

Readout is not generation: $SE(3)$ in the patch tokens of frozen encoders

Anonymous Authors

ANONYMOUS@OPENREVIEW.NET

Abstract

How much of a frozen vision transformer’s latent change under real camera motion is produced by the motion itself, and where does that structure live? It lives in the patch-token field, and it is computed rather than learned: resampling the field by the lattice homography of the known camera rotation, an operator with no fitted parameters, explains 16–37% of held-out change at a four-frame gap, and its plane-induced extension roughly half. With its one parameter frozen rather than tuned, that extension beats a learned flow head fitted on the same pairs in all fifteen comparisons across five encoders, and its margin widens where the learner is weakest: carried to a new scene, the head loses 0.232 in skill (95% CI [0.101, 0.379]); since the operator is fitted to nothing, that loss is also the growth of its advantage. The structure is easy to miss, because the standard probe is aimed exactly where it cancels. Mean pooling removes the geometric action up to boundary and density terms, so the pooled displacement is, to first order, exactly the change geometry does not produce, and every operator we test on it fails, as it must; run as a readout, the same probe reproduces the published linear null at short gaps and inverts to positive at long ones. What bounds the claim is the operator’s validity rather than the data: transport beats copying until the pose error reaches about 1.35 times the true rotation, past which it is worse than not transporting. How much of the structure an encoder has to begin with follows what it was trained to predict: a pre-registered ordering of six encoders held. Thresholds and controls were fixed before each measurement, and eleven of the twelve mechanisms we proposed did not survive them. The result is a division of labour: compute what the camera determines, and learn the remainder.

Keywords: equivariance, $SE(3)$, camera motion, frozen encoders, vision transformers, patch tokens, world models, pre-registration

1. Introduction

When a camera moves, the tokens of a frozen vision transformer move with it, and that part of the change does not have to be learned: it can be computed from the pose. We establish this in the simplest setting in which the ground truth is exact: a frozen encoder, real video, and a camera whose motion is known to motion-capture precision.

That structure is invisible to the instrument this question is usually asked with. A vision transformer represents a scene as a field of tokens on a lattice, but a probe on the pooled embedding sees only that field’s zero mode. The question has a two-by-two structure, which we state once because the rest of the paper follows it: two objects, the pooled vector and the token field, and two directions, readout, which decodes the pose from the latent change, and generation, which predicts the latent change from the pose. We measure three of the four cells. On the token field, generation succeeds with nothing to fit: resampling the field by the lattice homography of the known rotation, which we call the rotation warp, explains 16–37% of the held-out change at a four-frame gap, and its plane-induced extension roughly

half. On the pooled vector, generation fails for every operator we test, and Section 3 shows why it must. On the pooled vector, readout works once the frame gap is long enough: SeeSE3’s linear null at short gaps reproduces, and the same readout turns positive at long gaps. Readout is not generation, and the field is not the pool. The fourth cell is companion work.

The rest of the paper measures what that operator is worth and where it stops. **(1)** In a low-data, in-scene regime the one-parameter plane warp beats a matched learned flow head in all fifteen cells above the displacement floor, across five encoders, by a median of 0.141 in skill, and the rotation warp, with no fitted parameter, wins ten of the fifteen. **(2)** We derive and measure the readout–generation asymmetry on the pooled vector, and measure field generation at 42–52% of held-out change. **(3)** Carrying the learned head to a new scene costs it 0.232 in skill (95% CI [0.101, 0.379]); because the operator’s score does not depend on the training scene, this cost and the growth of the operator’s advantage are the same quantity. **(4)** We quantify three limits: a head supervised on the true flow delivers 76% of the required displacement at a 24-frame gap; the operator tolerates pose error up to about 1.35 times the true rotation; and the tokens’ orientation response is angle-specific, but its fitted linear form is not a representation of $SO(2)$. **(5)** A pre-registered ordering of six encoders by how movable their pretraining target is was confirmed, though at this n the target cannot be separated from token-grid resolution. **(6)** We report every mechanism we refuted and every claim we withdrew.

The setting is narrow — three sequences, one camera, frozen encoders, small learners — and every claim is scoped to it. What the narrowness buys is exactness: ground-truth pose, an operator with nothing to fit, and controls written before the numbers existed.

2. Setup

We use three TUM RGB-D sequences (Sturm et al., 2012) with motion-capture poses: two static desk scenes (fr2_desk and fr1_desk) and one with people walking through it (fr3_walking_xyz). Frames are paired at gaps of 1, 4 and 12. Each sequence is cut into ten contiguous time blocks, even blocks for fitting and odd for scoring, with a guard band of gap frames at every boundary. The encoders are frozen and off the shelf: DINOv2 (Oquab et al., 2023) (ViT-S/14 for the generation measurements of Section 3, ViT-B/14 at 448 px after), DINOv3 (Siméoni et al., 2025) ViT-B/16, and for the encoder sweep MAE (He et al., 2022), CLIP (Radford et al., 2021) and DeiT-III (Touvron et al., 2022). We use last-layer patch tokens reduced to 64 principal components fitted on training blocks only; Appendix F gives the protocols.

Throughout, a *cell* is one encoder $\times$ scene $\times$ gap combination, an *arm* is one predictor, and the *displacement floor* excludes cells whose median token displacement is below one token. We score every arm by identity-relative skill on held-out blocks, $\rho = 1 - \text{MSE}(\hat{z}, z_{t+\text{gap}}) / \text{MSE}(z_t, z_{t+\text{gap}})$, so that $\rho = 0$ is copy-last and $\rho < 0$ is worse than doing nothing. Skill is a fraction of held-out change, not a correlation.

Operators. For a rotation R with intrinsics K , the token lattice moves by the homography $H = KRK^{-1}$. The *rotation warp* (**warp_rot**) resamples the token field by H and has zero fitted parameters. The *plane warp* (**warp_plane**) adds the translation term of a plane-induced homography, in two forms: in Section 3 the plane’s normal and depth

are fitted on training pairs (three numbers); from Section 4 on the normal is fixed and the depth frozen at 2.0 m, a one-parameter operator whose parameter was never tuned. Every operator score here is a lower bound, for two reasons: the frozen depth is the best choice in none of the four cells of the depth panel (Appendix B), and resampling a field of approximately localised tokens is itself an approximation.

Learners. The matched learned arm is a convolutional head that emits a per-token displacement field applied by the same resampling as the warp; we call it a flow head, and it can express the warp. It replaced an additive head, a residual added to the identity, which proved a broken baseline (Section 4). Every arm has exactly zero skill at step 0, asserted at run time. The regime is small: heads have at most 1.4M parameters and are fitted on 250–390 pairs per scene, so Section 4 is a low-data, in-scene result; Sections 5 and 6 are where the difference stops being about data.

Floors, selection, intervals. We exclude cells below the displacement floor before any tally, on geometry fixed in advance; this removes every gap-1 cell and every walking-scene cell at gap 4. Where a learned competitor was selected across seeds, selection used the test split, which favours the learner and makes our margins conservative. Intervals are block bootstraps over time blocks. Predictions, thresholds and kill conditions were written before each measurement ran; the failures are in Section 10.

3. Where the group acts

The group acts on the token field, and mean pooling cancels almost all of what it does. We derive that, measure generation on both objects, then run the probe in the other direction. Write the token field as $z(x)$ and let $(T_g z)(x) = z(w_g(x))$ be the resampling of token positions that a camera motion g induces, with w_g the lattice homography; we call T_g the base action. Let $p(z) = \frac{1}{N} \sum_x z(x)$ be the mean-pooled vector. Then $p(T_g z) - p(z)$ is, up to interpolation, the sum of two terms: a boundary term from tokens that enter or leave the frame, and a density term from the Jacobian of the resampling, which departs from one only insofar as the homography is not a rigid shift of the lattice. For nearby frames both are small. If the observed change is the base action plus a residual that the base action does not induce, the pooled displacement is therefore, to first order, the pooled residual. An operator that generates latent change from the pose generates the base action, whose pooled effect is about zero, so the pooled displacement is exactly the part of the change it cannot reach. For mean pooling this holds by construction; for a CLS token it is an empirical question.

The measurement bears this out (Fig. 1, left). On a frozen DINOv2 ViT-S/14, with 2324, 603 and 848 pairs on the three sequences, the magnitude of the pooled displacement tracks ego-motion (correlation up to 0.84 at gap 12), but its direction is not predictable across time blocks: a shift, a first-order Lie operator fitted by ridge regression, and an unconstrained MLP all score held-out skill at or below zero at every gap, for patch-mean and CLS pooling alike. Because even the unconstrained MLP fails, what is missing is information, not function class. The pooled vector is the field’s zero mode.

On the token field, the same motion is generated with nothing to fit. On DINOv2 ViT-S/14 patch tokens at 448 px, the rotation warp explains a held-out fraction of the change that rises to 0.294 at gap 4 on the static desk and 0.374 on the walking scene; the fitted

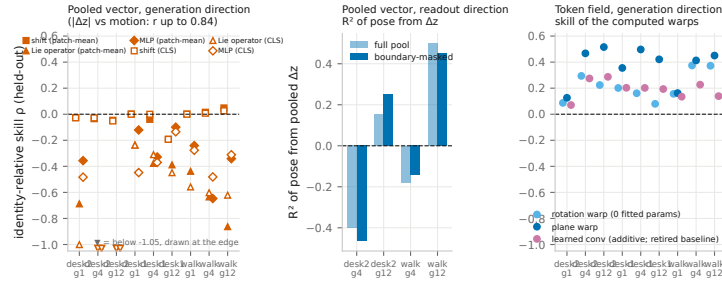


Figure 1: Same frames with motion-capture poses; frozen DINOv2 throughout (ViT-S/14 in the left and right panels, ViT-B/14 at 448 px, patch-mean, in the middle). Left, generation on the pooled vector: every predictor of the pooled displacement scores held-out $\rho \leq 0$, for patch-mean and CLS pooling alike. Middle, readout on the pooled vector: the linear pose-from- Δz R^2 is negative at gap 4 and positive at gap 12, and survives the boundary mask on both scenes. Right, generation on the token field: the rotation warp explains 0.16 to 0.37 of held-out change at gap 4 and the fitted plane warp about half.

plane warp of Section 2 reaches 0.42–0.52 on the static scenes and 0.41–0.45 with people walking through (Fig. 1, right). One detail matters later: the rotation warp peaks at gap 4 and falls by gap 12 (0.294 to 0.224 on one desk) while the plane warp keeps rising (0.467 to 0.516), rotation alone being increasingly mis-specified as the camera translates. The learned arm in that figure is the additive baseline we later retired; the fair comparison is Section 4.

Decoding pose from the pooled displacement is the other direction, and it has a different answer. SeeSE3 (Chen et al., 2026) measure this linear readout and report $R^2 = -0.28$ for DINOv2-B on TUM RGB-D with a linear probe on CLS or mean-pooled features. At a short gap we reproduce the null: R^2 is -0.400 and -0.178 on our two scenes (patch-mean, DINOv2 ViT-B/14 at 448 px; theirs is a CLS probe on ViT-B). At gap 12 the same readout turns positive, to $+0.155$ and $+0.500$, and the inversion is not an artefact of tokens leaving the frame: pooled only over tokens still in frame under the known warp it survives on both scenes, larger on the desk ($+0.717$ against $+0.555$) and smaller on the walking scene ($+0.591$ against $+0.678$; Fig. 1, middle, Appendix C). We have no mechanism for it, and the derivation above supplies none: it concerns generation, not readout. From here the paper stays on the field, and on generation.

4. Computed versus learned operators

Result. Across five frozen encoders, two scenes and two gaps, the one-parameter plane warp beats a flow head fitted on the same pairs in all fifteen cells above the displacement floor, by a median margin of 0.141 in skill (minimum 0.029, maximum 0.305). The rotation warp, with zero fitted parameters, wins ten of the fifteen and never loses by more than 0.05. Table A.1 lists every cell; Fig. 2 shows the paired margins with intervals.

Margins with intervals. On the two primary encoders, the mean paired margin of the plane warp over the flow head is 0.104 on DINOv2 (95% CI [0.049, 0.162]); three of six

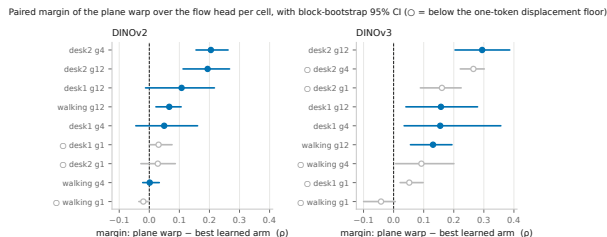


Figure 2: Per-cell paired margin of the one-parameter plane warp over the flow-output learned head, with block-bootstrap 95% intervals, on the two primary encoders. Cells below the one-token displacement floor are hollow. One DINOv2 cell straddles the floor between two displacement estimators and is excluded from every fifteen-cell tally; its margin is $+0.001$ (Appendix A).

block intervals exclude zero) and 0.184 on DINOv3 (95% CI $[0.138, 0.260]$; four of four). These margins are conservative: the competitor is the best of three seeds on the test split, worth 0.003 to 0.033 per cell in its favour (Section 2).

Why the additive baseline was replaced. The first version of this comparison used an additive head and found the learned arm near zero on DINOv3. That was the parameterisation, not the encoder: on the same cell the additive head scores 0.014 and the flow head 0.335, and the flow head wins all fifteen cells at fewer parameters. We no longer claim that learned predictors fail to recover this structure; the narrower claim is the one above, against the stronger baseline.

Operator quality versus the residual. With the rotation-only warp, a learned residual helps in one of four conditions, and restricting the operator to a learned subspace rescues the one failing cell. With the plane warp that cell rises from 0.224 to 0.516, the residual adds 0.007, 0.009, -0.001 and -0.014 , and the restriction hurts everywhere (0.347 against 0.474; 0.337 against 0.525). Fixing the operator was worth 0.29 in the worst cell; the best contribution we ever measured from a residual was 0.029. The structure appears in all five encoders and training objectives, and DINOv3 is the strongest column: its frozen-depth plane warp scores at least 0.509 in every cell above the floor.

5. Transfer across scenes

The operator’s advantage is widest exactly where a learned head is weakest, on a scene it was not fitted on. We fit each learned head on one scene and score it on the held-out blocks of another, for all nine ordered pairs of the three scenes and both gaps, on DINOv2 ViT-B. The analytic operator is fitted to nothing, so its score does not depend on which scene the learner saw; the learner’s transfer cost and the growth of the operator’s advantage are therefore the same quantity, which we report once. Carrying the flow head across scenes costs it 0.232 in skill (95% CI $[0.101, 0.379]$, a scene-level paired cluster bootstrap over six clusters of target scene $\times$ gap). The operator’s advantage over it grows by the same amount, from 0.103 in scene (95% CI $[0.044, 0.162]$) to 0.335 under transfer (95% CI $[0.226,$

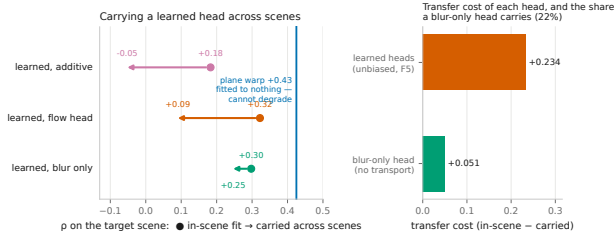


Figure 3: Transfer. Left: each learned head’s in-scene skill and its skill when carried to another scene; the analytic warp is fitted to nothing and cannot degrade. Right: each head’s transfer cost and the fraction a blur-only head carries; the remainder’s mechanism is open (§5).

0.449]); the additive head’s ratio is $1.97\times$ (95% CI [1.69, 2.30]; Fig. 3, right). Seed selection is train-side (Appendix D).

Candidate mechanisms. We pre-registered two explanations for the cost; both were refuted. It is not a scene-tuned denoiser: a symmetric-blur head that can smooth but cannot transport loses only 0.051 when carried across scenes, a fifth of the flow head’s cost (ratio 0.22, inside the pre-registered band). It is not parallax: the correlation between the cost and the target scene’s non-rotational motion fraction is negative in both arms ($r_s = -0.664$, $p = 0.018$; $r_s = -0.707$, $p = 0.010$, relative cost). We leave the mechanism open.

6. Reach and tolerance

A computed operator has a domain; we measure two of its edges.

Reach. At a 24-frame gap on the static desk, where the mean true displacement is 5.8 tokens against 1.2 at gap 4, a flow head trained directly on the true flow, converged at zero weight decay, delivers 76% of the required displacement and stays 0.159 below the warp. Weight decay is not the cause: removing it moves skill by +0.012 at gap 24 and -0.012 at gap 4, opposite signs inside seed noise. Optimisation budget accounted for about a fifth of what we had attributed to capacity; our earlier figure of 56% was under-trained and is withdrawn (Section 10). The head samples its displacement field globally, so its reach is unbounded and the shortfall is not a kernel-size artefact. One candidate remains untested: a 3×3 convolution has no position input while the flow of a rotation is position-dependent, so position can enter only through what the principal components retain of the positional embedding. We tried neither a coordinate channel nor a global-attention head at this gap. For this head the limit is representational, by elimination.

Tolerance. We inject a pose error of magnitude ε times the true inter-frame rotation ω and sweep ε . The error at which the rotation warp stops beating copy-last is consistent across cells (Fig. 9): $c = 1.35$ (95% CI [1.06, 1.59]), pooling the eleven of twelve cells whose curve crosses zero within the sweep. Over the nine well-specified cells (rotation fraction at least 0.60) the crossings average 1.54; in the two cells where the operator is already mis-specified they sit at 0.72 and 0.29, and we draw them apart. Past the crossing, mis-specified transport is worse than not transporting. One registered prediction missed:

the Spearman correlation between a cell’s rotation fraction and its rotation-warp skill, $r_s = +0.580$ ($p = 0.048$), fell below its registered bar of 0.6. We do not know why an operator 1.35ω from the truth still beats one ω from it; the one candidate, that the overshoot’s error is bounded by the field’s spatial autocorrelation while the identity’s is not, is untested. Whether the walking scene’s people are where the operator fails can be tested only in differential form; Appendix C gives the result and the reason.

7. Orientation content of tokens

The warp moves tokens; it does not turn them. The polar decomposition of its Jacobian gives a camera-induced local rotation $\theta(x)$ at every token, which the warp knows analytically and discards. Two questions follow: whether the tokens carry orientation at all, and whether correcting them for θ is worth anything.

They carry it, and angle-specifically. Rotating the image in-plane by $5\text{--}90^\circ$, un-rotating the token grid, and fitting a linear map from the original tokens to the rotated ones, a per-angle map recovers skill of 0.372 at 5° , rising to 0.641 at 90° , and the map fitted at 90° scores -0.561 when applied at 5° . A single-generator fit of the form $\exp(\theta A)$ trails the per-angle maps, at 0.121 at 5° and 0.545 at 90° (Fig. 10, right).

Correcting for θ , on this data, is worth nothing. A control that uses the same fitted generator with θ permuted across positions matches the gauge-correction arms to three decimals, leaving about 0.013 in one cell and nothing elsewhere (Fig. 10, left). The reason is the motion, not the encoder: desk video has a median gauge angle of $0.46\text{--}1.09^\circ$, and a re-test at 90th-percentile angles of $7.4\text{--}8.6^\circ$ found margins of 0.001–0.018 over the shuffled control, so the null is not for lack of power.

What the fit does show is that the response is not a group representation. The fitted generator has eigenvalues in conjugate pairs, but at non-integer frequencies (1.55, 0.75, 0.67, 0.42, ...) and with a real-to-imaginary ratio of 3.57, whereas a finite-dimensional representation of a compact group is equivalent to an orthogonal one, with a skew generator and integer frequencies. Two readings remain open: either the response is not a group action, because token content is not closed under rotation for an encoder trained on upright images, or the pipeline’s non-rotation components (resampling blur, borders, grid un-rotation) contaminate the fit. The control that would separate them was not run.

8. What sets the ceiling: the pretraining target, at $n = 6$

A token can only be as equivariant as the thing it was trained to predict. Rotating a real frame by a known angle gives exact ground truth for the lattice warp, so the equivariance ceiling, the warp’s skill under that rotation, is measurable per encoder with no training; before running it we ordered six checkpoints by how movable their pretraining target is: pixels (MAE), distilled features (DINOv2, DINOv3), invariant semantic targets (DeiT-III, CLIP), and a self-latent target (V-JEPA 2). The pre-registered ordering was confirmed (Fig. 4, Appendix G): at 5° the ceilings run from 0.374 (V-JEPA 2) to 0.758 (MAE), Table A.4 has the column, and the rank order matches the prediction (Spearman $+0.971$, $p = 0.001$ at the registered angle; the same at 10°). Among the four image checkpoints sharing a 28×28 token grid the ordering holds exactly ($n = 4$); at $n = 6$ there is no evidence for or against

token-grid resolution as the axis, and the separating test at larger n is companion work. Nor is there evidence for depth (a deeper DINOv2 has the same decay length and ceiling), temporal blending (unchanged on duplicated frames) or token delocalisation ($r_s = -0.400$, $p = 0.600$). The ceiling is measured under a synthetic pan of a static frame, so it bounds the rotation-only operator and is silent on translation; that is why it agrees with Section 4 on the family split but not within the top three (MAE leads the ceiling, DINOv3 real-motion skill).

The self-latent encoder (Assran et al., 2025) behaves as predicted: transport structure exists in its latents (0.26–0.36 above the floor), a learned blur alone beats the warp alone (0.375–0.415 against 0.310–0.358), and warp plus blur beats a 1.4M-parameter attention predictor in all four cells.

9. Related work

SeeSE3 (Chen et al., 2026) asks whether vision features reflect Euclidean space by decoding pose from the displacement of CLS or mean-pooled features with a linear probe and a trained adapter, and concludes positively once the adapter is used. Probes of 3D awareness (El Banani et al., 2024) find frozen-feature correspondence deteriorating with viewpoint; Section 6 measures that exactly.

Transport operators and Lie generators have been learned from image pairs for decades (Rao and Ruderman, 1998; Culpepper and Olshausen, 2009; Sohl-Dickstein et al., 2010); ours is computed and tested under transfer. Neural Isometries (Mitchel et al., 2024) learn a latent space whose encodings are related by a learned linear operator, training the encoder; flow-equivariant recurrent networks (Keller, 2025) build equivariance to one-parameter motion flows into the sequence model; here the flow is known and the field frozen. The Lie derivative (Gruver et al., 2022) finds that training and data matter more than architecture for learned equivariance; Section 8 adds the target.

Symmetric embeddings (Park et al., 2022), Neural Fourier Transform (Koyama et al., 2023) and class-pose decomposition (Marchetti et al., 2022) fit the encoder around a known operator so that it is valid by construction; we hold the encoder fixed and measure where one is valid. Steerers (Bökman et al., 2024) learn a linear $SO(2)$ action on descriptors and derive its orthogonality; Section 7 reports an encoder whose fitted response lacks that form. Known-operator learning bounds error assuming an exact operator (Maier et al., 2019); adjoint-range results (Lunz et al., 2021) say where forward-only correction can act. Frozen-DINO video predictors (Karypidis et al., 2024; Baldassarre et al., 2025) report copy-last baselines and no geometric-warp baseline (Šarić et al., 2020).

10. Refuted mechanisms and withdrawn claims

Twelve mechanisms were proposed in the course of this work, each with a control registered before it ran. Eleven were refuted by that control; the twelfth, that a token is only as equivariant as its pretraining target, is not refuted at $n = 6$. Six claims were withdrawn. Appendix D gives each mechanism with what refuted it, and each withdrawal with its replacement. What survives is the part of the latent change a camera-side operator can compute, and the boundary at which it stops.

References

- Mido Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv:2506.09985*, 2025.
- Federico Baldassarre et al. Back to the features: DINO as a foundation for video world models. *arXiv:2507.19468*, 2025.
- Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Steerers: A framework for rotation equivariant keypoint descriptors. In *CVPR*, 2024. *arXiv:2312.02152*.
- 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:2607.14228*, 2026.
- Benjamin J. Culpepper and Bruno A. Olshausen. Learning transport operators for image manifolds. In *NeurIPS*, 2009.
- 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. *arXiv:2404.08636*.
- Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The Lie derivative for measuring learned equivariance. *arXiv:2210.02984*, 2022.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.
- Efstathios Karypidis et al. DINO-Foresight: Looking into the future with DINO. *arXiv:2412.11673*, 2024.
- T. Anderson Keller. Flow equivariant recurrent neural networks. In *NeurIPS*, 2025. *arXiv:2507.14793*.
- Masanori Koyama, Kenji Fukumizu, Kohei Hayashi, and Takeru Miyato. Neural Fourier transform: A general approach to equivariant representation learning. *arXiv:2305.18484*, 2023.
- Sebastian Lunz, Andreas Hauptmann, Tanja Tarvainen, Simon Arridge, and Carola-Bibiane Schönlieb. On learned operator correction in inverse problems. *SIAM Journal on Imaging Sciences*, 14(1):92–127, 2021.
- Andreas K. Maier, Christopher Syben, Bernhard Stimpel, Tobias Würfl, Mathis Hoffmann, Frank Schebesch, Weilin Fu, Leonid Mill, Lasse Kling, and Silke Christiansen. Learning with known operators reduces maximum error bounds. *Nature Machine Intelligence*, 1, 2019. *arXiv:1907.01992*.
- Giovanni Luca Marchetti, Gustaf Tegnér, Anastasiia Varava, and Danica Kragic. Equivariant representation learning via class-pose decomposition. *arXiv:2207.03116*, 2022.

- Thomas W. Mitchel, Michael Taylor, and Vincent Sitzmann. Neural isometries: Taming transformations for equivariant ML. In *NeurIPS*, 2024. arXiv:2405.19296.
- Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. arXiv:2304.07193, 2023.
- Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan Willem van de Meent, and Robin Walters. Learning symmetric embeddings for equivariant world models. arXiv:2204.11371, 2022.
- Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- Rajesh P. N. Rao and Daniel L. Ruderman. Learning Lie groups for invariant visual perception. In *NeurIPS*, 1998.
- Josip Šarić, Marin Oršić, and Siniša Šegvić. Warp to the future: Joint forecasting of features and feature motion. In *CVPR*, 2020.
- Oriane Siméoni et al. DINOv3. arXiv:2508.10104, 2025.
- Jascha Sohl-Dickstein, Ching Ming Wang, and Bruno A. Olshausen. An unsupervised algorithm for learning Lie group transformations. arXiv:1001.1027, 2010.
- Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012.
- Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *ECCV*, 2022.

Appendix A. Per-cell tables

A.1 The audited encoder sweep (F4). Five encoders $\times$ two scenes $\times$ two gaps under the tight protocol, two seeds, best seed favouring the learner; the five walking-gap-4 cells fall below the one-token floor (0.86–0.98 tokens) and are excluded from every tally. Per-cell values, read from the sweep artifact at build time, are in the table below the summary (cells below the floor marked [†]). Summary against the flow head:

arm	vs flow head	margins (min / median / max)
warp_rot (zero fitted parameters)	10/15	−0.048 / +0.037 / +0.108
warp_plane (d = 2.0 frozen)	15/15	+0.029 / +0.141 / +0.305
flow head vs additive head	15/15	—

encoder	cell	disp. (tokens)	warp_rot	warp_plane ($d=2$)	flow head	additive head
DINOv2	desk2 g4	1.16	+0.341	+0.501	+0.323	+0.200
DINOv2	desk2 g12	3.24	+0.286	+0.531	+0.334	+0.182
DINOv2	walking g4 [†]	0.98	+0.394	+0.403	+0.397	+0.222
DINOv2	walking g12	2.21	+0.398	+0.427	+0.348	+0.160
DINOv3	desk2 g4	1.01	+0.416	+0.635	+0.379	+0.076
DINOv3	desk2 g12	2.83	+0.330	+0.652	+0.347	+0.065
DINOv3	walking g4 [†]	0.86	+0.439	+0.486	+0.416	+0.169
DINOv3	walking g12	1.94	+0.443	+0.509	+0.335	+0.014
MAE	desk2 g4	1.01	+0.394	+0.527	+0.400	+0.232
MAE	desk2 g12	2.83	+0.293	+0.513	+0.308	+0.177
MAE	walking g4 [†]	0.86	+0.356	+0.355	+0.365	+0.261
MAE	walking g12	1.94	+0.373	+0.358	+0.329	+0.177
CLIP	desk2 g4	1.01	+0.340	+0.427	+0.287	+0.261
CLIP	desk2 g12	2.83	+0.305	+0.477	+0.337	+0.186
CLIP	walking g4 [†]	0.86	+0.316	+0.329	+0.287	+0.245
CLIP	walking g12	1.94	+0.354	+0.373	+0.282	+0.182
DeiT-III	desk2 g4	1.01	+0.349	+0.436	+0.311	+0.260
DeiT-III	desk2 g12	2.83	+0.313	+0.489	+0.297	+0.201
DeiT-III	walking g4 [†]	0.86	+0.324	+0.334	+0.303	+0.244
DeiT-III	walking g12	1.94	+0.366	+0.370	+0.282	+0.184

Table A.1: the audited sweep, every cell; [†] below the one-token displacement floor, excluded from every tally; best of the three transport arms in bold.

A.2 The two primary encoders under the tight protocol (VALIDATE 2).

Additive (conv), flow-output (flow_conv), and attention (xf, flow_xf) learners beside the operators; the additive head’s near-zero on DINOv3 is the retired baseline.

encoder · scene · gap	warp_rot	warp_plane	conv	flow_conv	xf	flow_xf
DINOv2 static g4	.341	.501	.205	.294	.201	.306
DINOv2 static g12	.286	.531	.193	.327	.311	.364
DINOv2 walking g4	.394	.403	.203	.401	.203	.383
DINOv2 walking g12	.398	.427	.161	.320	.262	.375
DINOv3 static g4	.416	.635	.072	.392	.094	.359
DINOv3 static g12	.330	.652	.065	.337	.238	.473
DINOv3 walking g4	.439	.486	.002	.428	.087	.418
DINOv3 walking g12	.443	.509	.021	.383	.152	.402

A.3 The field-level measurement (Exp I). DINOv2 ViT-S/14 patch tokens at 448 px, 32×32 grid, 64 PCs, held-out time blocks; the learned conv is the additive baseline later retired.

sequence	gap	warp_rot	warp_plane	learned conv
static fr2	1	0.087	0.126	0.070
static fr2	4	0.294	0.467	0.275
static fr2	12	0.224	0.516	0.287
static fr1	1	0.202	0.355	0.203

sequence	gap	warp_rot	warp_plane	learned conv
static fr1	4	0.161	0.497	0.202
static fr1	12	0.080	0.421	0.193
walking	1	0.157	0.162	0.135
walking	4	0.374	0.413	0.227
walking	12	0.372	0.450	0.139

A.4 The ceiling at $n = 6$ (R1e). Exact-rotation ceilings per encoder, with the pre-registered rank. There is no evidence for depth as the axis (DINOv2-L and DINOv2-B have decay lengths of 57 and 58 px and ceilings of .614 and .591 at 5°), for delocalisation (r_s (decay length, ceiling) = -0.400 , $p = 0.600$), or for temporal blending (the ceiling measured on duplicated frames is unchanged).

encoder	target	rank	2°	5° (registered)	10°
MAE	pixels	5	+0.806	+0.758	+0.726
DINOv2	distilled feature	4	+0.535	+0.602	+0.507
DINOv3	distilled feature	4	+0.656	+0.582	+0.474
DeiT-III	class label	2	+0.624	+0.528	+0.468
CLIP	language	2	+0.553	+0.488	+0.459
V-JEPA 2	self latent	1	+0.319 (below floor)	+0.374	+0.409

A.5 The transportable subspace, corrected. A two-dimensional subspace of the 384-d token space, selected by a generalised eigenproblem on training blocks, against a matched-normalisation random control (random columns of the same eigenbasis, max of 20): real on **static_desk2** (.544 vs max .374 at g4; .504 vs .354 at g12) and on walking (.570 vs .469; .553 vs .469), failing on **static_desk** (.213 vs .254; .040 vs .110 $\pm$.105). The subspace carries ≈ 0.1 – 1.3% of total latent movement. The original control, random *Euclidean* directions, reproduced the full space at every row and was a tautology, not a control.

Appendix B. The depth panel

warp_plane’s one parameter, the plane depth d , swept after the fact on the two primary scenes. The pre-registered “flat” threshold was a spread ≤ 0.05 ; the observed maximum spread is 0.304, so the control fired and **warp_plane** is a one-parameter operator, never a zero-parameter one (that is **warp_rot**’s property alone). $d = 2.0$ is the argmax in none of the four cells, so every reported **warp_plane** number is a lower bound.

cell	$d=1.0$	$d=1.5$	$d=2.0$	$d=2.5$	$d=3.0$	$d=4.0$	spread	argmax
static	+.424	+.508	+.501	+.480	+.463	+.437	0.084	1.5
g4								
static	+.435	+.562	+.531	+.487	+.453	+.410	0.152	1.5
g12								

cell	d=1.0	d=1.5	d=2.0	d=2.5	d=3.0	d=4.0	spread	argmax
walking	+.222	+.352	+.403	+.423	+.431	+.434	0.212	4.0
g4								
walking	+.177	+.336	+.427	+.464	+.479	+.481	0.304	4.0
g12								

Appendix C. The gap inversion

Linear readout of pose from the pooled displacement, held-out R^2 , full pool and pooled only over tokens still in frame under the known warp. The inversion between gap 4 and gap 12 survives the mask on both scenes, larger on one and smaller on the other; it is not boundary turnover. We attach no mechanism: a mechanism is reported only once it has survived a registered control.

cell	full pool	boundary-masked	valid fraction
static_desk2 g4	−0.400	−0.465	0.96
static_desk2 g12	+0.155	+0.252	0.89
walking g4	−0.178	−0.142	0.96
walking g12	+0.500	+0.449	0.90
inversion, static_desk2 (g12 − g4)	+0.555	+0.717	
inversion, walking (g12 − g4)	+0.678	+0.591	

To be precise about the comparison with SeeSE3: their own recovery of pose comes from a trained nonlinear adapter, not from the frame gap, and their adapter value on TUM, 0.348, is an R^2 of pose from features, not comparable to any skill value in this paper.

Movers, in differential form. A mask on the warp residual (absolute intensity difference above 0.06 in de-normalised units) was meant to pick out tokens on people walking through the scene. A *static* scene registers a “mover” fraction of 0.33–0.53 under it, so the mask is substantially detecting warp error, not motion. What can be stated is a difference: the walking scene’s masked-residual ratio is $1.34\times$ against the static control’s $1.12\times$, with a registered bar of 1.15 that the control clears by only 0.03. The data do not support an absolute statement about mover tokens. A separate decay of +0.042 in the learner’s skill at gap 4, real and unexplained, stays open.

Appendix D. The refuted mechanisms, the freeze audit, and the withdrawals

The twelve mechanisms (Section 10). Each was proposed with a control registered before it ran, and each control could have failed.

proposed mechanism	what refuted it
token delocalisation	Spearman −0.400 with the ceiling, $p = 0.600$
low-rank Δz (V-JEPA)	its effective rank is higher, 50.1 vs 43.4
an optimisation horizon	true-flow supervision scores below unsupervised

proposed mechanism	what refuted it
additive parallax budget	over-predicts both held-out cells (+0.50, +0.84)
one knee in token units	spread 115% wider in tokens than in degrees
geometry traded away	blur keeps $\approx 24\%$ of the decay; clean targets $3\times$ worse
parallax (tested twice)	r_s inverted: $-0.694/-0.678$; $-0.664/-0.707$
flow transfers worse	$+0.235$ vs $+0.232$
transformer saturation	margins inverted: $+0.035$ V-JEPA, $+0.344$ image
weight decay limits reach	± 0.012 , opposite signs, inside noise
a scene-tuned denoiser	denoising carries at $0.22\times$ the cost
target sets the ceiling	not refuted at $n = 6$; registered before compute

The freeze audit. Every shipping script was read against the seven failure classes this programme actually hit before the numbers were frozen. Findings and dispositions:

check	finding	disposition
PCA / whitening fitted on training rows only	two early scripts fit on every-other-frame slices spanning held-out blocks	superseded: the sweep, the grid and the pooled analysis were all re-run
boundary guard band	absent in the same two scripts	superseded by the re-runs
zero skill at initialisation	one early conv not zero-initialised	the re-run asserts skill $< 1e-6$ per cell at step 0
learned-arm parameterisation	additive only — the refuted baseline	superseded by the flow-output arm, 15/15 grid
hidden fitted parameters	the plane depth is a parameter (spread 0.304)	“zero fitted parameters” restricted to warp_rot
selection on the test split	flattered the transfer drop by $+0.014$; flattered the learner only in the margins	transfer patched to train-side; margins kept conservative and documented

Withdrawals. (i) “*Learned ≈ 0 on DINOv3*” — the additive head scores $+0.014$ and the flow head $+0.335$ on the same cell; a $214\times$ change from output form alone, at fewer parameters (234k vs 305k). (ii) “*warp_rot wins 8/8*” — 5 of 8 under the tight protocol, 10 of 15 in the audited sweep. (iii) “*Margins of $+0.18$ to $+0.58$* ” — against the weakest baseline with the wrong variance component; replaced by $+0.104$ [$+0.049$, $+0.162$] and $+0.184$ [$+0.138$, $+0.260$]. (iv) “*A transfer ratio of $2.4-3.3\times$* ” — the summary filtered cells with a substring that matched both arms and printed their pooled mean $((0.242 + 0.103)/2 = 0.172$, reproduced to three decimals); replaced by per-arm cluster bootstraps, conv $1.97\times$ [1.69 , 2.30] and flow diff $+0.232$ [$+0.101$, $+0.379$]; a preflight now fails on any substring arm selection in the tree. (v) “*The conv head cannot represent large displacements, full stop*” at 56% — 1,500 steps at constant learning rate; at 4,000 steps with a cosine schedule and the same weight decay it is 0.71, and 0.76 at zero weight decay. (vi) “*39-51% of latent movement*” in a two-dimensional subspace — withdrawn: the quantity printed was the subspace’s skill, not its share of movement; the share is $\approx 0.1-1.3\%$, and the random-direction control was an identity. Test-side seed selection, where a learned competitor was

the quantity selected, biases every margin against the operator and is kept; where the selected quantity was ours (the transfer drop) it was replaced by train-side selection, which moved the transfer cost from +0.248 to +0.234. The +0.232 in §5 is the flow arm’s own drop with its cluster-bootstrap interval.

Appendix E. Pre-registrations and the claim $\rightarrow$ contract-slot map

Every number in this paper is a copy from the programme’s append-only record and resolves to a slot in a frozen contract that a verifier checks against the artifacts (100 slots; 82 checkable, all inside their frozen bounds; 18 recorded without a machine-checkable interval and printed as point values). The map from each printed claim to its slot identifier, the registered prediction and threshold, the control that could have failed, and the artifact will be released with the code and the frozen contract on acceptance; it could not be anonymised in time for review. Selected registrations, as written before compute:

- **Field-level group fraction** (Exp I): P1 warp beats identity; P2 `warp_rot` decays with gap where `warp_plane` does not; P3 walking lower than static — *not confirmed*; P4 learned fails at walking gap 12 — confirmed.
- **Encoder sweep** (F4): P1 `warp_plane` > flow head in every floored cell; P2 flow > additive in every floored cell. Both held at 15/15.
- **Transfer** (R5/F2): P1 advantage larger under transfer — held; P2 parallax — refuted twice; denoiser ratio ≤ 0.5 band — held (0.22).
- **Reach** (F1): P1 shortfall persists (≤ 0.70) / P2 closes (≥ 0.85 and gap ≤ 0.10) — *neither*; 0.76 sits between, and neither mechanism is claimed.
- **Tolerance** (R4): P3 crossing ratio with CI width < 1.0 — held ($c = 1.35$ [1.06, 1.59]); P1 Spearman(`rot_frac`, ρ_{rot}) > 0.6 — *missed* at +0.580 ($p = .048$).
- **Movers** (F3): walking > 1.15 $\times$ and static < 1.15 $\times$ — held by 0.03 on the control side.
- **Gauge** (L, O): gauge arms beat the θ -shuffled control — refuted, then refuted at larger angles.
- **Ceiling** (R1e): Spearman(registered rank, `ceiling@5°`) $\geq +0.70$ — held (+0.971); the ordering itself confirmed.

Appendix F. Protocols

Pairs and splits. Frames are associated to the nearest motion-capture pose within 50 ms. Pairs ($t, t + \text{gap}$) are formed within each of ten contiguous time blocks; even blocks train, odd blocks score; a guard band of gap frames on either side of every block boundary is excluded from pairing.

Tokens. Last-layer patch tokens of the frozen encoder, per-dimension standardised and reduced to 64 principal components, with the standardisation and the PCA fitted on training-block tokens only (a subsample of 60,000 tokens). All skills are computed in that 64-dimensional space.

Operators. `warp_rot` resamples the token field by the lattice homography $H = KRK^{-1}$ of the relative rotation, bilinear, with border replication. `warp_plane` uses the

plane-induced homography $K(R + tn^T/d)K^{-1}$; in §3 the plane (n, d) is fitted on training pairs, from §4 on n is fixed and $d = 2.0$ m.

Learners. Heads take the 64-PC token field and the relative pose as six broadcast input channels. The convolutional head is three 3×3 layers with 128 hidden channels; in *flow* mode it emits a two-channel displacement field applied by the same bilinear resampling as the warp, in *additive* mode a 64-channel residual added to the identity. The final layer is zero-initialised so skill is exactly zero at step 0, asserted at runtime. Training minimises mean squared error on the held-in tokens with AdamW (learning rate 10^{-3} , weight decay 10^{-4} , cosine schedule, batch 32, up to 200 epochs) with early stopping on a held-in validation split of the training blocks (patience 20); attention heads use learning rate 3×10^{-4} , batch 8, up to 150 epochs, patience 15. Two seeds in the sweep, three elsewhere.

Pooled arms (§3). The shift arm adds a fitted constant; the Lie arm is a first-order operator, linear in the pose, fitted by ridge with the penalty swept over $\{0.1, 1, 10, 100, 1000\}$ on an inner validation split of the training blocks; the MLP takes the pooled vector and the pose through one hidden layer of 256 GELU units.

Reach (§6). On the static desk sequence at gaps 4 and 24, DINOv2 ViT-B/14: the flow head is trained *directly on the true lattice flow* of the plane homography (mean squared error on the displacement field) with AdamW, learning rate 10^{-3} , 4,000 steps, cosine schedule, at weight decay 0 and at AdamW’s default 0.01 — a different setting from the unsupervised learners’ 10^{-4} , which is why Fig. 7 labels them so.

Intervals. Margins: paired cluster bootstrap over held-out time blocks, 2,000 resamples. Transfer: scene-level paired cluster bootstrap, six clusters (target scene $\times$ gap), 20,000 resamples.

Tolerance sweep. A rotation error of magnitude ε degrees about a uniformly random axis is composed with the true relative rotation before the warp is built; eight independent axis draws per ε over the grid $\{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 20\}^\circ$; the crossing of held-out skill through zero is found by linear interpolation between adjacent grid points and divided by the cell’s mean true rotation ω .

Ceiling. 200 frames of the static desk sequence; a synthetic camera pan of 2° , 5° or 10° about the vertical axis is applied to the image as the exact homography KRK^{-1} (bilinear, border replication) and both images are encoded; the ceiling is the identity-relative skill of the exact lattice warp of the original tokens against the tokens of the warped image. Cells whose mean displacement is below one token are excluded.

Fibre. 600 frames. In-plane rotation of the image by $5\text{--}90^\circ$ about its centre; the token grid of the rotated image is un-rotated so tokens are compared at corresponding scene points; a per-angle linear map, and a single-generator map $M(a) = \exp(aA)$ shared across angles, are fitted by least squares on training frames and scored on held-out frames; the cross-angle control applies the map fitted at one angle to every other angle. The gauge correction on TUM uses the local rotation $\theta(x)$ from the polar decomposition of the warp Jacobian; its control permutes θ across positions.

Appendix G. Figures

Every value in every panel of every figure is read from a frozen artifact.

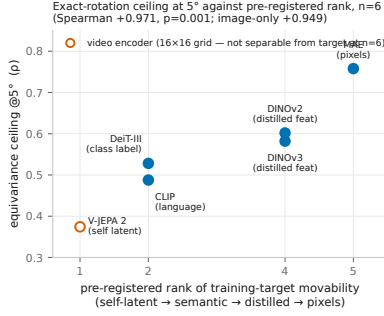


Figure 4: The exact-rotation ceiling at 5° against the pre-registered rank of each encoder’s pretraining target. The ordering was written before compute and was confirmed; at $n = 6$ there is no evidence for grid resolution as the axis and none against it.

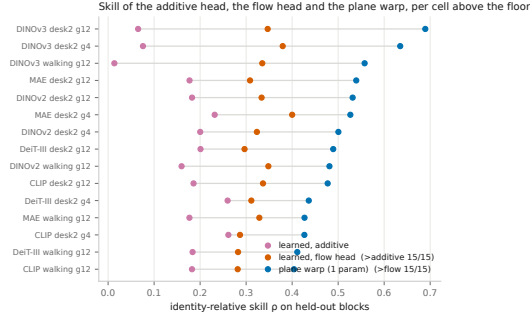


Figure 5: The function class: additive head → flow-output head → plane warp, per floored cell. The flow head beats the additive one in every cell; the warp beats the flow head in every cell.

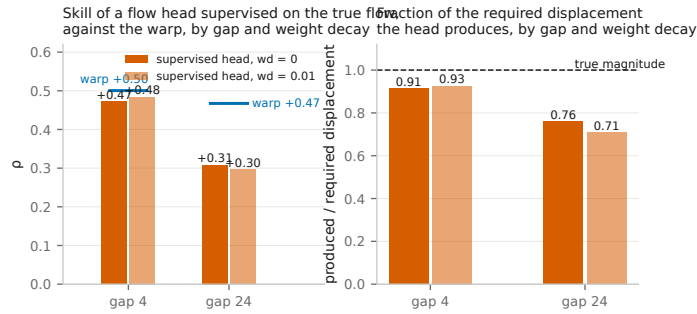


Figure 6: The reach bound at a 24-frame gap (mean true displacement 5.8 tokens): a flow head supervised on the true flow, at zero and at AdamW’s default weight decay, against the warp; right, the fraction of the required displacement it produces.

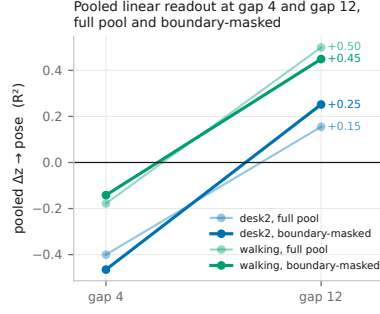


Figure 7: The gap inversion of the pooled linear readout under the boundary mask: it survives on both scenes, larger on one and smaller on the other. We have no mechanism for it (§3).

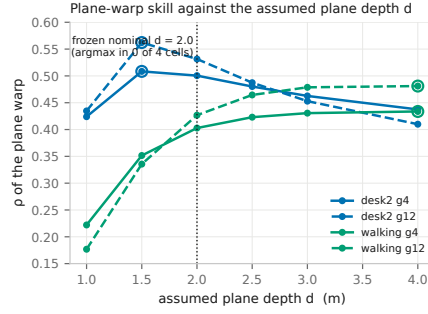


Figure 8: The depth panel: the plane warp’s skill against its one parameter. The frozen $d = 2.0$ is the best value in none of the four cells, so every reported plane-warp number is a lower bound.

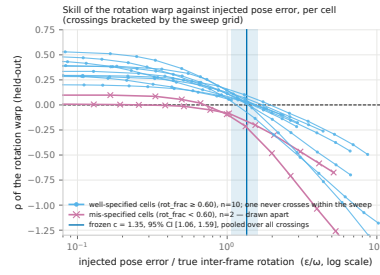


Figure 9: The tolerance law. Held-out skill of the rotation warp against injected pose error in units of the true inter-frame rotation, per cell; crossings are bracketed by the sweep grid. The ten well-specified curves (rotation fraction ≥ 0.60 ; one never crosses within the sweep) cross near the frozen constant $c = 1.35$ (band: 95% CI, pooled over all crossings); restricted to the well-specified cells the crossings average 1.54; the two mis-specified cells are drawn apart.

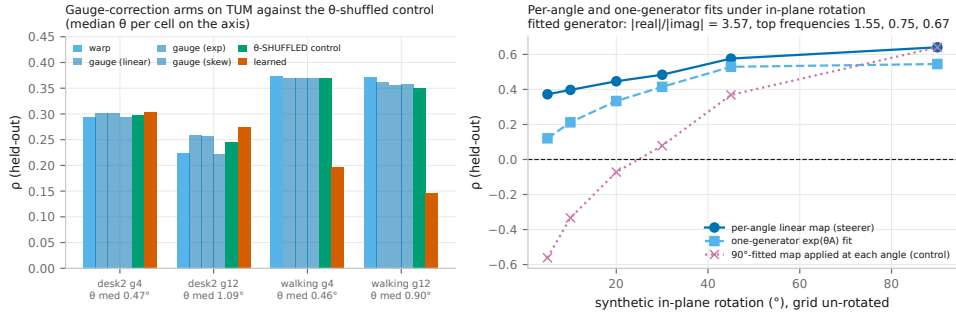


Figure 10: Orientation content. Left: correcting token content for the camera-induced local rotation on TUM is matched by the θ -shuffled control. Right: under synthetic in-plane rotation with the grid un-rotated, a per-angle linear map recovers orientation content, the one-generator fit trails it, and the map fitted at 90° fails at 5° . The fitted generator’s spectrum is printed, not interpreted.