ATLAS: ALIGNED TRANSPORT OF LATENT STRUCTURE FOR RELIABLE WORLD MODEL PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent world models rely on representation geometry for planning, yet regularizing the latent marginal alone does not determine the state-to-state relationships used for action selection. We show that this can cause planning-relevant novelty structure to be weakened as representations are transformed into the final latent used by the planner. We introduce Aligned Transport of Latent Structure (ATLAS), a training objective that explicitly preserves relational geometry while calibrating the global latent distribution. ATLAS transfers normalized pairwise structure from an informative encoder representation to the planning latent and uses Wasserstein embedding matching (WEMReg) to calibrate its marginal through one-dimensional Wasserstein-2 transport. Our analysis shows that relational preservation and marginal calibration impose non-redundant constraints, and connects finite-candidate planning stability to relational distortion, latent-scale mismatch, and prediction error. Instantiated in LeWM, ATLAS improves mean goal-reaching success across PushT, TwoRoom, and OGBench-Cube on both lower- and higher-novelty evaluation subsets, with the largest gain on higher-novelty TwoRoom episodes. Representation and rollout diagnostics further show stronger novelty-related structure in the planning latent, improved marginal calibration, and lower multi-step prediction error. Together, these results highlight preservation of planning-relevant latent geometry as an important ingredient for reliable world-model planning. Code is available at `https://anonymous.4open.science/r/atlas-world-model-72C4/`.

1 INTRODUCTION

Latent world models (WMs) plan by predicting where different candidate actions will lead and selecting the action whose predicted outcome is closest to a desired goal (Hafner et al., 2019; Hansen et al., 2024). The geometry of the learned representation space is therefore part of the planning problem: states that are behaviorally similar should remain appropriately close, while states that differ in important ways should remain distinguishable. This becomes especially important when planning encounters out-of-distribution (OOD) states that are rare or poorly represented in the offline training data.

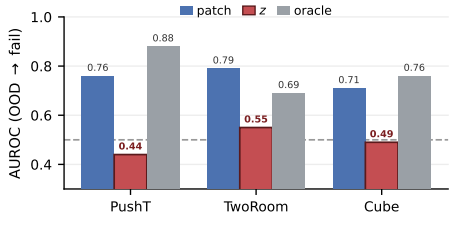


Figure 1: **An OOD-related diagnostic gap in latent planning on PushT.**

Many recent WMs learn these representations by predicting future latent features rather than reconstructing raw observations, an approach closely related to joint-embedding predictive architectures (JEPAs) (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). A central challenge in such models is preventing representation collapse, where different inputs are mapped to nearly identical features (Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Bardes et al., 2024). Existing approaches address this through anti-collapse regularization that constrains statistical properties of the latent representation, such as its variance, covariance, or overall distribution (Bardes et al., 2022; Balestriero & LeCun, 2025). For example, SIGReg encourages latent projections to follow a standard Gaussian distribution (Balestriero & LeCun, 2025). These constraints can keep the representation non-

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.

degenerate, but they do not specify which individual states should remain close or far apart. Two latent spaces can therefore have similarly well-behaved global distributions while preserving substantially different relationships between states. For planning, this distinction matters because these relationships directly determine how predicted outcomes are compared with the goal.

We observe this gap in LeWM (Maes et al., 2026b), a recent JEPA-style WM. We measure how well distance from the training data predicts downstream planning failure at different stages of the learned representation. This signal is substantially stronger in the encoder's mean-pooled patch features than in the final latent used for planning (Figure 1). On PushT (Florence et al., 2022), for example, failure-prediction AUROC decreases from 0.76 in the encoder representation to 0.44 in the planning latent. Thus, information indicating that a state is unfamiliar is present within the encoder but substantially weakened in the representation ultimately used for planning. This suggests that maintaining a well-behaved latent distribution alone does not ensure preservation of planning-relevant state relationships.

We therefore introduce Aligned Transport of Latent Structure (ATLAS), which explicitly targets two complementary properties of the planning representation. The first is relational preservation: state-to-state relationships present in an informative encoder representation should be retained in the planning latent. Our OOD-recovery objective therefore transfers normalized pairwise distances from the encoder's mean-pooled patch features into the planning representation. The second is marginal calibration: the overall latent distribution should remain well-behaved and resistant to collapse. For this purpose, we introduce Wasserstein embedding matching (WEMReg), which matches one-dimensional projections of the latent distribution to a standard Gaussian using Wasserstein-2 distance (Bonneel et al., 2015). Together with the predictive objective, ATLAS targets three complementary properties of a planning representation: preserving state relationships, maintaining a stable latent distribution, and accurately predicting future states.

Our theoretical analysis shows that relational preservation and marginal calibration impose non-redundant constraints on the latent representation. We further relate representation distortion and rollout prediction error to the stability of finite-candidate goal-conditioned planning. Across three benchmark datasets, ATLAS improves mean planning success over LeWM on both in-distribution (ID) and OOD episodes, with the largest improvement occurring on OOD episodes. Controlled ablations and representation- and rollout-level diagnostics further support the complementary roles of relational preservation and WEMReg (Sections 4 and 5). Overall, our contributions are:

- We show that a WM can retain useful OOD-related information in its encoder while losing much of that signal in the latent representation used for planning, exposing a gap between anti-collapse regularization and preservation of planning-relevant geometry.
- We introduce ATLAS, a latent WM that explicitly preserves planning-relevant relational structure while calibrating the global latent geometry through Wasserstein transport.
- We connect representation distortion and prediction error to planning stability, and evaluate ATLAS across manipulation and navigation tasks using controlled ablations and representation- and rollout-level diagnostics.

2 Related Work

Latent WMs for planning. A growing line of work learns WMs in a joint-embedding predictive latent space and plans by rolling that latent forward. LeWM trains a ViT encoder and an autoregressive latent predictor end-to-end from pixels and plans by latent Cross-Entropy search (Maes et al., 2026b). DINO-WM instead fixes a frozen DINOv2 encoder pretrained on external images and learns dynamics on top of it (Zhou et al., 2025), PLDM learns latent dynamics from reward-free offline data (Sobal et al., 2025), and FAST-WM is a faster LeWM variant (Gao & Xu, 2026). Where these methods change the encoder or the data, ATLAS changes only the training objective of a from-scratch model, which makes it a controlled modification rather than a different backbone.

A fundamental challenge in latent WM is that learned representations can collapse. We need an anti-collapse mechanism, so the encoder does not map every input to the same point. VICReg does this with explicit variance and covariance penalties on the embedding (Bardes et al., 2022), while LeJEPA regularizes the embedding toward an isotropic Gaussian through a characteristic-function

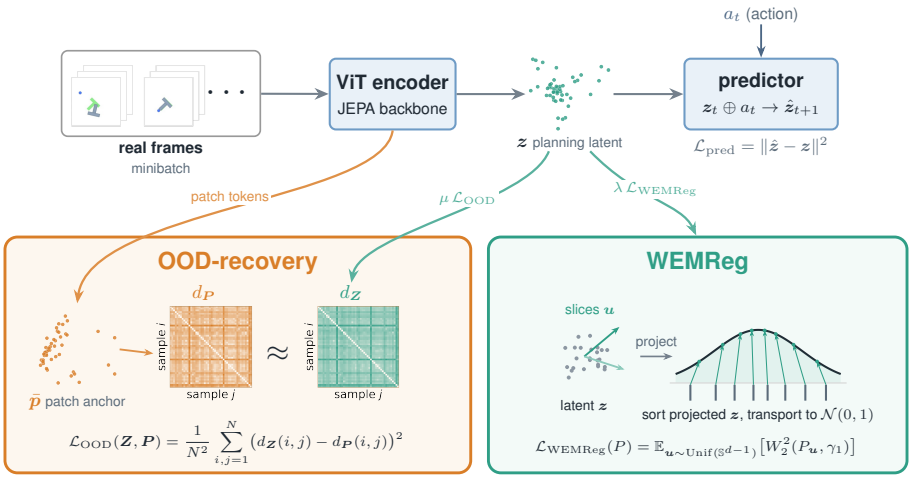


Figure 2: **Overview of ATLAS.**

normality test evaluated at a finite set of frequencies (Balestriero & LeCun, 2025), the SIGReg mechanism LeWM adopts (Maes et al., 2026b). Our diagnosis is that this per-sample Gaussianization, while effective at preventing collapse, also removes the OOD structure a planner needs. ATLAS keeps the anti-collapse goal but reaches it with a Wasserstein embedding matching regularization term, which we show is both lower in design constants and strictly more discriminative than finite-frequency matching (Section 3).

Optimal transport and sliced distances. Matching a distribution to a target through one-dimensional projections underlies the sliced Wasserstein distance (Bonneel et al., 2015), which replaces an intractable high-dimensional transport problem with an average of exact one-dimensional ones. We use this device in the specific form of matching each projected marginal to a standard Gaussian by its exact quantile transport, which yields a closed-form batch objective and, unlike a finite-frequency test, penalizes every projected quantile residual (Section 3.1).

OOD detection and relational distillation. Distance to a bank of training features is a standard nonparametric OOD signal: a k-nearest-neighbor distance in representation space separates in- from out-of-distribution inputs without a parametric density model (Sun et al., 2022). We use this score both as the diagnostic that reveals the washed-out planning latent (Section 1) and as the quantity our theory relates to planning stability (Section 3). To restore it we transfer the relational geometry of a stronger representation into a weaker one, in the spirit of relational knowledge distillation, which matches pairwise structure rather than absolute activations (Park et al., 2019). Here the roles are internal to a single model: the encoder's own patch tokens act as a fixed anchor whose pairwise-distance geometry the planning latent is taught to preserve.

3 ATLAS

Planning with a latent WM requires a representation that supports accurate prediction under candidate actions, remains well-scaled and non-degenerate, and preserves state-to-state relationships relevant to goal comparison. Standard marginal regularization addresses only the distributional requirement and does not determine which states should remain close or far apart.

ATLAS is designed to address these three requirements (Figure 2). The predictive objective learns latent dynamics, WEMReg calibrates the global distribution of the planning latent, and OOD-recovery preserves relational structure from an informative encoder representation. Following the latent predictive architecture of LeWM (Maes et al., 2026b), a trainable ViT encoder maps each observation o_t to a d-dimensional planning latent, $z_t = E_\theta(o_t)$, and a predictor models its evolution under actions:

$$\widehat{z}_{t+1} = P_\phi(z_{t-H+1:t}, a_{t-H+1:t}), \tag{1}$$

where the input contains the previous H latent–action pairs and a_t denotes the action at time t.

3.1 CALIBRATING THE LATENT DISTRIBUTION

A useful planning representation should remain well-scaled and non-degenerate. Existing anti-collapse methods impose this through distributional constraints on the latent space. We propose Wasserstein embedding matching (WEMReg), which directly compares the projected distributions (Tolstikhin et al., 2019; Kolouri et al., 2018). For a latent distribution P with finite second moments, we define

$$\mathcal{L}_{\mathrm{WEMReg}}(P) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{S}^{d-1})} W_2^2(P_{\boldsymbol{u}}, \mathcal{N}(0,1)), \tag{2}$$

where $\boldsymbol{u}$ is a randomly sampled unit direction and $P_{\boldsymbol{u}}$ denotes the distribution of the scalar projection $\boldsymbol{u}^\top \boldsymbol{z}$ for $\boldsymbol{z} \sim P$. Each direction provides a one-dimensional view of the d-dimensional latent distribution, and W_2^2 measures how far that projected distribution is from a standard Gaussian. Averaging over random directions therefore encourages Gaussian structure across arbitrary directions, rather than only along individual coordinates. In practice, we approximate the expectation using L random directions per minibatch, where L is the number of one-dimensional projections being averaged. For each direction, we sort the projected latent values and match them to the corresponding standard-Gaussian quantiles. The resulting one-dimensional Wasserstein cost is therefore determined by the discrepancy between the ordered projected samples and their Gaussian targets.

LeWM instead uses SIGReg for marginal regularization, which tests Gaussianity by matching characteristic functions at a finite set of frequencies. WEMReg targets the same marginal calibration objective, but compares the full projected distributions through Wasserstein distance. This distinction matters because agreement at finitely many characteristic-function frequencies does not uniquely determine a distribution: non-Gaussian discrepancies may remain undetected even when the mean and covariance are correct. We formalize this separation in Theorem 1; the precise finite-frequency objective and proof are provided in the Appendix C.

Theorem 1 (Finite-frequency ambiguity)**.** *For any finite set of characteristic-function frequencies with positive weights, there exists a non-Gaussian distribution P with zero mean and identity covariance that is indistinguishable from a standard Gaussian under the corresponding finite-frequency objective, while $\mathcal{L}_{\mathrm{WEMReg}}(P) > 0$.*

Thus, finite-frequency matching can miss non-Gaussian structure that remains visible to projected Wasserstein matching.

3.2 PRESERVING RELATIONAL GEOMETRY

Marginal calibration controls the overall latent distribution, but it does not determine which states should remain close or far apart. We therefore introduce a relational objective that transfers state-to-state geometry from an informative encoder representation to the planning latent.

For frame i, let $\bar{\boldsymbol{p}}_i = K^{-1}\sum_{k=1}^{K} \boldsymbol{p}_{i,k}$ denote the mean of its K final-layer patch tokens. As shown in Section 1, novelty measured in this representation is substantially more predictive of downstream planning failure than novelty measured in the LeWM planning latent. We therefore use the mean-pooled patch representation as a relational anchor. For a minibatch of N samples, let $\boldsymbol{Z} = [\boldsymbol{z}_i^\top]_{i=1}^N$ denote the planning latents and $\boldsymbol{P} = [\bar{\boldsymbol{p}}_i^\top]_{i=1}^N$ the corresponding patch representations. For either representation $\boldsymbol{X}$, we first standardize each coordinate by its batch standard deviation and compute pairwise Euclidean distances. We then divide all pairwise distances by their batch mean, removing differences in overall scale. Denoting the resulting normalized distance between samples i and j by $d_{\boldsymbol{X}}(i,j)$, we define

$$\mathcal{L}_{\mathrm{OOD}}(\boldsymbol{Z},\boldsymbol{P}) = \frac{1}{N^2}\sum_{i,j=1}^{N} \left(d_{\boldsymbol{Z}}(i,j) - d_{\boldsymbol{P}}(i,j)\right)^2. \tag{3}$$

When computing this loss, $\boldsymbol{P}$ is treated as stop-gradient. Thus, the patch representation provides a fixed target for each update, while still evolving over training through the shared encoder.

Intuitively, $\mathcal{L}_{\mathrm{OOD}}$ does not force the planning latent to reproduce the patch features themselves. Instead, it preserves their relative geometry: states that are far apart in the anchor representation are encouraged to remain far apart in the planning latent, and similarly for nearby states (Tung & Mori,

2019; Peng et al., 2019; Park et al., 2019). WEMReg and relational preservation therefore constrain different aspects of the representation.

Proposition 2 (Non-redundant constraints)**.** *For a fixed anchor, two representations can have the same WEMReg loss but different relational losses, or the same relational loss but different WEMReg values.*

Thus, neither objective determines the other: $\mathcal{L}_{\mathrm{WEMReg}}$ constrains the global latent distribution, whereas $\mathcal{L}_{\mathrm{OOD}}$ constrains normalized state-to-state geometry relative to the encoder anchor.

The complete ATLAS objective is $\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda\mathcal{L}_{\mathrm{WEMReg}} + \mu\mathcal{L}_{\mathrm{OOD}}$, where $\mathcal{L}_{\mathrm{pred}}$ is the latent prediction loss and $\lambda, \mu > 0$ control the two representation objectives. The three terms respectively target accurate dynamics prediction, marginal calibration, and preservation of relational geometry.

3.3 Finite-candidate planning stability

At test time, ATLAS uses the same goal-conditioned latent planner as LeWM. Given a goal image o_g, CEM searches over candidate action sequences and rolls each candidate forward for T prediction steps(De Boer et al., 2005). Candidates are scored by the distance between their predicted terminal latent $\widehat{\boldsymbol{z}}_T$ and the encoded goal $\boldsymbol{z}_g$,

$$|\widehat{\boldsymbol{z}}_T - \boldsymbol{z}_g|_2^2, \qquad \boldsymbol{z}_g = E_\theta(o_g). \tag{4}$$

CEM repeatedly refines its sampling distribution toward low-cost candidates, executes a prefix of the selected sequence, and replans from the next observation.

We now ask when this finite-candidate selection remains stable. Three factors can perturb the planner: inaccurate terminal predictions, distortion of state-to-state geometry, and mismatch in latent scale. Let ϵ_{rel} denote the maximum discrepancy between normalized pairwise distances in the planning latent and the patch-anchor representation, and let e denote the maximum terminal prediction error over the candidate set. We additionally use η to quantify latent-scale mismatch. Precise definitions of these quantities and the fixed normalization constants are given in Appendix C.1.

Theorem 3 (Planning stability)**.** *For a common reference bank, the mean k-nearest-neighbor novelty scores s computed in the planning and anchor representations satisfy*

$$|s_{\boldsymbol{Z}}(q) - s_{\boldsymbol{P}}(q)| \le \epsilon_{\mathrm{rel}}. \tag{5}$$

Moreover, for a fixed finite candidate set and bounded latent-scale mismatch, any candidate selected using predicted latent distance has anchor-geometry regret bounded by

$$2\left[\frac{e}{\ell} + (1+\eta)\epsilon_{\mathrm{rel}} + \eta M\right], \tag{6}$$

where ℓ is the latent distance scale and M bounds the candidate-to-goal distance in the anchor geometry.

The theorem connects the ATLAS objectives directly to planning stability. The prediction loss reduces terminal rollout error e, the relational objective reduces geometry distortion ϵ_{rel}, and WEMReg controls latent-scale mismatch η. Together, these terms bound how far the action selected using predicted latent distances can deviate from the action preferred under the anchor geometry. The kNN result additionally shows that reducing relational distortion preserves the novelty structure that motivates our relational objective. Full assumptions, definitions, and proofs are provided in Appendix C.1.

4 Experiments

Datasets. We evaluate on three offline control tasks spanning navigation and manipulation, with different observation structures and dynamics. **TwoRoom** (Maes et al., 2026a) is a 2D navigation task in which an agent must move between two rooms through a narrow opening to reach a target position. **PushT** (Florence et al., 2022; Chi et al., 2023) is a contact-rich 2D manipulation task that requires pushing and rotating a T-shaped block into a target pose. **OGBench-Cube** (Park et al., 2025) extends the evaluation to 3D robotic manipulation, where a robot arm moves a cube toward a

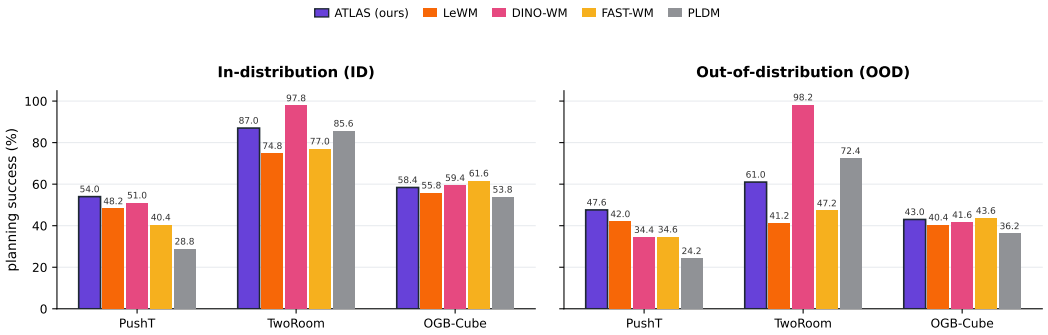


Figure 3: **Planning success at goal offset** 50 **on lower- and higher-novelty evaluation subsets.** We refer to the two median-split subsets as ID and OOD, respectively. Bars show mean success over five evaluation seeds. The split is determined by a model-independent expert-path novelty score based on distance to training states. Exact values are reported in Table 3.

target configuration. Together, the tasks probe latent planning across navigation, contact-rich planar manipulation, and visually richer robotic manipulation. Each model is trained on the corresponding offline demonstration data; dataset details are given in Appendix A.

Baselines. We compare ATLAS with four latent world models representing different representation-learning and regularization choices. **LeWM** (Maes et al., 2026b) is the closest controlled comparison: ATLAS uses the same general JEPA-style encoder–predictor setup while replacing the marginal regularizer and adding relational preservation. **DINO-WM** (Zhou et al., 2025) represents a different representation-learning regime, learning latent dynamics on top of a frozen DINOv2-S/14 encoder pretrained on large-scale external image data. Following the comparison protocol of Maes et al. (2026b), we remove proprioceptive inputs and retrain its trainable components on each task for 10 epochs. **FAST-WM** (Gao & Xu, 2026) is an efficiency-oriented LeWM variant and is evaluated from its released checkpoints with the original encoder and predictor. **PLDM** (Sobal et al., 2025) provides an alternative JEPA-style training objective based on VICReg-derived regularization; we retrain it on the same task data for the same 10-epoch budget. The comparison to LeWM therefore isolates the effect of the ATLAS training objective most directly, whereas the remaining baselines provide broader reference points with different encoders, objectives, or checkpoint provenance.

Settings. We evaluate all methods under a common goal-conditioned planning protocol, using the same task environments, planning objective, and success criterion. Each WM is trained offline and used at test time to score candidate action sequences through the terminal latent goal-matching objective in Section 3. We use receding-horizon CEM with 300 candidates, 30 refinement iterations, and 30 elites. ATLAS, LeWM, and DINO-WM are evaluated on identical sampled episodes within each evaluation condition; FAST-WM and PLDM use their released or saved episode draws from the same task distributions and are evaluated with the same protocol and episode-partitioning rule.

Planning success is defined as the fraction of episodes in which the simulator reaches the target configuration. Unless stated otherwise, the goal is selected 50 steps ahead along the demonstration trajectory. We evaluate 200 episodes for each of five evaluation seeds (42–46) and report the mean and standard deviation across these seeds. For ATLAS, we use λ=3.0 for WEMReg and μ=0.1 for OOD-recovery across all tasks. Complete training, evaluation, and planning hyperparameters are provided in Appendix A.

4.1 Main Results

Figure 3 reports planning success at goal offset 50. For each task and evaluation seed, we split the 200 episodes into lower- and higher-novelty halves based on their distance from the training-state support. Specifically, each episode receives a novelty score given by the mean k-nearest-neighbor distance between states along the expert trajectory and a fixed bank of training states, and episodes are split at the median. We refer to the two subsets as in-distribution (ID) and OOD, respectively. Full details and exact values are provided in Appendix A.1 and Table 3 in the Appendix.

Table 1: 2×2 ablation at goal offset 50 over marginal calibration (SIGReg vs. WEMReg) and relational preservation (OOD-recovery off vs. on). Planning success rate (%, mean $\pm$ std over 5 evaluation seeds); highest mean per task in bold.

Arm	PushT	TwoRoom	OGBench-Cube
LeWM (base)	45.10 ± 3.65	58.00 ± 3.36	48.10 ± 2.94
ATLAS w/o WEMReg	50.30 ± 3.85	71.80 ± 2.99	50.50 ± 2.55
ATLAS w/o OOD-recovery	47.00 ± 2.00	73.90 ± 2.46	46.90 ± 2.60
ATLAS	$\mathbf{50.80 \pm 0.68}$	$\mathbf{74.00 \pm 1.92}$	$\mathbf{50.70 \pm 2.69}$

Across all three tasks, ATLAS improves mean planning success over LeWM on both ID and OOD episodes. The gains are modest on PushT and OGBench-Cube and substantially larger on TwoRoom, where ATLAS improves from 74.8% to 87.0% on ID episodes and from 41.2% to 61.0% on OOD episodes. The largest improvement therefore occurs on higher-novelty TwoRoom episodes. Importantly, the gains are not restricted to OOD states: ATLAS also improves over LeWM on the lower-novelty subset in every task. Performance relative to the broader baseline set is task dependent. ATLAS achieves the highest mean success on both novelty subsets in PushT, while several methods are similar on OGBench-Cube. On TwoRoom, DINO-WM performs substantially better, but it uses a frozen DINOv2-S/14 encoder pretrained on roughly 142M external images, whereas ATLAS is trained from task data without external visual pretraining(Oquab et al., 2024). We therefore treat DINO-WM as a reference for the benefit of a strong pretrained representation rather than as a controlled objective-level comparison. Other baselines, including PLDM and FAST-WM, are competitive on particular tasks, underscoring that the most direct comparison for isolating the effect of ATLAS is with LeWM and the controlled ablations in Section 4.2.

4.2 Ablation study

We study the roles of marginal calibration and relational preservation with a 2×2 ablation over SIGReg, WEMReg $\times$ OOD-recovery off, on. All four variants use the same architecture, data, training budget, and planner, and differ only in these objective terms. LeWM corresponds to SIGReg without OOD-recovery, while ATLAS uses WEMReg together with OOD-recovery. Table 1 reports planning success at goal offset 50.

The two objectives affect planning differently across tasks. On TwoRoom, either WEMReg or OOD-recovery alone produces a large improvement over LeWM, while the full ATLAS objective performs similarly to the stronger single-component variant. On PushT, OOD-recovery provides the larger individual gain, and the full model performs best. On OGBench-Cube, OOD-recovery again improves over LeWM, whereas WEMReg alone does not; the full objective recovers the strongest mean performance. Overall, ATLAS attains the highest mean success on all three tasks, although its gain over the stronger single-component variant is small. These results suggest that WEMReg and OOD-recovery constrain different properties of the representation, but their downstream benefits need not be additive and can overlap in the representation bottlenecks they address. This is consistent with Proposition 2, which establishes that neither objective determines the other without implying additive improvements in planning performance.

5 Analysis of latent geometry and planning reliability

The planner in Section 3 selects actions by comparing predicted terminal latents with an encoded goal. Its behavior therefore depends on three properties highlighted by Theorem 3: preservation of state-to-state geometry, calibration of the latent distribution, and accuracy of multi-step latent prediction. We examine each property below. These analyses are intended as mechanism-oriented diagnostics rather than causal explanations of planning success.

ATLAS preserves more OOD-related structure in the planning latent. We first ask whether the failure-associated novelty signal identified in Section 1 remains detectable in the representation used for planning. Using the frozen-model diagnostic from Appendix B, we compute a k-NN novelty score at different representation depths and measure how well it predicts downstream plan-

ning failure. In LeWM, this signal weakens substantially from the patch representation to the final planning latent. Under ATLAS, the planning latent retains much more of this signal, reaching 0.75 compared with 0.77 in the patch representation. Similar improvements are observed on TwoRoom and OGBench-Cube (Figure 8 in Appendix).

This behavior is consistent with the relational objective in Section 3.2, which transfers normalized pairwise geometry from the patch representation to the planning latent without directly supervising OOD labels or novelty scores. Direct measurements of pairwise relational distortion across controlled variants are provided in Figure 11. On PushT, where LeWM shows the largest patch-to-latent degradation, OOD-recovery also produces the larger single-component gain in the ablation study.

The same pattern is visible qualitatively in Figure 4. States are colored by model-independent true-state novelty after projecting the planning latent to two dimensions using PCA. On PushT, ATLAS shows the clearest structure: the in-distribution (blue, low-novelty) states sit close to one another, while the out-of-distribution (red, high-novelty) states lie farther from the rest, so distance in the latent tracks novelty; under LeWM this distance separation is much weaker. On TwoRoom and OGBench-Cube the effect is less pronounced, though the ATLAS projections still place high-novelty states farther from the in-distribution cluster than LeWM does (Figure 9 in Appendix). Because these PCA plots are only two-dimensional projections, we use them as qualitative support for the stronger quantitative k-NN diagnostic rather than as a direct measure of relational preservation.

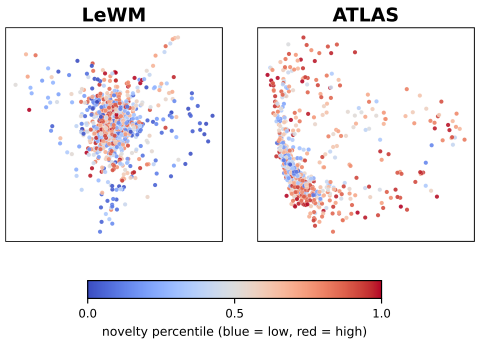


Figure 4: **Planning-latent geometry (PushT).** Two-dimensional PCA projections of z. Each point is a state, colored by its model-independent true-state novelty percentile.

WEMReg improves latent-marginal calibration. Relational preservation and marginal calibration act on different properties of the representation. The former preserves normalized pairwise structure, while the latter controls the global scale and shape of the latent distribution. In Theorem 3, this distinction appears through the scale-mismatch term η. We evaluate marginal calibration on fresh random-policy simulator frames, separate from the ID/OOD split used for planning evaluation. Figure 5 reports mean deviation, covariance error, projected distributional discrepancy, and projected discrepancy after whitening. Across the evaluated tasks, variants using WEMReg improve several of these diagnostics, with especially clear reductions in covariance and projected-distribution errors on TwoRoom. The effect is not uniform across every task and statistic, but overall the results indicate improved latent-marginal calibration under WEMReg.

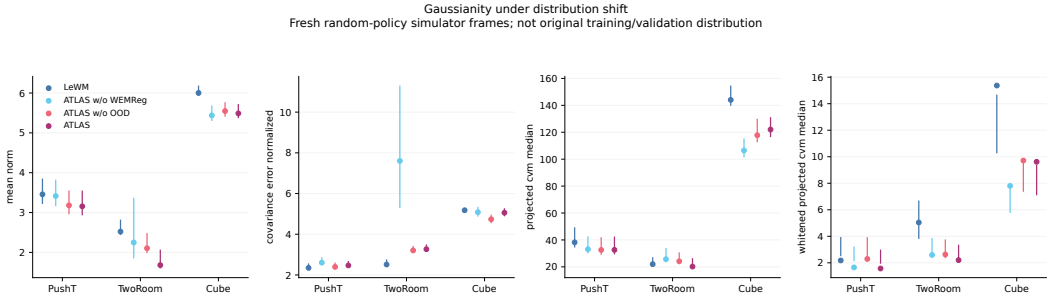


Figure 5: **Latent-marginal diagnostics under distribution shift.** Planning-latent diagnostics on fresh random-policy simulator frames. From left to right, the panels report mean norm, normalized covariance error, median projected CvM discrepancy, and median whitened projected CvM discrepancy.

ATLAS has lower multi-step latent prediction error. The third factor in Theorem 3 is terminal prediction error. Because CEM evaluates candidate actions after repeatedly applying the learned dynamics, small one-step errors can accumulate over the rollout horizon. We therefore measure normalized latent prediction error as the horizon increases on PushT and TwoRoom (Figure 6). ATLAS has lower error than LeWM at every measured horizon, with the gap generally increasing over longer rollouts. This shows that the full ATLAS model provides more accurate multi-step latent predictions on these shared-query evaluations.

Taken together, these diagnostics align with the three factors identified by the planning-stability analysis: OOD-recovery reduces relational distortion, WEMReg improves latent marginal calibration, and ATLAS exhibits lower multi-step prediction error. Additional diagnostics of the planning-stability analysis, including the bound-versus-regret comparison and the decomposition of the prediction, relational, and scale terms in Theorem 3, are provided in Figures 10 and Figure 12.

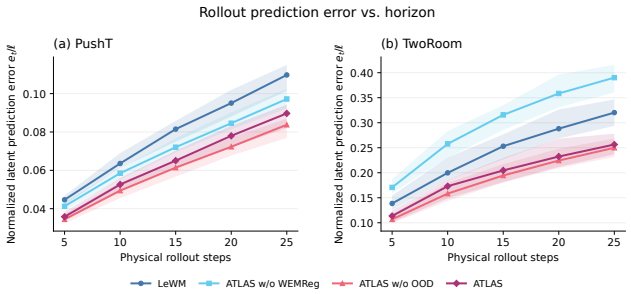


Figure 6: **Multi-step rollout prediction error.** Results on PushT (a) and TwoRoom (b) use 100 shared queries and 128 shared candidates per query. Lines show the across-query median of candidate-median latent prediction errors, normalized by the fixed checkpoint-specific reference-bank scale ℓ. Shading denotes 95% percentile intervals from 1,000 paired query-bootstrap resamples. One prediction step spans five physical steps.

Closed-loop trajectories illustrate the resulting behavior. Figure 7 shows a selected TwoRoom episode in which ATLAS succeeds while LeWM fails. ATLAS passes through the connecting opening and reaches the goal region, whereas LeWM remains on the opposite side of the wall. Analogous PushT and OGBench-Cube examples are shown in Figure 13. These trajectories are qualitative illustrations of the behavior underlying the aggregate success rates rather than representative estimates of typical behavior.

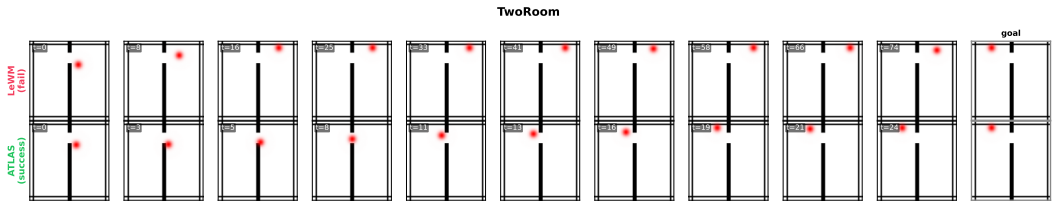


Figure 7: **Selected closed-loop planning trajectory on TwoRoom.** One episode in which LeWM fails and ATLAS succeeds (LeWM above ATLAS; frame indices indicate elapsed steps, goal in the rightmost column). The PushT and OGBench-Cube cases are shown in Figure 13 in Appendix.

6 Conclusion

We introduced ATLAS, a training objective for latent WMs that jointly targets relational preservation and marginal calibration. OOD-recovery preserves normalized state-to-state geometry relative to an internal patch anchor, while WEMReg calibrates the global latent distribution. Our analysis further connects these representation properties to finite-candidate planning through relational distortion, latent-scale mismatch, and terminal prediction error. Across PushT, TwoRoom, and OGBench-Cube, ATLAS improves mean goal-reaching success over LeWM on both lower- and higher-novelty evaluation subsets. Ablations show that relational preservation and marginal calibration provide distinct but partially overlapping benefits, while representation and rollout diagnostics show stronger novelty-related structure in the planning latent, improved marginal calibration, and lower multi-step prediction error. Together, these results support preserving planning-relevant latent geometry as a useful principle for improving the reliability of latent WM planning.

AI USE STATEMENT

Large language model based coding assistants were used by the authors during this project, in line with the ICLR policy on the use of large language models. Their use was limited to software engineering and presentation support: assisting with data-processing and plotting code, refactoring evaluation scripts, drafting and typesetting figures and tables, and drafting and copy-editing portions of the manuscript text. They were not used to generate research ideas, design the method, or produce experimental results. All experiments were run and verified by the authors, and the authors checked every reported number, claim, and figure against the underlying data and take full responsibility for the content of the paper.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

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

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021. doi: 10.1109/CVPR46437.2021.01549.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

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(1):19–67, 2005.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning (CoRL)*, 2022.

Yuntian Gao and Xiangyu Xu. Fast LeWorldModel. *arXiv preprint arXiv:2606.26217*, 2026.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cf73d57b6dcda32b293df7c2d5341f49-Paper-Conference.pdf.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model, 2018. URL https://arxiv.org/abs/1804.01947.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Lucas Maes, Quentin Le Lidec, Luiz Facury, Nassim Massaudi, Ayush Chaurasia, Francesco Capuano, Richard Gao, Taj Gillin, Dan Haramati, Damien Scieur, Yann LeCun, and Randall Balestriero. stable-worldmodel: A platform for reproducible world modeling research and evaluation. *arXiv preprint arXiv:2605.21800*, 2026a.

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*, 2026b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Vlad 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. *arXiv preprint arXiv:2502.14819*, 2025.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders, 2019. URL https://arxiv.org/abs/1711.01558.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning (ICML)*, 2025.

A EXPERIMENTAL SETUP DETAILS

Table 2 lists the training, model, and planning hyperparameters. All world models use the same architecture, the arms differ only in the anti-collapse regularizer (SIGReg vs. WEMReg) and the OOD-recovery weight μ. Values are shared across the three tasks unless noted.

Table 2: Training, model, and planning hyperparameters.

Group	Hyperparameter	Value
Model	Latent / embedding dimension	192
	Predictor	6-layer causal Transformer
	Attention heads / MLP dimension	16 / 2048
	History length	3 frames
	Frameskip	5
Training	Epoch budget (all tasks)	10
	Batch size	128
	Optimizer	AdamW
	Learning rate (cosine schedule)	5×10^{-5}
	Weight decay	1×10^{-3}
	Training seed	3072
Objective	WEMReg weight λ	3.0
	WEMReg projections	1024
	OOD-recovery weight μ	0.1 (ours); 0 (LeWM)
	Prediction loss	next-step latent MSE
Planning (CEM)	Candidates / iterations / elites	300 / 30 / 30
	Sampling variance scale	1.0
	Planning horizon	5
	Receding horizon (replan every)	5
	Evaluation budget	offset + 25 steps
	Episodes $\times$ seeds	200×5 (seeds 42–46)

Datasets. All models are trained on the offline demonstration datasets shipped with the STABLE-WORLDMODEL harness: **TwoRoom** and **PushT** (2D navigation and manipulation) and **OGBench-Cube** (3D robot-arm manipulation, from OGBench (Park et al., 2025)). Each dataset is split $90/10$ into training and validation clips, with the validation set used only to monitor training. Evaluation goals are sampled from the full set of valid episodes, not from a held-out test split, and every method is evaluated on 200 episodes per seed at goal offset 50.

A.1 IN-DISTRIBUTION / OUT-OF-DISTRIBUTION SPLIT

We split each evaluation condition (task, seed) into in-distribution (ID) and out-of-distribution (OOD) halves by how far each episode's expert path lies from the training distribution, using a fixed, model-independent rule. We draw a bank of 50,000 training states, standardize each coordinate by its bank mean and standard deviation, and fit a k-nearest-neighbor index with k=50. For each of the 200 episodes we average, over the states of its expert path (from the start row through the goal), the distance to the k nearest bank states, giving one OOD score per episode; episodes above the per-condition median are labeled OOD and the rest ID (100 each). The bank, standardization, and neighbor count are fixed and use no model, so the labels depend only on the states.

B OOD DIAGNOSTIC: PROTOCOL AND FULL RESULTS

This appendix details the probe used in Section 1 (and reused in Section 5) and reports its per-representation results at goal offset 50.

Representations probed. For a frozen model we score, for the same input frame, a set of representations drawn from the encoder at increasing depth: the `[CLS]` token at intermediate ViT blocks

Table 3: Main results at goal offset 50, split into in-distribution (ID) and out-of-distribution (OOD) episodes (%, mean ± std over 5 seeds). Best per column in bold; exact data of Figure 3.

	PushT		TwoRoom		OGBench-Cube	
Method	**ID**	**OOD**	**ID**	**OOD**	**ID**	**OOD**
ATLAS (ours)	$\mathbf{54.00 \pm 3.63}$	$\mathbf{47.60 \pm 2.58}$	87.00 ± 2.28	61.00 ± 2.19	58.40 ± 3.72	43.00 ± 4.94
LeWM	48.20 ± 4.62	42.00 ± 5.76	74.80 ± 2.79	41.20 ± 4.35	55.80 ± 4.31	40.40 ± 4.22
DINO-WM	51.00 ± 5.93	34.40 ± 4.67	$\mathbf{97.80 \pm 1.17}$	$\mathbf{98.20 \pm 0.98}$	59.40 ± 4.50	41.60 ± 7.28
FAST-WM	40.40 ± 5.08	34.60 ± 4.96	77.00 ± 4.00	47.20 ± 3.87	$\mathbf{61.60 \pm 4.63}$	$\mathbf{43.60 \pm 5.71}$
PLDM	28.80 ± 4.87	24.20 ± 5.19	85.60 ± 3.72	72.40 ± 1.85	53.80 ± 4.62	36.20 ± 6.97

(*L3, L6, L9*), the mean-pooled final-layer patch tokens (*patch*), the final `[CLS]` token before the projector (*cls*), and the projected `[CLS]` token that SIGReg regularizes and that the planner uses, i.e. the planning latent z. As a model-independent reference we also score an *oracle* representation built from the true environment state.

OOD score. Each representation is scored by a k-NN distance to a bank of N=1.5k training frames encoded once and standardized (k=50). For a query frame the score is the mean distance to its k nearest bank neighbors; a large distance means the frame is far from training support, i.e. OOD. The oracle uses the same k-NN construction on the raw true-state.

Metric: does the OOD score predict planning failure? We run the *unchanged* official planner (latent CEM) on n=200 evaluation episodes and log, at every step, each representation's OOD score; planning itself is untouched, only the logged signal differs. Per episode we aggregate the score over its last three steps and label the episode by whether the real simulator marks it a failure, and summarize each representation by $\mathrm{AUROC}(\text{OOD score} \rightarrow \text{failure})$ (chance $= 0.5$). The true-state oracle, computed from the ground-truth state, measures the *true* degree to which a model's failures are OOD-driven. Our aim is therefore not to raise a representation's AUROC as high as possible, but to make the planning latent's AUROC *agree* with the oracle's: a latent scoring far below the oracle is blind to OOD structure its failures actually carry, whereas one that matches the oracle reflects that structure as faithfully as the true state does.

Results. Table 4 reports this AUROC for every representation on the three tasks, at goal offset 50, for the baseline LeWM model, and Table 5 reports the same for ATLAS; both run on the same episode set under the same planner as our main results (n=200 episodes). For LeWM the OOD signal is present in the mid/late ViT features and the patch tokens (often 0.6–0.9) but drops sharply at the SIGReg-regularized planning latent z, to 0.44 on PushT (below chance) and to 0.55/0.49 on TwoRoom/Cube, in every case well below the patch anchor and the true-state oracle. This is the washout that the OOD-recovery term of Section 3 repairs, as Section 5 shows: under ATLAS z rises to 0.75/0.64/0.69, matching its own patch anchor and oracle.

Table 4: OOD diagnostic at goal offset 50 (n=200 episodes, baseline LeWM). $\mathrm{AUROC}(\text{OOD score} \rightarrow \text{planning failure})$ per representation (shallow → deep, ending at the planning latent z). "oracle" is the true-state reference. Chance $= 0.5$.

Task	**L3**	**L6**	**L9**	**patch**	**cls**	z **(SIGReg)**	**oracle**
PushT	0.59	0.81	0.83	0.76	0.72	0.44	0.88
TwoRoom	0.69	0.70	0.60	0.79	0.71	0.55	0.69
OGBench-Cube	0.82	0.86	0.85	0.71	0.79	0.49	0.76

Table 5: OOD diagnostic at goal offset 50 for **ATLAS** (n=200 episodes, same episode set and planner as Table 4). AUROC(OOD score → planning failure) per representation.

Task	L3	L6	L9	patch	cls	z	oracle
PushT	0.55	0.63	0.74	0.77	0.77	0.75	0.82
TwoRoom	0.30	0.38	0.61	0.65	0.69	0.64	0.59
OGBench-Cube	0.82	0.89	0.88	0.69	0.84	0.69	0.70

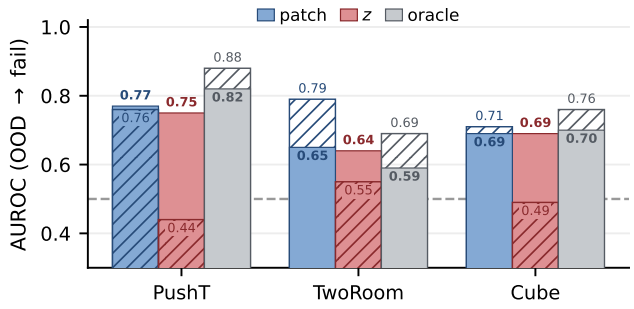


Figure 8: **OOD-related diagnostic signal across representations.** AUROC for predicting planning failure from a k-NN novelty score at goal offset 50. Hatched bars denote LeWM and solid bars ATLAS. Scores are computed from mean-pooled patch features, the planning latent z, and true environment states (oracle reference). This is a visualization of the per-representation values in Tables 4 and 5.

C Proofs for the Theoretical Analysis

C.1 Definitions and Setup

Distributional objectives. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote distributions with finite second moments, $\gamma_d = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, and $P_{\boldsymbol{u}}$ the distribution of $\boldsymbol{u}^\top \mathbf{z}$ for $\mathbf{z} \sim P$ and $\boldsymbol{u} \in \mathbb{S}^{d-1}$. Here W_2 is the quadratic Wasserstein distance, and WEMReg is the exact directional expectation in Eq. equation 2. For $\boldsymbol{Z} = [\boldsymbol{z}_1^\top; \dots; \boldsymbol{z}_N^\top]$, $N \geq 2$, set $P_{\boldsymbol{Z}} = N^{-1} \sum_i \delta_{\boldsymbol{z}_i}$ and $\mathcal{L}_{\mathrm{WEMReg}}(\boldsymbol{Z}) = \mathcal{L}_{\mathrm{WEMReg}}(P_{\boldsymbol{Z}})$. Its exact finite-batch decomposition is given in Lemma 4. For finitely many frequencies $t_j \in \mathbb{R}$ and weights $\omega_j > 0$, define

$$\mathcal{L}_{\mathrm{CF}}(P) = \mathbb{E}_{\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \sum_{j=1}^{J} \omega_j \left| \varphi_{P,\mathbf{u}}(t_j) - e^{-t_j^2/2} \right|^2, \qquad \varphi_{P,\boldsymbol{u}}(t) = \mathbb{E}_{\mathbf{z} \sim P}[e^{it\boldsymbol{u}^\top \mathbf{z}}]. \tag{7}$$

This includes finite-frequency Epps–Pulley quadrature up to a positive normalization factor (Balestriero & LeCun, 2025; Maes et al., 2026b). Theorem 1 concerns this finite-frequency objective, not integration over all frequencies.

Relational geometry. Let $\bar{\boldsymbol{p}}_i \in \mathbb{R}^m$ be the mean-pooled patch-token anchors and $\boldsymbol{P} = [\bar{\boldsymbol{p}}_1^\top; \dots; \bar{\boldsymbol{p}}_N^\top]$, held fixed when differentiating the relational loss. For $\boldsymbol{X} = [\boldsymbol{x}_1^\top; \dots; \boldsymbol{x}_N^\top]$, let $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\sigma}_{\boldsymbol{X}}$ be the empirical mean and coordinate-wise standard deviations, with $\boldsymbol{\mu}_{\boldsymbol{X}} = N^{-1} \sum_i \boldsymbol{x}_i$ and $(\boldsymbol{\sigma}_{\boldsymbol{X}})_j^2 = N^{-1} \sum_i ((\boldsymbol{x}_i)_j - (\boldsymbol{\mu}_{\boldsymbol{X}})_j)^2$. For $\varepsilon_0 \geq 0$, set

$$\boldsymbol{S}_{\boldsymbol{X}} = \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{X}} + \varepsilon_0 \mathbf{1}), \qquad c_{\boldsymbol{X}} = \frac{1}{N^2} \sum_{i,j} \|\boldsymbol{S}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_2, \qquad d_{\boldsymbol{X}}(i,j) = \frac{\|\boldsymbol{S}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_2}{c_{\boldsymbol{X}}}. \tag{8}$$

Here $\mathbf{1}$ is the all-ones vector. Assume positive diagonal entries of $\boldsymbol{S}_{\boldsymbol{X}}$ and $c_{\boldsymbol{X}} > 0$. The relational loss is

$$\mathcal{L}_{\mathrm{OOD}}(\boldsymbol{Z}, \boldsymbol{P}) = \frac{1}{N^2} \sum_{i,j} \big(d_{\boldsymbol{Z}}(i,j) - d_{\boldsymbol{P}}(i,j)\big)^2. \tag{9}$$

Evaluation and planning. Fix a finite pool $\mathbb{I} = \{1, \dots, N\}$ and normalizers shared across all queries and candidates, computed on this pool or imported from a fixed reference bank. Define

$$\epsilon_{\mathrm{rel}} = \max_{i,j \in \mathbb{I}} |d_{\boldsymbol{Z}}(i,j) - d_{\boldsymbol{P}}(i,j)|. \tag{10}$$

When the relational loss uses this same pool and these same normalizers, the maximum squared error is bounded by the sum, giving

$$\epsilon_{\mathrm{rel}} \le N \sqrt{\mathcal{L}_{\mathrm{OOD}}(\boldsymbol{Z}, \boldsymbol{P})}. \tag{11}$$

For $q \in \mathbb{I}$, a nonempty bank $\mathbb{B} \subseteq \mathbb{I}$, and $1 \le k \le |\mathbb{B}|$, the mean kNN score is

$$s_{\boldsymbol{X}}(q) = \min_{\substack{\mathbb{J} \subseteq \mathbb{B} \\ |\mathbb{J}| = k}} \frac{1}{k} \sum_{j \in \mathbb{J}} d_{\boldsymbol{X}}(q,j), \tag{12}$$

which is independent of tie breaking. For a nonempty candidate set $\mathbb{C} \subseteq \mathbb{I}$ and goal $g \in \mathbb{I}$, let $\boldsymbol{z}_i$ and $\widehat{\boldsymbol{z}}_i$ be the true and predicted terminal embeddings, respectively, and set

$$\begin{aligned} e &= \max_{i \in \mathbb{C}} \|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_i\|_2, \qquad M = \max_{i \in \mathbb{C}} d_{\boldsymbol{P}}(i,g), \\ \eta &= \frac{\max_j |(\boldsymbol{\sigma}_{\boldsymbol{Z}})_j - 1|}{1 + \varepsilon_0}, \qquad \ell = (1 + \varepsilon_0) c_{\boldsymbol{Z}}. \end{aligned} \tag{13}$$

Scope. The guarantees concern the evaluated pool and finite candidate set; small training losses alone do not control errors on unseen states. The objectives need not change e, ϵ_{rel}, and η independently or monotonically, and exact Gaussian matching and anchor geometry preservation need not be simultaneously feasible. Task-cost guarantees additionally require anchor alignment (Corollary 8).

C.2 Exact Finite-Sample Sliced Transport

Lemma 4 (Gaussian quantile-cell decomposition)**.** *Let $N \ge 2$, $x_{(1)} \le \cdots \le x_{(N)}$, and $\widehat{P} = N^{-1} \sum_i \delta_{x_{(i)}}$. With Φ the standard Gaussian cumulative distribution function, set*

$$m_i = N \int_{(i-1)/N}^{i/N} \Phi^{-1}(p)\, dp, \qquad v_N = \frac{1}{N} \sum_i m_i^2, \qquad \kappa_N = 1 - v_N. \tag{14}$$

Then

$$W_2^2(\widehat{P}, \gamma_1) = \kappa_N + \frac{1}{N} \sum_i (x_{(i)} - m_i)^2. \tag{15}$$

Moreover, $N^{-1} \sum_i m_i = 0$, $0 < v_N < 1$, and $\kappa_N > 0$.

Proof. Write $q = \Phi^{-1}$ and $I_i = ((i-1)/N, i/N]$. Since $\int_{I_i} (q(p) - m_i)\, dp = 0$, the quantile representation gives

$$\begin{aligned} W_2^2(\widehat{P}, \gamma_1) &= \sum_i \int_{I_i} (x_{(i)} - q(p))^2\, dp \\ &= \frac{1}{N} \sum_i (x_{(i)} - m_i)^2 + \sum_i \int_{I_i} (q(p) - m_i)^2\, dp. \end{aligned}$$

The final sum equals $\int_0^1 q(p)^2\, dp - v_N = 1 - v_N = \kappa_N$, while $N^{-1} \sum_i m_i = \int_0^1 q(p)\, dp = 0$. As q is nonconstant on every cell, $\kappa_N > 0$; as $m_N > 0$ for $N \ge 2$, $v_N > 0$. Thus $0 < v_N < 1$. □

Applying the lemma to sorted projections gives the exact batch WEMReg cost, with non-optimizable residual κ_N. This lower bound is attained by $x_{(i)} = m_i$ in one projection, but need not be simultaneously attainable in all directions of a finite high-dimensional point cloud.

C.3 Proof of Finite-Frequency Ambiguity

Proof of Theorem 1. Choose $J+3$ disjoint spherical shells of positive γ_d-measure. Their indicators span a $(J+3)$-dimensional space of bounded radial functions. The following $J+2$ homogeneous linear constraints therefore admit a nonzero solution r, normalized so that $\|r\|_\infty = 1$:

$$\begin{aligned} \int r\, d\gamma_d = 0, \qquad \int \|\boldsymbol{x}\|_2^2 r(\boldsymbol{x})\, d\gamma_d(\boldsymbol{x}) = 0, \\ \int \cos\big(t_j(\boldsymbol{x})_1\big) r(\boldsymbol{x})\, d\gamma_d(\boldsymbol{x}) = 0, \qquad j = 1, \dots, J. \end{aligned} \tag{16}$$

Define $dP = (1 + r/2)\, d\gamma_d$. The first constraint makes P a probability measure with $1/2 \le dP/d\gamma_d \le 3/2$, hence finite second moments. Also, $P \neq \gamma_d$ because r is nonzero on a set of positive measure. Orthogonal invariance and the second constraint give

$$\mathbb{E}_P[\mathbf{z}] = \mathbf{0}, \qquad \mathrm{Cov}_P(\mathbf{z}) = \frac{\mathbb{E}_P\|\mathbf{z}\|_2^2}{d} \boldsymbol{I}_d = \boldsymbol{I}_d.$$

Reflection symmetry eliminates the imaginary parts of the characteristic functions. The remaining constraints and orthogonal invariance give $\varphi_{P,\boldsymbol{u}}(t_j) = e^{-t_j^2/2}$ for every unit $\boldsymbol{u}$ and every j, so $\mathcal{L}_{\mathrm{CF}}(P) = 0$.

Finally, orthogonal invariance makes $W_2^2(P_{\boldsymbol{u}}, \gamma_1)$ independent of $\boldsymbol{u}$. If this value were zero, all unit projections would be standard Gaussian, so $\mathbb{E}_P[e^{\mathrm{i}\boldsymbol{v}^\top \mathbf{z}}] = e^{-\|\boldsymbol{v}\|_2^2/2}$ for every $\boldsymbol{v} \in \mathbb{R}^d$. Uniqueness of characteristic functions would imply $P = \gamma_d$, a contradiction. Thus $\mathcal{L}_{\mathrm{WEMReg}}(P) > 0$. □

C.4 WEMReg Controls the Normalization Mismatch

Lemma 5 (Quantitative moment calibration)**.** *Let $P \in \mathcal{P}_2(\mathbb{R}^d)$ have mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and set $s = \mathcal{L}_{\mathrm{WEMReg}}(P)$. Then $\|\boldsymbol{\mu}\|_2^2 \le ds$ and*

$$\|\boldsymbol{\Sigma} - \boldsymbol{I}_d\|_{\mathrm{F}} \le C_d(s), \tag{17}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and

$$C_d(s) = \sqrt{\frac{d(d+2)}{2}} \left[\sqrt{d}(1 + \sqrt{s}) + 1\right] \sqrt{s}. \tag{18}$$

Consequently, $\sigma_j = \sqrt{\boldsymbol{\Sigma}_{j,j}}$ satisfies $\max_j |\sigma_j - 1| \le C_d(s)$. For the reference distribution used in relational standardization, this gives

$$\eta \le \frac{C_d(s)}{1 + \varepsilon_0}. \tag{19}$$

Proof. Let $\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $q(\boldsymbol{u}) = \sqrt{\boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u}}$. Centering any coupling and applying the reverse triangle inequality in L^2 gives

$$W_2^2(P_{\boldsymbol{u}}, \gamma_1) \ge (\boldsymbol{u}^\top \boldsymbol{\mu})^2 + (q(\boldsymbol{u}) - 1)^2. \tag{20}$$

Since $\mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \boldsymbol{I}_d/d$, averaging proves $\|\boldsymbol{\mu}\|_2^2/d + \mathbb{E}(q(\mathbf{u}) - 1)^2 \le s$. Hence

$$\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{d}} = \|q\|_{L^2} \le 1 + \sqrt{s}. \tag{21}$$

For any symmetric $\boldsymbol{A}$, the spherical fourth-moment identity is

$$\mathbb{E}(\mathbf{u}^\top \boldsymbol{A} \mathbf{u})^2 = \frac{\mathrm{tr}(\boldsymbol{A})^2 + 2\|\boldsymbol{A}\|_{\mathrm{F}}^2}{d(d+2)}. \tag{22}$$

Taking $\boldsymbol{A} = \boldsymbol{\Sigma} - \boldsymbol{I}_d$ and using $q(\boldsymbol{u}) \le \sqrt{\mathrm{tr}(\boldsymbol{\Sigma})}$ yields

$$\begin{aligned} \mathbb{E}(\mathbf{u}^\top \boldsymbol{A} \mathbf{u})^2 &= \mathbb{E}\big[(q(\mathbf{u}) - 1)^2 (q(\mathbf{u}) + 1)^2\big] \\ &\le \left[\sqrt{d}(1 + \sqrt{s}) + 1\right]^2 s. \end{aligned} \tag{23}$$

Dropping the nonnegative trace term in Eq. equation 22 proves the covariance bound. Finally,

$$|\sigma_j - 1| = \frac{|\boldsymbol{\Sigma}_{j,j} - 1|}{\sigma_j + 1} \le |\boldsymbol{\Sigma}_{j,j} - 1| \le \|\boldsymbol{\Sigma} - \boldsymbol{I}_d\|_{\mathrm{F}},$$

which also proves the bound on η. □

The lemma applies to population or empirical reference distributions. Its s is the exact directional expectation; substituting a finite Monte Carlo estimate requires an additional integration-error bound.

C.5 Proof of Non-Redundancy

Proof of Proposition 2. Take $N = 3$ and $d = m = 1$. Gaussian symmetry gives $(m_1, m_2, m_3) = (-b, 0, b)$ with $b > 0$. Set $\boldsymbol{P} = (-b, 0, b)^\top$, $\boldsymbol{Z}^{\mathrm{p}} = (-b, b, 0)^\top$, and $\boldsymbol{Z}^{(c)} = c\boldsymbol{P}$ for $c > 0$. In one dimension, the positive standardization factor cancels, so

$$d_{\boldsymbol{X}}(i, j) = \frac{9|x_i - x_j|}{\sum_{a,b} |x_a - x_b|}.$$

The distances for pairs $(1,2), (1,3), (2,3)$ are respectively $\frac{9}{8}(1, 2, 1)$ for $\boldsymbol{P}$ and $\frac{9}{8}(2, 1, 1)$ for $\boldsymbol{Z}^{\mathrm{p}}$. Together with their identical sorted values and Lemma 4, this gives

$$\begin{aligned} &\mathcal{L}_{\mathrm{WEMReg}}(\boldsymbol{P}) = \mathcal{L}_{\mathrm{WEMReg}}(\boldsymbol{Z}^{\mathrm{p}}) = \kappa_3, \\ &\mathcal{L}_{\mathrm{OOD}}(\boldsymbol{P}, \boldsymbol{P}) = 0, \qquad \mathcal{L}_{\mathrm{OOD}}(\boldsymbol{Z}^{\mathrm{p}}, \boldsymbol{P}) = \frac{4}{9}\left(\frac{9}{8}\right)^2 = \frac{9}{16}. \end{aligned} \tag{24}$$

The two one-dimensional projection directions have equal transport costs by Gaussian symmetry. Positive scaling preserves all normalized distances, while the sorted scaled values are cm_i. Thus

$$\begin{aligned} &\mathcal{L}_{\mathrm{OOD}}(\boldsymbol{Z}^{(c)}, \boldsymbol{P}) = 0, \\ &\mathcal{L}_{\mathrm{WEMReg}}(\boldsymbol{Z}^{(c)}) = \kappa_3 + \frac{1}{3}\sum_i (cm_i - m_i)^2 = \kappa_3 + (c-1)^2 v_3. \end{aligned} \tag{25}$$

Taking $c \neq 1$ proves the second claim. The construction is valid for every $\varepsilon_0 \ge 0$ because all scalar standardization denominators are positive. □

C.6 Stability of Finite Minimization

Lemma 6 (Uniform cost perturbations)**.** *Let $\mathbb{C}$ be nonempty and finite and let $f, g : \mathbb{C} \to \mathbb{R}$ satisfy $\max_{i \in \mathbb{C}} |f(i) - g(i)| \le \epsilon$ for $\epsilon \ge 0$. Then*

$$\left|\min_{\mathbb{C}} f - \min_{\mathbb{C}} g\right| \le \epsilon, \qquad f(\widehat{i}) - \min_{\mathbb{C}} f \le 2\epsilon \quad (\widehat{i} \in \arg\min_{\mathbb{C}} g).$$

If f has a unique minimizer $i^\star$ with $f(i) - f(i^\star) > 2\epsilon$ for every $i \neq i^\star$, then every minimizer of g equals $i^\star$.

Proof. Taking minima in $f - \epsilon \le g \le f + \epsilon$ proves the first claim. For $i^\star \in \arg\min f$ and $\widehat{i} \in \arg\min g$,

$$f(\widehat{i}) \le g(\widehat{i}) + \epsilon \le g(i^\star) + \epsilon \le f(i^\star) + 2\epsilon.$$

The strict margin rules out $\widehat{i} \neq i^\star$. □

C.7 Proof of Reliability and Planning Stability

Proof of Theorem 3. For every k-element subset $\mathbb{J} \subseteq \mathbb{B}$, its mean distances under $d_{\boldsymbol{Z}}$ and $d_{\boldsymbol{P}}$ differ by at most ϵ_{rel}. Applying the first part of Lemma 6 to the finite family of these subsets proves Eq. equation 5.

For planning, put $a = 1 + \varepsilon_0$ and $\boldsymbol{v}_i = \boldsymbol{S}_{\boldsymbol{Z}}^{-1}(\boldsymbol{z}_i - \boldsymbol{z}_g)$. By definition, $\|\boldsymbol{S}_{\boldsymbol{Z}}/a - \boldsymbol{I}_d\|_{\mathrm{op}} = \eta$, so

$$\begin{aligned}\left|\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_g\|_2}{\ell} - d_{\boldsymbol{Z}}(i,g)\right| &= \frac{1}{c_{\boldsymbol{Z}}}\,\big|\|(\boldsymbol{S}_{\boldsymbol{Z}}/a)\boldsymbol{v}_i\|_2 - \|\boldsymbol{v}_i\|_2\big| \\ &\le \frac{\|(\boldsymbol{S}_{\boldsymbol{Z}}/a - \boldsymbol{I}_d)\boldsymbol{v}_i\|_2}{c_{\boldsymbol{Z}}} \le \eta\, d_{\boldsymbol{Z}}(i,g).\end{aligned} \tag{26}$$

Combining this with the reverse triangle inequality and $d_{\boldsymbol{Z}}(i,g) \le M + \epsilon_{\mathrm{rel}}$ gives, uniformly on $\mathbb{C}$,

$$\begin{aligned}\left|\frac{\|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_g\|_2}{\ell} - d_{\boldsymbol{P}}(i,g)\right| &\le \frac{e}{\ell} + \eta d_{\boldsymbol{Z}}(i,g) + \epsilon_{\mathrm{rel}} \\ &\le \frac{e}{\ell} + (1+\eta)\epsilon_{\mathrm{rel}} + \eta M = \epsilon_{\mathrm{plan}}.\end{aligned} \tag{27}$$

Since $\ell > 0$, minimizing $\|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_g\|_2^2$ is equivalent to minimizing $h(i) = \|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_g\|_2/\ell$. Lemma 6, applied to $f(i) = d_{\boldsymbol{P}}(i,g)$ and $h(i)$, proves the regret bound. □

Margins and exact preservation. Under the assumptions of Theorem 3, including $\eta < 1$, a unique anchor minimizer $i^\star$ is guaranteed to be selected whenever

$$d_{\boldsymbol{P}}(i,g) - d_{\boldsymbol{P}}(i^\star,g) > 2\epsilon_{\mathrm{plan}} \qquad (i \in \mathbb{C} \setminus \{i^\star\}), \tag{28}$$

by Lemma 6; this condition is sufficient, not necessary. Similarly, the kNN bound implies

$$s_{\boldsymbol{P}}(q_1) - s_{\boldsymbol{P}}(q_2) > 2\epsilon_{\mathrm{rel}} \quad \Longrightarrow \quad s_{\boldsymbol{Z}}(q_1) > s_{\boldsymbol{Z}}(q_2),$$

without requiring identical nearest-neighbor sets. If $\epsilon_{\mathrm{rel}} = 0$, all kNN scores agree on the pool. If additionally $e = \eta = 0$, every selected candidate is anchor-optimal, and equals the anchor minimizer when it is unique.

C.8 From One-Step Prediction to Terminal Error

Lemma 7 (Finite-horizon prediction error)**.** *Let $g : \mathbb{R}^d \times \mathbb{A} \to \mathbb{R}^d$ be the learned predictor. For each candidate i, let $\boldsymbol{z}_{i,0}, \dots, \boldsymbol{z}_{i,T}$ be its true encoded trajectory, and define $\widehat{\boldsymbol{z}}_{i,0} = \boldsymbol{z}_{i,0}$ and $\widehat{\boldsymbol{z}}_{i,t+1} = g(\widehat{\boldsymbol{z}}_{i,t}, \boldsymbol{a}_{i,t})$. Assume, for every relevant candidate and time,*

$$\|g(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t}) - \boldsymbol{z}_{i,t+1}\|_2 \le \epsilon_{\mathrm{dyn}},$$

and that $g(\cdot, \boldsymbol{a})$ is L-Lipschitz on all true and predicted states being compared, with $L \ge 0$. Then

$$\max_i \|\widehat{\boldsymbol{z}}_{i,T} - \boldsymbol{z}_{i,T}\|_2 \le \epsilon_{\mathrm{dyn}} \sum_{t=0}^{T-1} L^t.$$

Proof. The assumptions give $e_{t+1} \le Le_t + \epsilon_{\mathrm{dyn}}$ for $e_t = \max_i \|\widehat{\boldsymbol{z}}_{i,t} - \boldsymbol{z}_{i,t}\|_2$, with $e_0 = 0$. Iterating proves the claim for every $L \ge 0$, including $L = 1$. □

C.9 Connection to a Task-Aligned Terminal Cost

Corollary 8 (Task-cost regret)**.** *Under Theorem 3, suppose $J_{\mathrm{task}} : \mathbb{C} \to \mathbb{R}$ satisfies $|J_{\mathrm{task}}(i) - d_{\boldsymbol{P}}(i,g)| \le \epsilon_{\mathrm{anchor}}$ for every $i \in \mathbb{C}$, where $\epsilon_{\mathrm{anchor}} \ge 0$. Then*

$$J_{\mathrm{task}}(\widehat{i}) - \min_{i \in \mathbb{C}} J_{\mathrm{task}}(i) \le 2(\epsilon_{\mathrm{plan}} + \epsilon_{\mathrm{anchor}}).$$

Proof. By anchor alignment and the planning regret bound,

$$\begin{aligned}J_{\mathrm{task}}(\widehat{i}) - \min_{i \in \mathbb{C}} J_{\mathrm{task}}(i) &\le d_{\boldsymbol{P}}(\widehat{i}, g) - \min_{i \in \mathbb{C}} d_{\boldsymbol{P}}(i,g) + 2\epsilon_{\mathrm{anchor}} \\ &\le 2(\epsilon_{\mathrm{plan}} + \epsilon_{\mathrm{anchor}}).\end{aligned}$$

□

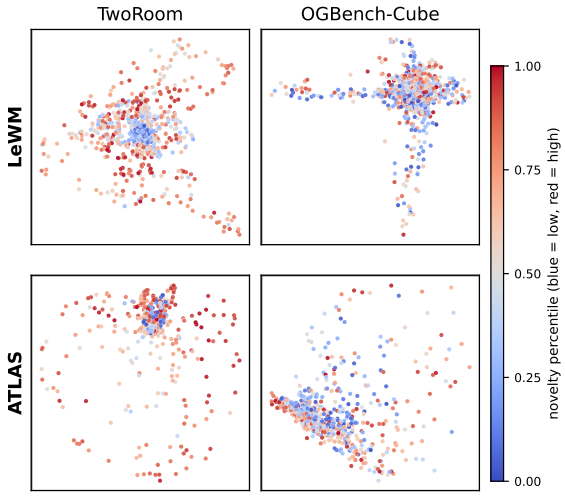


Figure 9: **Planning-latent geometry on TwoRoom and OGBench-Cube.** Two-dimensional PCA projections of z for LeWM (top) and ATLAS (bottom), companion to the PushT panels in Figure 4. Each point is a state, colored by its model-independent true-state novelty percentile, from lower novelty in blue to higher novelty in red. The separation between low- and high-novelty states is less pronounced than on PushT, but the ATLAS projections still place high-novelty states farther from the in-distribution cluster than LeWM does.

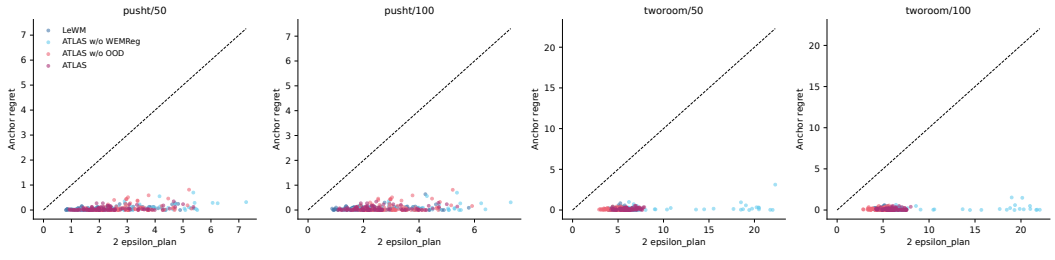


Figure 10: **Bound versus regret.** Each point is one arm-query evaluation; the dashed line is equality. The conservative bound and absent sufficient-margin events limit quantitative interpretation.

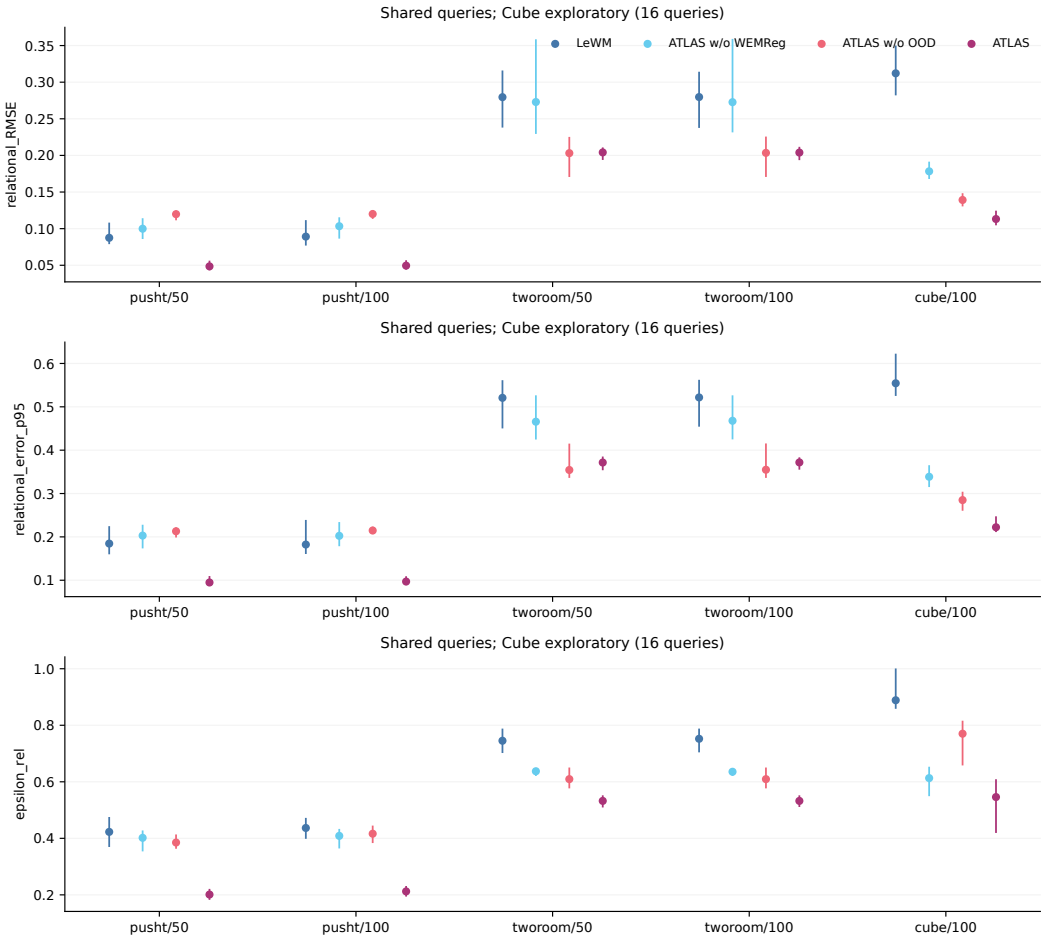


Figure 11: **Relational metrics.** Across-query medians and paired-bootstrap percentile intervals for within-query pairwise RMSE, 95th-percentile distortion, and maximum distortion. PushT/TwoRoom each use 100 queries, Cube 16 exploratory queries. All four arms use controlled checkpoints.

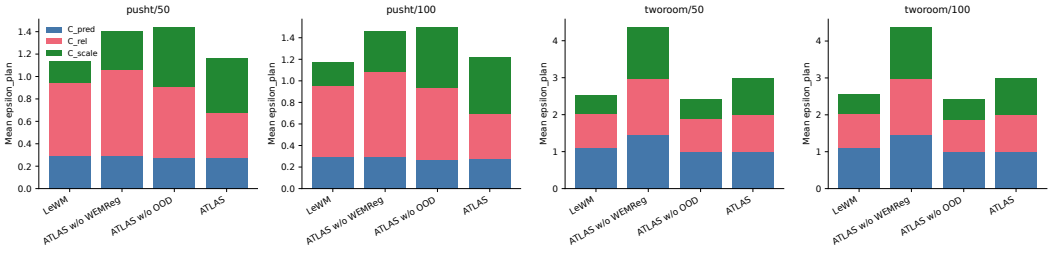


Figure 12: **Theorem 3 decomposition.** Bars show mean prediction, relational, and scale contributions to epsilon plan on the original queries. This is a structural decomposition, not a quantitative planning-success predictor.

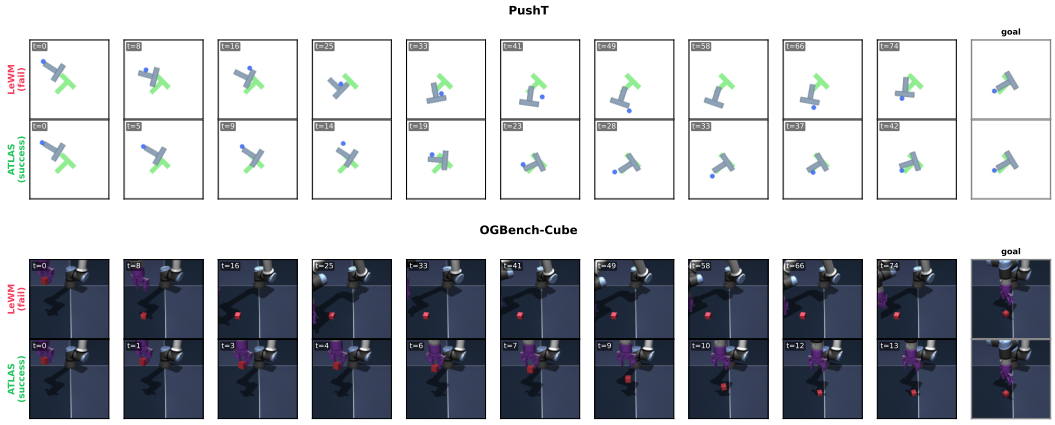


Figure 13: **Selected closed-loop planning trajectories on PushT (top) and OGBench-Cube (bottom).** One episode per task in which LeWM fails and ATLAS succeeds (LeWM above ATLAS in each panel; frame indices indicate elapsed steps, goal in the rightmost column). On PushT, ATLAS rotates and aligns the T-shaped block with the target pose; on OGBench-Cube, ATLAS carries the cube to the target configuration.