LEARNING PERSPECTIVE MATTERS IN LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent world models often organise prediction along different latent directions, yet the perspective governing these directions and their use is typically predefined rather than learned. This leaves an important design choice outside the learning process, even though both the choice of directions and their relative weighting can shape predictive computation. We introduce orthogonal subspace modulation, which jointly learns an orthonormal basis and adaptive gains conditioned on the current state and action. The basis determines where modulation acts, while the gains determine how strongly each direction is used. We formalise their interaction through Gain–Basis Coupling, showing that the latent perspective becomes functionally consequential when different subspaces receive different gains Across four visual-control tasks, learning this perspective improves control performance over strong latent world-model baselines. Detailed analyses further reveal the empirical validity of Gain–Basis Coupling, functional coordination among subspaces, and physical-dynamics accessibility from subspace combinations. These results show that learning perspective can improve how latent representations are organised and used for prediction and control.

1 INTRODUCTION

Latent predictive world models provide an increasingly practical basis for planning and control (Sobal et al., 2025; Zhou et al., 2025; Maes et al., 2026). By predicting representations rather than reconstructing observations (LeCun, 2022; Assran et al., 2023), these models learn compact latent dynamics from offline trajectories. Yet a useful latent representation need not have a unique orientation: an orthogonal change of basis preserves its information and Euclidean geometry while expressing it along different directions. This freedom is also reflected in identifiability analyses of LeJEPA (Balestriero & LeCun, 2025) and controlled world models, which recover latent states up to an orthogonal transformation under their respective assumptions (Klindt et al., 2026; Zhang et al., 2026). Such freedom becomes computationally meaningful once different latent directions are treated differently. Recent methods already exploit this structure: Sub-JEPA applies Gaussian regularisation within projected subspaces (Zhao et al., 2026), SD-JEPA assigns different objectives to predefined coordinate subspaces (Thil et al., 2026), and FiLM and Squeeze-and-Excitation modulate features along existing network axes (Perez et al., 2018; Hu et al., 2018). Yet the directions along which these operations act are largely predefined rather than selected jointly with their predictive use; Sub-JEPA, for example, primarily uses frozen random projections while also evaluating trainable projections in its regularisation pathway. We refer to the choice of latent orientation used for subsequent computation as the *latent perspective*. If perspective shapes how latent directions participate in prediction, fixing it in advance leaves a computational degree of freedom unexplored. This motivates our central question:

Can perspective be learned rather than predefined in latent world models?

We address this question with *orthogonal subspace modulation*. Figure 1 illustrates the motivation: nominally different predefined projections can exhibit similar responses to the same action perturbation, whereas our learned subspaces reveal more differentiated predictive responses. Our method parameterises the latent perspective by a complete orthonormal basis learned jointly with adaptive subspace gains. The perspective determines *where* modulation acts through complementary orthogonal subspaces, while a gating network conditioned on each subspace's coordinates and the

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.

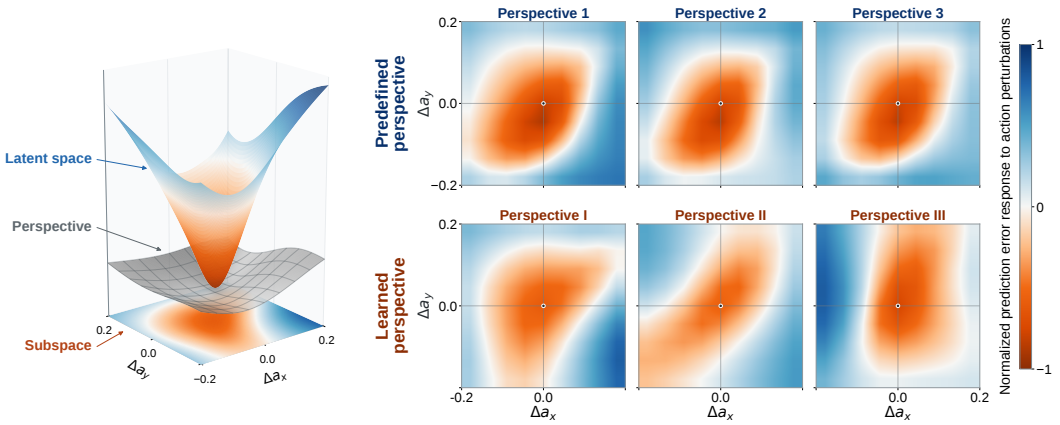


Figure 1: **Learning the latent perspective reveals more diverse subspace responses.** For the same PushT transition, the displayed predefined Sub-JEPA projections exhibit similar responses to action perturbations, whereas the learned perspective produces more varied predictive response patterns.

current action determines *how much* it is modulated. The resulting modulation is added residually before prediction, and all components are trained jointly under objective of the world model. Because the basis is complete and orthonormal, learning the perspective does not add, discard, or compress latent information; it changes the directions along which adaptive computation acts. The principle is simple: *learn where to modulate, adapt how much*. Figure 2 illustrates the mechanism.

Why should learning the perspective matter if an orthogonal change of orientation preserves the representation itself? We formalise the answer through *Gain–Basis Coupling*. For fixed realised gains, equal gains reduce recomposition to isotropic scaling, making it independent of the latent perspective. When the gains differ, however, the effect of modulation depends on the directions along which those gains act. More precisely, for perturbations that mix two subspaces, the local sensitivity of the realised modulation operator is weighted by the corresponding gain difference. Thus, the perspective becomes functionally consequential when the model treats its induced subspaces differently. This motivates a concrete empirical question: does end-to-end predictive learning discover both differentiated gains and useful directions along which to apply them?

Across PushT, Cube, Reacher, and TwoRoom, orthogonal subspace modulation improves closed-loop control over LeWorldModel, with controlled ablations showing an additional benefit from learning the basis rather than keeping it random and fixed. We then examine whether this benefit follows the mechanism predicted by Gain–Basis Coupling. The learned model develops stronger inter-subspace gain differentiation, while same-checkpoint interventions confirm that the effect of these gains depends on the directions along which they act. Beyond this coupling, learning the perspective also changes the organisation of predictive computation: the resulting subspace combinations exhibit functional coordination and physical-dynamics accessibility. Together, these results show that jointly learning where and how strongly to modulate can organise latent computation in ways that benefit both prediction and control.

Our main contributions are:

- **Learning the latent perspective.** We introduce *orthogonal subspace modulation*, which parameterises the latent perspective by a complete orthonormal basis and learns it jointly with adaptive state–action-dependent subspace gains.
- **Gain–Basis Coupling.** We characterise when the latent perspective becomes functionally consequential: equal gains make recomposition perspective-independent, while gain differences govern its sensitivity to changes in perspective.
- **Empirical evidence for learned predictive organisation.** Across four visual-control tasks and detailed PushT analyses, we show improved closed-loop control together with stronger gain differentiation, functional coordination across subspaces, and greater physical-dynamics accessibility at matched representation capacity.

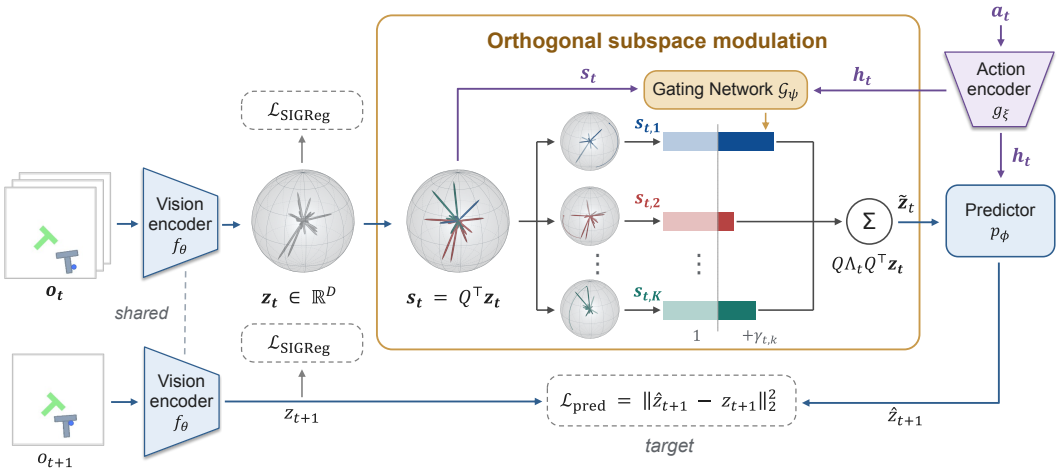


Figure 2: **Overview of orthogonal subspace modulation.** The learned latent perspective induces an orthogonal decomposition of the representation, while a gating network shared across subspaces assigns adaptive modulation strengths from the corresponding subspace coordinates and current action before future latent prediction.

2 Method

2.1 Problem Setting and Overview

We build on the latent predictive world-model setting of LeWorldModel (Maes et al., 2026) and consider reward-free visual trajectories of observations o_t and actions a_t. A vision encoder f_θ maps each observation to a D-dimensional latent $z_t = f_\theta(o_t)$, while an action encoder g_ξ produces an action embedding $h_t = g_\xi(a_t)$. Bold symbols $\mathbf{o}_t$, $\mathbf{a}_t$, $\mathbf{z}_t$, and $\mathbf{h}_t$ denote the corresponding ordered three-step histories ending at t. Given the modulated latent history $\widetilde{\mathbf{z}}_t$ and action-embedding history $\mathbf{h}_t$, the predictor estimates $\hat{z}_{t+1} = p_\phi(\widetilde{\mathbf{z}}_t, \mathbf{h}_t)$ with target $z_{t+1} = f_\theta(o_{t+1})$. Our question is whether predictive learning can use the freedom of an information-preserving orthogonal re-expression to discover a latent perspective that is useful for subsequent predictive computation. Learning this perspective does not add or discard latent information; instead, it determines the directions along which subspace modulation operates.

Orthogonal subspace modulation combines a learnable latent perspective with adaptive subspace gains. The perspective determines the subspaces along which modulation acts, while a gating network shared across subspaces determines how strongly each subspace is modulated from its coordinates and the current action. The same mechanism is applied at each of the three context positions to form $\widetilde{\mathbf{z}}_t$. Because the latent perspective is represented by a complete orthonormal basis, changing it re-expresses the full representation without removing any latent dimension. The perspective therefore determines *where* modulation acts, while the gains determine *how much* each subspace is modulated. All learnable components are optimised jointly end to end through the predictive objective. The following sections formalise the latent perspective, adaptive subspace gains, their gain–basis coupling, and the predictive learning objective.

2.2 Learnable Latent Perspective

We represent the latent perspective by a complete orthonormal basis $Q \in \mathbb{R}^{D\times D}$ satisfying $Q^\top Q = I_D$. This perspective induces an orthogonal decomposition of the latent space into K equal-dimensional subspaces, each of dimension $d = D/K$, with K chosen to divide D. Writing $Q = [Q_1, \ldots, Q_K]$ with $Q_k \in \mathbb{R}^{D\times d}$, the columns of Q_k form an orthonormal basis for the kth subspace and define its orthogonal projector $P_k = Q_k Q_k^\top$. For a latent $z_t \in \mathbb{R}^D$, its coordinates in the kth subspace are $s_{t,k} = Q_k^\top z_t$, so $P_k z_t = Q_k s_{t,k}$ and $z_t = \sum_{k=1}^K P_k z_t = \sum_{k=1}^K Q_k s_{t,k}$. Because Q is complete and orthonormal, the induced subspaces are mutually orthogonal and jointly span the full latent space, with $P_i P_j = 0$ for $i \neq j$ and $\sum_k P_k = I_D$. The coordinate change also

preserves the latent norm, $\|Q^\top z_t\|_2 = \|z_t\|_2$. The latent perspective therefore provides an exact, information-preserving re-expression of the representation rather than a dimensionality reduction. Learning this perspective changes how the latent space is organised into orthogonal subspaces and thereby determines the directions along which subsequent modulation acts. Its role is not to increase the dimensionality or information content of the representation, but to provide an explicit structured parameterisation through which the encoder, gating network, and predictor can jointly organise directional modulation during end-to-end predictive learning.

Trainable parameterisation. We obtain Q from an unconstrained matrix $U \in \mathbb{R}^{D\times D}$ through sign-corrected QR factorisation. This maintains orthogonality throughout training while allowing the latent perspective to be learned jointly end to end under the training objective. The full parameterisation is given in Appendix A.1.

2.3 Adaptive Subspace Gains

The learned latent perspective enables separate modulation of its constituent subspaces. A gating network $\mathcal{G}_\psi$, whose parameters are shared across all K subspaces, maps the subspace coordinates $s_{t,k} = Q_k^\top z_t$ and action embedding h_t to a scalar logit $\ell_{t,k} = \mathcal{G}_\psi(s_{t,k}, h_t)$. We convert this logit to a bounded residual modulation strength $\gamma_{t,k} = \alpha\,\sigma(\ell_{t,k}/\tau)$, where $\alpha = \sigma(\widehat{\alpha})$, and define the realised gain as $g_{t,k} = 1 + \gamma_{t,k}$. Thus, the gating network is shared, but each subspace receives its own realised gain. Smaller gains correspond to less additional modulation, while larger gains more strongly amplify the corresponding subspace. We use this bounded non-negative residual parameterisation in the main model; the complete gating-network architecture and gain parameterisation are provided in Appendix A.2.

Collecting the realised gains into $\Lambda_t = \mathrm{diag}(g_{t,1}I_d, \ldots, g_{t,K}I_d)$ gives the realised modulation operator

$$M_t = Q\Lambda_t Q^\top = \sum_{k=1}^{K} g_{t,k} P_k, \qquad \widetilde{z}_t = M_t z_t. \tag{1}$$

The latent perspective therefore determines the subspaces along which modulation acts, while the adaptive gains determine how strongly each subspace is modulated. Equivalently, the same operation can be written as $\widetilde{z}_t = z_t + \sum_{k=1}^{K} r_{t,k}$, where $r_{t,k} := \gamma_{t,k} P_k z_t$ is the residual correction contributed by subspace k. This form is used directly in our subspace-intervention analyses. Since all realised gains are positive, each realised modulation operator is symmetric positive definite and full rank. The corresponding operator properties, including invertibility of each realised linear operator, are detailed in Appendix A.2. We apply the same mechanism independently at each of the three context positions to obtain the modulated history $\widetilde{\mathbf{z}}_t$ used for future latent prediction.

2.4 Gain–Basis Coupling

An orthogonal change of perspective alone does not alter the information in the latent representation. It becomes functionally consequential when subsequent computation treats the induced subspaces differently. From equation 1, $M_t Q_k = g_{t,k} Q_k$, so the subspace spanned by Q_k is an eigensubspace of M_t with eigenvalue $g_{t,k}$. The latent perspective therefore determines the subspaces along which modulation acts, while the realised gains determine their amplification.

To separate global scaling from directional modulation, let $\bar{g}_t = K^{-1}\sum_{k=1}^{K} g_{t,k}$. Since $\sum_k P_k = I_D$,

$$M_t = \bar{g}_t I_D + \sum_{k=1}^{K} (g_{t,k} - \bar{g}_t) P_k. \tag{2}$$

The first term scales every latent direction equally, whereas the second captures directional modulation induced by differences between the subspace gains.

Isotropic special case. If all gains are equal, $g_{t,k} = \bar{g}_t$ for every k, the directional term vanishes and $M_t = \bar{g}_t I_D$, giving $\widetilde{z}_t = \bar{g}_t z_t$. For fixed realised gains, the modulation is then independent of the latent perspective. Thus, gain differentiation is necessary for the perspective to have an explicit

directional effect on the realised recomposition operator. We next quantify how this effect depends on the gain differences.

Theorem 1 (Gain–Basis Coupling). *Fix a context t and condition on the realised gains $\{g_{t,k}\}_{k=1}^{K}$. Let $Q(\epsilon) = Q\exp(\epsilon\Omega)$, where $\Omega^\top = -\Omega$, and define $M_t(\epsilon) = Q(\epsilon)\Lambda_t Q(\epsilon)^\top$. Then*

$$\left.\frac{\mathrm{d}M_t(\epsilon)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = Q[\Omega, \Lambda_t]Q^\top, \tag{3}$$

where $[\Omega, \Lambda_t] = \Omega\Lambda_t - \Lambda_t\Omega$. Partitioning Ω into $d \times d$ matrix blocks Ω_{ij} aligned with the K subspaces gives

$$\left\| \left.\frac{\mathrm{d}M_t(\epsilon)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} \right\|_F^2 = 2\sum_{i<j}(g_{t,i} - g_{t,j})^2 \|\Omega_{ij}\|_F^2. \tag{4}$$

The theorem makes the coupling explicit. For fixed realised gains, rotations confined to a single subspace do not change the recomposition operator, whereas perturbations that mix subspaces i and j are weighted by the corresponding gain difference. Thus, for a fixed cross-subspace perturbation, larger gain gaps produce greater local sensitivity of the realised modulation to the latent perspective. Equivalently, replacing Q_k by $Q_k R_k$ for any $R_k \in O(d)$ leaves $P_k = Q_k Q_k^\top$ unchanged, so, conditional on the realised gains, the recomposition operator depends on the subspaces rather than on the particular orthonormal coordinates chosen within them.

This analysis isolates the explicit geometric coupling while holding the realised gains fixed. In the full adaptive model, the latent perspective also affects the gains through the gate inputs $s_{t,k} = Q_k^\top z_t$, providing an additional adaptive coupling not captured by the fixed-gain analysis. We examine empirically whether predictive learning enters this coupled regime, whether the resulting subspaces exhibit functional coordination, and whether their combinations improve physical-dynamics accessibility in Secs. 3.1, 3.2, and 3.3. The proof is provided in Appendix A.3.

2.5 Predictive Learning

We follow the end-to-end representation-learning setup of LeWorldModel (Maes et al., 2026). Given the next latent target $z_{t+1} = f_\theta(o_{t+1})$, the primary predictive loss is

$$\mathcal{L}_{\text{pred}} \triangleq \|\hat{z}_{t+1} - z_{t+1}\|_2^2. \tag{5}$$

Following LeWorldModel (Maes et al., 2026), we additionally apply SIGReg to the encoder latents to prevent representation collapse; its definition and the complete training objective are given in Appendix A.4. Orthogonal subspace modulation introduces no semantic, physical-state, or task-reward supervision. Predictive gradients propagate through the predictor, gating network, QR-parameterised latent perspective, action encoder, and vision encoder, with all learnable components optimised jointly end to end. The latent perspective and adaptive subspace gains are therefore learned through representation-space prediction rather than externally specified subspace labels or physical annotations. Their role is inherently co-adaptive: the encoder determines what is represented, while the learned perspective and gating network determine how that representation is organised and modulated for prediction. A complete chain-rule derivation of these gradient paths is provided in Appendix A.5.

3 Experiments

We evaluate reward-free visual control on PushT (contact-rich pushing), Cube (3D object manipulation), Reacher (articulated reaching), and TwoRoom (navigation), following the evaluation setting of LeWorldModel (Maes et al., 2026). Our primary metric is success rate (SR), defined as the percentage of episodes that satisfy the task-specific success criterion within 50 environment steps. Planning uses the cross-entropy method (CEM). For locally evaluated methods, following LeWorldModel, we report mean SR $\pm$ standard deviation across six CEM planner seeds, with 50 episodes per seed and the same planner seeds used across methods. We use the same task datasets and CEM planning protocol as LeWorldModel, with latent mean-squared error (MSE) to the goal representation as the planning objective (Maes et al., 2026). Our model uses a $D = 192$ latent representation, with $K = 8$

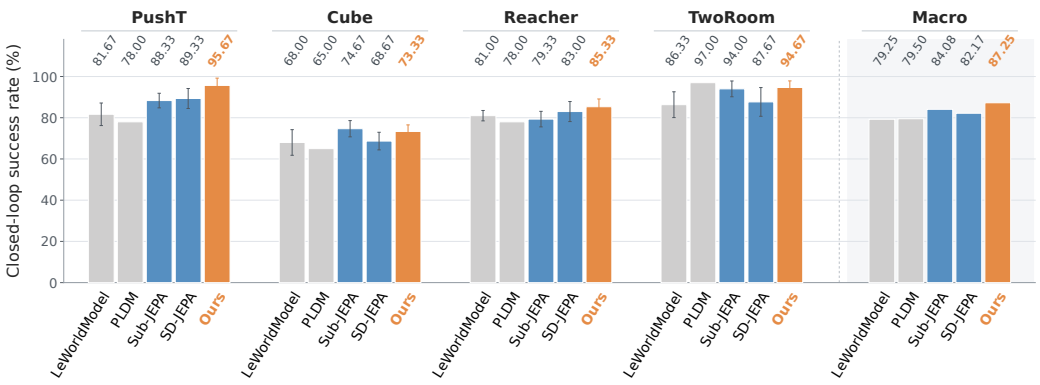


Figure 3: **Closed-loop visual-control performance.** Success rate (SR, %) on PushT, Cube, Reacher, and TwoRoom, together with the unweighted macro-average across tasks. LeWorldModel, Sub-JEPA, SD-JEPA, and Ours are evaluated locally under the common planning protocol; PLDM results are taken from LeWorldModel Maes et al. (2026).

subspaces of dimension 24 on PushT and $K = 16$ subspaces of dimension 12 on Cube, Reacher, and TwoRoom. A complete ablation over the number of subspaces is provided in Appendix B.1.

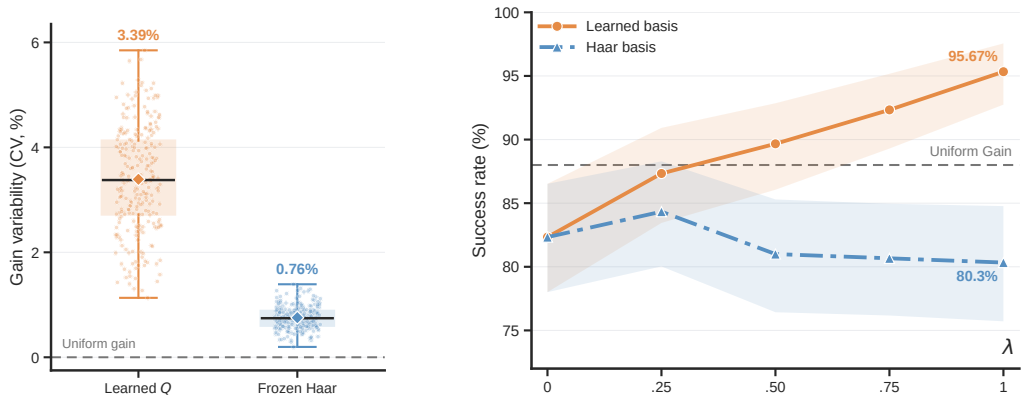
Figure 3 compares closed-loop control performance against latent predictive baselines and methods using predefined latent subspaces. Our model achieves a macro-average SR of 87.25%, exceeding LeWorldModel by 8.00 percentage points and Sub-JEPA by 3.17 points. It improves over LeWorldModel on all four tasks and obtains the highest reported SR on PushT and Reacher, while Sub-JEPA and PLDM obtain the highest reported SR on Cube and TwoRoom, respectively.

Having established the overall control performance, we next examine the role of the learned latent perspective in greater detail. The mechanistic analyses from this point onward focus on PushT, whose contact-rich dynamics provide directly interpretable object translation and rotation variables for controlled interventions and physical probing. The broader effect of perspective and gain design is evaluated across all four tasks in Appendix B.1. We first test the Gain–Basis Coupling predicted by Sec. 2.4. We then examine functional coordination across subspaces and physical-dynamics accessibility under matched representation capacity.

3.1 Does Predictive Learning Realise Gain–Basis Coupling?

Yes: predictive learning develops stronger inter-subspace gain differentiation, while same-checkpoint interventions confirm that the effect of these gains depends on the directions along which they act. Sec. 2.4 shows that, for fixed realised gains, the directional effect of the perspective vanishes when the gains are equal and increases with inter-subspace gain differences. We therefore examine whether such differentiation emerges during training and whether its utility depends on the learned directions. Full variant definitions and intervention details are provided in Appendix B.1.

Training produces the required gain differentiation. Across the four control tasks, Uniform Gain, Frozen Haar, and the full model achieve macro-average SRs of 80.50%, 82.75%, and 87.25%, respectively (Appendix B.1). Since these variants are trained independently, this ordering alone does not isolate the mechanism. We therefore inspect the realised gains directly on the common PushT evaluation contexts. We measure inter-subspace gain differentiation using the coefficient of variation (CV); its precise definition and computation across the fixed PushT evaluation cohort are provided in Appendix B.1. Figure 4a shows substantially stronger inter-subspace differentiation in the full model, with a mean CV of 3.39% compared with 0.76% for Frozen Haar. Uniform Gain has zero inter-subspace variation by construction. Joint predictive learning therefore produces the differentiated modulation required for the latent perspective to exert an explicit directional effect.



(a) Native inter-subspace gain differentiation.

(b) Same-checkpoint perspective intervention.

Figure 4: **Gain–Basis Coupling on PushT.** (a) Learning the perspective is accompanied by stronger native gain differentiation. (b) Under the same checkpoint and realised gains, performance under the learned perspective increasingly separates from that under a fixed Haar-random replacement as native gain differentiation is restored. The horizontal Uniform Gain line is an independently trained reference and is not the $\lambda = 0$ state of the intervened full-model checkpoint.

Differentiated gains are more effective on the learned perspective. We isolate this dependence within a single full-model checkpoint by interpolating its native gain differentiation from the isotropic case to its original strength. At each interpolation level, the resulting realised gains are held fixed while recomposition is evaluated using either the learned perspective or a fixed Haar-random replacement. The full intervention construction and closed-loop success-rate evaluation protocol are detailed in Appendix B.1. At zero differentiation, the two perspectives are equivalent in effect, as predicted by Sec. 2.4. As the native gain differences are restored, their control performance progressively separates (Fig. 4b). At full differentiation, the learned perspective reaches 95.67% SR, whereas the Haar-random replacement yields 80.3%. Because the checkpoint and realised gains are otherwise unchanged, the intervention shows that the utility of differentiated modulation depends strongly on the directions on which it acts.

Together, these results empirically validate Gain–Basis Coupling: predictive learning develops the gain differentiation required for the perspective to have a directional effect, and the utility of those differentiated gains depends strongly on the learned directions.

3.2 Does the Learned Latent Perspective Enable Functional Coordination Across Subspaces?

Yes: the learned subspaces remain constructively useful in combination, with every learned correction retaining additional predictive value even when the other seven corrections are already active. We use *functional coordination* to describe this persistent marginal predictive utility under composition. We evaluate it directly with leave-one-out (LOO) interventions on the same fixed cohort of 256 PushT transitions. For each subspace, its residual correction is removed jointly from all three history positions while the remaining seven corrections are kept at their native values. Full intervention definitions, exhaustive coalition analysis, and exact Shapley attribution are provided in Appendix B.2.

All learned subspaces retain marginal predictive utility. Figure 5 shows that removing any learned subspace correction increases prediction error, giving positive leave-one-out effects for all eight subspaces. The relative degradation ranges from 0.4% to 10.0%, with particularly strong effects for S2, S3, and S8. Thus, each learned correction retains additional predictive utility even when the remaining seven are already active. In contrast, Frozen Haar is positive for only six of eight subspaces: removing S5 or S6 slightly improves prediction, indicating that subspace-specific adaptive modulation alone does not guarantee constructive participation once the corrections are combined.

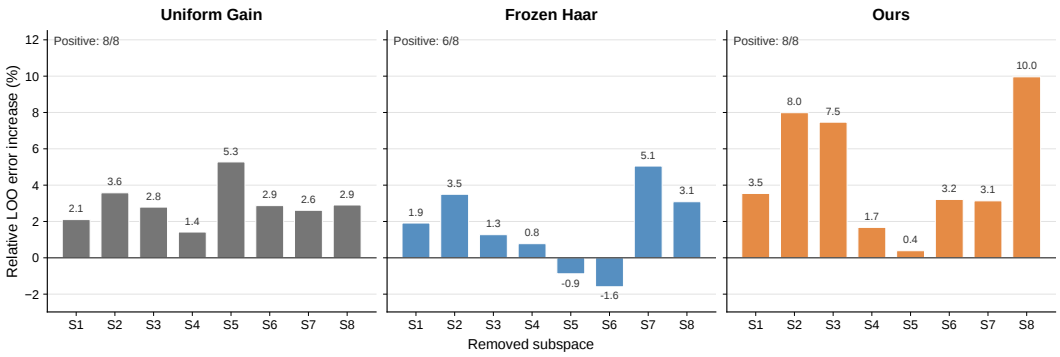


Figure 5: **Leave-one-out predictive contributions on PushT.** Bars show the relative increase in full-latent squared prediction error when one subspace correction is removed while the remaining seven stay active. Positive values indicate additional predictive utility.

The perspective allows useful gains to act directionally. Uniform Gain also yields positive leave-one-out effects for all eight diagnostic subspaces, showing that the context-dependent gating mechanism itself provides useful predictive modulation. However, Uniform Gain applies the same realised gain to every subspace within a context, making its native modulation isotropic and removing any directional role for the latent perspective. Its positive leave-one-out effects are therefore generally weaker than those of the full model in Fig. 5: the gain remains useful, but it cannot be differentially allocated across perspective-defined directions. By contrast, the full model combines a learned perspective with differentiated subspace gains, allowing useful modulation to be concentrated along directions where it contributes more strongly. This contrast complements the Gain–Basis Coupling analysis: effective gains alone are not sufficient to realise directional modulation; their effect depends on the subspace directions along which they act.

Coordination is distributed but non-uniform. The positive leave-one-out effects of the full model are distributed across all eight learned subspaces but vary substantially in magnitude, indicating that their predictive contributions are coordinated without being allocated uniformly. The broader intervention analysis in Appendix B.2 supports the same pattern: the full model remains constructive across all tested cohort-averaged subspace backgrounds, whereas Frozen Haar is positive in only 79.49% of configurations, and exact Shapley attribution assigns positive mean contribution to all eight learned subspaces but not to all Frozen-Haar subspaces.

Together, these results show that functional coordination emerges from the interaction between useful adaptive gains and the latent perspective on which they act. A shared isotropic gain can already improve prediction, but learning perspective-defined directions together with differentiated gains allows that modulation to be used more effectively across subspaces, reinforcing the empirical importance of Gain–Basis Coupling.

3.3 Does the Learned Latent Perspective Improve Physical-Dynamics Accessibility?

Yes, but selectively: the model trained with a learned latent perspective makes the manipulated object's dynamics, particularly its rotation, more linearly accessible under matched representation budgets, while pusher dynamics show little advantage. We evaluate this using physical readout from encoder representations of actual PushT transitions, probing pusher displacement, T-object displacement, and T-object rotation. At each feature budget from 48 to 192 dimensions, we evaluate all admissible subspace or projection combinations and summarise their linear readout rather than selecting the best-performing subset. Frozen Haar provides the closest architectural control because it retains subspace-specific adaptive modulation while fixing the orthogonal perspective, whereas Sub-JEPA provides a comparison with predefined projected subspaces. Uniform Gain is omitted because its native modulation is isotropic, so the latent perspective has no directional role in the forward computation; SD-JEPA is shown only as a full-latent reference. Because these models are trained independently, the comparison reflects the resulting representations and decompositions

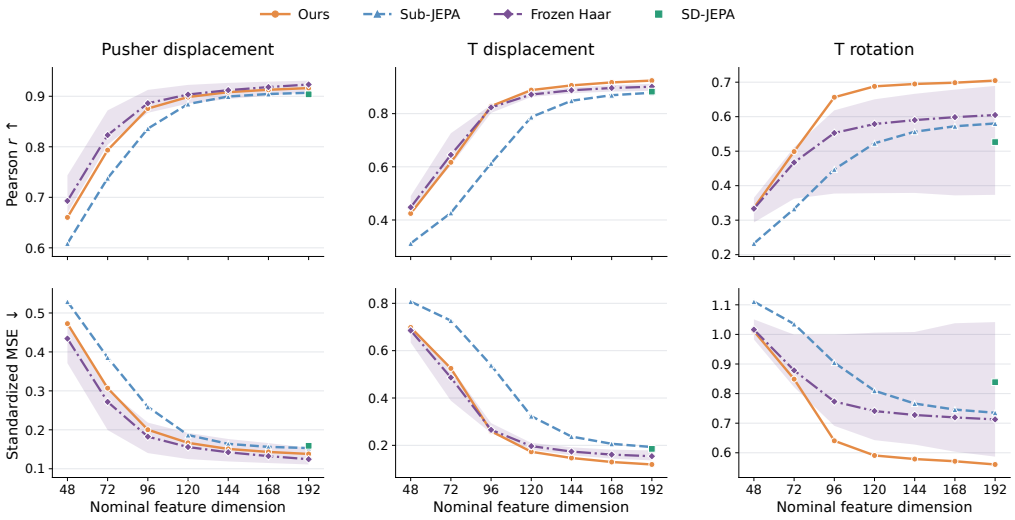


Figure 6: **Matched-capacity physical readout on PushT.** Higher Pearson r and lower standardised MSE indicate stronger linear accessibility. Frozen-Haar shading shows the range across independently trained checkpoints; SD-JEPA is shown only at the full 192-dimensional representation.

jointly rather than a post-hoc effect of changing the orthogonal basis alone. Full probing details are provided in Appendix B.3.

The strongest accessibility gains occur for object dynamics. Figure 6 shows the clearest separation for T-object rotation: Ours achieves consistently higher Pearson correlation and lower standardised MSE than both Sub-JEPA and Frozen Haar from moderate capacities onward. T-object displacement shows a similar but smaller advantage over Sub-JEPA, with Ours and Frozen Haar becoming close at several capacities. Pusher displacement is comparatively easy to recover for all models and approaches a common high-readout regime as capacity increases; Frozen Haar is slightly stronger at several lower and intermediate budgets. The accessibility improvement is therefore dynamics-dependent rather than uniform. In particular, end-to-end learning with the latent perspective produces its clearest advantage for the translation and, especially, rotation of the manipulated object.

4 Conclusion and Future Work

We introduced orthogonal subspace modulation, which learns not only a predictive representation but also the latent perspective from which that representation is decomposed and adaptively used. The learned orthogonal basis determines where modulation acts, while context-dependent subspace gains determine how strongly those directions are modulated. Our Gain–Basis Coupling analysis shows that the perspective becomes directionally consequential when realised gains differ across subspaces, and the experiments show that predictive learning both develops this differentiation and exploits the learned directions in practice. Across four reward-free visual-control tasks, the resulting model achieves a macro-average success rate of 87.25%. Further analyses reveal functional coordination across learned subspaces and greater physical-dynamics accessibility under matched representation budgets, with the clearest accessibility gains appearing for the dynamics of the manipulated object. These effects are not uniform across all physical variables, and our detailed mechanistic analyses focus on the relatively compact PushT environment. It therefore remains open how stable the learned perspective is under changes in dynamics, goals, embodiment, and observation statistics, whether it can be efficiently reconfigured without relearning the full representation, and whether similar organisation emerges in richer manipulation, navigation, multimodal, and multi-task world models. Together, our results suggest that how a latent representation is viewed and organised for downstream computation is itself an important part of learning a predictive world model.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. URL https://arxiv.org/abs/2511.08544.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model?, 2026. URL https://arxiv.org/abs/2605.26379.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

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, 2025. doi: 10.52202/085713-1465.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Subspace-decomposed JEPAs: Disentangling progression and content in latent world models, 2026. URL https://arxiv.org/abs/2605.31111.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models, 2026. URL https://arxiv.org/abs/2607.22430.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. Sub-JEPA: Subspace gaussian regularization for stable end-to-end world models, 2026. URL https://arxiv.org/abs/2605.09241.

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. URL https://proceedings.mlr.press/v267/zhou25t.html.

AI USE STATEMENT

Generative AI tools were used to assist with translation and language editing of author-written text, and to provide suggestions for software code editing and debugging. They were not used to generate experimental data, formulate the core research hypotheses or methodology, derive theoretical claims, or interpret experimental results.

All AI-assisted text was reviewed and revised by the authors. AI-assisted code suggestions were manually inspected, tested, and validated by the authors before being incorporated into the experimental pipeline. All reported experimental results and scientific claims were independently verified by the authors. The authors take full responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

A METHOD DETAILS

A.1 LATENT PERSPECTIVE PARAMETERISATION

We parameterise the latent perspective Q through an unconstrained matrix $U \in \mathbb{R}^{D\times D}$. At each forward pass, we compute the QR factorisation $U = \bar{Q}R$, where $\bar{Q}$ is orthogonal and R is upper triangular. Because the QR factorisation is ambiguous up to column signs, we impose the deterministic sign convention

$$D_R = \operatorname{diag}\big(\operatorname{sgn}_+(R_{11}), \ldots, \operatorname{sgn}_+(R_{DD})\big), \qquad \operatorname{sgn}_+(r) = \begin{cases} +1, & r \geq 0, \\ -1, & r < 0. \end{cases} \tag{6}$$

The latent perspective is then given by

$$Q = \bar{Q}D_R. \tag{7}$$

This fixes the column-sign convention while preserving $Q^\top Q = I_D$. Gradients are propagated through the QR factorisation, allowing the latent perspective to be learned jointly with the encoder, gating network, and predictor under the predictive objective.

A.2 ADAPTIVE SUBSPACE GAIN PARAMETERISATION

For each subspace, the gating network receives the local coordinates $s_{t,k} = Q_k^\top z_t$ together with the action embedding h_t. The action embedding is first linearly projected to the subspace dimension as $\bar{h}_t = W_a h_t + b_a \in \mathbb{R}^d$ and concatenated with the subspace coordinates to form $x_{t,k} = [s_{t,k}; \bar{h}_t] \in \mathbb{R}^{2d}$. A shared two-layer MLP produces a scalar logit for each subspace,

$$\ell_{t,k} = W_2\ \mathrm{SiLU}(\mathrm{LN}(W_1 x_{t,k} + b_1)) + b_2. \tag{8}$$

The MLP parameters are shared across all subspaces and use a hidden dimension of 256. On PushT, $d = 24$ and the action embedding has dimension 192, giving an action projection $\mathbb{R}^{192} \to \mathbb{R}^{24}$ and a 48-dimensional gating-network input.

The logit is converted to the bounded gate activation $m_{t,k} = \sigma(\ell_{t,k}/\tau)$, while the learned global scale is parameterised as $\alpha = \sigma(\widehat{\alpha})$. The residual modulation strength and realised subspace gain are therefore

$$\gamma_{t,k} = \alpha m_{t,k}, \qquad g_{t,k} = 1 + \gamma_{t,k} = 1 + \alpha\sigma\left(\frac{\ell_{t,k}}{\tau}\right). \tag{9}$$

For finite logits, $0 < \gamma_{t,k} < \alpha$ and $1 < g_{t,k} < 1 + \alpha$, approaching the corresponding closed bounds in the saturation limit. We use $\tau = 1$ in the main experiments. For the reported PushT checkpoint, $\widehat{\alpha} = -1.41528$, giving $\alpha = 0.19540$ and hence an upper gain bound of approximately 1.1954.

Gate budget. We apply a small regulariser to the mean sigmoid activation,

$$\mathcal{L}_{\mathrm{budget}} = \lambda_{\mathrm{budget}}(\bar{m} - \rho)^2, \qquad \bar{m} = \mathrm{mean}_{\mathrm{batch},\,t,\,k}\, m_{t,k}. \tag{10}$$

The main model uses $\rho = 0.3$ and $\lambda_{\mathrm{budget}} = 0.01$. The budget is applied to the sigmoid activation $m_{t,k}$ rather than directly to the residual modulation strength $\gamma_{t,k}$. Because the weighting coefficient is already included in the definition of $\mathcal{L}_{\mathrm{budget}}$, this term enters the complete training objective without an additional multiplicative factor.

Properties of the realised modulation operator. For any fixed set of realised gains, $\Lambda_t = \operatorname{diag}(g_{t,1}I_d, \ldots, g_{t,K}I_d)$ is positive definite. Since Q is orthogonal, $M_t = Q\Lambda_t Q^\top$ is symmetric positive definite and full rank, with inverse $M_t^{-1} = Q\Lambda_t^{-1}Q^\top$. Thus, each realised linear modulation operator is invertible and does not annihilate any latent direction. This pointwise invertibility does not imply that the full adaptive map $z_t \mapsto M_t(z_t, h_t)z_t$ is globally invertible, because the realised gains themselves depend on the input through the gating network.

A.3 PROOF OF THE GAIN–BASIS COUPLING THEOREM

For $Q(\epsilon) = Q\exp(\epsilon\Omega)$ with $\Omega^\top = -\Omega$, we have $Q(\epsilon)^\top = \exp(-\epsilon\Omega)Q^\top$. Therefore,

$$M_t(\epsilon) = Q\exp(\epsilon\Omega)\Lambda_t \exp(-\epsilon\Omega)Q^\top. \tag{11}$$

Differentiating at $\epsilon = 0$ gives

$$\left.\frac{\mathrm{d}M_t(\epsilon)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = Q(\Omega\Lambda_t - \Lambda_t\Omega)Q^\top = Q[\Omega, \Lambda_t]Q^\top. \tag{12}$$

Because Q is orthogonal, the Frobenius norm is invariant under orthogonal conjugation,

$$\left\| \left.\frac{\mathrm{d}M_t(\epsilon)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} \right\|_F^2 = \|[\Omega, \Lambda_t]\|_F^2. \tag{13}$$

Recall that $\Lambda_t = \mathrm{diag}(g_{t,1}I_d, \ldots, g_{t,K}I_d)$. Writing Ω in the same $K \times K$ matrix-block structure aligned with the subspaces, the (i, j) block of the commutator is

$$[\Omega, \Lambda_t]_{ij} = \Omega_{ij}g_{t,j} - g_{t,i}\Omega_{ij} = (g_{t,j} - g_{t,i})\Omega_{ij}. \tag{14}$$

The diagonal matrix blocks therefore vanish, and

$$\|[\Omega, \Lambda_t]\|_F^2 = \sum_{i \neq j} (g_{t,i} - g_{t,j})^2 \|\Omega_{ij}\|_F^2. \tag{15}$$

Since $\Omega^\top = -\Omega$, we have $\Omega_{ji} = -\Omega_{ij}^\top$ and hence $\|\Omega_{ji}\|_F = \|\Omega_{ij}\|_F$. Pairing the (i, j) and (j, i) terms yields

$$\left\| \left.\frac{\mathrm{d}M_t(\epsilon)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} \right\|_F^2 = 2\sum_{i<j} (g_{t,i} - g_{t,j})^2 \|\Omega_{ij}\|_F^2. \tag{16}$$

Thus, perturbations confined within a subspace do not change the realised recomposition to first order, whereas perturbations that mix subspaces are weighted by the corresponding gain differences. This proves Theorem 1.

A.4 Training Objective

Following LeWorldModel (Maes et al., 2026), we combine next-latent prediction with the Sketched Isotropic Gaussian Regularizer (SIGReg), originally introduced in LeJEPA (Balestriero & LeCun, 2025). Let $Z \in \mathbb{R}^{N \times D}$ denote a collection of encoder latents and let $u^{(m)} \in \mathbb{S}^{D-1}$, $m = 1, \ldots, M$, denote randomly sampled unit directions. SIGReg projects the latents onto these directions,

$$h^{(m)} = Zu^{(m)}, \qquad u^{(m)} \in \mathbb{S}^{D-1}, \tag{17}$$

and averages a univariate Epps–Pulley statistic across the resulting projections,

$$\mathcal{L}_{\mathrm{SIGReg}} = \mathrm{SIGReg}(Z) = \frac{1}{M}\sum_{m=1}^{M} T\left(h^{(m)}\right). \tag{18}$$

Following the formulation used by LeWorldModel (Maes et al., 2026), each statistic compares the empirical characteristic function of the projected samples with that of a standard Gaussian,

$$T\left(h^{(m)}\right) = \int_{-\infty}^{\infty} w(\omega) \left|\phi_N\left(\omega; h^{(m)}\right) - \phi_0(\omega)\right|^2 \mathrm{d}\omega, \tag{19}$$

where

$$\phi_N(\omega; h) = \frac{1}{N}\sum_{n=1}^{N} e^{i\omega h_n}, \tag{20}$$

ϕ_0 denotes the characteristic function of $\mathcal{N}(0, 1)$, and $w(\omega)$ is the Gaussian weighting function used by SIGReg.

The representation-learning component of the objective is

$$\mathcal{L}_{\mathrm{repr}} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{SIGReg}}\mathcal{L}_{\mathrm{SIGReg}}. \tag{21}$$

Including the gate-budget regulariser from equation 10, the complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{SIGReg}}\mathcal{L}_{\mathrm{SIGReg}} + \mathcal{L}_{\mathrm{budget}}. \tag{22}$$

Unless otherwise stated, we follow the optimisation hyperparameters and training protocol of LeWorldModel (Maes et al., 2026). All learnable components, including the vision encoder, action encoder, predictor, gating network, and QR-parameterised latent perspective, are optimised jointly end to end under equation 22. No semantic, physical-state, or task-reward supervision is used to learn the latent perspective or subspace gains.

A.5 End-to-End Prediction Gradients

We make explicit how the predictive signal in equation 5 reaches the adaptive gains and latent perspective. The derivation below considers one generic context position; the same computation is applied independently to all three context positions, and the resulting gradients are accumulated during back-propagation. Let

$$u_t := \frac{\partial \mathcal{L}_{\text{pred}}}{\partial \widetilde{z}_t} \in \mathbb{R}^D \tag{23}$$

denote the upstream gradient arriving at the modulated latent. Using the residual form $\widetilde{z}_t = z_t + \sum_{k=1}^{K} \gamma_{t,k} P_k z_t$, we obtain the following gradient paths.

Gradients to the adaptive gains. Holding z_t and Q fixed locally,

$$\frac{\partial \widetilde{z}_t}{\partial \gamma_{t,k}} = P_k z_t, \qquad \frac{\partial \mathcal{L}_{\text{pred}}}{\partial \gamma_{t,k}} = u_t^\top P_k z_t. \tag{24}$$

Let $\eta_{t,k} := u_t^\top P_k z_t$. Using $\gamma_{t,k} = \alpha m_{t,k}$, $m_{t,k} = \sigma(\ell_{t,k}/\tau)$, and $\alpha = \sigma(\widehat{\alpha})$, the chain rule gives

$$\frac{\partial \mathcal{L}_{\text{pred}}}{\partial m_{t,k}} = \alpha \eta_{t,k}, \qquad \frac{\partial \mathcal{L}_{\text{pred}}}{\partial \ell_{t,k}} = \frac{\alpha}{\tau} m_{t,k}(1 - m_{t,k}) \eta_{t,k}. \tag{25}$$

For a single context position, the learned global scale receives

$$\frac{\partial \mathcal{L}_{\text{pred}}}{\partial \widehat{\alpha}} = \alpha(1-\alpha) \sum_{k=1}^{K} m_{t,k} \eta_{t,k}, \tag{26}$$

with contributions summed over context positions and samples during training. Thus, the prediction error directly determines both the relative subspace activations and the global modulation scale.

Gradients to the latent perspective. The latent perspective affects modulation through two paths: directly through the projector $P_k = Q_k Q_k^\top$ and indirectly through the gating-network coordinates $s_{t,k} = Q_k^\top z_t$. Let

$$a_{t,k} := \left.\frac{\partial \mathcal{L}_{\text{pred}}}{\partial s_{t,k}}\right|_{\text{gate}} = \frac{\partial \mathcal{L}_{\text{pred}}}{\partial \ell_{t,k}} \nabla_{s_{t,k}} \ell_{t,k}. \tag{27}$$

For the kth subspace residual, differentiation of $P_k z_t = Q_k Q_k^\top z_t$ gives the direct projector contribution

$$\left.\frac{\partial \mathcal{L}_{\text{pred}}}{\partial Q_k}\right|_{\text{proj}} = \gamma_{t,k} \left[u_t s_{t,k}^\top + z_t (Q_k^\top u_t)^\top \right]. \tag{28}$$

The dependence of the gating network on $s_{t,k} = Q_k^\top z_t$ contributes

$$\left.\frac{\partial \mathcal{L}_{\text{pred}}}{\partial Q_k}\right|_{\text{gate}} = z_t a_{t,k}^\top. \tag{29}$$

Hence the predictive gradient reaching Q_k is

$$\frac{\partial \mathcal{L}_{\text{pred}}}{\partial Q_k} = \gamma_{t,k} \left[u_t s_{t,k}^\top + z_t (Q_k^\top u_t)^\top \right] + z_t a_{t,k}^\top. \tag{30}$$

The first term is the direct gradient through the subspace projector, while the second arises from the dependence of the adaptive gains on the perspective-defined gate inputs.

Gradients through the QR parameterisation. The matrices Q_k are not free parameters; they are induced by the unconstrained matrix U through the sign-corrected QR map described in Appendix A.1. Concatenating the subspace gradients gives

$$G_Q = \left[\frac{\partial \mathcal{L}_{\text{pred}}}{\partial Q_1}, \ldots, \frac{\partial \mathcal{L}_{\text{pred}}}{\partial Q_K} \right]. \tag{31}$$

Writing the QR-based map as $Q = \mathcal{Q}(U)$, the gradient with respect to the unconstrained parameter is

$$\nabla_U \mathcal{L}_{\text{pred}} = D\mathcal{Q}(U)^* \left[G_Q \right], \tag{32}$$

where $D\mathcal{Q}(U)^*$ denotes the adjoint of the differential of the QR-based parameterisation. In practice, this operation is evaluated by automatic differentiation through the QR factorisation. Away from sign changes of the diagonal entries of R, the sign correction is locally constant, so predictive gradients propagate continuously from Q to U.

Gradients to the latent encoder. The context latent z_t affects prediction both directly through the residual modulation and indirectly through the gating-network coordinates. Since $\nabla_{z_t}\gamma_{t,k} = Q_k \nabla_{s_{t,k}}\gamma_{t,k}$, the local predictive gradient reaching the context latent is

$$\frac{\partial \mathcal{L}_{\text{pred}}}{\partial z_t} = M_t u_t + \sum_{k=1}^{K} \eta_{t,k} \nabla_{z_t}\gamma_{t,k}. \tag{33}$$

The first term is the direct gradient through the realised modulation operator, while the second accounts for the input dependence of the adaptive gains.

The target branch is also trainable. For the per-sample squared prediction error,

$$\frac{\partial \mathcal{L}_{\text{pred}}}{\partial \hat{z}_{t+1}} = 2(\hat{z}_{t+1} - z_{t+1}), \qquad \frac{\partial \mathcal{L}_{\text{pred}}}{\partial z_{t+1}} = -2(\hat{z}_{t+1} - z_{t+1}). \tag{34}$$

Because both context and target latents are produced by the trainable vision encoder, their gradient contributions are accumulated in f_θ.

Gradients to the action pathway. The action embedding affects the loss through both the predictor input and the gating network. The gating-network contribution is

$$\left.\frac{\partial \mathcal{L}_{\text{pred}}}{\partial h_t}\right|_{\text{gate}} = \sum_{k=1}^{K} \eta_{t,k} \nabla_{h_t}\gamma_{t,k}, \tag{35}$$

which is combined with the direct predictor gradient through $\mathbf{h}_t$ and propagated further through the action encoder g_ξ.

Taken together, equation 24–equation 35 show that the future-latent prediction error reaches the adaptive gains, gating network, latent perspective, QR parameter U, vision encoder, action encoder, and predictor within a single end-to-end computational graph. SIGReg adds an auxiliary gradient to the encoder representation, while the small gate-budget regulariser adds an auxiliary gradient through the gating-network activations; neither introduces semantic or physical supervision.

B Additional Experimental Details and Analyses

B.1 Gain–Basis Coupling: Variants and Interventions

Independently trained variants. We compare three variants trained independently under the same world-model setting.

Uniform Gain. This variant retains context-dependent adaptive modulation but removes inter-subspace gain differentiation. At each context position, the gating network receives the full latent $z_t \in \mathbb{R}^{192}$ together with the 192-dimensional action embedding h_t. Their concatenation is processed by a two-layer MLP,

$$\ell_t = W_2\ \mathrm{SiLU}(\mathrm{LN}(W_1[z_t; h_t] + b_1)) + b_2,$$

with hidden dimension 256. The resulting scalar is converted using the same bounded positive residual parameterisation,

$$m_t = \sigma(\ell_t), \qquad \alpha_{\mathrm{u}} = \sigma(\widehat{\alpha}_{\mathrm{u}}), \qquad \gamma_t = \alpha_{\mathrm{u}} m_t, \qquad g_t = 1 + \gamma_t,$$

and the same realised gain is applied to the entire latent,

$$\widetilde{z}_t = g_t z_t.$$

Equivalently, under any orthogonal decomposition this operation is $Q(g_t I_D)Q^\top z_t = g_t z_t$, so the latent perspective has no directional role in the native forward computation. Uniform Gain therefore preserves context-dependent gating while removing subspace-specific gain allocation.

Frozen Haar. This variant retains the full subspace-specific gating mechanism and bounded gain parameterisation of Sec. 2.3, but replaces the learnable latent perspective with a Haar-random orthogonal basis that is fixed throughout training. The encoder, action encoder, gating network, and

Table 1: **Perspective and gain ablations.** Closed-loop SR (%) for independently trained variants. Entries report mean ± standard deviation across the same planner seeds used in the main evaluation. Macro is the unweighted average across the four tasks.

Variant	PushT	Cube	Reacher	TwoRoom	Macro
Uniform Gain	88.00 ± 6.43	71.33 ± 5.73	69.67 ± 3.73	93.00 ± 3.79	80.50
Frozen Haar	89.00 ± 2.77	69.33 ± 7.36	80.67 ± 4.85	92.00 ± 3.06	82.75
Learned Perspective	$\mathbf{95.67 \pm 3.54}$	$\mathbf{73.33 \pm 3.20}$	$\mathbf{85.33 \pm 3.77}$	$\mathbf{94.67 \pm 3.20}$	**87.25**

Table 2: **Ablation over the number of subspaces.** Closed-loop success rate (SR, %) with fixed latent dimensionality $D = 192$. Entries report mean ± standard deviation across the same CEM planner seeds used in the main evaluation. Bold values indicate the setting used for each task.

K	PushT	Cube	Reacher	TwoRoom
2	90.67 ± 5.62	69.00 ± 7.37	84.33 ± 5.59	93.67 ± 4.23
4	90.00 ± 4.47	72.67 ± 4.71	81.33 ± 8.77	92.67 ± 3.77
8	$\mathbf{95.67 \pm 3.54}$	70.00 ± 6.11	84.33 ± 7.06	93.67 ± 4.38
16	90.67 ± 4.11	$\mathbf{73.33 \pm 3.20}$	$\mathbf{85.33 \pm 3.77}$	$\mathbf{94.67 \pm 3.20}$
32	91.33 ± 4.85	69.67 ± 5.09	81.67 ± 4.23	94.00 ± 3.46
64	92.00 ± 3.06	69.67 ± 7.78	84.33 ± 2.92	93.33 ± 4.99

predictor remain trainable. Frozen Haar therefore preserves adaptive subspace-specific modulation while removing learning of the perspective itself.

Learned Perspective. The full model uses the same subspace-specific gating mechanism as Frozen Haar, but learns the orthogonal basis jointly with the encoder, action encoder, gating network, and predictor through the QR parameterisation described in Sec. 2.2 and Appendix A.1. It therefore learns both where modulation acts and how strongly each induced subspace is modulated.

Table 1 reports closed-loop control performance for the three independently trained variants across all four tasks.

Number of subspaces. Following the subspace-count range considered in Sub-JEPA (Zhao et al., 2026), we vary the number of subspaces over $K \in \{2, 4, 8, 16, 32, 64\}$ while keeping the latent dimensionality fixed at $D = 192$, so that each subspace has dimension $d = D/K$. Table 2 reports the corresponding closed-loop control performance. Performance remains relatively stable over a broad range of decompositions, although the preferred granularity is task dependent. PushT achieves its strongest result with $K = 8$ ($d = 24$), whereas Cube, Reacher, and TwoRoom attain their highest SRs with $K = 16$ ($d = 12$). We therefore use $K = 8$ on PushT and $K = 16$ on the remaining tasks throughout the main experiments.

Gain differentiation. We quantify native inter-subspace gain differentiation as $\mathrm{CV}_t = \mathrm{Std}_k(g_{t,k}) / \mathrm{Mean}_k(g_{t,k})$. The statistic is computed across subspaces within each state–action context and then averaged over the same fixed set of 256 PushT contexts used throughout the mechanistic analyses. Uniform Gain has $\mathrm{CV}_t = 0$ by construction. The full model and Frozen Haar yield mean CVs of 3.39% and 0.76%, respectively, as reported in Fig. 4a.

Same-checkpoint perspective intervention. To isolate the dependence of differentiated gains on the latent perspective, we intervene on a single trained full-model checkpoint. Its native realised gains are interpolated as

$$g_{t,k}^{(\lambda)} = \bar{g}_t + \lambda \, (g_{t,k} - \bar{g}_t) \,, \qquad \lambda \in [0, 1], \tag{36}$$

and the resulting gains are then held fixed during recomposition. We compare the learned latent perspective with a single Haar-random orthogonal replacement sampled once and kept fixed throughout the intervention. The gating network is not re-evaluated after replacing the perspective, so the two conditions use exactly the same realised gain pattern and differ only in the subspace directions on which those gains act.

The interpolation scales every pairwise gain difference as $g_{t,i}^{(\lambda)} - g_{t,j}^{(\lambda)} = \lambda(g_{t,i} - g_{t,j})$. Hence $\lambda = 0$ removes all inter-subspace gain differentiation and yields the isotropic case $M_t = \bar{g}_t I_D$, whereas $\lambda = 1$ restores the checkpoint's native gain pattern. This intervention therefore directly controls the gain differences appearing in Theorem 1 without retraining or recomputing the gating network.

B.2 Functional Coordination Across Subspaces on PushT

Evaluation cohort. All predictive-intervention analyses use the same fixed cohort of 256 PushT transitions across model variants. The cohort is episode-unique and balanced across four motion regimes—stationary, translation-only, rotation-only, and coupled translation–rotation—with 64 transitions in each regime. The same cohort is also used for the primary PushT physical-dynamics readout.

Unified intervention setup. Each subspace is treated as one intervention unit across the full three-step context. For a trained checkpoint, we first compute the native residual correction $r_{t-j,k}$ of subspace k at each context position $t - j$, with $j \in \{0, 1, 2\}$. For an active-subspace set $S \subseteq \mathcal{K}$, where $\mathcal{K} = \{1, \ldots, K\}$, the intervened latent is

$$\tilde{z}_{t-j}^{(S)} = z_{t-j} + \sum_{k \in S} r_{t-j,k}. \tag{37}$$

The original latent is always retained; the intervention only removes or restores the checkpoint-native residual corrections. Model parameters, gating-network outputs, and residual values remain fixed throughout the comparison. For transition n, the resulting full-latent sum of squared prediction errors (SSE) is

$$\mathcal{E}_n(S) = \left\| \hat{z}_{n,t+1}^{(S)} - z_{n,t+1} \right\|_2^2. \tag{38}$$

Leave-one-out contributions. The main text reports leave-one-out (LOO) predictive contributions. For subspace k, we remove its correction while all remaining corrections stay active:

$$\Delta_{n,k}^{\mathrm{LOO}} = \mathcal{E}_n(\mathcal{K} \setminus \{k\}) - \mathcal{E}_n(\mathcal{K}). \tag{39}$$

A positive value means that subspace k provides additional predictive utility in the presence of the other $K - 1$ corrections. The cohort-averaged absolute LOO contribution is

$$\overline{\Delta}_k^{\mathrm{LOO}} = \frac{1}{N} \sum_{n=1}^{N} \Delta_{n,k}^{\mathrm{LOO}}, \qquad N = 256. \tag{40}$$

For Fig. 5, we report the cohort-aggregated relative increase by normalising with the cohort-mean full-coalition error $\overline{\mathcal{E}}(\mathcal{K}) = N^{-1} \sum_{n=1}^{N} \mathcal{E}_n(\mathcal{K})$:

$$\delta_k^{\mathrm{LOO}} = 100 \, \frac{\overline{\Delta}_k^{\mathrm{LOO}}}{\overline{\mathcal{E}}(\mathcal{K})} = 100 \, \frac{\sum_{n=1}^{N} \Delta_{n,k}^{\mathrm{LOO}}}{\sum_{n=1}^{N} \mathcal{E}_n(\mathcal{K})}. \tag{41}$$

Thus, $\overline{\Delta}_k^{\mathrm{LOO}}$ is the absolute increase in full-latent SSE, whereas δ_k^{LOO} is the relative quantity reported in the main text.

Exhaustive conditional contributions. LOO tests the strongest coalition background, in which all other corrections are active. To examine functional coordination more broadly, we additionally evaluate each subspace under every configuration of the remaining seven subspaces. For $k \notin S$, its conditional contribution is

$$\Delta_{n,k}(S) = \mathcal{E}_n(S) - \mathcal{E}_n(S \cup \{k\}), \qquad S \subseteq \mathcal{K} \setminus \{k\}. \tag{42}$$

A positive value means that activating subspace k further reduces prediction error given the corrections already active in S. With eight PushT subspaces, each tested subspace is evaluated under all $2^7 = 128$ configurations of the remaining seven. We visualise the cohort mean $\overline{\Delta}_k(S) = N^{-1} \sum_{n=1}^{N} \Delta_{n,k}(S)$. The positive-combination statistic is the percentage of the resulting 8×128 subspace–configuration pairs satisfying $\overline{\Delta}_k(S) > 0$.

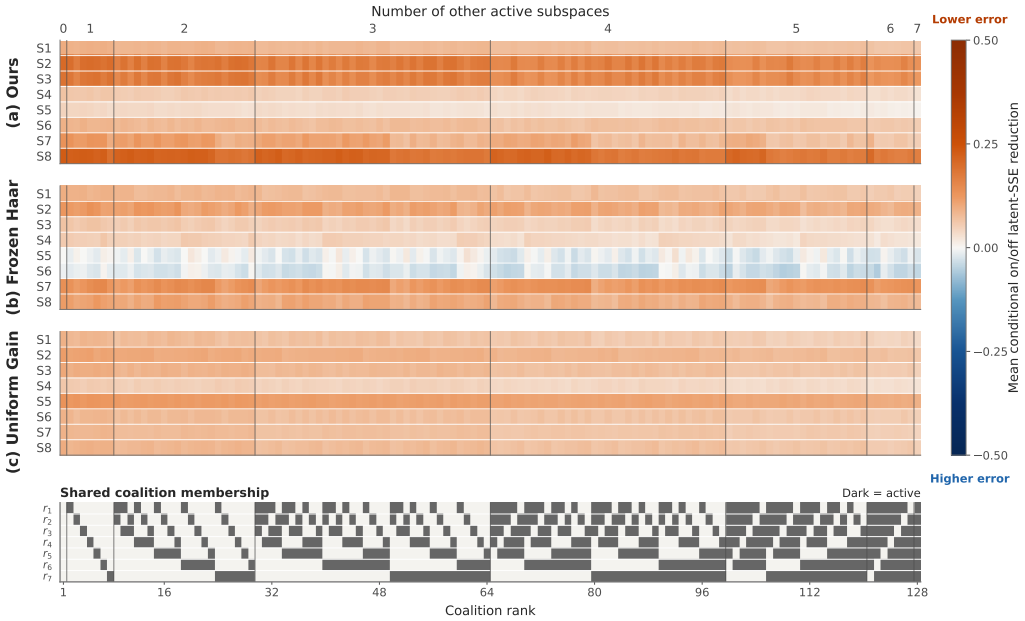


Figure 7: **Exhaustive conditional predictive contributions on PushT.** Each cell reports the cohort-mean marginal reduction in full-latent SSE obtained by activating one subspace correction under a fixed configuration of the remaining seven. Positive values indicate reduced prediction error. The rightmost group, with all seven other subspaces active, corresponds to the same leave-one-out background used in the main text before relative normalisation.

Uniform-Gain diagnostic. Uniform Gain uses a context-dependent gating mechanism, but applies a single realised gain uniformly across all subspaces within each context. Its native modulation is therefore isotropic, so the latent perspective has no directional role in the complete forward operation. The positive LOO effects of Uniform Gain in Fig. 5 show that this shared context-dependent gain is itself predictively useful. However, because the gain cannot be differentiated across perspective-defined directions, the corresponding LOO effects are generally weaker than those of the full model. For the exhaustive analysis in Fig. 7, we decompose the shared residual post hoc using the orthogonal coordinates stored in the Uniform-Gain checkpoint and selectively activate the resulting subspace components. These interventions should therefore be interpreted as a diagnostic decomposition of a useful shared residual rather than evidence of learned directional coordination. This contrast provides an additional view of Gain–Basis Coupling: useful adaptive gains can improve prediction even when applied isotropically, while differentiated gains can exploit the learned perspective to allocate modulation more effectively across subspaces.

Exact Shapley attribution. The exhaustive conditional analysis retains the dependence of each contribution on the coalition background. To obtain one aggregate attribution per subspace, we additionally compute its exact Shapley contribution for every transition,

$$\phi_{n,k} = \sum_{S \subseteq \mathcal{K} \setminus \{k\}} \frac{|S|!\,(K - |S| - 1)!}{K!} \Delta_{n,k}(S). \tag{43}$$

This averages the marginal contribution of subspace k over all positions at which it can enter the active set. By the Shapley efficiency property, $\sum_{k=1}^{K} \phi_{n,k} = \mathcal{E}_n(\varnothing) - \mathcal{E}_n(\mathcal{K})$, so the subspace attributions jointly account for the all-off versus all-on prediction improvement. We report the mean Shapley contribution over the 256 transitions.

Functional coordination across coalition backgrounds. The main-text LOO analysis shows that all eight learned subspaces retain positive marginal predictive utility when the other seven corrections are already active. The exhaustive analysis in Fig. 7 extends this result across all tested

Table 3: **Mean exact Shapley contributions on PushT.** All denotes the mean all-off versus all-on reduction in full-latent SSE. Subspace indices are local to each model and are not aligned across variants. Absolute magnitudes are interpreted within each trained model rather than as calibrated cross-model error scales.

Variant	S1	S2	S3	S4	S5	S6	S7	S8	All
Uniform Gain	0.0636	0.0893	0.0762	0.0435	0.1118	0.0699	0.0679	0.0736	0.5958
Frozen Haar	0.0726	0.1059	0.0495	0.0398	−0.0098	−0.0237	0.1325	0.0974	0.4641
Ours	0.0779	0.1680	0.1572	0.0480	0.0270	0.0728	0.0930	0.1975	0.8414

coalition backgrounds: the full model is constructive for all cohort-averaged subspace–configuration pairs, whereas Frozen Haar is positive in only 79.49% of them. Exact Shapley attribution likewise assigns positive mean contribution to all eight learned subspaces, while Frozen Haar contains two negative mean contributions. Uniform Gain is also positive across the diagnostic decompositions, reflecting the usefulness of its shared context-dependent gain rather than native subspace-specific coordination. Together, these results show that the full model combines useful adaptive gains with learned perspective-defined directions to produce more consistently constructive coordination across subspaces. These quantities characterise persistent marginal predictive utility; they do not imply statistical independence, disentanglement, or super-additive synergy.

B.3 Physical-Dynamics Accessibility on PushT

Probe construction and data. We evaluate physical-dynamics accessibility using Ridge probes fitted to frozen encoder representations. For PushT, one model step corresponds to five raw environment frames. Each probe input is the latent change between two independently encoded simulator observations, $\Delta z^{(a)} = f_\theta(o_{t+5}^{(a)}) - f_\theta(o_t)$, where a indexes an action branch. The targets are pusher displacement, T-object displacement, and T-object rotation, with rotation represented by $(\sin \Delta\theta, \cos \Delta\theta)$. Predictor outputs, adaptive gains, and predicted future latents are not used to construct the probe features. The discovery set contains 512 episodes with 16 action branches per episode, giving $8{,}192$ fitting samples. A separate episode-disjoint confirmation set is used for evaluation, with the primary results reported on the same fixed, phase-balanced cohort of 256 PushT transitions used in Appendix B.2.

Matched-capacity fitting and evaluation. Feature and target coordinates are standardised using discovery-set statistics only. Ridge regularisation is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ by minimising standardised validation MSE. For each model, target, and representation budget, one regularisation coefficient is selected from a fixed set of calibration combinations and is then shared across all combinations at that budget. Each final probe is refitted on the complete discovery set. For Ours and Frozen Haar, features are formed by selecting coordinates from their respective orthogonal subspaces; Sub-JEPA uses combinations of its predefined projections. We enumerate all admissible combinations at each representation budget and report the median across combinations rather than selecting an oracle subset. Standardised MSE is averaged over target coordinates, while Pearson r is computed per coordinate and then averaged. SD-JEPA is included only as a full-192-dimensional reference and is not part of the matched-capacity combination analysis.

On PushT, Frozen Haar is aggregated over four independently trained checkpoints. Each checkpoint first yields a median across admissible combinations; the reported centre is the mean of these four checkpoint medians, and the shaded region spans their minimum and maximum. Because the compared models are trained independently, the resulting readout differences reflect the representations and decompositions produced jointly through training rather than the post-hoc effect of applying a different orthogonal basis to an otherwise fixed representation.

Qualitative image-space attribution. We complement the matched-capacity readout with gradient-times-input attribution on PushT. Figure 8 shows attribution patterns along executions that start from the same initial state and goal but subsequently follow each model's own planned trajectory. LeWorldModel maintains relatively broad attribution over the T object and pusher across much of

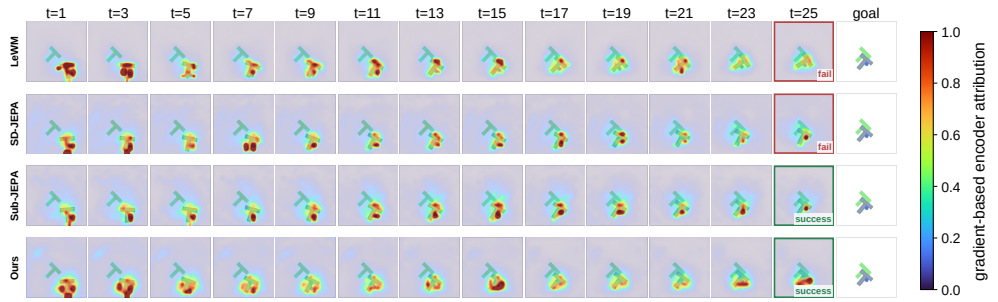


Figure 8: **Spatial attribution on PushT executions.** Gradient-times-input encoder attribution is shown on a shared scale for LeWorldModel, SD-JEPA, Sub-JEPA, and Ours. All methods start from the same initial state and goal and subsequently follow their respective planned executions.

the execution. Ours exhibits more motion-dependent localisation: during translation, attribution tends to concentrate near the moving side of the T object and follows its translational progression; during rotational phases, localised hotspots occasionally emerge near the upper-left arm of the T, consistent with the apparent rotational pivot. When T-object motion is weak, the attribution of Ours also becomes weaker rather than remaining broadly distributed over the object. These observations are qualitatively consistent with the stronger object-dynamics accessibility in the main analysis, but are descriptive rather than causal evidence about the features used for prediction or control.