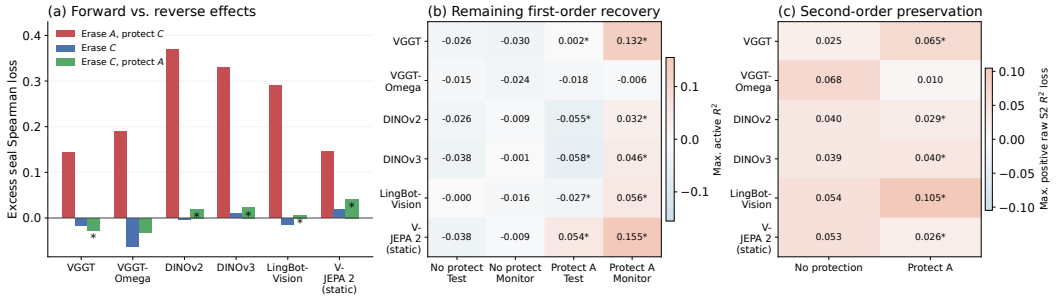

Figure 24: **Reverse-order ablation.** (a) Retrained seal loss beyond each path's matched-random control. (b) Maximum active first-order test and validation R^2: raw C, without or with protection of A. (c) Largest positive raw second-order group R^2 loss. Asterisks mark endpoints that miss the recorded cleanliness tolerance. Both reverse paths target the same raw first-order family. Results are point estimates at each path's endpoint; removed ranks differ.



Figure 25: **Single-target versus family erasure.** Each independent path starts from the original features and targets one second-order group while protecting the first-order family. Panels show remaining seal Spearman, maximum remaining second-order test R^2, and seal loss beyond matched random erasure. H is the three-coordinate matrix group. These six-encoder paths can retain the selected target and remove fewer directions than the family path.

All models apply PCA, including the 1,024-dimensional outputs for VGGT-Omega and DINOv3. Different target-selection schedules and endpoint ranks limit direct effect-size comparisons.

Excess seal losses are -0.063–0.018 for unprotected first-order removal and -0.033–0.041 with second-order protection, compared with 0.144–0.369 for the forward path. Preserving the second-order decoder span can retain first-order-predictive directions. Five protected reverse paths miss the family-cleanliness tolerance; their task effects measure partial first-order removal.

Fig. 4 displays the protected reverse trajectory. Fig. 24 additionally compares the unprotected path and audits each endpoint. Unprotected reverse erasure has smaller excess seal loss than forward second-order erasure in all six models.

Each single-target path restarts at the original representation and repeatedly fits only one of the eight second-order target groups, with up to six local rounds. All paths retain full-first-order-family protection. Fig. 25 reports all six models and every target, including weak or incomplete removals. The matrix-entry group is multivariate; the other seven groups are scalar. Each path has random controls matched to its own removed rank.

B.12 One-shot probe-subspace erasure

We adapt GeoSubProbe's one-shot probe-weight SVD construction (Zhou et al., 2026) to the cross-task question studied here. Its original subspace diagnostic evaluates fixed heads; our adaptation adds first-order-family protection and retrained seal prediction. We retain the main six-model split, frozen global–local features, fixed preprocessing and ridge-selection protocol. The sole discovery

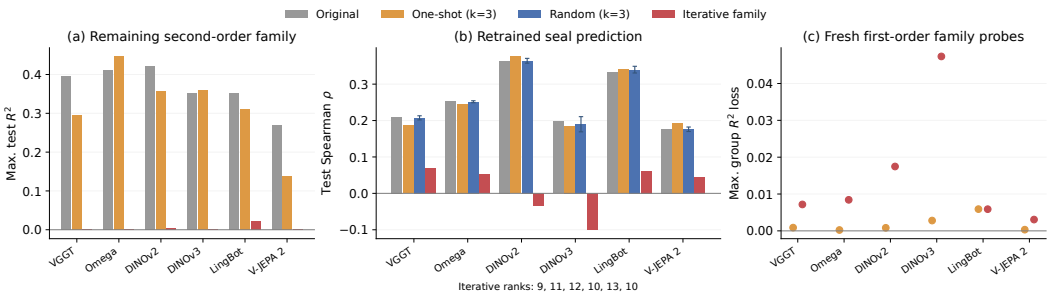

Figure 26: **One-shot matrix-probe erasure leaves geometry and seal scores predictable.** A GeoSubProbe-inspired single decoder-SVD removes three directions under the same first-order-family protection as the main experiment. (a) Full second-order-family prediction with refitted probes. (b) Retrained seal performance; the random control matches the one-shot rank, with SD over eight projections. (c) Maximum fresh first-order-group loss. The iterative-family reference removes 9–13 directions across all eight second-order groups.

target is the raw scale-aware matrix group $(r_c H_{11}, r_c H_{12}, r_c H_{22})$, with training-fitted coordinate standardization. Derived descriptors are audited but do not define the intervention.

Starting at $Q_0 = \varnothing$, we fit one matrix decoder on $D_0 = X(I - P_C P_C^\top)$ and compute

$$(I - P_C P_C^\top) W_H = U \Sigma V^\top, \qquad X_{\text{one}} = X(I - U_k U_k^\top). \tag{22}$$

We remove the entire numerically nonzero decoder span, at most three directions; all six fits have rank three. This single fit fixes both the basis and removed rank. Fresh A/C probes and a new seal head are fitted on the erased features. Eight random projections match rank and the training-covariance criterion within the same admissible complement.

One-shot erasure changes retrained seal Spearman by -0.016 to 0.023 (original minus erased), with excess loss over matched random ranging from -0.017 to 0.021. Maximum remaining second-order-family test R^2 is 0.137–0.447, and a refitted probe still predicts the targeted matrix ($R^2 = 0.091$–0.330). Maximum fresh first-order-group loss is below 0.006 for every model.

Geometry remains predictable after one-shot erasure. Removing one fitted decoder span leaves directions available to a new probe. The iterative-family endpoints produce larger seal losses after removing 9–13 directions across all eight groups. Target coverage and removed rank both differ, so the contrast combines these changes with iteration.

B.13 Geometry supervision and transfer to seal prediction

Experimental setup. We test whether second-order geometry supervision improves DINOv2's geometry prediction and transfers to seal ranking. Adaptation updates the local encoder branch. Fresh readouts evaluate both outcomes.

Data and splits. The experiment uses 13 scenes, 188 frames, and 11,280 qualified query points from the main training and validation partitions. The original 13 test scenes and 4,792 test points remain excluded. Three deterministic outer folds hold out entire scenes and balance frame counts. Within each outer training pool, an inner scene split selects ridge penalties. Encoder adaptation uses that full pool, including the inner probe-validation scenes. Outer held-out scores select neither checkpoints nor penalties.

Inputs and trainable parameters. Features concatenate a frozen 1,024-dimensional global branch and a 1,024-dimensional query-crop branch. Each crop spans 128 original pixels and is resized to 224×224. We sample its feature map bilinearly at the query. Adaptation updates DINOv2 blocks 22–23 (zero-indexed) and the final normalization; the global branch and cached local prefix remain fixed. Evaluation uses all 2,048 coordinates, without PCA.

Fold	Train scenes	Train frames	Train points	Held-out scenes	Held-out points
1	9	124	7,386	4	3,894
2	9	124	7,671	4	3,609
3	8	128	7,503	5	3,777

Table 8: Scene-grouped folds shared across supervision arms and seeds. Held-out frame counts are 64, 64, and 60.

Supervision	Raw A mean R^2	Seal ρ	Seal gain [95% interval]
None (frozen)	0.328	0.523	—
Gap ($\Delta\lambda$)	0.360	0.537	$+0.014$ [$+0.007$, $+0.018$]
Family	0.372	0.534	$+0.011$ [$+0.005$, $+0.015$]

Table 9: **Geometry supervision with baseline-matched seal regularization.** DINOv2 results averaged across three scene-grouped folds and three seeds. Adapted seal probes reuse each fold's frozen-baseline ridge penalty. Geometry scores measure raw second-order targets. Intervals use 1,000 paired held-out-scene bootstrap draws stratified within fold. Independent-regularization results appear in the text.

Supervision and optimization. The two supervision arms use either the raw eigenvalue gap $\Delta\lambda$ or all eight raw second-order groups. Coordinates are standardized on outer training data. For family supervision, each group with q_j coordinates is multiplied by $\sqrt{10/(8q_j)}$, giving equal total weight to each group in coordinate-averaged squared error. The geometry head uses a ridge penalty selected on the inner scene split. After refitting on all outer training scenes, its weights remain frozen during adaptation. The objective combines geometry-prediction MSE with local-feature anchoring MSE. The anchoring term has coefficient one and is normalized by the mean squared frozen local activation on outer training points. Query losses are point-weighted across frames.

Each run uses AdamW (Loshchilov & Hutter, 2017) with learning rate 2×10^{-6}, weight decay 0.01, gradient-norm clipping at one, crop batches of 16, and accumulation over eight frames. All runs end at epoch 10. Three seeds (20260916–20260918) vary training order while keeping the folds and pretrained initialization fixed, yielding 18 fold–arm–seed fits. Seal labels are used only for downstream readout fitting and evaluation.

Readout evaluation. Fresh geometry and seal probes use an inner-training feature scaler and ridge candidates $\{10^{-5}, 10^{-4}, \ldots, 10^{3}\}$. Penalties minimize inner-validation standardized MSE; weights are then refitted on all outer training points. We evaluate two seal-readout protocols. With *baseline-matched regularization*, adapted and frozen features use the frozen baseline's fold-specific penalty, $\alpha = 1, 10, 10$. With *independently selected regularization*, each feature condition selects its own penalty using the same inner-validation rule. Ridge is normalized by feature covariance. Equal α can therefore yield different absolute penalties and effective model capacities after adaptation. The raw-family geometry score averages the eight group R^2 scores, then folds and seeds.

Results. Geometry prediction improves in every fold and seed. Mean raw-family R^2 rises from 0.328 to 0.360 with gap supervision and 0.372 with family supervision. Seal transfer depends on the readout. With baseline-matched regularization, Spearman rises from 0.523 to 0.537 and 0.534, respectively (Tab. 9). Independently selected penalties yield 0.516 and 0.512, with gain intervals of $[-0.022, 0.006]$ and $[-0.026, 0.003]$. The geometry gains persist across these runs; the seal gains occur only under baseline-matched regularization.

C PHYSION++: DATA, READOUTS, AND INTERVENTIONS

This section describes the expanded Physion++ readout protocol, with ten spatial target coordinates and four motion control readouts.

C.1 EXPERIMENTAL PURPOSE

We test how a frozen video representation supports future contact through current spatial configuration and relative motion. The target A comprises relative position, distance, and target/zone scales (ten continuous coordinates; no target type); C comprises relative velocity coordinates and speed; B is future target–zone contact. We measure spatial predictability after motion-protected erasure and the accompanying change in contact prediction. Decoded spatial and motion channels test recovery. Joint erasure tests the combined representation without protection.

C.2 DATA SOURCES

We use the released Physion++ RGB videos and accompanying simulator metadata (Tung et al., 2023): object identities and scales, positions and velocities, camera matrices and object bounds, the prediction cutoff, and target–zone contact annotations. The encoder receives RGB frames and the rendered pair cue. Simulator metadata supplies the A/C/B labels and cue geometry; its numerical values are excluded from the feature vector.

The expanded supervised readout combines 706 compatible original readout trials with 4,997 additional dynamics-training trials; the original 1,081 test trials are unchanged. Of 5,008 expansion candidates, 11 are removed by state/RGB-key duplicate screening. The nine scenario families are bouncy platform, bouncy wall, cloth hanging, cloth hitting, friction collision, friction platform, mass collision, mass dominoes and mass water push. Coverage is uneven: the expansion adds only 34 mass-collision and one mass-water-push trial, versus hundreds in most other scenarios. The expanded protocol increases supervised readout data relative to the official small-readout setting.

C.3 TRIAL CONSTRUCTION AND DATA PREVIEW

Validity, temporal alignment and labels. The construction excludes trials marked invalid, missing sampled-frame metadata, incompatible scale metadata, nonfinite targets, and trials whose required visual cue cannot be formed. Let t_c be the supplied prediction cutoff and $t_0 = t_c - 1$. The last visible state's target and zone determine

$$p = p_i(t_0) - p_j(t_0), \qquad v = v_i(t_0) - v_j(t_0), \tag{23}$$

$$A = [p, \|p\|, s_i, s_j], \qquad C = [v_x, v_y, v_z, \|v\|]. \tag{24}$$

A has ten continuous coordinates. Target type is excluded from A discovery, probe audits, restoration channels, and reverse-protection decoders. C has four readouts of three physical velocity degrees of freedom.

B uses the released target–zone contact outcome. Trials already contacting at t_0 are excluded; when a first-contact timestamp is provided, a positive outcome must occur after t_0. All input indices are strictly smaller than t_c. A and C use only the last visible state. The encoder receives the observed history.

Observed input and pair cue. Eight frames sample the early observed phase and four sample the late pre-cutoff phase, with exact indices stored per trial. A separate cue at the first late frame marks camera-projected object bounds: target red, zone yellow. This bounds-based rendering follows the official role-cue convention and conveys both pair identity and visible spatial information. The motion control family comprises translational velocity and speed; force, friction, and mass are outside its coverage.

Splits and shared features. The expanded readout uses 5,703 trials in 130 configuration groups. A label-blind hash of group identity and seed 20260908 assigns 4,586 trials in 98 groups to discovery training and 1,117 trials in 32 groups to validation. All 1,081 test trials remain fixed, with 542 paired

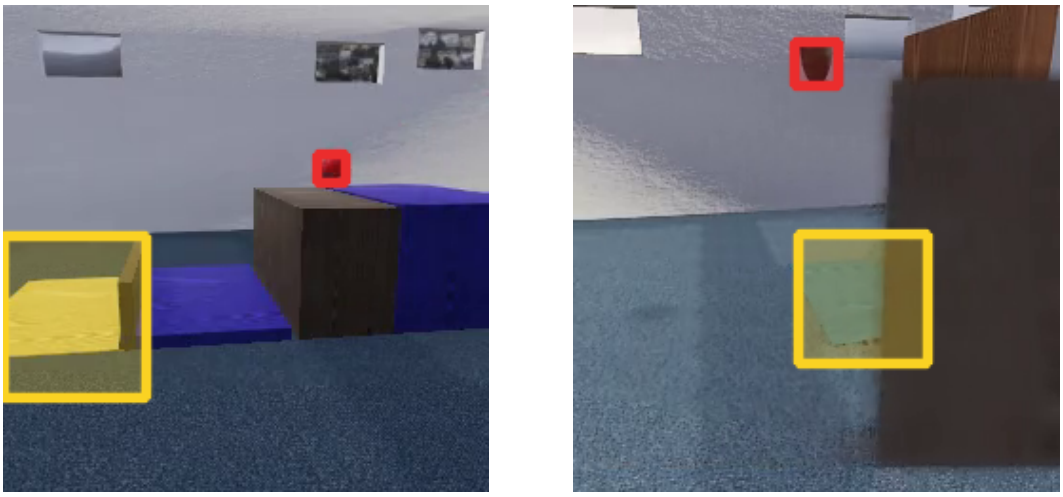

Figure 27: **Real observed Physion++ pair cues.** Two test-trial previews from the expanded analysis set. Red identifies the target and yellow the zone. The observed-frame overlays illustrate the query-pair specification.

groups used for bootstrap resampling. The training-fitted 512-dimensional PCA map is applied consistently to all examples. Trials sharing a configuration remain grouped despite the larger sample count.

Partition	Trials	Grouping units
Discovery training	4,586	98 configurations
Discovery validation	1,117	32 configurations
Readout total	5,703	130 configurations
Fixed test	1,081	542 paired groups

Table 10: Physion++ grouping. Discovery partitions are configuration-disjoint. Paired test groups define the resampling unit and are not the same unit as readout configurations. Duplicate screening is not proof of object- or configuration-disjointness between the released readout and test sets.

C.4 Frozen feature extraction and probing

For VGGT, VGGT-Omega, DINOv2, DINOv3 and LingBot-Vision, frames are encoded independently and patch tokens are spatially pooled. The readout concatenates eight early vectors, one cue vector and four late vectors, preserving order: $13 \times 1024 = 13{,}312$ dimensions. These are global frame features, unlike SuctionNet's query-point interpolation. Image-encoder layer/component choices follow Tab. 6. Square resizing and the respective model preprocessing remain fixed.

V-JEPA 2 processes early, cue, and late phases separately, with temporal attention within each phase. The cue image is repeated to form one tubelet. Spatially pooled tubelet vectors yield four early, one cue, and two late vectors: $7 \times 1024 = 7{,}168$ dimensions. We use only the frozen encoder. Its late-window feature block includes temporal context. Preprocessing also differs from the image models' square resizing, so comparisons across models include input and readout differences.

Training-fitted raw-channel standardization and the training-fitted 512-dimensional PCA map define a common space for A, C and B. The transform is fixed throughout erasure and restoration and is applied identically to readout and test rows. Ridge probes measure component-wise continuous R^2 for all ten A coordinates and four C readouts. A logistic head measures contact AUROC, supplemented by accuracy and balanced accuracy at probability 0.5. Negative R^2 is retained. The weak v_z baseline limits assessment of its preservation.

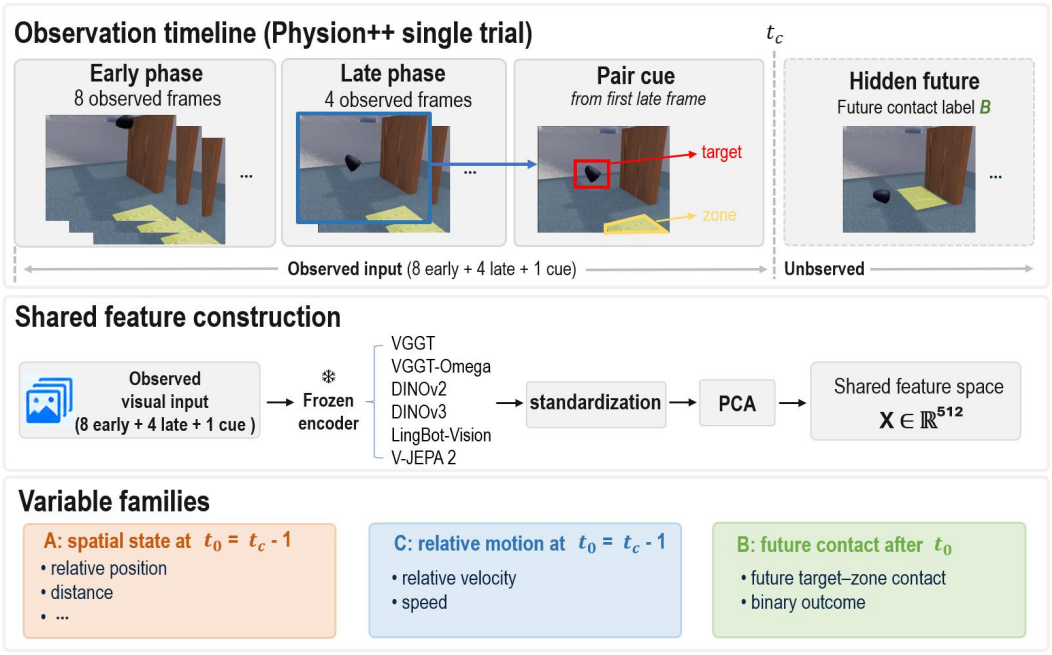

Figure 28: Physion++ observation and feature construction. The future lies outside the encoder input. Metadata constructs present-state supervision, the visible pair cue, and the future contact label.

C.5 Iterative erasure protocol

This instantiates the fixed-sweep policy in Alg. 1 and Tab. 3, reusing SuctionNet's direction-discovery routine. Discovery fits on $D_k = X(I - P_C P_C^\top - Q_k Q_k^\top)$; evaluation uses $X_k = X(I - Q_k Q_k^\top)$. Here k denotes erased rank, distinct from the last visible frame t_0.

The 10-dimensional A family contains relative position, distance, and two three-dimensional scale vectors. The four C readouts are relative velocity coordinates and their norm. Ridge penalties are selected from $\{0.01, 1, 100\}$ on the discovery split, then decoders are refitted on all readout examples. A four-output ridge decoder supplies its full feature-side span for motion protection. Fresh component probes select penalties independently and need not coincide with this protected decoder. Target discovery and protection both use the full video feature.

We traverse relative-position coordinates, distance, target scales, and zone scales in the fixed order. Each scalar permits up to six rank-one updates. A nonpositive active-target monitor on D_k, or a blocked direction, advances to the next scalar; earlier scalars are not revisited. The endpoint is the end of this one sweep. Protected paths permit at most $10 \times 6 = 60$ updates. Joint removal uses raw A followed by $(v_x, v_y, v_z, \|v\|)$, with empty protection, for at most $14 \times 6 = 84$ updates. B and test metrics never select an endpoint. All targets are audited after the sweep on X_k to measure remaining predictability. Monitors use models refitted on the readout data.

The same ten spatial coordinates are used for raw and residual-target discovery, joint erasure, restoration, and reverse protection. These paths share the split, feature transform, continuous-label scaling, motion protector, and contact-head regularization. The scalar discovery routine uses a numerical variance floor; the separate probe evaluation marks constant-target scores as undefined.

Contact-head weights are refitted after erasure. Logistic regularization remains fixed across ranks. Fifty random projections match removed rank and the implemented training-covariance spectrum criterion; protected controls also respect the motion span. Excess-loss intervals use 1,000 paired test-group bootstrap draws. The main figures show protected raw-A and reverse raw-C erasure with retrained heads; the appendix also includes fixed-head trajectories.

Rank-matched protection comparison. For Fig. 35b, we retain only the A-target prefix of the unprotected joint sweep, ending before the first motion-target attempt. The discovery routine processes one current scalar at a time and does not use later targets to choose an A direction; this prefix is therefore the corresponding unprotected A-only run. We match it to the protected A trajectory at each model's largest common recorded rank: 37, 40, 32, 37, 42 and 35 in Tab. 6's order. Feature preprocessing, sample splits, raw A scaling, ridge grids, scalar order and local budgets are identical. No test value selects the comparison rank. Natural endpoints in panel (a) need not equal these matched ranks. Different local stopping times can yield different active targets at the same rank; the contrast characterizes suppression under the implemented finite-budget procedure.

Fresh motion probes at the same ranks show that unprotected A erasure reduces v_x, v_y and speed test R^2 from their original baselines by 0.181–0.396, 0.782–1.049 and 0.560–0.742, respectively, across models. With protection, the largest absolute change over all motion components is 0.0068. The initially negative v_z scores remain negative in both paths and cannot establish useful variable prediction. These component audits qualify the stronger unprotected spatial suppression: additional A removal accompanies collateral motion loss.

C.6 Restoration protocol

Five-fold configuration-group cross-fitting constructs decoded readout channels from original pre-erasure features; test channels use decoders fitted on all readout examples. Restoration-decoder ridge penalty is fixed at 100, with output scaling fitted within each fold. Decoder weights and output scaling are cross-fitted; upstream PCA, hyperparameter selection, and erasure discovery remain fixed.

Decoded or GT A, C and A+C channels are appended to erased features; A denotes raw spatial descriptors in both protected and joint paths. Appended channels are standardized using readout data only; the erased feature block is unchanged. Every contact head is retrained with the fixed logistic penalty. Each channel has five controls that shuffle entire vectors independently within readout and test splits. This matches dimension and marginal scale but breaks configuration/source correlations as well as sample alignment. GT channels supply privileged test-time information. The OOF training channels and full-fit test channels can differ in distribution; effective regularization is discussed in Sec. A.

C.7 Per-model erasure trajectories

Figs. 29 to 34 show protected-A and joint-A+C trajectories for all six models. A/C and active-target scores audit prediction on the evaluated features; discovery monitors use D_k. Contact heads are fitted at recorded checkpoints, normally every eight ranks and at the endpoint. Lines connect these measurements. The final rank marks the reported endpoint, and random comparisons appear in the main contribution figure and Fig. 37.

Motion-protected plots audit the raw spatial and motion families; joint plots remove both with no protected span.

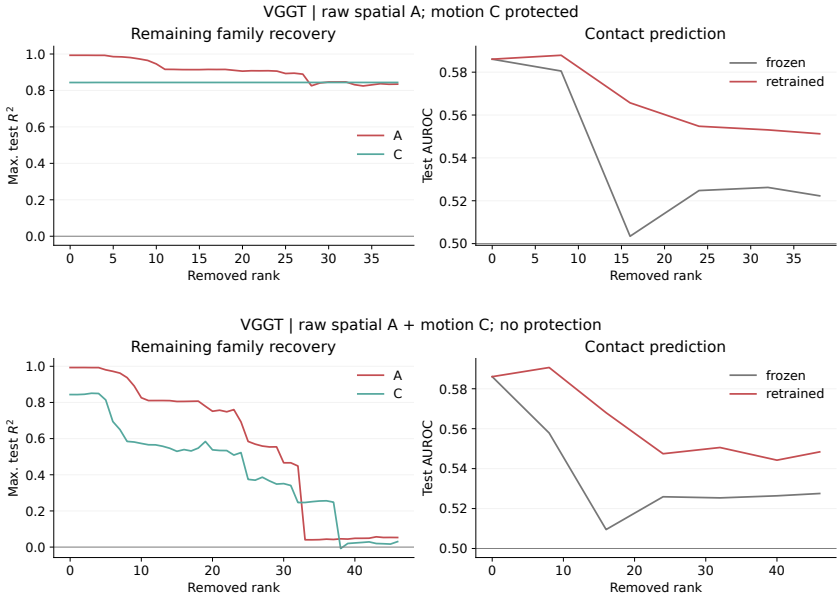

Figure 29: **VGGT: Physion++ fixed-sweep trajectories.** Top: raw spatial A with motion protection. Bottom: joint raw spatial–motion erasure. Left: test probe performance on evaluated features. Right: frozen and retrained contact heads. The final rank marks the fixed-sweep endpoint. Curves show point estimates at recorded checkpoints.

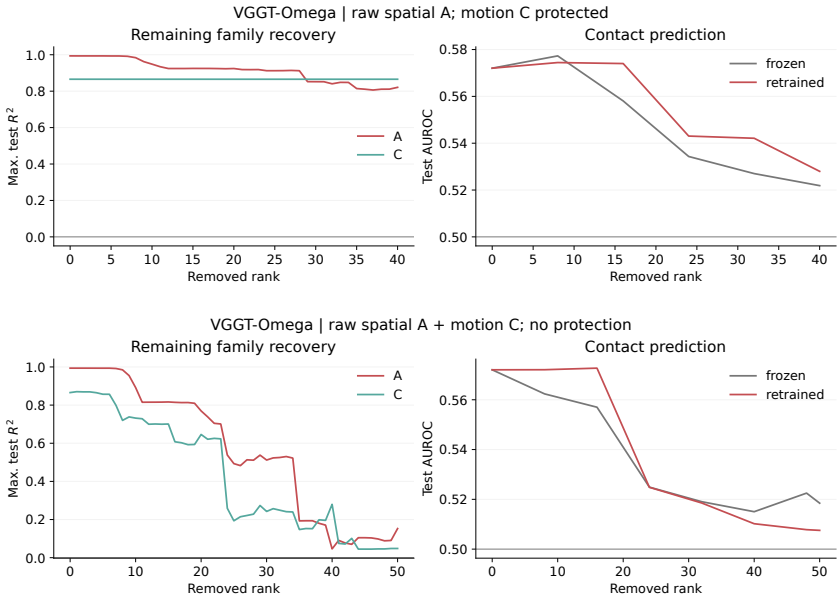

Figure 30: **VGGT-Omega: Physion++ fixed-sweep trajectories.** Top: raw spatial A with motion protection. Bottom: joint raw spatial–motion erasure. Left: test probe performance on evaluated features. Right: frozen and retrained contact heads. The final rank marks the fixed-sweep endpoint. Curves show point estimates at recorded checkpoints.

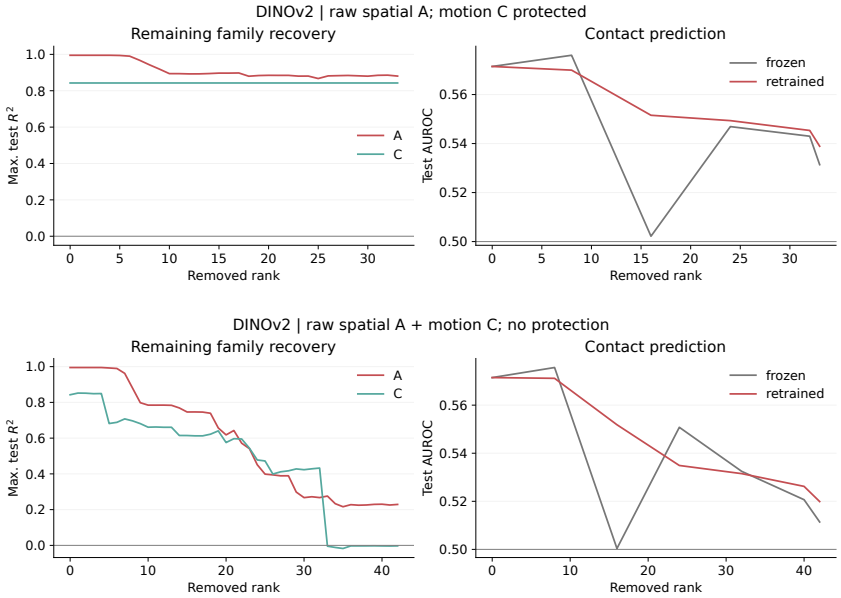

Figure 31: **DINOv2: Physion++ fixed-sweep trajectories.** Top: raw spatial A with motion protection. Bottom: joint raw spatial–motion erasure. Left: test probe performance on evaluated features. Right: frozen and retrained contact heads. The final rank marks the fixed-sweep endpoint. Curves show point estimates at recorded checkpoints.

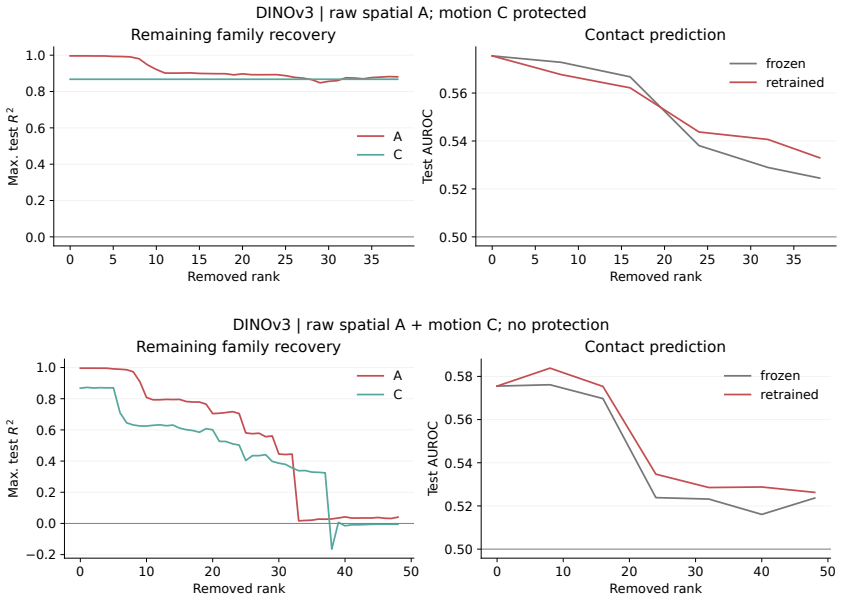

Figure 32: **DINOv3: Physion++ fixed-sweep trajectories.** Top: raw spatial A with motion protection. Bottom: joint raw spatial–motion erasure. Left: test probe performance on evaluated features. Right: frozen and retrained contact heads. The final rank marks the fixed-sweep endpoint. Curves show point estimates at recorded checkpoints.

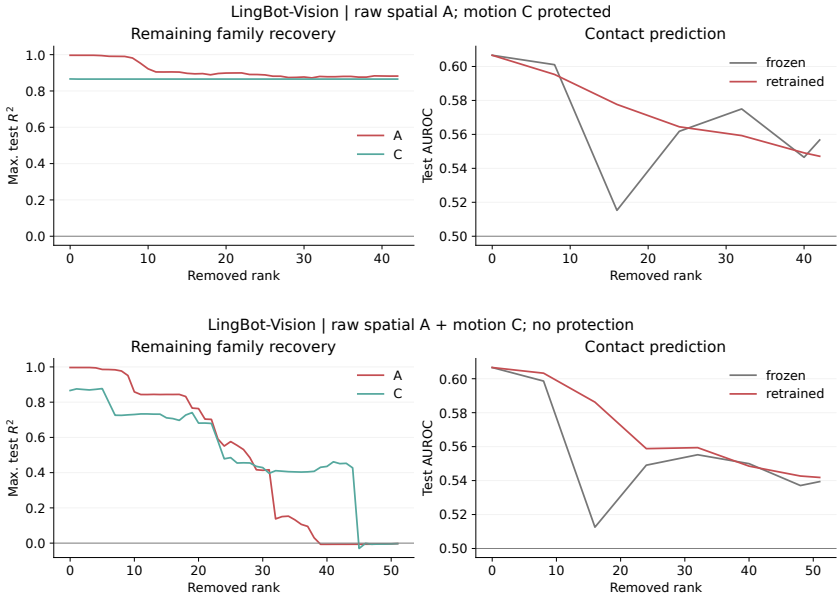

Figure 33: **LingBot-Vision: Physion++ fixed-sweep trajectories.** Top: raw spatial A with motion protection. Bottom: joint raw spatial–motion erasure. Left: test probe performance on evaluated features. Right: frozen and retrained contact heads. The final rank marks the fixed-sweep endpoint. Curves show point estimates at recorded checkpoints.

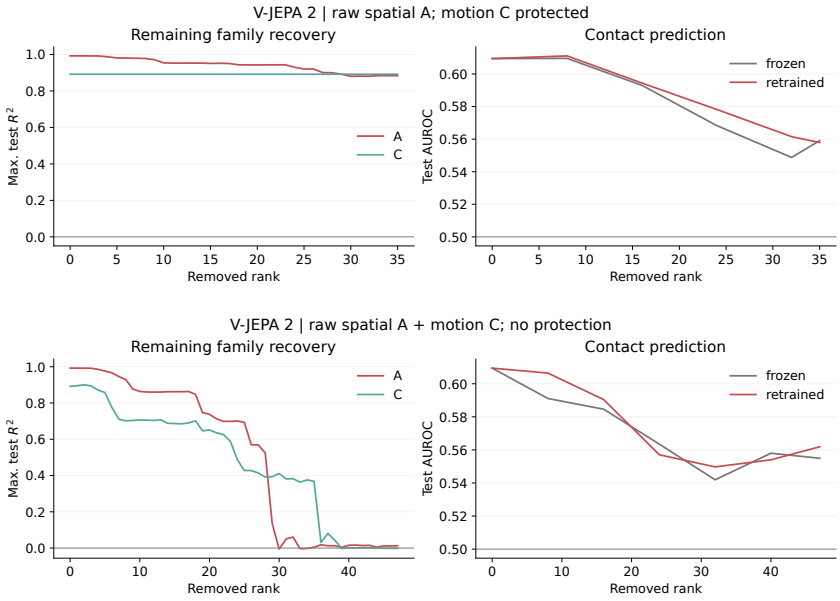

Figure 34: **V-JEPA 2: Physion++ fixed-sweep trajectories.** Top: raw spatial A with motion protection. Bottom: joint raw spatial–motion erasure. Left: test probe performance on evaluated features. Right: frozen and retrained contact heads. The final rank marks the fixed-sweep endpoint. Curves show point estimates at recorded checkpoints.

C.8 REVERSE MOTION ERASURE WITH SPATIAL PROTECTION

We retain the A/C names used in the main text but reverse their intervention roles: raw C is the active target family, and the full feature-side span of a 10-output A decoder is protected. Preprocessing, splits, the ridge grid and contact-head regularization match the forward intervention. The same discovery routine traverses $(v_x, v_y, v_z, \|v\|)$ in order, allowing up to six rank-one updates per scalar (24 total), advancing on a nonpositive or blocked active-target monitor. We report the final rank after the fixed sweep. A/C variable prediction is audited with refitted probes at every rank. Contact AUROC and 50 covariance-spectrum-matched random controls in the spatial-protected complement are evaluated every eight ranks and at the endpoint, including rank zero.

Results and interpretation. The reverse sweeps remove 21–24 directions. Motion remains readily predictable: maximum test R^2 falls from 0.843–0.892 to 0.758–0.814. Maximum spatial test R^2 changes by less than 0.00005, with a largest absolute component-wise change of 0.0057. Contact effects are small and heterogeneous. Retrained AUROC changes range from -0.0113 to $+0.0084$ relative to the original value. Four endpoint paired-group 95% excess-loss intervals include zero; the VGGT interval is entirely negative and the VGGT-Omega interval entirely positive. These task effects accompany partial motion removal.

The reverse target/control trajectories are shown in Fig. 5; the corresponding task trajectories and path-matched random controls appear in Fig. 42.

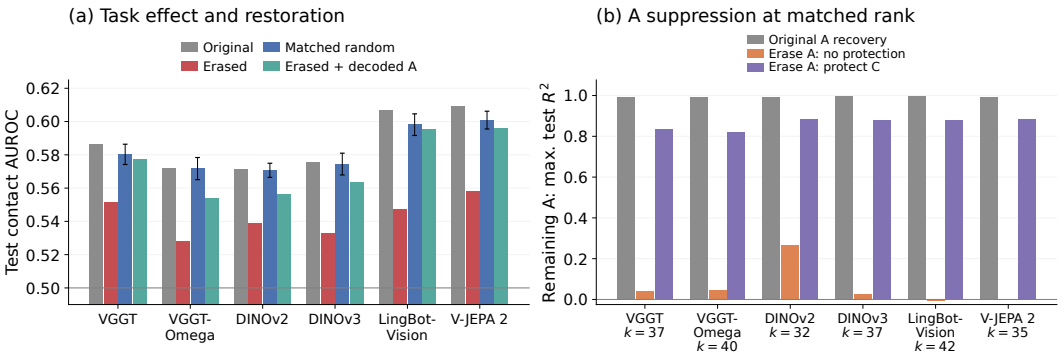

Figure 35: **Contact contribution, restoration, and rank-matched separability.** (a) Retrained contact AUROC for original features, motion-protected A erasure, matched-random erasure, and erased features plus decoded spatial A. Decoded channels are configuration-group cross-fitted; error bars show standard deviations across 50 random projections. (b) Remaining maximum spatial test R^2 with and without C protection at each model's largest common recorded A-only erasure rank k. Gray bars show probe performance on original features. The unprotected path uses only the A-target prefix of the fixed joint sweep, before any C target is erased. Full restoration controls appear in Fig. 36.

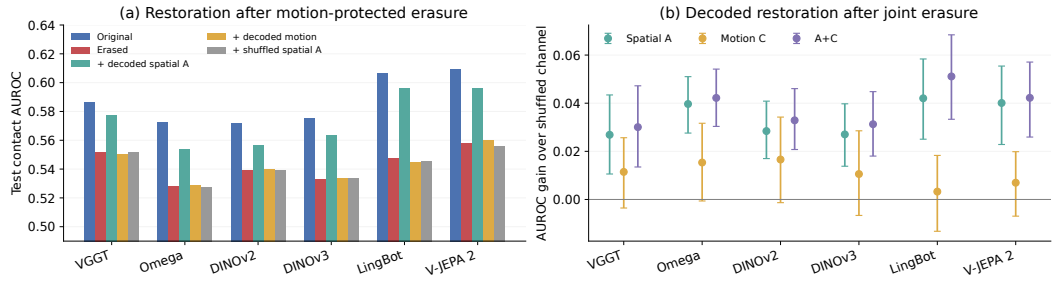

Figure 36: **Restoration through decoded spatial and motion channels.** (a) Absolute contact AUROC after motion-protected spatial erasure, with original and erased references. Shuffled bars average five whole-channel permutations. (b) After joint raw A+C erasure, gain from each decoded channel over its own same-dimensional shuffled control. Readout channels are configuration-group cross-fitted; test channels use a full-readout decoder. All task heads are retrained. Error bars in (b) show paired-group 95% intervals; (a) shows point estimates. Privileged GT results appear in Fig. 39.

C.9 Additional task and restoration results

Motion protection limits spatial suppression at matched rank. At the largest common A-only rank for each encoder (32–42 directions), maximum spatial test R^2 is 0.821–0.886 with motion protection but -0.006–0.267 without it (Fig. 35b). The unprotected comparison uses the fixed joint sweep's spatial-target prefix, before motion-target erasure. Both paths share the features, splits, target order and local budget. Removing protection also reduces probe performance for relative v_x, v_y, and speed; component-wise details are given in Sec. C. Stronger spatial suppression thus accompanies greater motion loss under this protocol.

Decoded spatial channels recover part of the lost task signal. Decoded spatial descriptors improve contact AUROC after motion-protected erasure and outperform shuffled channels in all six encoders (Fig. 35a; Fig. 36). Decoded motion adds little. Training channels use configuration-group cross-fitting, and each contact head is retrained. The full restoration figure also includes combined channels and joint-erasure results. Joint-path gains are measured against each channel's own matched shuffle.

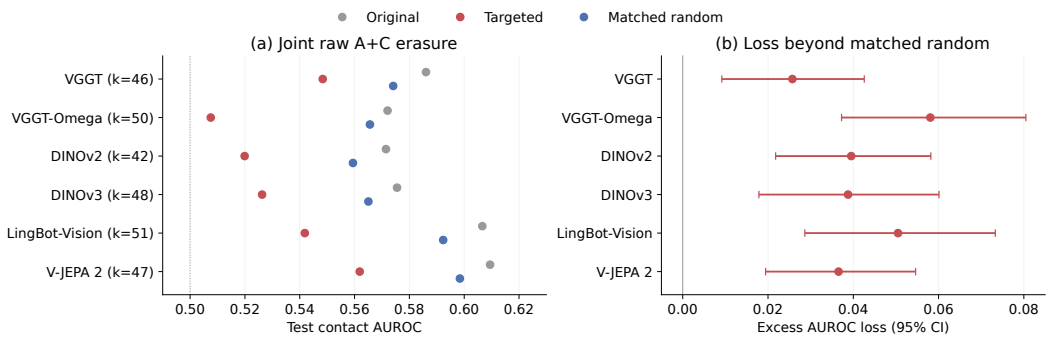

Figure 37: **Joint spatial–motion contribution.** (a) Contact AUROC with retrained heads at fixed-sweep endpoints; k is the removed rank from 512 dimensions. Random scores average 50 same-rank covariance-spectrum-matched projections. (b) Excess loss, random minus targeted AUROC, with 95% intervals from 1,000 paired test-group bootstrap samples.

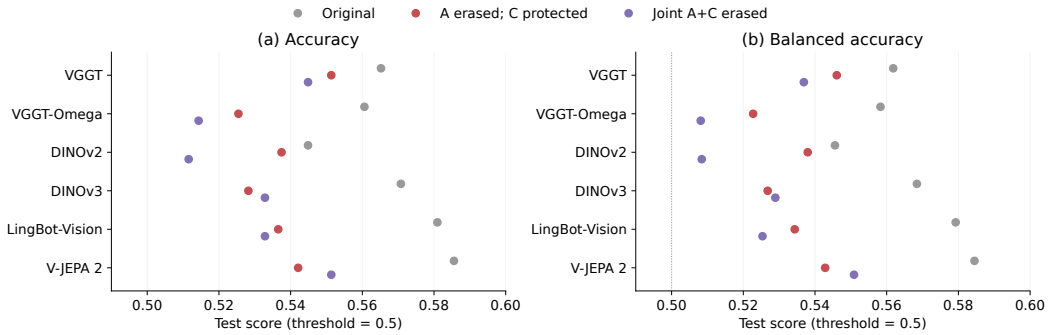

Figure 38: **Threshold-based contact prediction.** Test accuracy and balanced accuracy before erasure, after motion-protected spatial erasure, and after joint raw A+C erasure. All heads are retrained; the probability threshold is fixed at 0.5 without test-based tuning. Points show absolute estimates.

Fig. 37 reports joint-erasure contact AUROC and excess loss beyond matched random projections. Fig. 38 complements ranking performance with fixed-threshold classification metrics. Fig. 39 shows every aligned decoded and GT restoration condition on a shared absolute-AUROC scale.

C.10 Control coverage versus decoder preservation

The protected path retains the full relative-velocity-and-speed decoder span, with fresh probes auditing all four motion readouts. The reverse path protects a decoder of the same ten-coordinate spatial family used throughout the study. These audits distinguish preservation of the fitted decoder from predictability under refitted readouts. Forward and reverse paths differ in target counts, protected spans, and endpoint ranks. Their task effects describe responses to these respective interventions.

C.11 Reporting conventions

All main results use the same expanded spatial-A/minimal-motion-C protocol. Raw-A protected and joint-A+C paths share their baseline and fixed preprocessing. Restoration is evaluated at each path's own recorded endpoint. Spatial-probe summaries aggregate the ten continuous A coordinates. Captions specify whether error bars show random-projection standard deviations or paired-group confidence intervals.

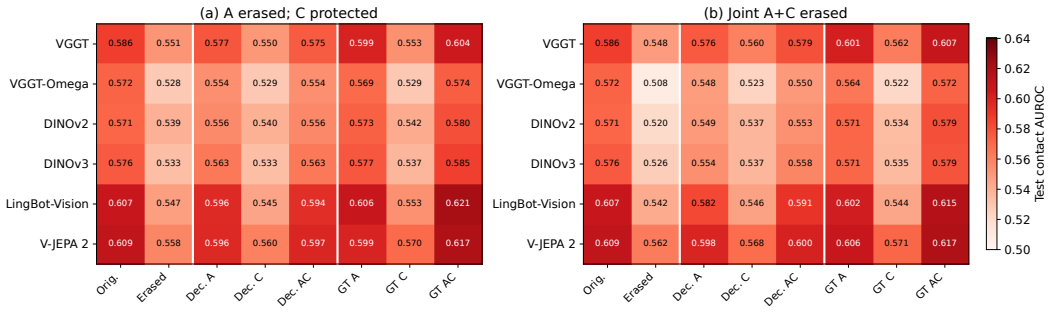

Figure 39: **Complete aligned restoration results.** Test AUROC with retrained heads after (a) motion-protected spatial erasure and (b) joint raw A+C erasure. Both panels share a color scale; larger values are deeper red. A denotes raw spatial targets in both paths. Decoded (Dec.) readout channels are group-cross-fitted; GT channels supply privileged supervision. Cells show absolute AUROC point estimates. Matched-shuffle contrasts appear in Fig. 36.

C.12 Unprotected bidirectional erasure in the main figures

The two paths in Fig. 8 start independently from the same original preprocessed features and set P_C empty. Spatial-only erasure uses the ten spatial targets in their fixed order; motion-only erasure uses the four motion targets in their fixed order. Both retain six local rounds per scalar and the original nonpositive-monitor/numerical-feasibility stopping rules. Both paths use the same splits, target definitions, discovery penalty grids, preprocessing and task-head penalty as the protected runs. No test metric selects a direction, rank, or endpoint.

The spatial-only trajectory is the prefix of the fixed joint A–C sweep ending immediately before the first C-target attempt. Fixed-order discovery makes the spatial prefix independent of the later C stages. Motion-only results use a separate unprotected C sweep starting from the original features. The spatial endpoint ranks are 37, 40, 32, 37, 44 and 35; motion endpoints are 23, 23, 21, 21, 22 and 24, in the encoder order VGGT, VGGT-Omega, DINOv2, DINOv3, LingBot-Vision and V-JEPA2. Contact performance is evaluated at the recorded checkpoints, including the final spatial-target rank, using refitted logistic heads.

For the same-rank comparison with protection, we use the largest rank recorded in both members of each pair, chosen from checkpoint availability alone. Spatial paired ranks are 37, 40, 32, 37, 42 and 35; motion paired ranks are 23, 23, 21, 21, 22 and 23 in the same encoder order. Hence LingBot-Vision's spatial comparison and V-JEPA2's motion comparison precede their unprotected endpoints. Both figures retain the full available trajectories.

The shared full-space random reference uses 50 nested projection draws and the union of the two unprotected paths' task ranks. Each draw uses nested prefixes of an orthonormal Gaussian basis. To extend a basis to the maximum required rank, independent Gaussian directions are twice projected onto its orthogonal complement, orthonormalized and appended, using seed $21260924 + r$. The construction preserves the nested Haar-uniform distribution and is independent of labels. Bands show projection SD; Sec. F distinguishes this reference from the protected path-matched controls.

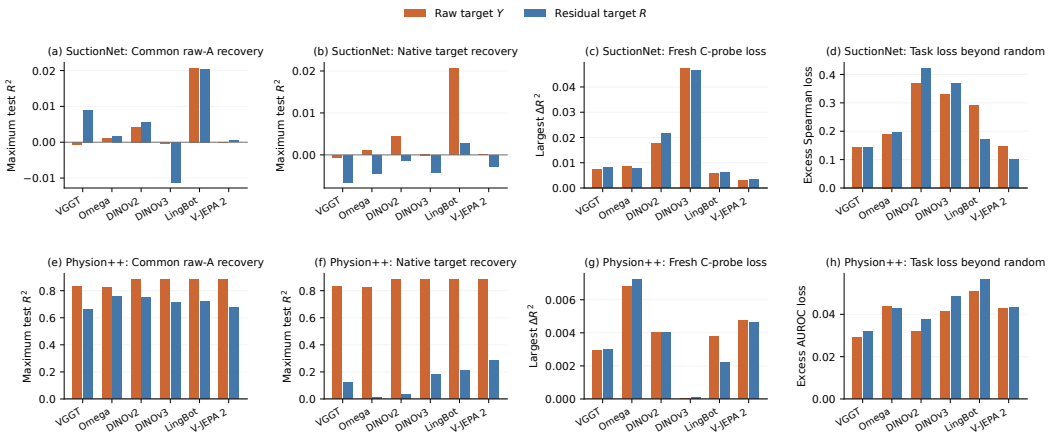

Figure 40: **Raw versus residualized target formulations under the same protected control space.** Rows compare SuctionNet and Physion++ using six encoders. The raw-A panel audits the same raw target family for both paths; residualized targets are shown only where the corresponding channel was actually fitted. C loss measures collateral change in fresh control probes, and task loss is computed relative to matched-rank random projections. Endpoint ranks differ; the equal-rank audit is reported separately.

D Raw and Residual Target Formulations

The main manuscript reports raw-target interventions. This appendix compares them with the linear residual-target formulation while keeping the feature extraction, split, PCA map, control protection and downstream evaluation fixed within each dataset. Changing the target changes the decoder directions and endpoint rank; we therefore compare the resulting suppression patterns and task effects.

SuctionNet. Under full first-order-family protection, both formulations reduce their native second-order targets to near-zero maximum test R^2. Fresh first-order losses remain small relative to the unprotected path. Excess seal loss is positive in all six models for both formulations. The task effect is thus consistent across these target parameterizations, despite differences in their regression targets and endpoint ranks.

Physion++. Both Physion++ formulations use the same ten continuous spatial coordinates, excluding target type. Residual-target paths leave maximum residual-target test R^2 of 0.011–0.281, while their maximum raw spatial-A test R^2 remains 0.661–0.758. Raw-target paths retain maximum raw-A test R^2 of 0.821–0.884 at their own endpoints. Native residual and raw R^2 scores use different labels. The common raw-A audit supplies the shared comparison space. Fresh motion-probe losses remain small in both paths. The result distinguishes suppression of the fitted residual from suppression of the original spatial family under the same measured motion protection.

Removability under protection depends on the target formulation. Writing $A = \widehat{m}(C) + [A - \widehat{m}(C)]$ makes the distinction explicit: residualized erasure targets the remainder of the training-fitted linear regression, whereas raw erasure targets the original measurements. Both paths constrain removal through the same protected decoder span. Equal removed rank controls dimensionality, while target covariance and discovered directions can still differ. The main Physion++ analysis uses raw targets to track predictability of the original spatial measurements.

E Discussion and Limitations

SuctionNet and Physion++ show different patterns of task contribution and control preservation. On SuctionNet, erasing a small second-order-aligned subspace reduces seal ranking with limited loss in first-order probe performance. Decoded second-order descriptors recover performance. On Physion++, motion protection limits spatial suppression, yet erasing spatial-predictive directions lowers contact prediction. Relaxing protection strengthens spatial suppression and also reduces motion predictability. The achievable selectivity depends on the representation and variable families.

Scope of attribution. Attribution is defined within the preprocessed feature space. PCA, the measured target families, and the probe class determine its scope. Protection preserves selected decoder outputs; high-rank erasure may also affect unmeasured task information. Restoration adds an explicit descriptor channel and can change effective regularization. These conditions limit how specifically a task effect can be assigned to the target family.

Data and evaluation. SuctionNet evaluates analytical seal quality on a qualified subset. Physion++ uses simulated contact outcomes, modest baseline readouts, and a motion family restricted to translational velocity and speed. Restoration cross-fits target decoders with upstream preprocessing and erasure discovery held fixed. Discovery monitors include refitted examples.

F Shared Random Baselines and Path-Matched Controls

The bidirectional task plots use a common *unprotected* random-erasure reference in each encoder's fixed preprocessed feature space. For each repetition, we draw an independent standard Gaussian matrix and obtain an orthonormal basis by reduced QR. Its prefixes define nested Haar-uniform random subspaces at the union of the two targeted paths' recorded task ranks. We use eight repetitions for SuctionNet and 50 for Physion++, with seed $20260924 + r$ for zero-based repetition r. Direction sampling is independent of labels, test scores, protected spans, and target-specific covariance criteria. Features, splits and task-head fitting match the corresponding targeted experiment. SuctionNet retains validation-selected ridge refitting and Physion++ retains its fixed logistic penalty with refitted weights. Bands show standard deviations across random projections.

The shared baseline matches removed rank and samples the full feature space. Path-specific controls additionally match the implemented training-covariance criterion and respect the corresponding protected decoder span. We use these path-specific controls for all excess-loss contrasts and report each path separately because its admissible subspace differs.

G Further Connections to Probing, Concept Erasure, and Physical Reasoning

Decodability depends on the evaluation question. Information-theoretic probing estimates target information available in a representation; a more expressive probe can tighten that estimate (Pimentel et al., 2020). Minimum-description-length probing measures the cost of learning and encoding a readout (Voita & Titov, 2020). Both address accessibility. Task contribution requires intervention evidence. Kumar et al. (2022) further show that a probe can exploit concept-associated features without reliably identifying those that constitute the concept, complicating detection and removal. We assess accessibility with linear probes and relate it to task performance through erasure, control audits, and refitted readouts.

Concept erasure and the scope of removal. Linear Adversarial Concept Erasure formulates removal as a minimax problem against a linear predictor (Ravfogel et al., 2022a); kernel-space erasure extends the approach to selected nonlinear classes (Ravfogel et al., 2022b). The predictor class defines the removal criterion. Log-linear guardedness also distinguishes direct concept prediction from downstream restrictions and indirect recovery (Ravfogel et al., 2023). Our experiments measure finite-family suppression with refitted linear probes while protecting selected control-decoder spans. Single-target, one-shot, and protected/unprotected comparisons test how these choices affect physical-task prediction.

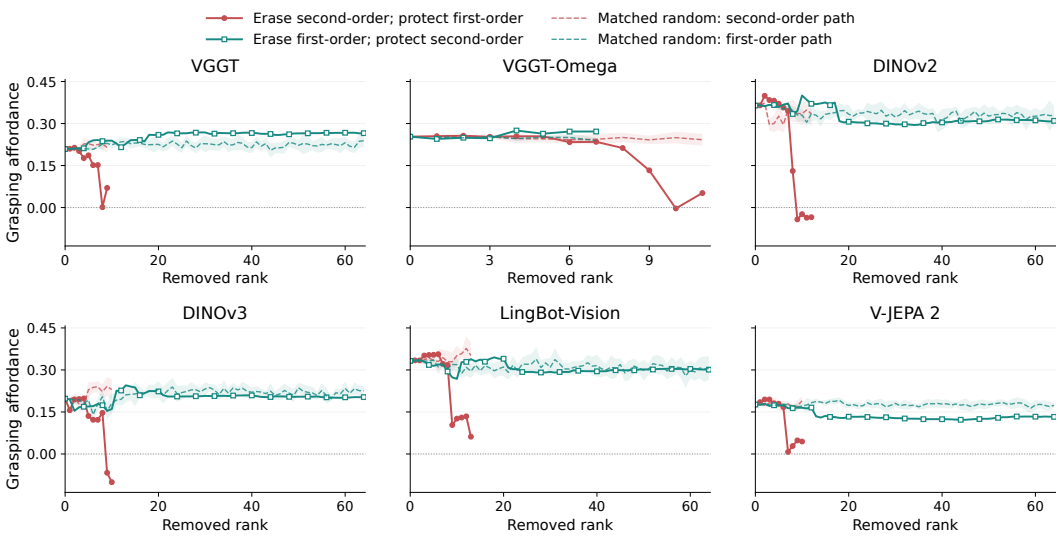

Figure 41: **Path-matched SuctionNet controls.** Each targeted trajectory is paired with its own protected matched-random mean and projection SD (eight repetitions). Colors identify the erased family; dashed lines show its corresponding control. The shared unprotected reference appears in Fig. 6.

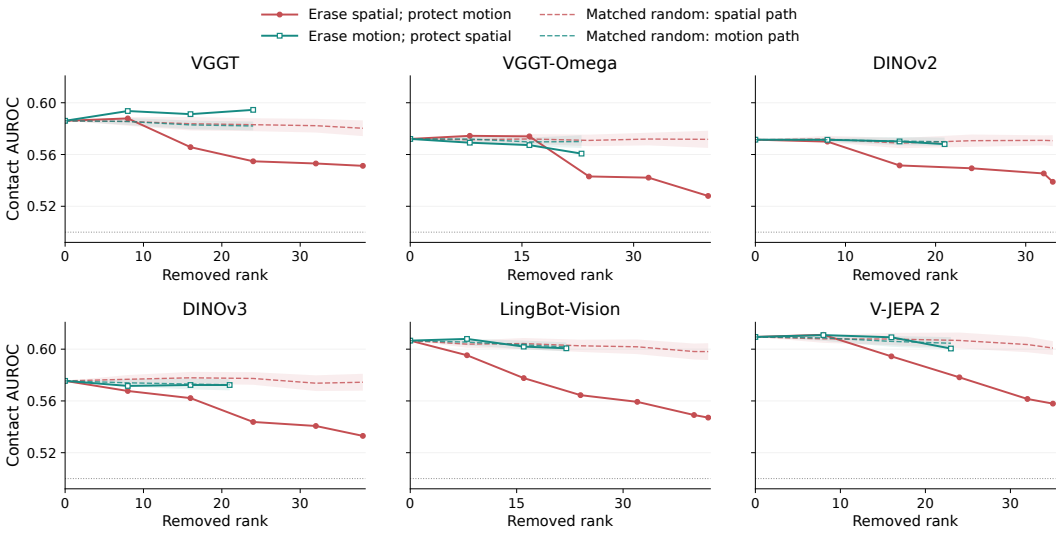

Figure 42: **Path-matched Physion++ controls.** Solid lines show protected spatial or motion erasure; same-color dashed curves and bands show the corresponding protected matched-random mean $\pm$ SD over 50 projections. Each random control respects the same protected decoder span as its targeted path. Within-path excess loss uses these protected controls.

Representation interventions and causal abstraction. Causal abstraction aligns neural representations with variables in a high-level causal model and tests their counterfactual correspondence through interchange interventions (Geiger et al., 2021). Causal representation learning seeks to discover such variables from observations (Schölkopf et al., 2021). Our targets are specified physical measurements. We erase directions predictive of those measurements and evaluate the task performance supported by the remaining features. Random and reverse-role controls contextualize the task effect; target and control probes measure the information affected.

Physical prediction and the information supporting it. PLATO evaluates intuitive physics through object-centric representations and violation-of-expectation tests (Piloto et al., 2022). Graph Network-based Simulators learn dynamics by passing messages over particle-based states (Sanchez-Gonzalez et al., 2020). Our experiments examine which geometric, spatial, and motion measurements in frozen visual features support task readouts. This analysis links representation content to physical-task performance.