LATENTGAUGE: PLANNING-GEOMETRY ATTACKS IN A MEASURE-PRESERVING BLIND SPOT

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent world models plan by comparing imagined futures with an encoded goal. Accurate prediction and a regular latent distribution constrain the representation, while planning also depends on the geometry delivered to its scorer. Following the planning equations, we identify an inference-time degree of freedom: a shared coordinate map inserts a metric into the action-side Gauss–Newton matrix while the predictive model stays fixed. We introduce LATENTGAUGE, a planning-geometry attack that searches within Gaussian-measure-preserving radial transformations under a displacement budget. Probability conservation determines the admissible construction; spatially varying rotation changes its geometry; an exact cost-shift identity connects that change to candidate ranking. For CEM, the native elite boundary determines calibration, and a finite-round bound links exact score changes to successive mean and variance updates. A shared scoring operation and one constrained ranking objective make this chain executable without modifying the encoder or predictor. A shared kernel extension specifies the radial profile; feasible trust-region optimization fits it under the exact displacement constraint. We derive the closed-form map, its induced metric and spectrum, and the separation between displacement and local distortion. Controlled Gaussian and affine-model calculations exhibit an exact ranking reversal within an independently integrated budget and reproduce the geometric invariances. The resulting construction links distribution-preserving geometry to candidate selection and native planning updates.

1 INTRODUCTION

Latent world models plan by comparing predicted futures with an encoded goal. DINO-WM uses pretrained features (Zhou et al., 2025); LeWorldModel (LeWM) learns the encoder and predictor from pixels with prediction and Gaussian regularization (Maes et al., 2026). Temporal straightening shapes representation geometry to improve planning (Wang et al., 2026).

At zero prediction risk, a representable measure-preserving conjugacy retains prediction and the reference marginal while Euclidean goal comparisons can change. This separates *predictive equivalence* from *scoring geometry*. One shared inference-time adapter maps frozen predictions to scoring coordinates (Section 3). Gaussian marginals constrain single states (Balestriero & LeCun, 2025); identifiability uses temporal or controlled joint information to recover coordinates up to an orthogonal map under stated assumptions (Klindt et al., 2026; Zhang et al., 2026).

Our analysis separates distribution, movement, and geometry. In the controlled Gaussian audit, Gaussian-preserving remapping leaves diagnostics within resampling variability, while matching displacement leaves local distortion uncontrolled (Section 5.4). The planning equations insert a position-dependent metric between the frozen rollout Jacobian and its transpose, preserving the optimization–controllability correspondence. Training-time straightening shapes the representation; inference-time remapping changes the scorer's metric. For CEM, crossing the elite boundary changes the fitting sample; reordering elite members leaves its moments unchanged. A finite-round bound tracks subsequent mean and variance changes (Proposition 1 and Theorem 6).

We ask: *which shared coordinate changes alter planning selection and updates while preserving the latent target law under a displacement budget?* Probability conservation and nonzero local strain

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.

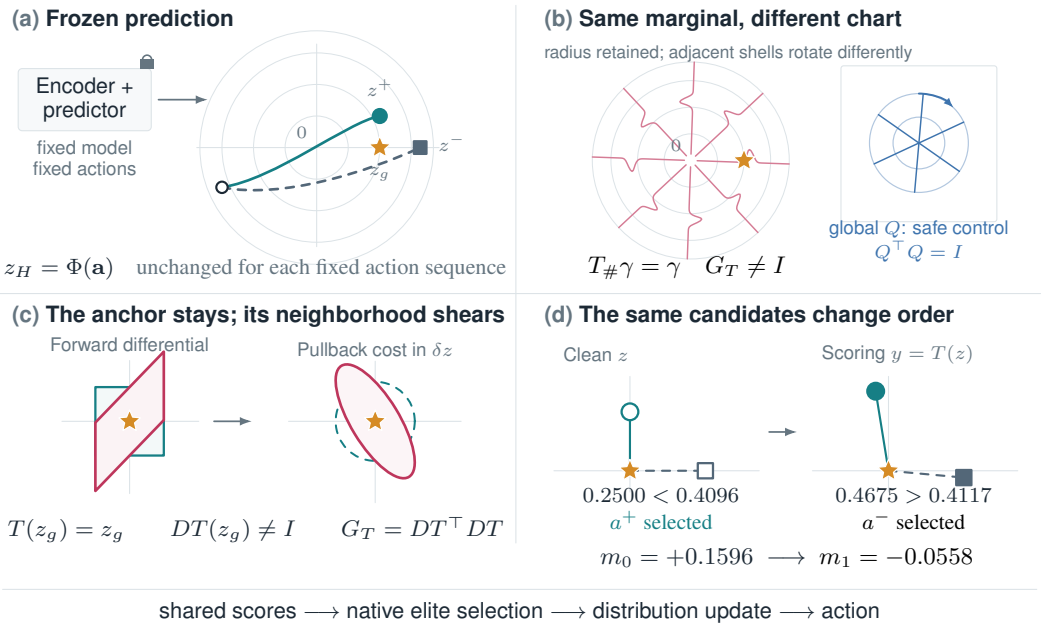

Figure 1: **Same prediction, same marginal, different ruler.** Fixed-action rollouts give unchanged endpoints (a). Radial mapping preserves the reference Gaussian; global rotation is the inert control (b). A fixed goal acquires sheared local geometry (c), reversing the same candidates' ranking (d). The footer follows native planning. Paths are schematic; geometry uses the exact witness in (89).

motivate a closed-form radial rotation. Native elite boundaries supply calibration pairs; a shared kernel extends their phases, and feasible trust-region fitting optimizes exact candidate scores.

Our contributions are: (1) We formulate inference-time planning interventions inside a measure-preserving coordinate space. (2) We connect predictive coordinate freedom, local action geometry, and exact candidate costs to a margin-resolved bound on successive CEM updates. (3) We construct a shared scoring map whose native-boundary calibration feeds kernel extension and budgeted profile fitting. (4) We validate the construction through exact ranking, displacement, and invariance calculations and specify paired planning traces for task-level evaluation.

2 Related Work

World models and planning geometry. DINO-WM plans in pretrained features (Zhou et al., 2025); PLDM learns predictive representations from reward-free offline data (Sobal et al., 2025). LeWM learns an encoder and dynamics with Gaussian regularization and plans with CEM elite moments (Maes et al., 2026). Temporal straightening shapes trajectories (Wang et al., 2026); LeJEPA targets isotropic Gaussian representations (Balestriero & LeCun, 2025). Temporal pairing and controlled dynamics support coordinate identification under their respective assumptions (Klindt et al., 2026; Zhang et al., 2026).

Coordinate freedom. Measure-preserving latent changes underlie non-identifiability (Locatello et al., 2019), Gaussian-preserving spirals (Klindt et al., 2026), and lower-cost normalizing-flow rearrangements (Morel et al., 2023). Gauge-dependent metrics, pullback geometry, kernel interpolation, and the continuity equation supply related tools (Cain, 2026; Khan & Storkey, 2023; Schölkopf et al., 2001; Villani, 2009). We connect this freedom to exact scoring after frozen prediction.

Adversarial interventions. Latent search generates natural adversarial inputs (Zhao et al., 2018); feature objectives guide transferable perturbations (Inkawhich et al., 2019). TRAP targets imagined-trajectory rankings with deployment-time visual triggers and a fixed model/training pipeline (Duan et al., 2026); WMAttack searches world-model attack configurations (Guo et al., 2026).

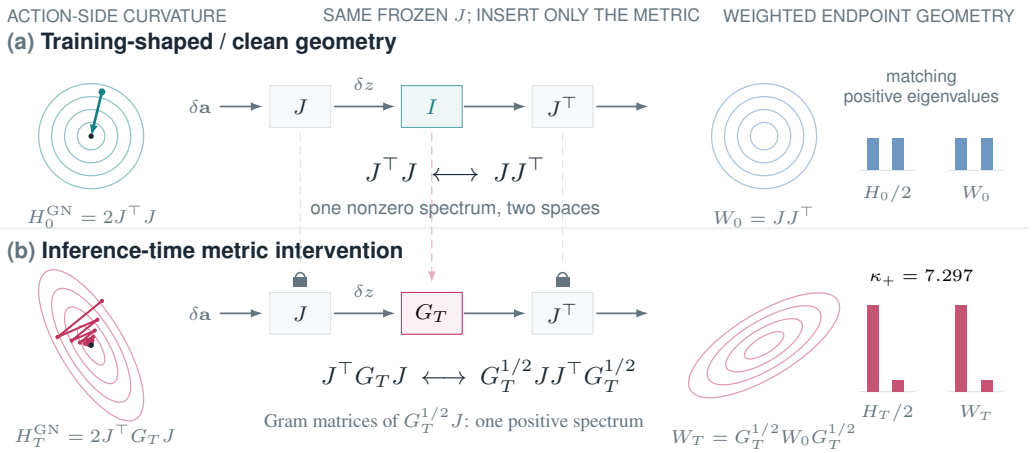

Figure 2: **Training-time geometry and an inference-time metric share the planning structure.** With J fixed and $JJ^\top = I_2$, mapping changes active conditioning from 1 to 7.29695. Contours, descent iterates, positive spectra, and endpoint ellipses depict this subspace's weighted Gram geometry. Optimizer and exact ranking effects are evaluated separately for the same map.

3 PRELIMINARY

A world model encodes $z_t = h(o_t) \in \mathbb{R}^d$ and predicts $\widehat{z}_{t+1} = F(z_t, a_t)$. For a fixed initial state, write $\Phi(\mathbf{a}) = z_H^{\mathbf{a}}$ and z_g for the predicted endpoint and encoded goal. LeWM regularizes latents toward $\gamma = \mathcal{N}(0, I_d)$ (Maes et al., 2026).

Coordinate freedom. At zero prediction risk, a conjugated pair representable by the model class,

$$\widetilde{h} = T \circ h, \qquad \widetilde{F}(y, a) = T\big(F(T^{-1}y, a)\big)\,, \tag{1}$$

retains exact prediction and, if $h(o_t) \sim \gamma$ and $T_{\#}\gamma = \gamma$, its marginal, although distances can change (Locatello et al., 2019; Klindt et al., 2026). The training-side identifiability arguments are in Section C. The attacker reads calibration endpoints and goals and fits one shared differentiable bijection *after* prediction. The encoder, predictor, observations, goal input, and scalar scorer stay fixed; the map is reused across test goals, candidates, and calls. Fixed-action rollouts are unchanged; later CEM proposals and MPC actions may change.

Planning geometry. For $J = D_{\mathbf{a}}\Phi$, clean and mapped costs give

$$C(\mathbf{a}) = \|\Phi(\mathbf{a}) - z_g\|^2\,, \qquad H_0^{\mathrm{GN}} = 2J^\top J, \tag{2}$$

$$C_T(\mathbf{a}) = \|T(\Phi(\mathbf{a})) - T(z_g)\|^2\,, \qquad G_T = DT^\top DT, \tag{3}$$

$$H_T^{\mathrm{GN}} = 2J^\top G_T(\Phi(\mathbf{a}))J. \tag{4}$$

$J^\top J$ shares its positive spectrum with $JJ^\top$, the finite-horizon controllability Gramian for linear dynamics (Wang et al., 2026); the weighted correspondence persists (Figure 2, Equation (32), and Section B.1). Controlled Gaussian diagnostic means 0.40853 and 0.40910 lie within clean resampling variability, yet matched-displacement distortion varies by orders of magnitude (Sections 5.4 and F and Theorem 4).

CEM inclusion. CEM fits its sampling distribution to the k lowest-cost candidates; reordering these members preserves their moments (Section E.3).

Proposition 1 (CEM boundary selection)**.** *For N fixed candidates, $1 \le k < N$, and distinct clean and mapped costs, let E, E_T be their lowest-cost k index sets. With parentheses denoting clean rank,*

$$C_T(\mathbf{a}_{(k+1)}) < C_T(\mathbf{a}_{(k)}) \quad \Longrightarrow \quad E_T \neq E. \tag{5}$$

Boundary gaps, score shifts, and rollout sensitivity bound the ensuing single- and finite-round moment updates within one planning call (Section B.5 and Theorem 6).

Algorithm 1 Shared Gaussian-preserving scoring, SCORE

Require: $\{z_i\}_{i=1}^N,\ z_g,\ \alpha,\ A$
1: $\widetilde{z}_g \leftarrow e^{\alpha(\|z_g\|^2)A} z_g$
2: **for** $i = 1, \ldots, N$ **do**
3: $\quad \widetilde{z}_i \leftarrow e^{\alpha(\|z_i\|^2)A} z_i$
4: $\quad C_i^T \leftarrow \|\widetilde{z}_i - \widetilde{z}_g\|^2$
5: **end for**
6: **return** $(C_i^T)_{i=1}^N$

4 METHOD

4.1 PROBLEM FORMULATION AND OVERVIEW

LATENTGAUGE fits one shared map from frozen predicted endpoints and goals to the planner's scorer (Figure 1). It preserves the reference law under an average displacement budget:

$$T_{\#}\gamma = \gamma, \qquad D_\gamma(T) := \mathbb{E}_{Z\sim\gamma} \|T(Z) - Z\|^2 \le \epsilon^2. \tag{6}$$

The calibration bank $\mathcal{B}$ records clean candidate pools: their action sequences, scores, endpoints, goals, and native elite counts. CEM supplies adjacent pairs across its elite boundary; GD-MPC retains clean top-one comparisons. For each pair, define $m_i = C(\mathbf{a}_i^-) - C(\mathbf{a}_i^+) > 0$ and $m_i^T = C_T(\mathbf{a}_i^-) - C_T(\mathbf{a}_i^+)$. Fitting minimizes their mapped softplus margins; a negative CEM boundary margin changes the elite set (Proposition 1). Algorithm 1 scores with the shared map; Algorithm 2 calibrates its plane and kernel, fits the profile, and removes a score-invariant common rotation. The fitted map stays fixed at inference, while native proposals and actions respond to its scores.

4.2 SHARED MEASURE-PRESERVING SCORING

Conservation determines the construction. Gaussian probability conservation requires $\operatorname{div} v - z^\top v = 0$; changing geometry additionally requires nonrigid motion. A radius-dependent rotation satisfies both and integrates exactly because its radius is constant:

$$v(z) = \alpha_c(\|z\|^2)Az, \qquad T_c(z) = e^{\alpha_c(\|z\|^2)A} z, \qquad A^\top = -A. \tag{7}$$

Skew symmetry gives $z^\top v = \operatorname{div} v = 0$. Shell-dependent angles create shear while conserving probability (Figure 3 and Section A.1). Such rotations appear in coordinate-identification examples (Klindt et al., 2026); here we fit the profile to exact planner costs.

Proposition 2 (Shared measure-preserving map)**.** *For $\alpha_c \in C^2$ and skew-symmetric A,* (7) *is a radius- and Gaussian-preserving bijection with inverse $T_c^{-1}(y) = e^{-\alpha_c(\|y\|^2)A} y$. Its spatially varying metric depends on α_c'.*

Calibration determines one shared profile. Let $s_1, \ldots, s_m$ be the distinct positive squared radii in the calibration triples. A smoothness kernel derived from the Gaussian reference operator extends their phases:

$$\alpha_c(s) = \sum_{j=1}^m c_j K(s, s_j). \tag{8}$$

For prescribed site values, this is the minimum-norm interpolant in the specified kernel space (Schölkopf et al., 2001); the attack fits those values through c (Section D.1). The leading two-dimensional block of the calibration loss's skew first variation selects a unit rank-two A once (Section D.2). Fixed vectorization and its inverse handle spatial features (Section D.6). Algorithm 1 applies this same profile and generator to every endpoint and goal; the Jacobian and spectrum are derived in Section A.2.

The site values determine how observed shells rotate, and the kernel fixes their extension to unobserved radii. This separates the finite ranking problem from the smooth coordinate map used at deployment: every scored point uses the same function, including points outside the calibration bank. Evaluation uses one radial function and two plane projections even for a high-dimensional feature tensor.

Algorithm 2 Planner-aligned kernel-profile fitting

Require: $\mathcal{B}$, ϵ; fixed tolerances in Appendix D.4
$\{(z_i^+, z_i^-, z_{g,i})\}_{i=1}^n \leftarrow \textsc{Pairs}(\mathcal{B})$
$(A, \{s_j\}, \mathbf{K}) \leftarrow \textsc{Calibrate}(\{(z_i^+, z_i^-, z_{g,i})\})$
$c \leftarrow 0,\ B \leftarrow I,\ \Delta \leftarrow 1$
while call allowance remains **do**
  $(C_i^+, C_i^-) \leftarrow \textsc{Score}(\{z_i^+, z_i^-\}, z_{g,i}, \alpha_c, A), \quad i = 1, \dots, n$
  $(g, a, h) \leftarrow (\nabla_y \mathcal{L}, \nabla_y h, h), \quad y = \mathbf{K}^{1/2} c$
  if the KKT residual meets tolerance **then break**
  end if
  $p \leftarrow \arg\min_{a^\top p + h \le 0,\ \|p\| \le \Delta} \ g^\top p + \frac{1}{2} p^\top B p$
  $(y', \mathrm{ok}) \leftarrow \textsc{Restore}(y + p, h, \Delta)$
  $(y, B, \Delta) \leftarrow \textsc{Accept}(y, y', B, \Delta, \mathrm{ok}, \mathcal{L})$
  $c \leftarrow \mathbf{K}^{-1/2} y$
end while
$b \leftarrow -\arg \mathbb{E}_{S \sim \chi^2_{d+2}} e^{\mathrm{i}\alpha_c(S)}$
return $T(z) = e^{[\alpha_c(\|z\|^2) + b]A} z$

4.3 Planner-Aligned Kernel Fitting

Exact scores define the fitting problem. Calibration triples stay fixed during fitting and jointly determine the plane and kernel centers (Section G.4). For $z_H = \Phi(\mathbf{a})$ and $\Delta\alpha = \alpha_c(\|z_g\|^2) - \alpha_c(\|z_H\|^2)$, orthogonality gives

$$C_T(\mathbf{a}) - C(\mathbf{a}) = 2 z_H^\top [I - e^{\Delta\alpha A}] z_g. \tag{9}$$

Equal endpoint and goal radii give zero shift; a common angle b leaves every score unchanged. Writing $D_{\min}(c) = \min_b D_\gamma(e^{bA} \circ T_c)$, we solve

$$\min_c \ \mathcal{L}(c) := \frac{1}{n} \sum_i \mathrm{softplus}(m_i^{T_c}) \quad \text{s.t.} \quad D_{\min}(c) \le \epsilon^2. \tag{10}$$

Feasible fitting and exact rotation removal. Algorithm 2 starts at the identity and applies feasible trust-region SQP (Wright & Tenny, 2004). With $h = D_{\min} - \epsilon^2$, $\mathbf{K} = (K(s_j, s_k))_{j,k}$, and $y = \mathbf{K}^{1/2} c$, exact loss and budget derivatives define each local step. Normal correction enforces feasibility; true loss reduction controls acceptance and the trust radius (Section D.4). The kernel norm scales steps rather than adding a constraint. Finally, $M_c = \mathbb{E}_{S \sim \chi^2_{d+2}} e^{\mathrm{i}\alpha_c(S)}$ gives $b^* = -\arg M_c$ and $D_{\min} = 4(1 - |M_c|)$. Returning profile $\alpha_c + b^*$ minimizes displacement within its score-equivalent rotation class (Section D.3).

5 Experiments

Tables 1–2 compare task performance, shared-map controls, and calibration rules. Published protocols define the victim settings; controlled calculations separately examine the geometric mechanism.

5.1 Experimental Setup

Benchmarks and victim models. We retain three protocol families: DINO-WM's navigation and manipulation tasks, Temporal Straightening's four open-loop/MPC tasks, and LeWM's Two-Room, Reacher, PushT, and OGBench-Cube tasks (Zhou et al., 2025; Wang et al., 2026; Maes et al., 2026; Park et al., 2025). Each keeps its source input, feature layout, objective, and planner. The matched checkpoint settings are LeWM/PushT/CEM and spatial DINO+projector+straightening/PushT/GD-MPC. Same-capacity DINO projectors with and without straightening connect training-time geometric shaping to the inference-time intervention. Complete source tables, including PLDM and deformable-object tasks, remain in Section H. Gaussian-regularized LeWM is the reference-aligned primary setting; DINO-WM and straightened DINO retain their role as extended representation settings.

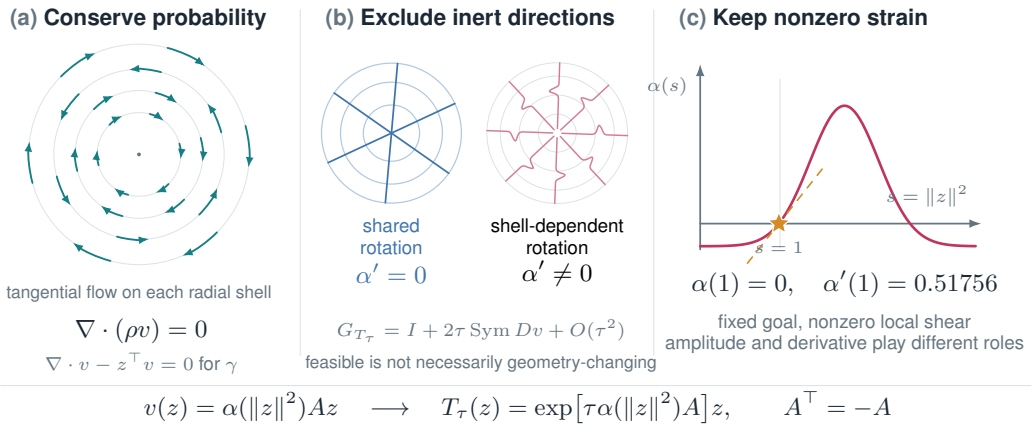

Figure 3: **Conservation defines admissible directions; strain selects effective ones.** (a) Radial flow is tangent to Gaussian shells. (b) Constant profiles preserve the metric; shell-dependent angles create strain. (c) $\alpha(1) = 0$ and $\alpha'(1) \neq 0$ fix the goal while shearing its neighborhood. Radius conservation gives (7).

Metrics and controls. We record residual success (SR), paired success reduction, and clean-success-to-failure transitions. FlipRate measures clean-positive top-one replay reversals; Boundary FlipRate measures native CEM inclusion-boundary reversals. Elite replacement is the changed fraction of k selected candidates. Mean shift and first-action difference use Euclidean norms; clean-coordinate regret evaluates the new selection under the original score. Statistical visibility is held-out two-sample detection at nominal 1% FPR (Gretton et al., 2012). Reference displacement and learned-feature movement/detection are assessed separately.

Implementation and evidence. The checkpoint protocol uses 50 held-out goal pairs, three attack-calibration seeds, and available source model seeds. Top-one replay caches 1,024 legal action sequences per goal; CEM uses native iteration pools and their inclusion-boundary pairs (Section G.4). The calibrated plane, kernel extension, and feasible SQP in Section D use $\epsilon = 0.07$ and 3,000 forward-equivalent calls. One map transfers from calibration to held-out goals with fixed numerical defaults across tasks. Table 1 organizes clean and mapped rows by victim protocol. Table 2 compares shared-map attacks and calibration rules under matched protocols. Controlled calculations separately isolate the geometric mechanisms (Sections G and I).

5.2 Main Comparison

Tasks and representations. Table 1 presents 20 clean settings and six mapped rows in three protocol blocks. DINO-WM's model and encoder tables are joined through their common evaluation protocol. The straightening block aligns four MPC tasks, while the LeWM block aligns victims with complete task coverage. Row averages stay within each four-task protocol.

Shared-interface controls. Table 2 compares one targeted method with four controls. The common orthogonal map tests exact geometric invariance. The random-profile control samples coefficients on the same calibrated plane; random-direction search uses the same ranking target, kernel coordinates, nonlinear-budget correction, and objective-call allowance as fitted LatentGauge. All methods share candidate and goal representations; proposal and rejected-trial rules are in Section D.5.

5.3 Planning Mechanism Analysis

Calibration at the native boundary. Table 2B compares top-one and elite-boundary calibration through the same construction and solver. Boundary calibration raises elite replacement from 9% to 22% and reduces residual SR from 71.04% to 62.40% under the same allowance. The selected members and their action differences determine the same-pool update (Equation (42)); Theorem 6 extends this relation through coupled candidate resampling. Each trace records clean boundary gaps,

Table 1: **Task and representation comparison.** Block A uses fractions (0–1); Blocks B–C use percentages. Avg. is the mean of the four displayed task scores.

A. DINO-WM protocol: SR (0–1)					
Model / representation	Maze	Wall	Reach	PushT	Avg.
IRIS	0.74	0.04	0.18	0.32	0.32
DreamerV3 (offline)	1	1	0.64	0.30	0.735
TD-MPC2 (offline)	0	0	0	0	0
R3M	0.94	0.34	0.40	0.42	0.525
ResNet	0.98	0.12	0.06	0.20	0.34
DINO CLS	0.96	0.58	0.60	0.44	0.645
DINO-WM	0.98	0.96	0.92	0.90	0.94
DINO-WM $+ T_{\mathrm{LG}}$	0.735	0.624	0.644	0.54	0.6358
B. Temporal Straightening: SR (%), GD/MPC					
Model / representation	Wall	UMaze	Medium	PushT	Avg.
Global DINO CLS	66.67	82.67	67.50	28.10	61.24
Global DINO + projector	76.06	79.33	46.03	11.33	53.19
Global DINO + projector + TS	56.67	96.76	78.30	14.67	61.6
Global ResNet	10.65	96.24	91.33	29.33	56.89
Global ResNet + TS	100	86.67	98.81	48.67	83.54
Spatial DINO (raw)	76.67	80.67	76.67	66.05	75.02
Spatial DINO + projector	90.67	81.33	96.67	78.67	86.84
Spatial DINO + projector + TS	100	100	98.67	85.33	96
Spatial ResNet	6.67	66.15	57.33	70.67	50.21
Spatial ResNet + TS	100	98.67	99.33	91.33	97.33
DINO + projector $+ T_{\mathrm{LG}}$	56.22	52.86	61.87	47.20	54.54
DINO + projector + TS $+ T_{\mathrm{LG}}$	72.48	74.64	69.07	58.02	68.55
C. LeWorldModel protocol: SR (%)					
Victim / source-native input	Two-Room	Reacher	PushT	Cube	Avg.
DINO-WM (source input)	100	79	74	86	84.75
PLDM	97	78	78	65	79.5
LeWM	87	86	96	74	85.75
DINO-WM $+ T_{\mathrm{LG}}$	68.99	56.09	45.88	56.76	56.93
PLDM $+ T_{\mathrm{LG}}$	71.78	60.06	53.04	45.50	57.6
LeWM $+ T_{\mathrm{LG}}$	60.90	62.78	62.40	44.40	57.62

TS: temporal straightening; T_{LG}: the shared interface transformation. Each block retains its own tasks, representations, and scoring convention.

exact score changes, and mean/variance differences, yielding a round-by-round propagation envelope. Native first-action and task metrics follow the source execution rule (Section G.4).

One map, fixed futures, exact reversal. Figure 4 uses the frozen affine model and canonical profile in Section F: the same predicted endpoints reverse ordering within the integrated budget. Exact nonlinear and local quadratic contours distinguish finite-cost and local-geometric effects.

Local geometry and finite candidate costs. The fixed goal's metric and canonical active action-side quadratic share their conditioning. Direct score calculations separately verify (9); Section 5.4 isolates the associated selection and invariance mechanisms.

Table 2: **One shared interface, one attack family.** Shared-map controls (A) and calibration rules (B) under matched victim protocols.

A. Shared-map comparison

	Exact γ	LeWM / PushT (CEM)			DINO+TS / PushT (GD/MPC)		
Shared map	law	SR ↓	Flip ↑	TPR ↓	SR ↓	Flip ↑	TPR ↓
Identity	Yes	96	0	1	85.33	0	1
Orthogonal Q	Yes	96	0	8	85.33	0	10
Random kernel profile	Yes	88.32	14	12	79.36	12	14
Random-direction search	Yes	77.76	32	13	70.82	29	15
LatentGauge (pairwise)	Yes	62.40	56	15	58.02	50	18

SR: residual task success; Flip: fixed-bank top-one replay reversal; TPR: learned-feature two-sample detection at nominal 1% FPR. All three are percentages. All kernel-profile rows share the calibration-defined plane. Panel B measures the native boundary separately. See Appendix I for cell provenance.

B. Calibration-rule comparison: LeWM / PushT / native CEM

Calibration	Boundary flip (%) ↑	Elite repl. (%) ↑	Mean shift ↑	Action diff. ↑	SR (%) ↓
Top-one	28	9	0.041	0.028	71.04
Elite boundary	56	22	0.093	0.063	62.40

Both calibration rules use the same pools, native elite count, $\epsilon = 0.07$, and 3,000-call allowance. Mean/action norms use the source action units; moments are paired within one candidate pool. Each calibration rule determines its own triples, plane, and kernel.

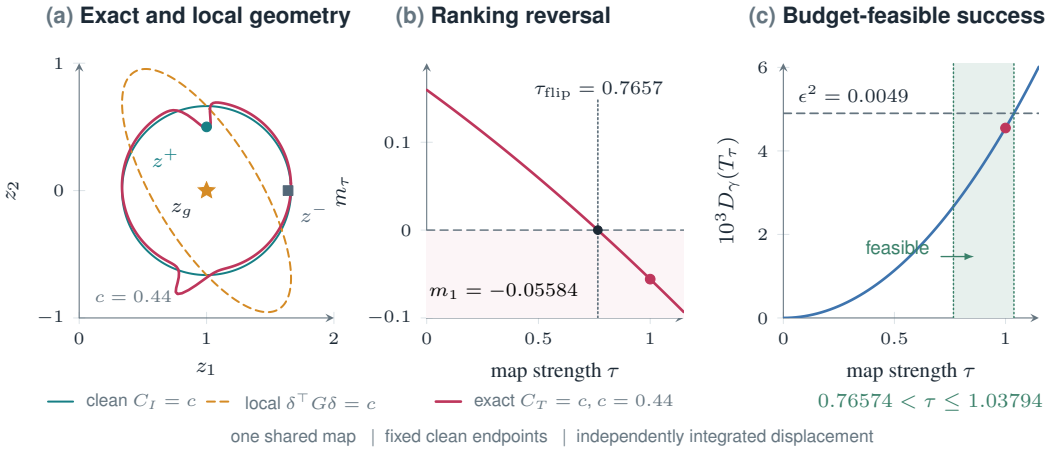

Figure 4: **One exact witness links geometry, ranking, and budget.** (a) Clean, local-quadratic, and exact nonlinear contours are distinguished at level 0.44. (b) The fixed-candidate margin crosses zero at $\tau = 0.765739$. (c) Independent integration gives $D_\gamma(T_1) = 0.004548694 < 0.0049$. Shading marks the intersection of strict reversal and budget feasibility within $\tau \in [0, 1.15]$.

5.4 Controlled Mechanism Checks

Selection depends on the inclusion boundary. We keep six two-component action sequences fixed, assign clean scores $0.1, 0.2, \ldots, 0.6$, and fit the mean and covariance of the three lowest-cost candidates. Swapping the first two scores reverses one ranking within the elite set; swapping the third and fourth crosses its boundary. Table 3 shows that only the boundary swap changes the fitting sample and its moments. These assigned score permutations isolate the selection rule in Proposition 1; the candidate coordinates and a 1,000-pool update-identity check are given in Section E.3.

Propagation through successive CEM rounds. On the frozen affine endpoint model, we compare identity, a common 0.7-radian rotation, and the canonical radial map. Each uses 64 seeded candidate

Table 3: Fixed-pool score-permutation diagnostic. Each condition swaps two adjacent score values in the same six-candidate pool, with $k = 3$ and direct moment fitting.

Score swap	Replaced (%)	Mean shift	Covariance shift
Within elite	0.00	0.000000	0.000000
Across boundary	33.33	0.360555	0.405161

Table 4: Exact controls on the same two endpoints. The normalization rows use cosine cost, for which this pair already prefers a^-; other rows use squared Euclidean cost.

Scoring interface	Clean margin	Mapped margin	Ordering
Identity	0.159600	0.159600	preserved
Shared rotation, 0.7 radians	0.159600	0.159600	preserved
Radial map T_1	0.159600	-0.055839	reversed
Normalize before T_1	-0.105573	-0.105573	preserved
Normalize after T_1	-0.105573	-0.202213	preserved

streams and eight diagonal-Gaussian CEM updates, with 64 candidates, eight elites, and four action dimensions. Paired runs share initial moments and base noises. All candidate, replacement, and moment envelopes in Theorem 6 hold across the 1,536 checked rounds; the maximum same-pool update-identity residual is 2.39×10^{-16}. The coupled clean and mapped runs select different elite index sets in 476 of 512 rounds under the radial map, and in none under identity or common rotation. These fixed-model checks verify the update identities and inequality coverage; per-round quantities and the protocol are in Section B.6.

Normalization order selects the scoring geometry. Table 4 evaluates the same two endpoints under exact interface controls. A common rotation preserves Euclidean ordering; the radial map reverses it. Normalizing before mapping places every point on one shell, where the shared rotation preserves all distances. Normalizing afterward retains the original shell-dependent angles and changes the cosine margin. For this pair both cosine margins favor a^-, so that change preserves the cosine ordering. The contrast isolates the role of relative shell angles in (9); the full protocol and additional paired checks are in Sections E and F.

Independent marginal and displacement checks. Across 20 independent Gaussian audit batches ($N = 1024$, $d = 64$), the mean paired statistic change is 0.000575, with 95% Student interval $[-0.000466, 0.001616]$. In a separate frequency sweep calibrated to displacement 0.0625, median $\|G_T - I\|_F$ grows from 0.4073 to 5.573×10^4, while independent probe displacement stays within $[0.06207, 0.06402]$. The audit samples test marginal agreement; the sweep measures geometry at matched calibration movement. Seeds, the full frequency grid, and the analytical sweep family are given in Section F.

6 Conclusion

LatentGauge identifies a planning-geometry degree of freedom between frozen prediction and downstream scoring. The planning equations locate the metric; probability conservation determines the admissible space; a closed-form radial map makes exact ranking optimization possible under a displacement budget. The shared-interface comparison gives residual success of 62.40% on LeWM and 58.02% on DINO+TS, below equal-budget random-direction search by 15.36 and 12.80 percentage points. Controlled calculations connect one shared transformation to local action geometry and finite candidate comparisons. The native CEM boundary connects exact score changes to elite selection, while a margin-resolved propagation bound carries mean and variance differences through later sampling rounds. Prediction, marginal validity, and the geometry used for decisions are distinct interface properties. The diagnostic follows the planner's computation: endpoint–goal relations determine scores, boundary comparisons determine the fitting sample, and elite actions determine later proposals.

AI USE STATEMENT

Generative AI tools were used only to polish the presentation of this paper, namely to edit wording, grammar, and LaTeX layout and to refine figures. They were not used for research ideation, theoretical derivations, experiment design, implementation, data analysis, or the generation of results or scientific claims. All AI-assisted edits were reviewed by the authors, who take full responsibility for the content of this paper.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.1090/S0002-9947-1950-0051437-7.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. doi: 10.48550/arXiv.2511.08544.

Jericho Cain. Gauge freedom and metric dependence in neural representation spaces. *arXiv preprint arXiv:2603.06774*, 2026. doi: 10.48550/arXiv.2603.06774.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005. doi: 10.1007/s10479-005-5724-z.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017.

DLMF. NIST Digital Library of Mathematical Functions. Release 1.2.8 of 2026-09-15, 2026. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. Chapter 18, Sections 18.3 and 18.8.

Siyuan Duan, Ke Zhang, and Xizhao Luo. TRAP: Tail-aware ranking attack for world-model planning. *arXiv preprint arXiv:2605.01950*, 2026. doi: 10.48550/arXiv.2605.01950.

T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. doi: 10.1093/biomet/70.3.723.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Zhixiang Guo, Siyuan Liang, Shi Fu, Cheng Guo, Andras Balogh, Mark Jelasity, and Dacheng Tao. WMAttack: Automated attack search for adversarial evaluation of world-model agents. *arXiv preprint arXiv:2605.23220*, 2026. doi: 10.48550/arXiv.2605.23220.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nick Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Nathan Inkawhich, Wei Wen, Hai (Helen) Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00723.

Asif Khan and Amos Storkey. Adversarial robustness of VAEs through the lens of local geometry. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8954–8967. PMLR, 2023.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, 2026. doi: 10.48550/arXiv.2605.26379.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. doi: 10.48550/arXiv.2603.19312.

Guillaume Morel, Lucas Drumetz, Simon Benaïchouche, Nicolas Courty, and François Rousseau. Turning normalizing flows into Monge maps with geodesic Gaussian preserving flows. *Transactions on Machine Learning Research*, 2023.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006. doi: 10.1007/978-0-387-40065-5.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025.

Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, Jan Achterhold, Joerg Stueckler, Michal Rolinek, and Georg Martius. Sample-efficient cross-entropy method for real-time planning. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1049–1065. PMLR, 2021.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Computational Learning Theory*, volume 2111 of *Lecture Notes in Computer Science*, pp. 416–426. Springer, 2001. doi: 10.1007/3-540-44581-1_27.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 43905–43941. Curran Associates, Inc., 2025. doi: 10.52202/085713-1465.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Ying Wang, Oumayma Bounou, Gaoyue Zhou, Randall Balestriero, Tim G. J. Rudner, Yann LeCun, and Mengye Ren. Temporal straightening for latent planning. In *International Conference on Machine Learning*, 2026.

Grady Williams, Andrew Aldrich, and Evangelos A. Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 40(2): 344–357, 2017. doi: 10.2514/1.G001921.

Stephen J. Wright and Matthew J. Tenny. A feasible trust-region sequential quadratic programming algorithm. *SIAM Journal on Optimization*, 14(4):1074–1105, 2004. doi: 10.1137/S1052623402413227.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026. doi: 10.48550/arXiv.2607.22430.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025.

SUMMARY OF APPENDIX

Section	Supplementary material
A	Probability conservation, exact geometry, and ranking guarantees
B	Action geometry and finite-round CEM propagation
C	Prediction, temporal pairing, and coordinate selection
D	Kernel extension, plane selection, and constrained fitting
E	Supplementary mechanism protocols
F	Numerical inputs, seeds, and implementation details
G	Checkpoint settings, scoring interfaces, and metric definitions
H	Results reported in prior work
I	Numerical provenance
J	Code and numerical artifacts

A MEASURE-PRESERVING CONSTRUCTION AND EXACT SCORING

A.1 FROM DENSITY CONSERVATION TO SYMMETRIC STRAIN

Invertible density models such as Real NVP evaluate transported densities through the change-of-variables formula (Dinh et al., 2017). Here the prescribed density is stationary, yielding the conservation condition below. Let $\rho_\tau = (T_\tau)_\# \rho$ denote the transported density. The continuity equation gives $\partial_\tau \rho_\tau = -\operatorname{div}(\rho_\tau v)$. If $\operatorname{div}(\rho v) = 0$ and the flow is regular and complete, the stationary solution is $\rho_\tau = \rho$. Since

$$\operatorname{div}(\rho v) = \rho\{\operatorname{div} v + (\nabla \log \rho)^\top v\}, \qquad \nabla \log \rho(z) = -z, \tag{11}$$

the admissible directions are

$$\operatorname{Stab}(\gamma) = \{T : T_\# \gamma = \gamma\}, \qquad \mathcal{T}_\gamma = \{v : \operatorname{div}(\rho v) = 0\} = \{v : \operatorname{div} v - z^\top v = 0\}. \tag{12}$$

Here the maps are bijective and the fields are sufficiently regular and complete. The density evolves according to

$$\partial_\tau \rho_\tau + \operatorname{div}(\rho_\tau v) = 0. \tag{13}$$

This converts preservation of a probability law into a local structural condition on its flow.

The first-order Fréchet expansion gives $DT_\tau = I + \tau Dv + O(\tau^2)$. Multiplying by its transpose gives

$$T_\tau(z) = z + \tau v(z) + O(\tau^2), \qquad G_{T_\tau}(z) = I + 2\tau \operatorname{Sym} Dv(z) + O(\tau^2), \tag{14}$$

where $\operatorname{Sym} Dv = (Dv + Dv^\top)/2$. Conservation controls probability flux, whereas symmetric strain controls first-order geometry. In particular,

$$\|G_{T_\tau} - I\|_F^2 = 4\tau^2 \|\operatorname{Sym} Dv\|_F^2 + O(\tau^3). \tag{15}$$

An integrable remainder bound permits the expansion under expectation. A useful constructive family outside the radial case is obtained from a smooth antisymmetric matrix field S: define $(\rho v)_i = \sum_j \partial_j S_{ij}$. Mixed derivatives are symmetric in (i, j) whereas S is antisymmetric, so $\operatorname{div}(\rho v) = 0$. Choosing a complete and computationally tractable field gives a further constructive family.

Theorem 3 (Marginal blindness)**.** *Fix T with $T_\# \gamma = \gamma$ independently of an audit sample $Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} \gamma$. For a statistic using only that sample and independent randomness, its distribution is identical under (Z_i) and $(T(Z_i))$. Upstream prediction residuals are unchanged pointwise.*

Equality of the joint iid laws proves the statistical statement. Realized values on a paired batch can differ while retaining the same law. Independent audit samples and the Gaussian reference are the assumptions. Clean–mapped correspondences and temporal pairs supply additional information beyond a marginal sample.

A.2 Proof of the radial construction

For $v(z) = \alpha(s)Az$, $s = z^\top z$, skew symmetry yields $z^\top Az = 0$ and $\operatorname{tr} A = 0$. Consequently

$$\operatorname{div} v = 2\alpha'(s)z^\top Az + \alpha(s)\operatorname{tr} A = 0, \quad \frac{d}{d\tau}\|z(\tau)\|^2 = 0. \tag{16}$$

The radius is constant, so the flow has the closed form $T_\tau(z) = e^{\tau\alpha(s)A}z$ and is complete. Its inverse is $e^{-\alpha(s)A}z$. Writing $R(s) = e^{\alpha(s)A}$ locally, A commutes with its exponential, and

$$dT = R(s)[I + 2\alpha'(s)Azz^\top]dz. \tag{17}$$

The matrix determinant lemma gives $\det(I + 2\alpha'(s)Azz^\top) = 1 + 2\alpha'(s)z^\top Az = 1$, and $\det R = e^{\alpha \operatorname{tr} A} = 1$. Radius preservation gives $\rho(Tz) = \rho(z)$; together these facts prove $T_\#\gamma = \gamma$ by change of variables. Consequently, with $s = \|z\|^2$,

$$DT(z) = e^{\alpha(s)A}[I + 2\alpha'(s)Azz^\top], \tag{18}$$

$$G_T(z) = I + 2\alpha'(s)(Azz^\top - zz^\top A) + 4\alpha'(s)^2\|Az\|^2 zz^\top. \tag{19}$$

The factorized Jacobian separates a local orthogonal rotation from a shear. A constant profile has $G_T = I$; nonzero radial derivative supplies the nonorthogonal factor.

For $z \neq 0$ and $Az \neq 0$, let $q = 2\alpha'(s)\|z\|\|Az\|$ within this calculation and use the orthonormal vectors $e_1 = z/\|z\|$ and $e_2 = Az/\|Az\|$. On their span the nonorthogonal factor of DT is

$$B_z = \begin{bmatrix} 1 & 0 \\ q & 1 \end{bmatrix}, \qquad B_z^\top B_z = \begin{bmatrix} 1+q^2 & q \\ q & 1 \end{bmatrix}. \tag{20}$$

The characteristic polynomial is $\lambda^2 - (2+q^2)\lambda + 1$, giving

$$\lambda_\pm = \frac{2 + q^2 \pm |q|\sqrt{q^2+4}}{2}, \quad \lambda_+\lambda_- = 1, \quad \|G_T - I\|_F = |q|\sqrt{q^2+2}. \tag{21}$$

The orthogonal complement has identity metric. Its two-plane condition number is $\lambda_+^2 \sim q^4$. A smooth pullback of Euclidean space by a diffeomorphism is intrinsically flat. The quantities affected here are coordinate chord costs, coordinate trajectory curvature, and the optimization geometry read by a fixed Euclidean scorer.

A.3 Displacement–distortion separation

Theorem 4 (Displacement–distortion separation)**.** *For $d \geq 2$ and every $\epsilon > 0$, there exist smooth Gaussian-preserving radial maps T_k satisfying $D_\gamma(T_k) \leq \epsilon^2$ uniformly in k and $\mathbb{E}_\gamma\|G_{T_k} - I\|_F^2 \to \infty$.*

For $T_k(z) = e^{a\sin(k\|z\|^2)A}z$, integration along the rotation yields

$$\|T_k(z) - z\| \leq |a|\|Az\|, \quad D_\gamma(T_k) \leq a^2\mathbb{E}\|AZ\|^2 = a^2\|A\|_F^2. \tag{22}$$

Choose a nonzero rank-two A and $a \neq 0$ small enough for the specified budget. The exact metric formula gives $\|G - I\|_F^2 = q^2(q^2+2) \geq 2q^2$, where $q = 2ak\cos(kS)\sqrt{S}\|AZ\|$ and $S = \|Z\|^2$. Conditional spherical symmetry gives $\mathbb{E}[\|AZ\|^2 \mid S] = S\|A\|_F^2/d$. Hence

$$\mathbb{E}\|G - I\|_F^2 \geq \frac{8a^2k^2\|A\|_F^2}{d}\mathbb{E}[S^2\cos^2(kS)]. \tag{23}$$

The final expectation tends to $\frac{1}{2}\mathbb{E}[S^2] > 0$: write $\cos^2(kS) = (1 + \cos(2kS))/2$ and apply the Riemann–Lebesgue lemma to the integrable weighted chi-square density. This proves unbounded expected distortion under a uniform budget. It is a statement about a sequence whose derivatives grow; a fixed smoothness bound excludes the construction. The numerical experiment separately matches an empirical displacement estimate at each frequency and reports independent test displacement.

A.4 Exact costs, normalization order, and stability

Orthogonality of each radial rotation preserves the two individual squared norms. Its cross term is $\langle T(z_H), T(z_g)\rangle = z_H^\top e^{\Delta\alpha A} z_g$, proving (9). If the two radii agree, then $\Delta\alpha = 0$ and the pair's cost is invariant. Equal radii of two candidates alone are not enough unless their relation to the goal is also controlled.

Adding a constant b to the profile applies a common rotation after T: $T_{\alpha+b} = e^{bA} T_\alpha$. Candidate distances therefore stay unchanged, although displacement from the original coordinates can change. The fitted family imposes no goal-specific anchoring; $\alpha(1) = 0$ belongs only to the fixed witness.

For cosine scoring of nonzero vectors,

$$C_{\cos,T} - C_{\cos} = \frac{z_H^\top [I - e^{\Delta\alpha A}] z_g}{\|z_H\|\|z_g\|}. \tag{24}$$

A radius-preserving map can therefore change cosine similarity across shells. Let $N(z) = z/\|z\|$. Then $T(N(z)) = e^{\alpha(1)A} N(z)$ is one common rotation, so scoring after $T \circ N$ is invariant for this family. In contrast, $N \circ T$ retains the original shell-dependent angles. This normalization identity applies to the radial family in (7).

For a profile with $|\alpha'| \le L$ between the two squared radii,

$$|C_T - C| \le 2\|z_H\|\|z_g\| \min\{2, \|A\|_2 L\, |\|z_H\|^2 - \|z_g\|^2|\}. \tag{25}$$

The fundamental theorem of calculus gives $\|I - e^{uA}\|_2 \le |u|\|A\|_2$, while orthogonality also gives a bound of two. The scalar mean-value bound gives $|\Delta\alpha| \le L|\|z_g\|^2 - \|z_H\|^2|$. Substitution into (9) proves the claim.

Proposition 5 (A pairwise ranking certificate)**.** *Let $e(z) = T(z) - z$, $r_a = \|z_H^a - z_g\|$, and $b_a = 2r_a\|e(z_H^a) - e(z_g)\| + \|e(z_H^a) - e(z_g)\|^2$. Then $|C_T(a) - C(a)| \le b_a$. The condition $m_i > b_{a_i^-} + b_{a_i^+}$ certifies preservation of the strict clean ordering.*

To prove the certificate, let $d_a = z_H^a - z_g$ and $\Delta e_a = e(z_H^a) - e(z_g)$. Expanding the square gives

$$C_T(a) - C(a) = 2d_a^\top \Delta e_a + \|\Delta e_a\|^2. \tag{26}$$

Cauchy–Schwarz yields the bound in Proposition 5, and subtraction for the two candidates proves $m_i^T \ge m_i - b_{a_i^-} - b_{a_i^+}$. A positive lower bound certifies preservation. A nonpositive bound leaves the ranking to direct evaluation.

B Connections to Planning Geometry

This section expands the planning identities used in Section 3, following nonlinear least squares and the geometry-to-conditioning analysis of Nocedal & Wright (2006); Wang et al. (2026). The interface map and the frozen endpoint sensitivity remain distinct throughout.

B.1 The spectral bridge and the full Hessian

For $r_T(\mathbf{a}) = T(\Phi(\mathbf{a})) - T(z_g)$, the chain rule gives $D_{\mathbf{a}} r_T = DT(\Phi)J$. Therefore

$$\nabla C_T = 2J^\top DT(\Phi)^\top r_T, \tag{27}$$

$$\nabla^2 C_T = 2J^\top G_T(\Phi) J + 2\sum_{j=1}^{d} (r_T)_j \nabla^2 (r_T)_j. \tag{28}$$

The first term is the Gauss–Newton matrix. It equals the full Hessian when the residual vanishes or the residual map is affine. At other points, the second term supplies the residual-weighted correction. The SVD of $G_T^{1/2} J$ proves that its action-side and endpoint-side Gram matrices share their positive eigenvalues. Rank-deficient action directions remain null directions; an ordinary full-matrix condition number would then be infinite. We use the ratio of largest to smallest *positive* eigenvalue on the identifiable active subspace.

For $z_{t+1} = Mz_t + Ba_t$, the endpoint Jacobian is $J_H = [M^{H-1}B, \dots, B]$, and

$$J_H J_H^\top = \sum_{j=0}^{H-1} M^j B B^\top (M^\top)^j. \tag{29}$$

The displayed sum is the finite-horizon controllability Gramian. In a nonlinear world model, $JJ^\top$ instead describes local endpoint sensitivity along the chosen rollout. This correspondence is exact algebra; improved conditioning requires additional spectral assumptions. Similarly, if $v_t = z_{t+1} - z_t$, then

$$v_{t+1} - v_t = (M - I)v_t + B(a_{t+1} - a_t) \tag{30}$$

in the noiseless linear case. Small turning constrains $(M - I)v_t$ only after controlling the action-difference term, and only on visited directions. The visited subspace and action-difference term specify when turning constrains the dynamics.

B.2 Local pullback and the weighted spectral correspondence

For a C^2 map with bounded second derivative near z_g,

$$\|T(z_g + \delta) - T(z_g)\|^2 = \delta^\top G_T(z_g)\delta + O(\|\delta\|^3). \tag{31}$$

The leading quadratic determines local lengths. At any action sequence, the Gauss–Newton matrix evaluates the metric at its predicted endpoint, as in (4). With all metrics evaluated at the same point,

$$\mathrm{spec}_+(J^\top G_T J) = \mathrm{spec}_+(G_T^{1/2} J J^\top G_T^{1/2}). \tag{32}$$

Here spec_+ denotes the positive eigenvalues with multiplicity. The SVD proof uses the two Gram matrices of $G_T^{1/2}J$. This correspondence gives two views of one sensitivity operator: action-side optimization and weighted endpoint geometry.

B.3 From temporal straightness to controlled dynamics

For latent velocities $v_t = z_{t+1} - z_t$ with positive lengths $r_t = \|v_t\|$, write their cosine as $\mathcal{C}_t$. The exact second-difference identity is

$$\|v_{t+1} - v_t\|^2 = (r_{t+1} - r_t)^2 + 2r_t r_{t+1}(1 - \mathcal{C}_t). \tag{33}$$

Angular bending and speed variation are separate terms. For $z_{t+1} = Mz_t + Ba_t + \xi_t$, subtraction gives

$$v_{t+1} - v_t = (M - I)v_t + B\Delta a_t + \Delta\xi_t. \tag{34}$$

Thus, for $\widehat{v}_t = v_t / r_t$,

$$\|(M - I)\widehat{v}_t\| \le \frac{\sqrt{(r_{t+1} - r_t)^2 + 2r_t r_{t+1}(1 - \mathcal{C}_t)} + \|B\|_2 \|\Delta a_t\| + \|\Delta\xi_t\|}{r_t}. \tag{35}$$

To pass from visited directions to an operator bound, collect $V = [\widehat{v}_{t_1}, \dots, \widehat{v}_{t_k}]$ and $E = (M - I)V$. On their span,

$$\|(M - I)P_{\mathrm{span}\, V}\|_2 \le \|E\|_2 \|V^\dagger\|_2. \tag{36}$$

Small angular change, controlled action/noise increments, and well-excited directions therefore constrain the local dynamics jointly.

For full-row-rank B and $\|M - I\|_2 \le \delta < 1$, the finite-horizon Gramian satisfies

$$\lambda_{\min}(J_H J_H^\top) \ge \sigma_{\min}(B)^2 \sum_{j=0}^{H-1} (1 - \delta)^{2j}, \tag{37}$$

$$\lambda_{\max}(J_H J_H^\top) \le \sigma_{\max}(B)^2 \sum_{j=0}^{H-1} (1 + \delta)^{2j}. \tag{38}$$

Their ratio bounds its condition number; the action-side positive spectrum gives the same ratio. These assumptions explain the route from straightening to a conditioning bound. At inference, LatentGauge keeps this endpoint map fixed and introduces the separate weight in (32).

B.4 Coordinate curvature and optimization iterates

For a smooth curve $z(s)$ and $\widetilde{z} = T(z)$,

$$\dot{\widetilde{z}} = DT(z)\dot{z}, \qquad \ddot{\widetilde{z}} = D^2T(z)[\dot{z},\dot{z}] + DT(z)\ddot{z}. \tag{39}$$

A nonlinear map can bend a straight coordinate trajectory. This concerns coordinate curves and Euclidean chord costs; a smooth pullback of Euclidean space remains intrinsically flat.

In the canonical active quadratic, both eigenmodes begin with equal error magnitude and use the optimal constant step. If the positive condition number is κ, the two mode multipliers are $\pm(\kappa-1)/(\kappa+1)$. The clean isotropic quadratic reaches its minimum in one step. These optimizer iterates and the fixed action rollouts are different objects, linked by the same endpoint sensitivity.

B.5 CEM selection, distribution fitting, and native actions

CEM iteratively fits a sampling distribution to elite candidates (de Boer et al., 2005). For real-time planning, iCEM improves sampling efficiency through temporally correlated samples and elite reuse (Pinneri et al., 2021). The identities below isolate the selected-sample update for a fixed candidate pool. Fix one candidate pool $\{\mathbf{a}_j\}_{j=1}^N$, its predicted endpoints, and an elite count $1 \le k < N$. Assume distinct clean and mapped costs, and let E and E_T index their respective lowest-cost k candidates. Then

$$E_T \neq E \quad \Longleftrightarrow \quad \min_{j\notin E} C_T(\mathbf{a}_j) < \max_{i\in E} C_T(\mathbf{a}_i). \tag{40}$$

If $E_T = E$, every member precedes every nonmember, so the inequality is false. If the inequality holds, a nonmember precedes a member and E cannot occupy the first k positions. Conversely, when the sets differ, any newly selected index has lower mapped cost than any index it replaces. This proves (40) and Proposition 1. The clean adjacent pair $(k, k+1)$ has the smallest clean cost difference across the inclusion boundary; its mapped reversal is a directly measurable sufficient event.

The selected sample determines the update. LeWM's CEM fits the sampling moments to its low-cost elite candidates (Maes et al., 2026). Write the native mean update with smoothing coefficient $\eta \in (0,1]$ as

$$\boldsymbol{\mu}^+ = (1-\eta)\boldsymbol{\mu} + \frac{\eta}{k}\sum_{i\in E}\mathbf{a}_i. \tag{41}$$

The direct empirical-mean update uses $\eta = 1$. Comparing two score assignments on the same pool and old mean cancels the common terms:

$$\boldsymbol{\mu}_T^+ - \boldsymbol{\mu}^+ = \frac{\eta}{k}\left(\sum_{j\in E_T\setminus E}\mathbf{a}_j - \sum_{i\in E\setminus E_T}\mathbf{a}_i\right). \tag{42}$$

For one entering candidate j and one exiting candidate i, this is $\eta(\mathbf{a}_j - \mathbf{a}_i)/k$. Multiple replacements contribute vectorially. The empirical second moment also has an exact set-difference form,

$$R_{E_T} - R_E = \frac{1}{k}\left(\sum_{j\in E_T\setminus E}\mathbf{a}_j\mathbf{a}_j^\top - \sum_{i\in E\setminus E_T}\mathbf{a}_i\mathbf{a}_i^\top\right), \quad R_E = \frac{1}{k}\sum_{i\in E}\mathbf{a}_i\mathbf{a}_i^\top. \tag{43}$$

Subtracting the corresponding outer products of empirical means yields the covariance change. We record the source implementation's variance smoothing, clipping, and numerical floor along with these raw moments.

Connection to distribution fitting. For fixed positive-definite Σ, the Gaussian mean fitted to the selected sample solves

$$\operatorname*{argmin}_{\nu}\ \frac{1}{2k}\sum_{i\in E}(\mathbf{a}_i-\nu)^\top\Sigma^{-1}(\mathbf{a}_i-\nu) = \frac{1}{k}\sum_{i\in E}\mathbf{a}_i. \tag{44}$$

Its gradient is $\Sigma^{-1}(\nu - k^{-1}\sum_{i\in E}\mathbf{a}_i)$ and its Hessian is $\Sigma^{-1} \succ 0$. Thus the scorer controls which samples determine the fitted action distribution. Combined with (9), this gives the chain from shared radial phases to exact costs, selected samples, and proposal updates. The parallel gradient-planning path uses the same scorer through (32).

Native action extraction. At the final CEM update, an implementation that executes the leading action block of the final mean has $\Delta a_{\text{exec}} = P_0(\boldsymbol{\mu}_T^+ - \boldsymbol{\mu}^+)$, where P_0 selects that block. For a source implementation that executes its best sample or another native output, we apply that same extraction function to the two scored planning states and log its outputs. The mechanism trace therefore follows the implementation's actual score-to-action path.

B.6 Finite-round CEM propagation through the inclusion boundary

The same-pool identities extend to successive resampling rounds by separating score-induced replacement from movement of the candidate pool. We analyze one planning call at a fixed initial observation and goal, so Φ and C are the same frozen functions throughout. The following finite-pool result applies to diagonal Gaussian CEM with uniform elite moments, native mean/variance smoothing, and a positive variance floor. These update choices are recorded with the planner, as in Section G.4.

Boundary occupancy controls replacements. For clean costs $c_1, \dots, c_N$, let E index the lowest k, and define

$$\Psi(\delta) = \frac{1}{k} \min\left\{ \#\{i \in E : c_i \geq c_{(k+1)} - 2\delta\},\ \#\{j \notin E : c_j \leq c_{(k)} + 2\delta\} \right\}. \tag{45}$$

Here and below costs are strictly ordered. If $\max_j |\widetilde{c}_j - c_j| \leq \delta$, the selected set $\widetilde{E}$ satisfies $|\widetilde{E} \setminus E|/k \leq \Psi(\delta)$. Indeed, each entering j and exiting i obeys $\widetilde{c}_j < \widetilde{c}_i$, hence $c_j - c_i < 2\delta$. Every entering index lies in the second set of (45), and every exiting index lies in the first. In particular, a clean boundary gap exceeding 2δ preserves the elite set. This count retains the actual discontinuous selection rule.

Coupled candidate generation. Let $\mathcal{A} \subset \mathbb{R}^p$ be a bounded closed convex action-sequence set with $\|a\|_2 \leq M$ and diameter D. For runs $r \in \{0, T\}$, generate candidates from the same base noises:

$$a_{t,j}^r = \Pi_{\mathcal{A}}\Big(\mu_t^r + \sqrt{\nu_t^r} \odot \xi_{t,j}\Big), \qquad \nu_t^r \geq \underline{\sigma}^2 \mathbf{1}, \qquad \beta_t = \max_j \|\xi_{t,j}\|_\infty. \tag{46}$$

The square root is componentwise. Write $\overline{a}_t^r$ and V_t^r for the elite mean and coordinatewise variance. The native updates are

$$\mu_{t+1}^r = (1-\eta_t)\mu_t^r + \eta_t \overline{a}_t^r, \qquad \nu_{t+1}^r = \mathrm{clip}_{[\underline{\sigma}^2, \overline{\nu}]}\big((1-\zeta_t)\nu_t^r + \zeta_t V_t^r\big), \tag{47}$$

where $0 < \eta_t, \zeta_t \leq 1$ are the recorded planner coefficients; an absent upper cap is $\overline{\nu} = +\infty$. The attacked run uses the same fixed T at every round.

Let C be L_C-Lipschitz on $\mathcal{A}$. For the terminal squared-distance scorer, a sufficient constant is

$$L_C = 2 \sup_{a \in \mathcal{A}} \|\Phi(a) - z_g\| \sup_{a \in \mathcal{A}} \|D_a \Phi(a)\|_2. \tag{48}$$

Thus the same rollout sensitivity that appears in (4) controls how resampled actions change the clean cost. Define the direct score intervention on the attacked pool by

$$b_t = \max_j |C_T(a_{t,j}^T) - C(a_{t,j}^T)|. \tag{49}$$

It is computed by (9); a valid upper bound may be substituted. All quantities are local to this planning call.

Theorem 6 (Finite-pool propagation envelope)**.** *Under* (46)–(47)*, let* Ψ_t *be* (45) *for the clean round-*t *pool. Start with nonnegative bounds* $U_0 \geq \|\mu_0^T - \mu_0^0\|_2$ *and* $W_0 \geq \|\nu_0^T - \nu_0^0\|_2$*, and recursively compute*

$$R_t = U_t + \frac{\beta_t}{2\underline{\sigma}} W_t, \qquad P_t = \Psi_t(b_t + L_C R_t), \tag{50}$$

$$\begin{aligned} U_{t+1} &= (1-\eta_t)U_t + \eta_t(R_t + DP_t), \\ W_{t+1} &= (1-\zeta_t)W_t + 4M\zeta_t(R_t + DP_t). \end{aligned} \tag{51}$$

Then, at every round, $\|\mu_t^T - \mu_t^0\|_2 \leq U_t$*,* $\|\nu_t^T - \nu_t^0\|_2 \leq W_t$*, and* $|E_t^T \setminus E_t^0|/k \leq P_t$*. The envelope includes both the exact scoring intervention and the changed candidate distribution.*

Proof. Nonexpansiveness of projection and the componentwise bound $|\sqrt{x} - \sqrt{y}| \le |x - y|/(2\underline{\sigma})$ give $\max_j \|a^T_{t,j} - a^0_{t,j}\| \le R_t$. Therefore

$$|C_T(a^T_{t,j}) - C(a^0_{t,j})| \le b_t + L_C R_t.$$

The boundary-occupancy result bounds the fraction of changed indices by P_t. To compare elite means, first move the candidates at fixed indices, contributing at most R_t, and then pair each entering index with an exiting one in the clean pool, contributing at most D per replacement. Hence $\|\overline{a}^T_t - \overline{a}^0_t\| \le R_t + DP_t$. For coordinatewise second moments, $\|x \odot x - y \odot y\|_2 \le 2M\|x - y\|_2$ on $\mathcal{A}$; the corresponding difference is at most $2M(R_t + DP_t)$. Subtracting the squared elite means contributes at most the same amount, so $\|V^T_t - V^0_t\| \le 4M(R_t + DP_t)$. Finally, variance clipping is nonexpansive. Applying the native smoothing updates proves (51); induction proves the result. □

Accumulation and exact stability. Put $\mathbf{e}_t = (U_t, W_t)^\top$ and

$$A_t = \begin{pmatrix} 1 & \eta_t\beta_t/(2\underline{\sigma}) \\ 4M\zeta_t & 1 - \zeta_t + 2M\zeta_t\beta_t/\underline{\sigma} \end{pmatrix}, \qquad q_t = DP_t \begin{pmatrix} \eta_t \\ 4M\zeta_t \end{pmatrix}. \tag{52}$$

For R rounds the envelope unrolls as

$$\mathbf{e}_R = A_{R-1} \cdots A_0 \mathbf{e}_0 + \sum_{t=0}^{R-1} A_{R-1} \cdots A_{t+1} q_t, \tag{53}$$

with empty products equal to the identity. The terms q_t quantify boundary entry and the products transport it through later mean/variance resampling. With identical initial states, if each successive clean boundary gap exceeds $2b_t$, induction gives identical selected sets and moments at every round. The bound measures an upper envelope; the signed update identity in (42) retains cancellation between replacements. The round index here counts CEM iterations within one planning call.

Evaluation from existing traces. The paired trace records b_t, clean boundary costs, base-noise maxima, both sampling moments and both selected sets. These supply the envelope without an additional optimizer, loss or tunable coefficient. For mean-prefix execution, the final action discrepancy is at most U_R because the prefix projection has operator norm one. Other output rules use the logged native action extraction. Empirical evaluation reports the observed moment differences alongside the envelope and the associated boundary counts.

C Coordinate Selection Before Deployment

The prior-work chain explains why latent planning comes to rely on a particular coordinate geometry. Prediction, Gaussian regularization, and temporal pairing impose different constraints on that geometry. The inference-time construction in the main text acts after the learned representation has been fixed.

C.1 Prediction, anti-collapse, and finite marginal diagnostics

LeWM jointly learns an encoder and an action-conditioned predictor with a prediction loss and Gaussian regularization (Maes et al., 2026; Balestriero & LeCun, 2025):

$$z_t = h(o_t), \qquad \widehat{z}_{t+1} = F(z_t, a_t), \qquad \mathcal{L}_{\text{train}} = \mathbb{E}\|\widehat{z}_{t+1} - z_{t+1}\|^2 + \lambda\,\text{SIGReg}(Z). \tag{54}$$

A constant encoder admits zero prediction error if the predictor returns the same constant. Matching an isotropic Gaussian excludes that fully collapsed representation. SIGReg uses sketched characteristic-function discrepancies over projected latents. For a sample $z_1, \ldots, z_N$ and a unit direction u, the empirical characteristic function is

$$\widehat{\varphi}_u(t) = \frac{1}{N} \sum_i e^{\mathrm{i}tu^\top z_i}. \tag{55}$$

It is compared with $e^{-t^2/2}$ and aggregated over directions and frequencies. The exact finite-grid diagnostic implemented for this manuscript is specified in Section F; Theorem 3 gives the sampling-law statement under the Gaussian reference.

C.2 PREDICTIVE COORDINATE FREEDOM

At zero prediction risk, suppose $h(o_{t+1}) = F(h(o_t), a_t)$ almost surely and $h(o_t) \sim \gamma$. If the model class represents the conjugated pair

$$\widetilde{h} = T \circ h, \qquad \widetilde{F}(y, a) = T(F(T^{-1}(y), a)), \tag{56}$$

then every Gaussian-preserving bijection T retains zero prediction risk and the target marginal. Indeed, $\widetilde{F}(\widetilde{h}(o_t), a_t) = T(h(o_{t+1})) = \widetilde{h}(o_{t+1})$. Nonlinear coordinate changes can nevertheless change distances and temporal bending. This connects the argument to distribution-preserving non-identifiability (Locatello et al., 2019) and to the Gaussian-preserving spiral used by Klindt et al. (2026).

At nonzero error, a transformed squared residual is locally weighted by G_T. Inference-time LATENT-GAUGE instead keeps h, F unchanged, so the original residual is computed before the shared map. An equivalent passive coordinate change would also transform the score through T^{-1}; the studied interface retains the Euclidean score on the delivered coordinates.

C.3 A RESTRICTED PREDICTOR LINKS ERROR TO TEMPORAL CURVATURE

Consider a locally affine physical path $x, x+u, x+2u$ and a C^3 encoder. A first-order predictor uses $h(x) + Dh(x)u$. With $e = h(x+u) - h(x) - Dh(x)u$ and $\Delta^2 z = h(x+2u) - 2h(x+u) + h(x)$, Taylor expansion gives

$$e = \tfrac{1}{2} D^2 h(x)[u, u] + O(\|u\|^3), \tag{57}$$

$$\Delta^2 z = D^2 h(x)[u, u] + O(\|u\|^3), \tag{58}$$

$$\|e\|^2 = \tfrac{1}{4} \|\Delta^2 z\|^2 + O(\|u\|^5). \tag{59}$$

If consecutive latent speeds equal r, (33) also yields $\|e\|^2 = \frac{1}{2} r^2 (1 - \mathcal{C}_t) + O(\|u\|^5)$. The same encoder-Hessian term drives prediction residual and coordinate turning. Under the stated predictor restriction, simpler dynamics therefore favor lower-curvature coordinates. The restriction is essential: an expressive predictor may represent the nonlinear conjugated dynamics in (56).

C.4 TEMPORAL PAIRING AND ORTHOGONAL IDENTIFICATION

For the Gaussian transition $Z' = \varrho Z + \sqrt{1 - \varrho^2}\, \eta$, with independent standard-Gaussian Z, η and $0 < \varrho < 1$, a normalized multivariate Hermite basis obeys

$$\mathbb{E}[H_k(Z') \mid Z] = \varrho^{|k|} H_k(Z). \tag{60}$$

If a centered scalar coordinate has coefficients b_k in this orthonormal basis, its variance is $\sum_{|k| \geq 1} b_k^2$ and its cross-time covariance is $\sum_{|k| \geq 1} b_k^2 \varrho^{|k|}$. At fixed variance, alignment favors degree one because $\varrho^{|k|} < \varrho$ for $|k| > 1$. Under the full assumptions of Klindt et al. (2026), the optimal Gaussian representation is identified up to an orthogonal map. Their radius-dependent spiral illustrates why the marginal alone is insufficient and temporal alignment supplies additional information.

Controlled identifiability further relies on predictable-representation spectral separation and nondegenerate action excitation (Zhang et al., 2026). These hypotheses concern the representation and transition learned from temporally paired, action-conditioned data. A downstream adapter constrained only by its output marginal is a different object. Thus the inference-time intervention complements, rather than reverses, the training-side coordinate-selection argument.

C.5 TRANSFER OF STRAIGHTNESS THROUGH NEAR-ORTHOGONAL COORDINATES

A shared orthogonal map preserves adjacent-velocity cosine exactly. For the approximate case, suppose $\widetilde{z}_t = Q z_t + r_t$, $Q^\top Q = I$, and $\|r_t\| \leq \varepsilon$. If both clean velocity lengths are at least $\nu > 2\varepsilon$, then

$$|\cos(\widetilde{v}_t, \widetilde{v}_{t+1}) - \cos(v_t, v_{t+1})| \leq \min\{2, 8\varepsilon/\nu\}. \tag{61}$$

To see this, each velocity perturbation is at most 2ε. The bound $\|(x+e)/\|x+e\| - x/\|x\|\| \leq 2\|e\|/\|x\|$ changes each normalized velocity by at most $4\varepsilon/\nu$; the change of their inner product is bounded by the sum. This supplies the finite-error link between coordinate identification and the straightness observed by a downstream reader.

D KERNEL EXTENSION AND CONSTRAINED PROFILE FITTING

This section specifies the shared radial function, its calibration plane, and the single constrained solver used in Algorithm 2. The calibration triples and the original displacement budget determine the optimization problem. All numerical defaults below are shared across tasks. For CEM, the triples are produced by the native-boundary rule in Section G.4; these same triples determine both the kernel sites and the first-variation plane.

D.1 THE RADIAL SMOOTHNESS MODEL AND ITS KERNEL

Why an extension rule is needed. Exact candidate scores depend on finitely many squared radii. Smooth bumps supported in shrinking neighborhoods of these radii can keep prescribed calibration phases while their integrated displacement tends to zero. Indeed, if the support is U, then

$$D_\gamma(T_\alpha) \le \frac{8}{d}\mathbb{E}_{S\sim\chi^2_d}[S\mathbf{1}_{S\in U}] \longrightarrow 0. \tag{62}$$

We therefore specify a shared smooth extension from calibration phases to the full radial domain. For each set of site values, kernel interpolation specifies its extension between observations; the actual displacement remains the attack constraint.

Reference-weighted differential operator. Let μ denote the probability density of χ^2_{d+2}. The size-biased radial law appears directly in the displacement identity:

$$D_\gamma(T_\alpha) = 4\mathbb{E}_\mu[1-\cos\alpha(S)]. \tag{63}$$

On mean-zero functions in $L^2(\mu)$, use the nonnegative self-adjoint Laguerre operator

$$\mathsf{L}f = -\frac{1}{\mu(s)}\frac{d}{ds}\{2s\mu(s)f'(s)\} = -2sf'' - (d+2-s)f'. \tag{64}$$

Its natural realization has zero endpoint flux. For a unit rank-two skew generator, direct integration gives

$$\mathbb{E}_\gamma\|\operatorname{Sym} Dv\|_F^2 = 2\mathbb{E}_\mu[2S\alpha'(S)^2] = 2\langle\alpha,\mathsf{L}\alpha\rangle_{L^2(\mu)}. \tag{65}$$

The first equality follows from $\|\operatorname{Sym} Dv\|_F^2 = 2\alpha'(S)^2S\|AZ\|^2$ and $\mathbb{E}[\|AZ\|^2 \mid S] = 2S/d$. Thus L measures radial variation through the same strain that changes the local scoring geometry.

We choose the squared-operator interpolation energy $\|\mathsf{L}f\|^2_{L^2(\mu)}$. With $a = d/2+1$, its normalized eigenfunctions and eigenvalues are (DLMF)

$$\phi_r(s) = \sqrt{\frac{r!\,\Gamma(a)}{\Gamma(r+a)}}\,L_r^{(a-1)}(s/2), \qquad \mathsf{L}\phi_r = r\phi_r. \tag{66}$$

The mean-zero space and its kernel are

$$\mathcal{H} = \left\{ f = \sum_{r\ge1} a_r\phi_r : \sum_{r\ge1} r^2a_r^2 < \infty \right\}, \qquad K(s,t) = \sum_{r\ge1}\frac{\phi_r(s)\phi_r(t)}{r^2}, \quad s,t>0. \tag{67}$$

The constant mode is handled separately by the common rotation. The norm is an interpolation criterion; Algorithm 2 constrains the exact displacement integral. Local elliptic regularity and the spectral gap give a bounded H^2 norm on every compact interval inside $(0,\infty)$, so point evaluation there is continuous. This places the extension in the reproducing-kernel Hilbert-space setting (Aronszajn, 1950).

Green-function evaluation and the origin. The following equivalent construction specifies the kernel without choosing a spectral truncation. Write $F_\mu(s) = \int_0^s \mu(u)\,du$ and $p(s) = 2s\mu(s)$. For each positive site t, define

$$H_t(s) = \int_t^s \frac{F_\mu(u) - \mathbf{1}_{u\ge t}}{p(u)}\,du, \tag{68}$$

$$G(s,t) = H_t(s) - \mathbb{E}_{U\sim\mu}H_t(U), \tag{69}$$

$$K(s,t) = \int_0^\infty G(u,s)G(u,t)\mu(u)\,du. \tag{70}$$

Here $\mathsf{L}G(\cdot,t) = \delta_t/\mu - 1$ and G has zero mean; consequently K is the kernel of L^{-2}. The Green function is symmetric by self-adjointness. An equivalent column evaluation solves

$$\partial_s K(s,t) = -\frac{\int_0^s G(u,t)\mu(u)\,du}{p(s)}, \qquad \int_0^\infty K(s,t)\mu(s)\,ds = 0. \tag{71}$$

These one-dimensional integrals fix the additive constants and endpoint conditions. For $s < t$, $F_\mu(s)/p(s) = 1/(d+2) + O(s)$ at zero. Thus $G(\cdot,t)$ is regular there, and

$$\partial_s K(0,t) = -G(0,t)/(d+2). \tag{72}$$

The same expansion gives a finite second derivative. At a positive site, G is continuous with a first-derivative jump; (71) therefore gives a C^2 kernel section. Each finite combination used by the method extends to $s = 0$ by its right limit and produces a C^2 radial map on $\mathbb{R}^d$. Positive sites suffice: the zero vector is fixed by every radial rotation and its angle does not affect any candidate–goal cost. The remaining tail grows at most polynomially in $\log s$ and is integrable against μ.

Minimum-norm extension from the data. Sort and deduplicate the positive squared radii of all calibration endpoints and goals to obtain $s_1, \dots, s_m$. Define $\mathbf{K}_{jk} = K(s_j, s_k)$. For prescribed site values u, the extension is

$$\alpha_c(s) = \sum_j c_j K(s, s_j), \qquad c = \mathbf{K}^{-1}u, \qquad \|\alpha_c\|_{\mathcal{H}}^2 = c^\top \mathbf{K} c. \tag{73}$$

The Gram matrix is positive definite for distinct positive sites: smooth mean-zero test functions can assign arbitrary values at finitely many such sites. For any competing interpolant $f = \alpha_c + f_\perp$, the reproducing property gives $f_\perp(s_j) = 0$ and $\|f\|_{\mathcal{H}}^2 = \|\alpha_c\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2$. This proves the minimum-norm extension claim (Schölkopf et al., 2001). The attack optimizes within this finite interpolation family, evaluating its original nonlinear displacement at every accepted iterate. Center count is determined by the calibration radii; the kernel and smoothness order remain fixed.

D.2 A CALIBRATION-DEFINED ROTATION PLANE

We determine A once from the first variation of the exact ranking loss. Use the normalized degree-one radial mode $\psi(s) = (s - d - 2)/\sqrt{2(d+2)}$. For an endpoint x and goal g, set $\delta\psi_{xg} = \psi(\|g\|^2) - \psi(\|x\|^2)$. Under the probe profile $t\psi$, differentiation of (9) at $t = 0$ gives

$$\frac{d}{dt} C_{t\psi,A}(x,g)\bigg|_0 = -2\delta\psi_{xg}\, x^\top A g. \tag{74}$$

Let σ be the logistic function. For the calibration triples, form the skew operator

$$S = \frac{1}{n}\sum_i \sigma(m_i)\big[\delta\psi_{z_i^- z_{g,i}}(z_i^- z_{g,i}^\top - z_{g,i}(z_i^-)^\top) - \delta\psi_{z_i^+ z_{g,i}}(z_i^+ z_{g,i}^\top - z_{g,i}(z_i^+)^\top)\big]. \tag{75}$$

The loss derivative is $-\langle S, A\rangle_F$. If $\varsigma = \|S\|_2 > 0$, choose a unit leading eigenvector u of $-S^2$, put $v = Su/\varsigma$, and set

$$A = vu^\top - uv^\top. \tag{76}$$

Then $u^\top v = 0$, $\|A\|_2 = 1$, and $\langle S, A\rangle_F = 2\varsigma$, the maximum over unit rank-two skew matrices. To see the maximum, write an arbitrary such matrix as $ba^\top - ab^\top$ for orthonormal a, b: its inner product with S is $2b^\top Sa \le 2\|S\|_2$. The kernel interpolant of ψ has the same calibration values, so the same decreasing first variation is available in the coefficient space. Its actual displacement is $O(t^2)$ near the identity.

For a repeated leading singular value, project the first canonical basis vector with nonzero projection onto the leading eigenspace and normalize it to define u; the sign of its first nonzero coordinate is positive. If $S = 0$, use $u = e_1, v = e_2$ as the fixed fallback. The optimization and its stopping test then operate on the same exact objective. The low-rank outer-product form in (75) permits matrix–vector products directly from cached features. The plane is computed only on calibration data and shared by the target method and its kernel-profile controls.

D.3 Exact common-rotation elimination and budget evaluation

For any fixed orthonormal u, v and A in (76),

$$\|T_\alpha(Z) - Z\|^2 = 2\{(u^\top Z)^2 + (v^\top Z)^2\}\{1 - \cos\alpha(S)\}. \tag{77}$$

Spherical conditioning gives expected plane energy $2S/d$, proving

$$D_\gamma(T_\alpha) = \frac{4}{d}\mathbb{E}_{S\sim\chi^2_d}[S\{1-\cos\alpha(S)\}] = 4\mathbb{E}_{S\sim\mu}[1-\cos\alpha(S)]. \tag{78}$$

A common angle b cancels from every endpoint–goal phase difference. Write

$$M(c) = \mathbb{E}_\mu e^{\mathrm{i}\alpha_c(S)}, \quad b^*(c) = -\arg M(c), \quad D_{\min}(c) = 4\{1 - |M(c)|\}. \tag{79}$$

Indeed, $D_\gamma(T_{\alpha_c+b}) = 4[1 - \mathrm{Re}(e^{\mathrm{i}b}M(c))]$, whose minimum over b is (79). For $M(c) = 0$, set $b^* = 0$; every common angle has displacement four. The main budget $\epsilon = 0.07$ keeps feasible iterates in $|M(c)| \geq 1 - \epsilon^2/4 > 0$. The original displacement condition is therefore smooth throughout the fitted feasible region. Common-angle elimination is applied during optimization and to the returned profile.

Derivatives and numerical evaluation. Let $k(s) = (K(s, s_j))_{j=1}^m$. Then

$$\nabla_c M = \mathrm{i}\mathbb{E}_\mu[k(S)e^{\mathrm{i}\alpha_c(S)}], \qquad \nabla_c D_{\min} = -\frac{4}{|M|}\,\mathrm{Re}(\overline{M}\nabla_c M). \tag{80}$$

The same kernel values are reused by the score and derivative calculations. For numerical budget evaluation, split one-dimensional adaptive quadrature at the positive calibration radii and the mode of μ. Use absolute tolerances 10^{-10} and 10^{-12} in two passes. Evaluate the centered quantity $\mathbb{E}_\mu(e^{\mathrm{i}\alpha_c} - 1)$ to preserve accuracy near the identity. If centered integration stops at $s_{\max}$, the omitted contribution is bounded by $2\Pr_\mu(S > s_{\max})$; choose this tail allowance below 10^{-12}. The estimated error δ_M combines quadrature error, pass discrepancy, kernel-evaluation error, and the tail allowance. The displacement check uses

$$D_{\mathrm{up}} = 4\left[1 - \max\{0, |\widehat{M}| - \delta_M\}\right] \leq \epsilon^2. \tag{81}$$

Here the quadrature and kernel terms are numerical error estimates, recorded with each check. Independent higher-precision evaluation verifies the returned profile. Reference-law preservation holds algebraically for every coefficient vector. Kernel integrals use (70) or (71) with the same two-pass accuracy rule; near the origin use the regular endpoint expansions. Exact duplicates are merged, and ill-conditioned Gram systems are recomputed at higher precision with interpolation residual checked against 10^{-10}.

D.4 Feasible trust-region sequential quadratic programming

Constrained adversarial optimization makes the attack objective and admissible perturbations explicit (Madry et al., 2018). In this formulation, the variables are shared radial-profile coefficients and feasibility is measured by the integrated displacement. The finite problem minimizes the exact loss in (10) under $h(c) = D_{\min}(c) - \epsilon^2 \leq 0$. We adopt a feasible trust-region SQP scheme (Wright & Tenny, 2004; Nocedal & Wright, 2006). Use $y = \mathbf{K}^{1/2}c$ so that Euclidean steps in y measure the norm of the profile update. This scales the solver; it imposes no upper bound on the interpolation norm. In the following equations, gradients are with respect to y.

Quadratic subproblem. At a feasible iterate define $g_t = \nabla\mathcal{L}(y_t)$, $a_t = \nabla h(y_t)$, and a positive-definite quasi-Newton matrix B_t. Solve

$$\min_p \; g_t^\top p + \tfrac{1}{2}p^\top B_t p \quad \text{s.t.} \quad h(y_t) + a_t^\top p \leq 0, \quad \|p\|_2 \leq \Delta_t. \tag{82}$$

The zero step is feasible. Both the budget-normal and budget-tangential directions are available to the subproblem. The derivatives here are those of the attack loss and displacement with respect to c; $J^\top G_T J$ retains its distinct role as the action-side planning geometry.

Feasibility correction and acceptance. Start with $y' = y_t + p$. When the exact numerical budget test fails, apply normal corrections

$$y' \leftarrow y' - \frac{h(y') + 2\delta_h}{\|\nabla h(y')\|^2} \nabla h(y'), \tag{83}$$

where δ_h is the current budget-error allowance. Re-evaluate the original constraint after each correction. A zero or undefined normal, a nonreducing violation, or a corrected step exceeding $2\Delta_t$ rejects the trial and halves the trust radius. A correction is retained only when it passes (81).

For the corrected step $s = y' - y_t$, form the predicted decrease $P_t = -g_t^\top s - \frac{1}{2} s^\top B_t s$ and ratio $R_t = [\mathcal{L}(y_t) - \mathcal{L}(y')]/P_t$. Accept when $P_t > 0$ and $R_t \geq 1/4$. Otherwise keep y_t and halve Δ_t. After an accepted step with $R_t \geq 3/4$ and $\|p\| \geq 0.9\Delta_t$, double Δ_t; otherwise retain it. Thus every accepted step satisfies the actual budget and reduces the actual ranking loss.

Curvature and termination. Initialize $y_0 = 0$, $B_0 = I$, and $\Delta_0 = 1$. After acceptance, use a Powell-damped BFGS update on Lagrangian-gradient differences, with the nonnegative multiplier from (82). Specifically, for $s = y_{t+1} - y_t$ and $z = \nabla(\mathcal{L} + \lambda_t h)(y_{t+1}) - \nabla(\mathcal{L} + \lambda_t h)(y_t)$, set $\theta = 1$ if $s^\top z \geq 0.2 s^\top B_t s$, and otherwise $\theta = 0.8 s^\top B_t s / (s^\top B_t s - s^\top z)$. With $\bar{z} = \theta z + (1 - \theta) B_t s$, update

$$B_{t+1} = B_t - \frac{B_t s s^\top B_t}{s^\top B_t s} + \frac{\bar{z}\bar{z}^\top}{s^\top \bar{z}}. \tag{84}$$

Zero-length steps retain B_t. The KKT test uses primal feasibility, $\lambda \geq 0$, complementarity, and $\|\nabla\mathcal{L} + \lambda \nabla h\|_\infty \leq 10^{-6}$. For an inactive constraint take $\lambda = 0$; near the boundary use the nonnegative least-squares multiplier. Numerical feasibility/complementarity tolerance is 10^{-10}. Stop at this test, at $\Delta_t < 10^{-10}$, or at the 3,000-call allowance, and return the current feasible iterate. A QP residual tolerance of 10^{-10} fixes the inner solve. These constants are numerical rules shared by all tasks.

Every objective trial, rejected proposal, and restoration attempt consumes an evaluation slot. A forward/backward evaluation consumes two forward-equivalent slots. Budget quadrature and kernel setup are additionally timed and counted in the work log; exhaustion during an inner solve returns the previous feasible iterate. Calibration-plane work is shared across the kernel-method comparison. The method uses the exact ranking objective and one displacement constraint, with a fixed interpolation model and one solver.

D.5 Matched random-direction controls

The random controls use the same calibration sites, Gram scaling, selected plane, common-angle elimination, and budget test. At current y and radius Δ, draw $\xi \sim \mathcal{N}(0, I_m)$ and $U \sim \mathrm{Unif}[0, 1]$ independently, and form

$$r = \Delta U^{1/m} \xi / \|\xi\|. \tag{85}$$

Project r onto the linearized-feasible trust ball by minimizing $\frac{1}{2}\|p - r\|^2$ subject to the constraints in (82). Apply the same normal correction and exact numerical budget check as the target solver. This is a local random-direction proposal in the same interpolation space.

For the random-profile control, start at zero with $\Delta = 1$ and return the first feasible proposal, halving Δ after failure; the loss does not select its coefficients. For random-direction search, accept a feasible proposal when the exact calibration loss decreases. Double the radius after an accepted step with $\|p\| \geq 0.9\Delta$; halve it on failure and otherwise retain it. Keep the current iterate when a trial is rejected. Both controls use the same numerical stopping floor and evaluation accounting as the target method. Each failed attempt is charged; an exhausted allowance returns the current feasible map. The shared plane uses the calibration ranking for all kernel-profile methods, while the random-profile coefficients are sampled without loss selection.

D.6 Spatial-feature interface and operation count

Apply C-order vectorization to the feature axes of one predicted endpoint, and identically to its goal. Candidate, batch, and time axes remain separate. Restore the original shape before invoking the native scorer. Thus a $14 \times 14 \times 384$ endpoint gives $d = 75{,}264$, a $14 \times 14 \times 8$ endpoint gives $d = 1{,}568$, and a global 1×384 representation gives $d = 384$. Vectorization preserves Frobenius

distance. The original sum or mean reduction is retained; its fixed normalization multiplies costs and derivatives. Masks, channel weights, and any separate uncertainty or aggregated score retain their source definitions and are logged with the victim interface. The exact Euclidean identities apply to the squared-distance component under its stated normalization; native composite scores are evaluated directly.

For the selected plane, with $q = u^\top z$ and $r = v^\top z$, apply an angle ϑ by

$$e^{\vartheta A} z = z + [(\cos\vartheta - 1)q - \sin\vartheta\, r]u + [\sin\vartheta\, q + (\cos\vartheta - 1)r]v. \tag{86}$$

The squared radius and plane projections cost $O(d)$, followed by m kernel-section evaluations and an $O(m)$ weighted sum. Calibration stores the $m \times m$ Gram matrix and its factorization; the plane operator is evaluated using its cached low-rank factors. Record the actual center count and kernel-evaluation work separately from model rollout time. The returned object contains the calibration radii, coefficients, fixed kernel definition, plane vectors, common angle, and original tensor-shape convention.

Kernel evaluation on the positive radial axis covers all finite deployment radii. The zero-vector value uses the endpoint continuation above. All-equal scored radii produce a common rotation and invariant scores; empty positive-site sets return the identity. Calibration-site deduplication, radius statistics, and Gram/interpolation residuals are logged before fitting. The independent endpoint and goal audits in Section G retain their original data splits.

E Supplementary Mechanism Protocols

This section specifies the inputs and statistics used by the controlled checks in Section 5.4. The scripts and stored records accompany these protocols.

E.1 Marginal Distribution and Displacement

The Gaussian audit uses paired clean and mapped samples from independent seeds. Its statistic, sample sizes, projection grid, and interval calculation are specified in Section F. For the frequency sweep, each profile is calibrated to the same empirical movement and evaluated on an independent probe sample; `data/distortion.csv` retains the complete grid and measurements.

E.2 Exact Ranking and Normalization Controls

The same goal and two endpoints are scored under identity, common rotation, the radial map, and the two normalization orders in Table 4. Euclidean and cosine comparisons retain their respective clean margins. The analytical identities are in Section A.4; `scripts/reproduce.py` regenerates the fixed-system checks from the inputs in Section F.

E.3 Fixed-pool CEM selection protocol

The diagnostic uses six two-component action sequences, $(-.9, .2)$, $(-.4, .8)$, $(.1, -.5)$, $(.7, .4)$, $(.9, -.8)$, and $(-.6, -.6)$, with clean scores $0.1, 0.2, \ldots, 0.6$. Select three candidates and fit their empirical mean and full covariance with divisor $k = 3$. Swap either the first two scores or the third and fourth, keeping all candidate coordinates fixed. Report the changed member fraction, the Euclidean norm of the mean difference, and the Frobenius norm of the covariance difference (Table 3).

The same script evaluates the set and update identities in Equations (40) and (42) on 1,000 random pools using seed 20260925, varying pool size, action dimension, elite count, and smoothing. Inputs and records are in `data/cem_identity_check.json`; `scripts/check_cem_identities.py` regenerates them. The coupled multi-round protocol and its theoretical envelope are in Section B.6.

F CONTROLLED NUMERICAL PROTOCOL

Executable entry point. Run `python scripts/reproduce.py` with NumPy and SciPy. The script writes the numerical CSV inputs and a complete `data/results.json`. Then run `python scripts/figure_data.py` to regenerate exact inverse-image contours and local-optimizer coordinates. Every main figure is editable TikZ artwork sharing this canonical model. Compilation requires pdfLaTeX and BibTeX; the numerical inputs are included. The mechanism script completes without downloading a checkpoint. The local quadratic starts with equal magnitudes in the two eigenmodes; clean and mapped problems share this initial vector and each uses its optimal constant step. Full trajectories and formula-level validation are regenerated by the scripts.

Gaussian study. Generator seed 1701 produces, in order, a calibration array of shape 16384×64, a probe array of shape 8192×64, and 32 unit-normalized directions. For each $k \in \{0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$, a scalar root solver calibrates c in

$$\alpha_{c,k}(s) = c \sin\Big(k(s-64)/\sqrt{128}\Big) \tag{87}$$

to average calibration displacement 0.25^2. Only the first two coordinates rotate; all 64 contribute to the radius. The metric is evaluated from its exact formula on the independent probe array. Calibration equality is not labeled a population equality.

Finite audit. The fixed $k = 10$ map is audited with new Gaussian arrays from seeds 3000 through 3019. For $N = 1024$, projected samples $u_j^\top Z_i$, grid $t_\ell \in [-4, 4]$ with 41 equally spaced points, and normalized $w_\ell \propto e^{-t_\ell^2/2}$, the statistic is

$$S_N = \frac{N}{32} \sum_{j=1}^{32} \sum_{\ell} w_\ell \left| \frac{1}{N} \sum_{i} e^{\mathrm{i} t_\ell u_j^\top Z_i} - e^{-t_\ell^2/2} \right|^2 . \tag{88}$$

This finite-grid statistic instantiates the characteristic-function discrepancy idea of Epps & Pulley (1983). The implemented statistic is the specified finite-grid Epps–Pulley-style diagnostic. The paired-change interval uses $t_{19,0.975}$ times its sample standard deviation divided by $\sqrt{20}$.

Two-dimensional witness. The frozen endpoint model is $\Phi(\mathbf{a}) = c_0 + J\mathbf{a}$, with $J = 2^{-1/2}[I_2 - I_2]$, $c_0 = z_g = (1, 0)$, endpoints $z^+ = (1, 0.5)$ and $z^- = (1.64, 0)$, and fixed actions $\mathbf{a}^\pm = J^\top(z^\pm - c_0)$. The analytical witness uses

$$\alpha(s) = 0.22 \left[e^{-((s-1.25)/0.185)^2} - e^{-((1-1.25)/0.185)^2} \right],$$
$$T_\tau(z) = e^{\tau\alpha(\|z\|^2)A} z, \quad A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad 0 \le \tau \le 1.15. \tag{89}$$

This fixed profile is used for the exact geometry illustrations and has $\alpha(1) = 0$. Its displacement is integrated using $S \sim \chi^2_2$:

$$D_\gamma(T_\tau) = \int_0^\infty s\{1 - \cos[\tau\alpha(s)]\} e^{-s/2}\, ds. \tag{90}$$

The integration upper limit is 70, absolute tolerance 10^{-12}, and maximum 1,200 subintervals. The omitted tail is at most $2\int_{70}^\infty s e^{-s/2} ds = 288 e^{-35} < 1.82 \times 10^{-13}$. The solver's numerical error estimate must be below 10^{-9}. Breakpoints at $1, 1.25, 1.5, 2$ resolve the localized bump. The exact scalar crossing is found by a bracketed root solver on $[0, 1]$, and the budget boundary is solved on $[1, 1.2]$. This script evaluates the fixed analytical family; the fitted-profile procedure is specified separately in Section D.

Certificate and identities. Seed 8080 produces 2,000 independent triples of Gaussian goal and endpoint pairs. Each pair is sorted in clean coordinates, then evaluated under the fixed witness map. We record certified orderings, actual flips, and violations of the certificate. An additional 4,000 endpoint–goal evaluations check the exact cost-shift identity. The declared fixed-map distribution determines the interpretation of these checks.

Table 5: Evidence and numerical provenance.

Category	Provenance and use
Analytical	Gaussian-preserving construction, exact metric and cost shift, spectral correspondence, displacement separation, normalization identities, pairwise certificate.
Controlled numerical	Gaussian audit, frequency and shell sweeps, fixed affine witness, local quadratic descent, 2,000-pair certificate test, and coupled CEM updates. CSVs and scripts specify these calculations.
Checkpoint evaluation	Mapped task success and shared-map comparisons in Tables 1 and 2; calibration, evaluation, and metric definitions are specified in Section G.
Literature-reported	Clean task scores and external attack results retain source table, setting, units, and uncertainty type.

G Checkpoint Settings and Planner Interfaces

This section specifies the checkpoint evaluations in Tables 1 and 2, their source representation and planner conventions, and the shared-coordinate scoring interfaces. The Gaussian and affine-model checks in Sections 5.4 and F separately isolate the geometric and selection mechanisms.

G.1 Protocol fingerprints and literature reuse

Each source result retains its paper version and table, task and dataset, visual/proprioceptive input, encoder/projector, feature layout, score, horizon, replanning interval, candidate count, optimizer, seeds, metric denominator, and uncertainty convention. The three blocks in Table 1 preserve these protocols, and each row average covers four same-unit tasks within one block.

Within DINO-WM, Tables 1 and 2 share an evaluation protocol; the repeated DINO-WM/DINO-patch row is deduplicated. The Temporal Straightening block selects the MPC columns of its Table 1; the complete open-loop/MPC, long-horizon, and combined-cost matrices remain separate in Section H. LeWM Figure-6 means are aligned by task and model, retaining the pixels-only DINO inputs where specified. Its Tables 5 and 10 remain separate because their input descriptions and dispersion conventions differ. Cross-paper PushT values of 90, 66, 74, and 92 consequently refer to distinct source settings.

TRAP Tables 1 and 2 share victim/task identities under one patch protocol, so their attack outcomes and evaluation costs can be read together. Mean per-episode relative drop and the ratio of mean returns retain their different denominators. WMAttack retains its own search budget and score normalization. These external attack results appear under their source protocols in Section H. The complete source locations and settings are in `experiment_data/literature_ledger.csv` and `sources_and_tables.json`.

G.2 Representation and planner interfaces

Gaussian-regularized LeWM supplies the reference-aligned model family. DINO-WM, straightened DINO, and PLDM use their source representations. The scoring interface retains the original feature scale without whitening or Gaussianization; its exact preservation guarantee concerns the reference law γ.

DINO-WM. The source uses pretrained DINOv2 visual features (Oquab et al., 2024), a learned predictor, goal-conditioned scoring, and CEM/MPC. Maze, Wall, Reach, and PushT use 50 evaluation goals; Rope and Granular use ten instances and Chamfer distance. Its offline DreamerV3 (Hafner et al., 2025) and TD-MPC2 (Hansen et al., 2024) rows are source adaptations. The shared-coordinate interface maps the predicted endpoint and goal after each fixed action sequence has been rolled out. Feature layout and norm reductions follow the source; Section D.6 specifies vectorization and tensor restoration.

Temporal Straightening. The source Table-1 GD planner uses 50 goal pairs per data-sampling seed. Raw spatial DINO features have shape $14 \times 14 \times 384$; projected DINO and spatial ResNet features have shape $14 \times 14 \times 8$; global features have shape 1×384. Projector-only and projector-

plus-straightening use the same capacity. Table-2 long-horizon goals are 50 environment steps away. Spatial cost and spatial plus 0.1 aggregated cost are distinct scoring settings, each retaining its original feature shape and reduction.

LeWM. The source encoder, predictor, goal files, input modality, and planner configuration define each setting. PushT Table 5 uses 50 goals reachable within 25 steps and a 50-step execution budget. Figure-6 task means and Table-10 solver results retain their respective inputs. The solver table compares CEM, SGD, RMSProp, and Adam with the source hyperparameters.

PLDM. The independent PLDM paper's Two-Room task uses 64×64 observations and up to 200 steps. Its native model predictive path integral (MPPI) controller (Williams et al., 2017) uses a sum of goal-distance terms and an uncertainty cost. The shared-coordinate interface is *goal-cost-only*: one map transforms every scored state and goal, while the upstream uncertainty term retains its native value. PLDM retains its native cumulative norm-plus-uncertainty score; the terminal squared-distance Gauss–Newton expression describes the local diagnostic in Section 3. The source no-door-passing dataset retains its own clean result.

G.3 Shared-map calibration and matched controls

The checkpoint evaluation uses 50 held-out goals and three attack-calibration seeds, with source model seeds where available. Top-one replay uses 1,024 legal action sequences per goal; native CEM retains its original pools and elite count. The displacement budget is $\epsilon = 0.07$, with 3,000 forward-equivalent evaluations per method. The construction, reference-budget integral, and solver defaults are specified in Section D. Calibration determines one plane and one kernel profile, which are then fixed across scoring calls. Model data, calibration, task evaluation, and any detector calibration are separated by episode or layout. CEM, GD/MPC, and MPPI retain their native candidate counts, horizons, replanning schedules, and score definitions.

A matched map comparison uses identical cached endpoints, goals, and candidate actions. Random profiles and random-direction search share the target method's calibration sites, plane, kernel coordinates, nonlinear feasibility correction, and call allowance. The proposal distribution, rejected-trial accounting, and stopping rules are given in Sections D.4 and D.5. Reference displacement D_γ and displacement under a learned endpoint or goal law are different quantities; each expectation is taken under its stated distribution.

G.4 Native-boundary calibration and planning diagnostics

Calibration bank and pair extraction. The calibration bank retains the native CEM iterations of the first planning call for each calibration episode in chronological order. Each entry contains the complete action pool, clean scores, endpoints, goal, old proposal moments, and elite count. Episode identities and pool ordering are fixed before coefficient fitting.

Algorithm 2, line 1, sorts a pool by clean cost with candidate index as the stable tie-breaker. It takes positions k and $k+1$ and retains their triple when the gap exceeds $10^{-8}(1 + |C(\mathbf{a}^-)| + |C(\mathbf{a}^+)|)$. Its output is $(z_i^+, z_i^-, z_{g,i})$ with the source action indices. The retained triples determine the plane and kernel sites before fitting. GD-MPC uses the clean best candidate and the next strictly higher-cost candidate under the same tolerance.

Matching calibration rules. A top-one versus elite-boundary comparison uses the intersection of pools with valid pairs under both rules. Goals, model, budget, solver, and native elite count are shared; each rule constructs its plane and kernel from its own triples. For $k = 1$, the rules coincide. Within a rule, the random controls share its triples and interpolation coordinates as specified in Section D.5.

Selection and native updates. For clean and mapped scoring of one fixed pool, the selected sets E, E_T define elite replacement and mean shift by

$$\frac{|E_T \setminus E|}{k}, \qquad \|\boldsymbol{\mu}_T^+ - \boldsymbol{\mu}^+\|_2. \tag{91}$$

The signed mean difference is given by (42); entering and exiting action vectors determine its direction and cancellation. Boundary FlipRate is the fraction of valid native boundary pairs with

negative mapped margin. Top-one replay uses its own pair and denominator. An action discrepancy applies the planner's native extraction rule at the end of a planning call; intermediate rounds supply proposal-moment diagnostics.

Finite-round propagation. The coupled diagonal-CEM process in Equations (46) and (47) uses common base noises, the same action domain, and the same smoothing and variance-floor rules. Clean boundary gaps and occupancy counts, exact score changes, and paired candidate displacements determine the envelope in Equations (50) and (51). Its cost-Lipschitz constant bounds the scorer over the action domain; a maximum over finitely sampled derivatives alone does not establish that domain-wide bound. Within-pool identities isolate direct scoring effects, while moment differences between the coupled runs also include propagation through resampling.

G.5 Metric definitions and uncertainty

Paired task utility. For clean and attacked success indicators $s_i^0, s_i^T \in \{0, 1\}$ on the same goal and seed, define

$$\mathrm{SR}_T = \frac{100}{n}\sum_i s_i^T, \quad \Delta\mathrm{SR} = \frac{100}{n}\sum_i (s_i^0 - s_i^T), \quad \mathrm{Harm}_{|\mathrm{success}} = \frac{100\sum_i s_i^0(1-s_i^T)}{\sum_i s_i^0}. \tag{92}$$

The final ratio requires a positive clean-success denominator. Clean-failure-to-success transitions contribute with the opposite sign to $\Delta\mathrm{SR}$, whereas conditional harm counts only clean successes. These paired metrics depend on the joint outcomes, which a source-reported clean mean alone does not determine. Absolute return differences remain defined when a relative-return denominator is zero.

Ranking and geometry. Pairwise FlipRate uses clean-positive margins. Top-one replay compares the clean best candidate with the next strictly higher-cost candidate; native CEM calibration uses its inclusion-boundary pair. The relative tie tolerance is $10^{-8}(1+|C(a^-)|+|C(a^+)|)$. Clean-coordinate regret evaluates the mapped selection using the original scorer. Ambient distortion $\|G_T - I\|_F$ and action-side positive-spectrum conditioning measure different objects: the latter depends on the restriction of the metric to directions reached by J.

Marginal detection. An RBF-kernel MMD statistic compares a learned law q with $T_{\#}q$ (Gretton et al., 2012). Its bandwidth and rejection threshold are determined from independent calibration samples. The audit uses 2,048 samples per batch and 500 test batches at a nominal 1% false-positive rate; true-positive rates refer to this fixed test. Endpoint and goal distributions define separate comparisons. Episode-block sampling accounts for temporal dependence. This learned-law comparison differs from the Gaussian characteristic-function diagnostic in Section F, which tests reference-law agreement. Clean–mapped correspondences or temporal pairs provide additional information beyond a marginal sample.

Source uncertainty. Finite-run uncertainty can alter conclusions drawn from aggregate reinforcement-learning scores (Agarwal et al., 2021); source dispersion and task averages therefore retain their distinct meanings here. The Temporal Straightening tables report standard deviations; PLDM Table 2 reports standard errors over three seeds. LeWM Table 10 leaves its dispersion type unspecified, while Table 5 calls it variance. The source-reported entries retain those labels. Main-table row averages are descriptive means of source task scores; the corresponding uncertainties depend on their joint sampling structure. The controlled numerical studies retain their own seeds and aggregation rules in Sections B.6 and F.

H Results Reproduced in Prior Work

The following tables retain the published measurement contexts. They reuse the verified literature inventory while joining compatible rows and avoiding artificial missing cells. Tables with different objectives or metrics are not one leaderboard. Source metadata and every original numerical field remain in the accompanying CSV/JSON ledger.

H.1 DINO-WM: JOINT MODEL AND REPRESENTATION TABLE

Table 6: Original DINO-WM Tables 1–2 merged and deduplicated (Zhou et al., 2025).

Model / representation	Maze	Wall	Reach	PushT	Rope	Granular
IRIS	0.74	0.04	0.18	0.32	1.11	0.37
DreamerV3 (offline adaptation)	1.00	1.00	0.64	0.30	2.49	1.05
TD-MPC2 (offline adaptation)	0.00	0.00	0.00	0.00	2.52	1.21
R3M	0.94	0.34	0.40	0.42	1.13	0.95
ResNet	0.98	0.12	0.06	0.20	1.08	0.90
DINO CLS	0.96	0.58	0.60	0.44	0.84	0.79
DINO-WM	0.98	0.96	0.92	0.90	0.41	0.26

First four metrics are SR (0–1, higher is better); last two are Chamfer distance (lower is better), not percentages. Offline DreamerV3 and TD-MPC2 adaptations are not their original online agents. DINO-WM and DINO patch are the same shared row and appear once. The source tables report no uncertainty.

H.2 STRAIGHTENING: COMPLETE SHORT- AND LONG-HORIZON COMPARISONS

Table 7: Complete Temporal Straightening Table 1 (Wang et al., 2026). O: open-loop; M: MPC. Every cell is the source mean ± standard deviation over three data-sampling seeds (SR, %).

Representation	Wall O	Wall M	UMaze O	UMaze M	Medium O	Medium M	PushT O	PushT M
CLS	28.67 ±12.68	66.67 ±10.50	25.33 ±0.94	82.67 ±9.98	20.00 ±8.16	67.50 ±3.54	19.33 ±8.22	28.00 ±1.63
Global DINO+proj	28.67 ±0.94	76.00 ±4.90	34.67 ±1.89	79.33 ±2.49	18.00 ±1.63	46.00 ±3.27	2.00 ±1.63	11.33 ±3.40
Global DINO+proj+TS	42.00 ±3.27	56.67 ±4.11	38.67 ±3.40	96.00 ±0.00	22.67 ±5.73	78.00 ±2.83	5.33 ±3.40	14.67 ±0.94
Global ResNet	4.67 ±3.40	10.00 ±1.63	82.00 ±8.49	96.00 ±2.83	66.00 ±2.83	91.33 ±0.94	2.00 ±2.83	29.33 ±3.40
Global ResNet+TS	84.00 ±7.12	100.00 ±0.00	52.00 ±6.53	86.67 ±0.94	54.00 ±7.12	98.00 ±0.00	19.33 ±3.40	48.67 ±4.99
Spatial DINO raw	52.67 ±5.73	76.67 ±6.18	35.33 ±4.11	80.67 ±6.18	40.83 ±10.07	76.67 ±5.14	56.00 ±4.32	66.00 ±4.90
Spatial DINO+proj	80.00 ±7.12	90.67 ±3.77	44.00 ±7.12	81.33 ±6.80	72.00 ±4.32	96.67 ±0.94	70.00 ±1.63	78.67 ±0.94
Spatial DINO+proj+TS	90.67 ±0.94	100.00 ±0.00	94.00 ±1.63	100.00 ±0.00	82.67 ±3.77	98.67 ±0.94	77.33 ±6.18	85.33 ±4.99
Spatial ResNet	1.33 ±1.89	6.67 ±1.89	14.67 ±4.99	66.00 ±9.09	18.67 ±4.11	57.33 ±4.71	71.33 ±7.36	70.67 ±10.50
Spatial ResNet+TS	84.67 ±2.49	100.00 ±0.00	64.67 ±8.38	98.67 ±1.89	80.67 ±0.94	99.33 ±0.94	70.67 ±0.94	91.33 ±2.49

The projector-only and projector+TS rows use the same $14 \times 14 \times 8$ representation; raw DINO uses $14 \times 14 \times 384$. Global settings are 1×384. Their dimensionality and scoring layouts are not pooled. Source PDF p.8.

Table 8: Temporal Straightening Table 2: both scoring blocks in one table (Wang et al., 2026).

Representation / score	PushT O	PushT M	Medium O	Medium M
Raw spatial DINO / spatial	3.33 ± 2.36	27.33 ± 6.66	35.00 ± 2.35	65.33 ± 3.13
Spatial DINO+proj; no / spatial	6.67 ± 3.77	26.67 ± 9.98	60.00 ± 3.27	72.00 ± 0.00
Spatial DINO+proj; yes / spatial	13.33 ± 3.77	24.00 ± 6.53	68.00 ± 8.64	88.00 ± 3.27
Spatial ResNet; no / spatial	13.33 ± 3.77	29.33 ± 9.43	14.67 ± 6.80	48.00 ± 9.80
Spatial ResNet; yes / spatial	10.67 ± 4.99	33.33 ± 4.99	76.00 ± 6.53	98.67 ± 1.89
Spatial DINO+proj; yes / combined	20.00 ± 0.00	33.33 ± 4.16	66.67 ± 7.57	92.00 ± 5.29
Spatial ResNet; yes / combined	13.33 ± 4.62	36.00 ± 5.29	68.67 ± 4.16	98.67 ± 1.15

Goals are 50 environment steps away. Combined score adds 0.1 times aggregated-feature cost to spatial cost. Source mean ± std, SR (%). Spatial and combined rows remain separate scoring conditions.
DINO+projector+TS has PushT MPC 24.00 versus 26.67 without TS in the spatial block; this unfavorable case is retained.

H.3 LeWM: COMMON TASK INVENTORY AND SOLVER SETTINGS

Table 9: Complete printed means in LeWM Figure 6 (Maes et al., 2026).

Task / model and input	SR (%)
Two-Room / LeWM	87
Two-Room / DINO-WM + proprioception	100
Two-Room / PLDM	97
Two-Room / DINO-WM pixels	100
Two-Room / GCBC	100
Two-Room / GCIQL	100
Two-Room / GCIVL	100
Two-Room / Random	0
Reacher / LeWM	86
Reacher / PLDM	78
Reacher / DINO-WM	79
Reacher / Random	10
PushT / LeWM	96
PushT / DINO-WM + proprioception	92
PushT / PLDM	78
PushT / DINO-WM pixels	74
PushT / GCBC	75
PushT / GCIQL	20
PushT / GCIVL	33
PushT / Random	2
Cube / LeWM	74
Cube / PLDM	65
Cube / DINO-WM	86
Cube / GCBC	84
Cube / GCIQL	64
Cube / GCIVL	56
Cube / Random	48

Only numbers printed in the figure are transcribed; no uncertainty is reconstructed from error bars. Proprioception is retained where explicitly labeled. Reacher and Cube DINO inputs retain the source convention. This task-by-model inventory explains the source selection in main Table 1.

Table 10: LeWM Table 10, with its original reported dispersion (Maes et al., 2026).

Optimizer	LeWM SR	PLDM SR
CEM	96.00 ± 2.83	78.00 ± 5.00
SGD	26.00 ± 4.32	4.67 ± 0.06
RMSProp	67.33 ± 2.49	49.33 ± 8.26
Adam	84.00 ± 7.12	80.00 ± 3.27

PushT success rates in percent. The table does not identify the $\pm$ dispersion as standard deviation or standard error; it is copied without reinterpretation.

Table 11: LeWM Table 5 retained independently (Maes et al., 2026).

Model	PushT SR
DINO-WM	92.00 ± 1.63
PLDM	78.00 ± 5.00
LeWM	96.00 ± 2.83

Same 50 goals; target reachable within 25 steps; execution budget 50. The source caption calls its dispersion "variance". Input modality is not restated, so this DINO row is retained separately from the pixels-only Figure-6 baseline.

H.4 PLDM: DATA SUPPORT IS A DISTINCT EXPERIMENTAL FACTOR

Table 12: PLDM Table 2, source-native Two-Room evaluation (Sobal et al., 2025).

Method	Good data	No door passing
CRL	89.30 ± 0.70	14.70 ± 4.10
GCBC	86.00 ± 2.00	8.40 ± 1.20
GCIQL	98.00 ± 0.90	99.60 ± 0.40
HILP	100.00 ± 0.00	100.00 ± 0.00
HIQL	96.40 ± 1.30	26.30 ± 5.60
PLDM	97.80 ± 0.70	34.40 ± 2.70

SR (%), mean $\pm$ standard error over three seeds; 64×64 observations and up to 200 environment steps. PLDM uses MPPI with accumulated goal distance and uncertainty. No-door data is a dynamics-support stress test, not an interface attack.

H.5 EXTERNAL ATTACKS: JOINED SOURCE TABLES, DISTINCT ACCESS MODEL

Table 13: TRAP Tables 1–2 joined by victim and task: attack outcomes (Duan et al., 2026).

Victim / task	Random drop	Random ASR	TRAP drop	TRAP ASR
DreamerV3 / Crafter	-0.10 ± 3.50	43.90 ± 8.40	63.20 ± 0.90	98.10 ± 0.80
DreamerV3 / cheetah-run	-0.40 ± 4.20	45.80 ± 6.30	22.80 ± 2.30	99.60 ± 0.50
DreamerV3 / dog-run	-0.30 ± 3.80	50.00 ± 6.10	9.80 ± 3.80	77.20 ± 3.70
DreamerV3 / humanoid-walk	3.10 ± 6.40	65.30 ± 5.50	69.80 ± 0.70	100.00 ± 0.00
DreamerV3 / walker-walk	-0.20 ± 5.80	48.00 ± 7.20	18.50 ± 1.50	100.00 ± 0.00
DreamerV3 / Seaquest	97.20 ± 1.90	100.00 ± 0.00	97.50 ± 0.80	100.00 ± 0.00
DreamerV3 / Pong	151.80 ± 11.60	100.00 ± 0.00	164.50 ± 8.30	100.00 ± 0.00
DreamerV3 / Breakout	97.10 ± 1.40	100.00 ± 0.00	98.80 ± 0.70	100.00 ± 0.00
DreamerV3 / Space Invaders	90.40 ± 3.00	100.00 ± 0.00	93.30 ± 1.80	100.00 ± 0.00
TD-MPC2 / hopper-hop	51.20 ± 4.00	95.00 ± 1.70	99.80 ± 0.30	100.00 ± 0.00
TD-MPC2 / cheetah-run	9.30 ± 5.30	90.00 ± 2.30	98.40 ± 0.40	100.00 ± 0.00
TD-MPC2 / walker-walk	1.90 ± 3.20	64.00 ± 2.50	92.40 ± 0.90	100.00 ± 0.00

Mean drop and ASR are percentages with source standard deviations. ASR means triggered return is below paired clean return; it is not candidate FlipRate. Patch area ratio is 0.09 and pixel ℓ_∞ budget is 64. Atari returns can be negative, allowing relative drop above 100%. The source task is walker-walk, not walker-run.

Table 14: The same TRAP join: clean return, patched return, and evaluation cost (Duan et al., 2026).

Victim / task	Clean return	Patched return	Seconds	Time ratio
DreamerV3 / Crafter	12.91 ± 1.88	4.61 ± 2.22	588.60	0.812
DreamerV3 / cheetah-run	889.44 ± 23.91	656.14 ± 186.41	1808.10	1.010
DreamerV3 / dog-run	368.63 ± 40.53	330.15 ± 32.36	3131.90	1.006
DreamerV3 / humanoid-walk	800.00 ± 34.73	231.77 ± 57.55	2216.20	1.009
DreamerV3 / walker-walk	966.63 ± 15.93	787.79 ± 49.17	1683.20	1.015
DreamerV3 / Seaquest	6611.20 ± 1255.86	165.20 ± 79.73	11386.20	0.177
DreamerV3 / Pong	20.31 ± 0.88	-13.09 ± 3.44	2225.20	0.960
DreamerV3 / Breakout	293.50 ± 153.64	1.95 ± 1.63	2535.70	0.166
DreamerV3 / Space Invaders	8787.85 ± 6591.83	586.10 ± 136.06	2765.90	0.159
TD-MPC2 / hopper-hop	304.23 ± 12.94	0.00 ± 0.02	3663.40	1.270
TD-MPC2 / cheetah-run	635.28 ± 27.16	10.38 ± 4.29	3690.50	1.202
TD-MPC2 / walker-walk	951.23 ± 25.44	72.71 ± 23.38	3845.00	1.214

Returns use source mean $\pm$ std; time is the total for 100 episodes. Early termination can reduce total runtime; ratios below one do not demonstrate faster per-step inference. The ratio of mean returns need not reproduce the mean per-episode return drop. Rows remain external references, not shared-interface comparators.

Table 15: WMAttack Table 1, retained under its own search protocol (Guo et al., 2026).

Victim / search method	Norm. drop	Action flip	Utility	Minutes
DreamerV3 Atari / Claudini	0.497	0.400	-0.132	77.20
DreamerV3 Atari / Random	0.446	0.377	-0.163	73.40
DreamerV3 Atari / WMAttack	1.034	0.660	0.430	106.20
DreamerV3 DMC / Claudini	0.319	0.705	-0.137	55.40
DreamerV3 DMC / Random	0.352	0.721	-0.123	59.00
DreamerV3 DMC / WMAttack	0.682	0.800	0.174	120.20

Version 2, 5 September 2026. DreamerV3 Atari aggregates 26 tasks and four attack families; DMC aggregates 20 tasks and five families. Source-specific normalization, utility and action-flip definitions are retained. No conversion to LatentGauge pairwise FlipRate is made.

I Numerical Provenance

This manuscript combines task results, controlled calculations, and source-reported benchmarks. Controlled calculations use explicit Gaussian samples or the fixed affine endpoint model, with inputs, seeds, and outputs stored alongside their scripts. Source-reported task and attack scores retain their original paper, table, setting, units, and uncertainty convention. The literature ledger preserves 15 protocol blocks, 95 settings, and 326 reported numerical fields.

Task and attack results. Tables 1 and 2 report clean and mapped task scores, shared-map controls, and calibration comparisons under the protocols in Section G. Row averages summarize the tasks within each protocol block. Original source tables remain in Section H; exact mechanism calculations retain their recorded inputs and outputs.

Controlled calculations. `data/results.json` stores the Gaussian audit, displacement sweep, and fixed-map checks. The exact contours and optimization coordinates are regenerated by `scripts/figure_data.py`. The fixed-pool selection checks are in `data/cem_identity_check.json`, and `data/cem_propagation_check.json` contains every round of the coupled CEM runs. Final-round moment differences are averaged over the 64 independent streams for each map; changed-set counts use all 512 rounds for that map. These denominators describe different quantities and are kept separate.

Source-reported calculations. The original fields are retained in `experiment_data/literature_ledger.csv` and `sources_and_tables.json`. Source task scores retain their original units and precision; within-protocol row averages in Table 1 are descriptive arithmetic on those values. A row average carries no newly estimated uncertainty. The full source tables in Section H retain their published uncertainty labels and task-specific metrics.

J Code and Numerical Artifacts

The accompanying scripts regenerate the Gaussian and affine-model calculations and the exact contour geometry. Numerical figure inputs and LaTeX style files are included, so the manuscript compiles with pdfLaTeX and BibTeX. The numerical inputs, seeds, and execution records are specified in Section F; Sections B.6 and E.3 identify the fixed-pool and multi-round CEM checks.

The source-result ledger records paper versions, tables, metrics, units, and uncertainty conventions for Section H. The shared-map fitting algorithm and its numerical defaults are specified in Section D, and the checkpoint settings, source interfaces, and metric definitions are in Section G.