# READABLE COORDINATES, DISTRIBUTED CONCEPTS: CONCEPT SUPERVISION IN DREAMERV3

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Concept supervision gives a world model a vocabulary: named coordinates report facts such as whether an agent has wood or is running low on energy. Understanding what these reports reveal requires connecting the readable interface to the representation behind it. We study this connection in DreamerV3 on Craftax-Classic through a paired comparison of concept supervision, localization regularization, and sparse-representation objectives. Across five starting checkpoints, supervised scalar readouts consistently outperform their raw-coordinate controls, while the same concepts remain highly predictable from the surrounding state. The most revealing differences emerge beneath the aggregate scores: objectives with similar overall readability expose different inventory, spatial, and physiological variables, producing distinct views of the agent's state alongside different training-return profiles. Two analytical results sharpen this interpretation. Identical decoding measurements can accompany different causal roles, and shared temporal structure can preserve predictability even after labels are reassigned between episodes. Together, these findings characterize concept supervision as a way to construct readable interfaces within distributed representations. They show how the value of a named coordinate depends on which concept it exposes, how reliably it reports that concept, and what the available evidence establishes about its relationship to the wider model.

## 1 INTRODUCTION

Concept supervision promises a readable interface to a world model's internal knowledge. Interpreting that interface requires understanding how its named coordinates relate to the wider representation in which they are embedded. A readout labeled *has wood*, for example, gives a direct way to inspect an inventory fact. Its accuracy describes the quality of that access; understanding the interface also requires examining which other concepts it exposes reliably and how much of the same information is available elsewhere in the state. These questions connect the design of a readable interface to the representation it helps us inspect.

Concept bottleneck models place named variables directly on the prediction path (Koh et al., 2020), and concept-based reinforcement-learning agents use them for inspection and policy revision (Ye et al., 2025; Delfosse et al., 2024). Coordinate supervision trains a small set of latent coordinates to predict specified concepts, creating named readouts within the world model's larger learned state. The remaining coordinates have no predefined semantic assignments. Localization penalties additionally act on concept associations in the remaining coordinates. Comparing these objectives allows us to examine how training shapes both the named readouts and their relationship to the surrounding state.

We study this relationship in DreamerV3 (Hafner et al., 2025) on Craftax-Classic (Matthews et al., 2024). Five checkpoints are each continued under six conditions: an unpenalized control, three nested concept objectives, an in-loop sparse autoencoder, and a self-ablation mask. We measure designated-coordinate readability, concept predictability from the remaining state, and paired training return. These measurements characterize the learned interface in the context of the agent's representation and observed training performance.

Across all five starting checkpoints, each supervised interface improves on its paired raw-coordinate control, while complement probes retain high predictive accuracy. More revealingly, similar aggregate readability scores conceal substantial differences between concepts. Although the A-read and A-full



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.

conditions (defined in Section 3.2) achieve comparable overall scores, A-read provides more accurate readouts of wood possession and low health, whereas A-full provides more accurate readouts of low energy and proximity to water. Their training-return profiles also differ. The contribution is thus a characterization of which variables each interface makes accessible and where those variables remain predictable, extending the account beyond a single readability score.

Two analytical observations clarify the limits of this interpretation. First, the same decoding performance can arise whether a designated coordinate directly influences an output or the model bypasses it in favor of information elsewhere. Second, reassigning labels between episodes need not eliminate predictability when the episodes share temporal patterns. Combined with the paired comparison, these observations distinguish what named readouts make accessible from whether the world model actually uses those coordinates in its computations.

## 2 Concept access and information distribution

Let the latent state be $s_t = (u_t, r_t)$, where $u_t$ contains designated concept coordinates and $r_t$ is the remaining state. For a binary concept $c_{t,k}$, a specified scalar readout $q_k(u_{t,k})$ measures access through the designated coordinate; a separately fitted probe $p_k(r_t)$ measures access through the complement. Table 1 relates the quality of the named readout to concept recoverability from the surrounding state. Neither is an estimator of all information in the representation: the scores depend on the readout class, state partition, and evaluation distribution.

Table 1: Interpreting a named readout in context. The two measurements describe access through the interface and concept recoverability from the remaining state.

| Dimension | Question | Measurement |
|---|---|---|
| Coordinate readability | Can the specified scalar readout recover a named concept from its assigned coordinate? | Held-out axis AUROC |
| Residual predictability | Can a fitted predictor recover that concept from the state outside the designated block? | Held-out complement AUROC |

**Interpreting a readout in context.** The designated readout $q_k$ and complement probe give complementary views of the same concept. The first measures how reliably the interface exposes it; the second situates that access within the wider state. In a recurrent world model, inventory, recent observations, spatial context, and the stochastic state can each support concept prediction. A complement probe measures recoverability from the remaining coordinates; it does not count how many independent copies exist. We use *distributed information* in this operational sense: the concept is recoverable outside its designated block, without assuming a particular internal coding geometry.

**Decoding does not identify causal use.** Consider a non-degenerate binary variable $C$ and two representations with $U = R = C$. Let one system produce $Y_1 = U$ and the other $Y_2 = R$. Their observational distributions over $(U, R, Y)$ are identical. Every decoding statistic computed from those distributions therefore agrees, including perfect readout from both $U$ and $R$. Under $\mathrm{do}(U = c')$, however, $Y_1 = c'$ while $Y_2 = C$ (Figure 1). Thus identical readability and residual predictability can accompany different causal roles. Conversely, removing every residual copy does not force an output to use the designated one. These elementary examples make no assumption about a particular learning algorithm.

This distinction is consistent with the separation of accessible and used information in probing (Belinkov, 2022; Ravichander et al., 2021) and of availability and assignment across world-model pathways (Vakalis, 2026b). It also distinguishes our measurements from causal abstraction, where correspondence between neural and abstract interventions is part of the criterion itself (Geiger et al., 2022; 2024). The example is an interpretive result, not an empirical intervention on the trained agents.

**Context constrains the meaning of localization.** Suppose the residual reconstructs a context variable exactly and linearly, $V = w^\top R + b$. Then

$$\mathrm{Cov}(V, C) = w^\top \mathrm{Cov}(R, C). \tag{1}$$

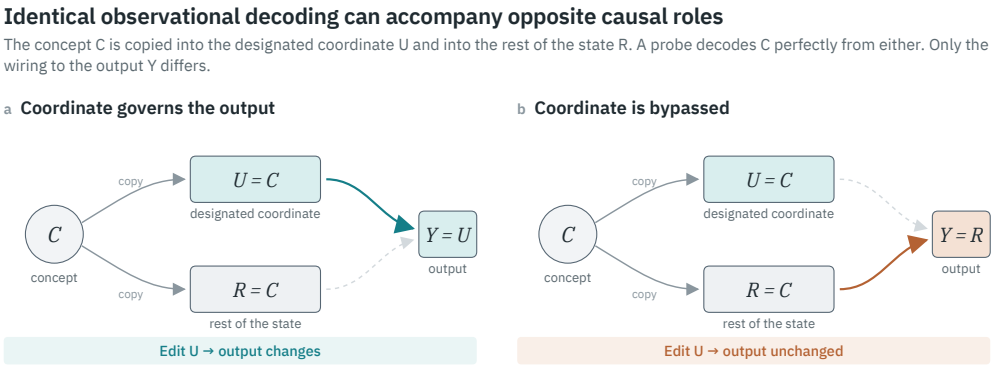



Figure 1: Identical concept access with different causal roles; an analytical illustration. Both systems have $U = R = C$, so either part of the representation perfectly reveals the concept. Editing $U$ changes $Y$ on the left and leaves it unchanged on the right. Solid colored arrows indicate the active output dependency; pale dashed arrows show the unused alternative.

If $V$ and $C$ are correlated, exact linear residual reconstruction of $V$ is incompatible with $\mathrm{Cov}(R, C) = 0$. The claim is conditional: context could instead be represented in the designated block or decoded nonlinearly. The identity explains why residual predictability is a property of a chosen partition and decoding class, rather than an unconditional measure of semantic purity.

## 3 Experimental Design

### 3.1 Agent and Paired Conditions

We use the public PyTorch implementation of DreamerV3, with a 29M-parameter world model, $64 \times 64$ pixel observations, and 17 actions. Its recurrent state is $h_t \in \mathbb{R}^{1024}$, and its stochastic state $z_t$ comprises 32 categorical variables of 32 classes. Fourteen binary concepts derived from the simulator supply training labels; eight additional concepts are recorded only for evaluation (Appendix A). Concept labels enter the penalties, not the image encoder or reconstruction targets.

Five checkpoints trained for $10^7$ environment steps are each continued for $2 \times 10^6$ steps under all six conditions. Every continuation uses fresh replay and reinitialized world-model optimizer state. The starting checkpoint is the unit of replication, giving five paired replicates. A separate five-seed unpenalized group trained to $10^7$ steps provides descriptive context; it is not the denominator for continuation-return comparisons.

### 3.2 Concept and Sparsity Objectives

With $D = \{0, \ldots, K-1\}$ and $K = 14$, the designated block is $u_t = h_t[D]$ and the complement is $r_t = [h_t[K{:}], z_t]$, with $d = 2034$ coordinates. Three nested objectives act on this partition.

**A1: coordinate supervision.** A learned scalar calibration predicts each concept from its assigned coordinate with binary cross-entropy. A1 couples the representation to a named readout while leaving the world model's and actor's use of that coordinate unconstrained.

**A2: residual regularization.** A normalized cross-covariance penalty discourages linear associations between the concepts and $r_t$. It is inspired by the population condition underlying LEACE (Belrose et al., 2023), rather than applying its closed-form erasure transformation. The complement excludes the entire designated block, so associations between sibling concept coordinates are outside this penalty.

**A3: readout regularization after ablation.** The designated block is zeroed, $z_t$ is retained, and the model takes one imagination step under the recorded action. A divergence loss draws the resulting concept readouts toward their running base rates. This is a training objective on the next-state reporter, not supervision of a semantic change in images, rewards, or actions.

The three supervised conditions are A-read (A1), A-loc (A1+A2), and A-full (A1+A2+A3). They form a nested comparison, not a full factorial design. Condition B jointly trains a TopK sparse autoencoder on $h_t$, with 2048 features, 32 active features, and AuxK, through a reconstruction-based penalty (Gao et al., 2025). Condition S adds prediction-head losses after retaining the 128 largest-magnitude coordinates of $h_t$, inspired by self-ablation (Ferrao et al., 2025); the recurrent model itself is not rerun under the mask. The sixth condition is the unpenalized continuation.

All penalties enter $L_{\mathrm{WM}} + \sum_j \mu_j P_j$. A multiplicative controller adapts their weights, with staged release of A2 and A3 and a brake based on predictive loss. These are finite regularizers, not hard constraints on information content. Appendix D specifies the losses and controller. The comparison concerns these implemented training procedures at a common continuation horizon.

### 3.3 Measurement and aggregation

Each of the 35 final checkpoints is evaluated on a rollout corpus of at most 300 episodes or $1.2 \times 10^5$ steps. Class-balanced, regularized logistic probes are fitted on 70% of episodes and scored on the remaining 30%; full probes read $h_t$ and complement probes read $r_t$. The trained scalar readouts use the same held-out episodes. AUROC uses average ranks for tied predictions. Whole-corpus axis scores are secondary records, and raw held-out predictions are retained.

In the unpenalized, B, and S conditions, raw coordinates serve as axis controls. They do not have the concept-specific learned scalar calibrations of A-read, A-loc, and A-full. The axis comparison therefore measures the trained interface as a whole, including its readout orientation; it is not a comparison of identically calibrated probes or of all information in the respective states. In particular, it does not measure readability of B's learned SAE features.

We summarize each checkpoint by the median score over its available trained concepts, then take the median across checkpoints. Undefined one-class comparisons are omitted, with their denominators reported in Appendix B. Pairing matches starting weights; the policies can subsequently visit different states even with matched environment seeds. Training return is reported separately as the mean of checkpoint-paired ratios to the unpenalized continuation, using the final 30 logged values for each run. It is a training-telemetry measure, not return from the probe-evaluation rollouts.

## 4 Readable coordinates within a distributed state

### 4.1 Consistent access through the supervised interface

Table 2 and Figure 2 summarize all conditions. The three supervised interfaces have median axis AUROC $0.842$, $0.834$, and $0.853$, compared with $0.493$ for the raw-coordinate continuation control. Each supervised condition exceeds its paired control at all five starting checkpoints. For A-read, the median within-checkpoint increase is $0.360$, with a range of $0.289$–$0.456$ (Appendix C). This consistency concerns the reported interface comparison, including its learned calibration.

Median complement AUROC is $0.999$–$1.000$ across these conditions. The supervised interface therefore offers direct access to concepts that also remain highly predictable from the surrounding state. This residual measurement places the readout improvement in context: the named coordinates are one source of concept access within the learned representation. The sparse-objective conditions B and S have raw-axis medians near $0.5$; without concept assignment, these objectives do not supply the same named-coordinate interface.

### 4.2 The concept profile matters

Similar condition medians conceal different concept profiles. Under A-read, median axis AUROC is $0.989$ for *has wood*, $0.912$ for *wood at least three*, and $0.937$ for *health low*. Under A-full, these scores are $0.614$, $0.615$, and $0.586$. Conversely, *energy low* is more readable under A-full ($0.966$)

Table 2: Concept access and training return, five checkpoints per condition. AUROC entries are medians of checkpoint medians over available trained concepts on held-out episodes. Axis scores use learned scalar readouts in A-read/A-loc/A-full and raw coordinates otherwise. Return is the mean checkpoint-paired training-return ratio to the unpenalized continuation. The 10M control is descriptive.

| Condition | Axis | Residual | Permuted | Return ratio |
|---|---|---|---|---|
| 10M control | 0.497 | 1.000 | 0.609 | – |
| Continuation | 0.493 | 1.000 | 0.597 | 1.000 |
| A-read | 0.842 | 1.000 | 0.610 | 0.996 |
| A-loc | 0.834 | 0.999 | 0.613 | 0.874 |
| A-full | 0.853 | 0.999 | 0.617 | 0.897 |
| B (SAE) | 0.512 | 1.000 | 0.609 | 0.975 |
| S (mask) | 0.505 | 1.000 | 0.635 | 0.750 |

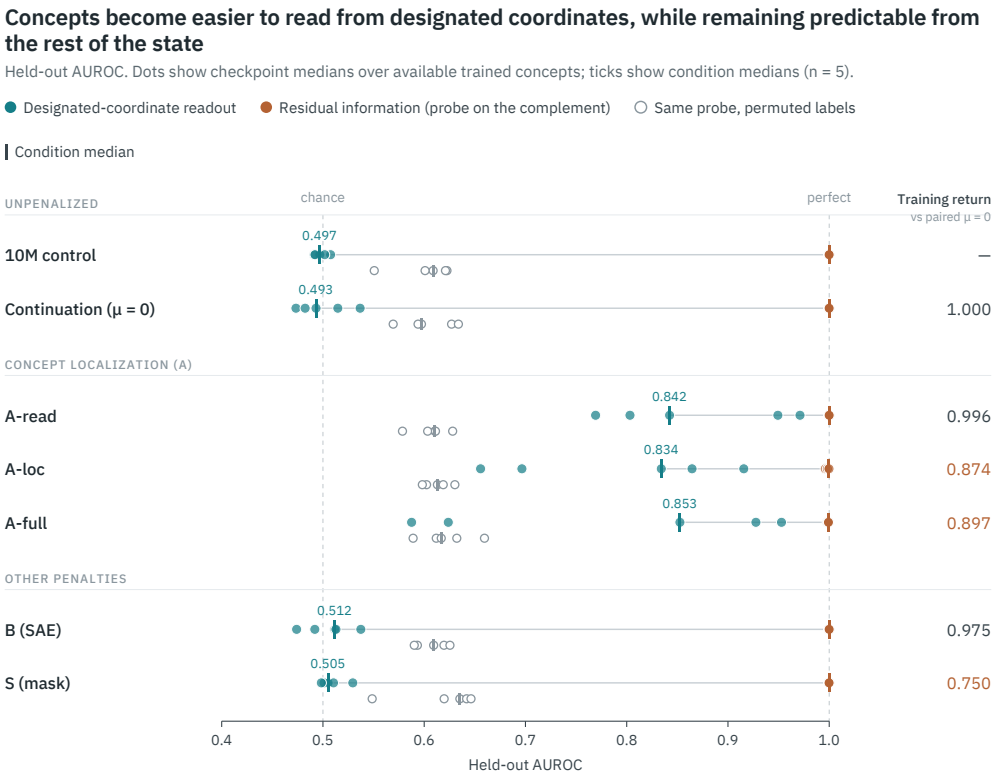



Figure 2: Coordinate readability and residual predictability. Each point is a checkpoint's median held-out AUROC over available trained concepts, with ticks at condition medians. Open circles show complement probes fitted to episode-reassigned labels. The right column gives paired training-return ratios. The near-ceiling residual aggregate does not imply identical scores for every concept. Dashed lines mark chance and perfect ranking.

than A-read (0.871), as is *near water* (0.895 versus 0.784). The objectives change the balance of readable concepts; their effect is not captured by a uniform increase or decrease in one aggregate score.

Residual predictability also varies by concept. For A-loc, complement medians for *near table* and *near tree* are approximately 0.934 and 0.940. Their lower values relative to the condition aggregate show why a near-ceiling median is not a claim of perfect recovery for every variable. AUROC

measures ranking, so even similar scores need not imply similar calibration or latent geometry. Appendix B provides the concept-level results.

### 4.3 TRAINING-RETURN PROFILES

A-read's mean paired training-return ratio is $0.996$. The ratios are $0.874$ for A-loc, $0.897$ for A-full, $0.975$ for B, and $0.750$ for S. Thus the readout-only procedure combines the observed readability gain with a return point estimate close to its paired control. Adding the other penalties produces distinct concept profiles and lower observed training returns in this comparison.

These ratios describe the implemented procedures, not an intrinsic cost of distributed or localized representations. Five paired checkpoints do not establish equivalence for a ratio near one, and the conditions do not trace an optimized performance frontier. The optimized model loss is excluded from this comparison because it contains differently weighted penalties; it is not a common measure of predictive quality across conditions.

## 5 TEMPORAL STRUCTURE IN EPISODE-REASSIGNED LABELS

The complement diagnostic reassigns whole label sequences between episodes, repeating or truncating each sequence to the destination length. Its condition medians range from $0.597$ to $0.635$; for A-full, the median score for *is dark* is $0.896$. These are measurements under a particular reassignment operation, not a universal chance baseline.

The distinction follows from a simple construction. Suppose equal-length episodes share a schedule $c_{e,t} = g(t)$, and the representation records time $t$. For any permutation $\pi$ of episode identities,

$$c_{\pi(e),t} = g(t) = c_{e,t}. \tag{2}$$

Reassignment then leaves every label unchanged, so the representation predicts the permuted labels as well as the original ones. A synchronized day–night cycle is an intuitive instance of this structure. The construction proves that episode reassignment alone does not imply independence of representations and labels; it does not identify the cause of the measured darkness score.

The diagnostic disrupts episode identity while retaining some within-episode structure. With unequal episode lengths, repetition and truncation can also alter prevalence and alignment. Its interpretation is consequently tied to those transformations. In Figure 2, the open circles characterize predictability after reassignment; they are not subtracted from the original scores or treated as an information-free floor. This resolves the role of the control without equating temporal predictability with concept-specific storage.

## 6 DISCUSSION

The learned interfaces are best understood through the concept profiles they expose and their relationship to the surrounding state. Coordinate supervision provides named access to selected variables, and the complement measurements show that those variables remain predictable elsewhere. Together, these observations describe how the interface sits within the representation. They also make the concept-level results consequential: an interface that exposes inventory reliably offers a different view of the agent's state from one that better exposes energy or nearby water, even when their aggregate scores are similar.

This interpretation explains why the condition ordering depends on the measurement. Aggregate axis AUROC, the profile of individual concepts, and training return answer different questions. A-full has a slightly higher aggregate axis median than A-read, but A-read has higher medians for several inventory and vital-state concepts and a higher paired training-return ratio. There is no single ordering that summarizes all three properties. The result is a characterization of what each objective makes accessible and its observed training-return profile.

The analytical results delimit what this characterization says about the model's computations. Equation 1 connects residual concept correlations to retained context under stated assumptions, while Figure 1 shows why even exact observational agreement leaves causal use unidentified. The episode-reassignment analysis further shows how shared temporal structure can sustain predictability after

labels are reassigned. A named readout therefore supports a specific interpretation: it reports a concept under the evaluation distribution, while its role in generating predictions or actions remains unestablished by decoding alone.

## 7 Related work

**Concept interfaces and leakage.** Concept bottlenecks (Koh et al., 2020) and concept whitening (Chen et al., 2020) constrain prediction routes or align latent axes with concepts. The meaning of these interfaces also depends on the information they carry: concept predictions can contain unintended information (Havasi et al., 2022), and a component predicting one concept may encode others (Almudévar et al., 2026). Parisini et al. (2026) distinguish concepts-task and interconcept leakage. Our complement measurement concerns another location for information: the state outside the designated concept block. The distinction between a concept's assigned output and its information elsewhere is central to the present comparison.

**Concepts in agents and world models.** LICORICE and SCOBots use concept-based interfaces in reinforcement learning (Ye et al., 2025; Delfosse et al., 2024). A causal audit of quantized delta representations studies concept access and interventions in a transformer world model on Crafter (Xia & Jiang, 2026); here the representation and environment are DreamerV3 and Craftax-Classic, and the empirical comparison concerns training objectives. Contextual Dreamer models separate observed context from evolving state for generalization (Prasanna et al., 2024). Recent world-model analyses distinguish information available in a state from its assignment to computational pathways and propagation by native dynamics (Vakalis, 2026b;a). Our analytical construction isolates the observational part of that distinction.

**Intervention-aware representations.** Interchange intervention training and distributed alignment search establish correspondences between neural and abstract interventions (Geiger et al., 2022; 2024). IntCEM trains concept embeddings to receive edits (Espinosa Zarlenga et al., 2023), while causally reliable concept bottlenecks incorporate causal structure (De Felice et al., 2025). Intervention choice and relations between concepts affect their interpretation (Shin et al., 2023; Singhi et al., 2024); RAVEL distinguishes intended changes from preserved attributes (Huang et al., 2024). Subspace interventions can activate dormant or divergent computations (Makelov et al., 2024; Grant et al., 2026). These works explain why intervention evidence and observational decoding establish different properties, rather than supplying an intervention claim for the present agents.

**Causal representations and temporal dynamics.** Paired interventions support weakly supervised causal representation learning (Brehmer et al., 2022) and intervention-centered benchmarks such as CausalTriplet (Liu et al., 2023). CITRIS, iCITRIS, and BISCUIT identify causal components from temporal or interactive data under their respective assumptions (Lippe et al., 2022; 2023b;a); causal dynamics learning connects transition structure to state abstraction (Wang et al., 2022). Object-latent masking in Causal-JEPA (Nam et al., 2026), counterfactual retrieval in a CausalVAE world model (Ding et al., 2026), and dynamics-effective latent directions in a controlled recurrent model (Liu & Chen, 2026) address different relations between representation and dynamics. Those identification or counterfactual results are not implied by concept-label supervision alone.

**Sparse and recurrent representations.** TopK sparse autoencoders (Gao et al., 2025) and self-ablation (Ferrao et al., 2025) motivate the two sparsity comparisons. Sparse memory circuits in DreamerV3 have been connected to behavioral edits (Sobotka et al., 2025), while mechanistic analysis has identified emergent planning in a recurrent agent (Bush et al., 2025). Language-model studies examine recurrent writes (Young, 2026), feature-specific persistence (Luo et al., 2026), and self-repair after ablation (Rushing & Nanda, 2024). Our raw-coordinate comparisons do not evaluate the quality of learned SAE features or establish a recurrent recovery mechanism. They isolate whether these training objectives supply the same named-coordinate access as explicit concept supervision.

## 8 SCOPE AND LIMITATIONS

The empirical claims concern one environment, architecture, implementation, and continuation horizon, with five paired starting checkpoints. Concepts are simulator-defined and incomplete; the study does not assess usefulness to human users. Learned scalar readouts and high-dimensional complement probes have different capacities, and the raw-coordinate controls lack learned concept calibration. Readability comparisons therefore concern the implemented interfaces, not equal-capacity information estimators.

Evaluation corpora depend on each policy, and episode-level confidence intervals are not reported. Training-return ratios summarize telemetry rather than matched evaluation rollouts. The experiment measures observational concept access and residual predictability; it does not measure semantic editing, behavioral steering, or the causal role of the learned coordinates. The analytical construction establishes a distinction between these properties, not the causal behavior of the trained agents. These boundaries define the claims made here.

## 9 CONCLUSION

A readable concept interface gains meaning from the variables it exposes and its relationship to the wider world-model state. In DreamerV3 on Craftax-Classic, supervised scalar readouts consistently improve on paired raw-coordinate controls across five starting checkpoints, while the remaining state retains high concept predictability. Similar aggregate readability scores conceal different concept profiles, accompanied by different training-return profiles. These findings characterize the particular view of the agent's state that each objective makes available.

The analytical results establish the limits of that view: identical decoding measurements can accompany different causal roles, and episode reassignment can preserve shared temporal predictability. Taken together, the study explains how named coordinates provide access to concept information within a recurrent world model and how to interpret that access without attributing an unmeasured computational role to the readouts.

### AI USE STATEMENT

OpenAI Codex was used during manuscript preparation to assist with organization and language editing, figure preparation, and code. This assistance included improving the clarity and coherence of the argument. No new experimental data were generated through this editorial process. The authors retain full responsibility for the accuracy of the reported results, the validity of the references, and the interpretations and conclusions presented in the paper.

### REPRODUCIBILITY STATEMENT

The tables and figures are generated from 35 checkpoint evaluation records and the corresponding training summaries. Appendix C specifies the paired readability calculation; Appendices A, D, and E give concept definitions, training objectives, and evaluation details. Raw held-out predictions are retained with the evaluation outputs.

## REFERENCES

Antonio Almudévar, José Miguel Hernández-Lobato, and Alfonso Ortega. There was never a bottleneck in concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2506.04877.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.03819.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html.

Thomas Bush, Stephen Chung, Usman Anwar, Adrià Garriga-Alonso, and David Krueger. Interpreting emergent planning in model-free reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/cc4d9cfc45325e460b455a820d5f212c-Abstract-Conference.html.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:772–782, 2020. arXiv:2002.01650.

Giovanni De Felice, Arianna Casanova Flores, Francesco De Santis, Silvia Santini, Johannes Schneider, Pietro Barbiero, and Alberto Termine. Causally reliable concept bottleneck models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/db251964772b285a7aaf2d974f7e5f49-Abstract-Conference.html.

Quentin Delfosse, Sebastian Sztwiertnia, Mark Rothermel, Wolfgang Stammer, and Kristian Kersting. Interpretable concept bottlenecks to align reinforcement learning agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7b76eea0c3683e440c3d362620f578cd-Abstract-Conference.html.

Ziyi Ding, Xianxin Lai, Weiyu Chen, Xiao-Ping Zhang, and Jiayu Chen. CausalVAE as a plug-in for world models: Towards reliable counterfactual dynamics. *arXiv preprint arXiv:2604.07712*, 2026. URL https://arxiv.org/abs/2604.07712. Version 1, April 2026.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/770cabd044c4eacb6dc5924d9a686dce-Abstract-Conference.html.

Jeremias Ferrao et al. Self-ablating transformers: More interpretability, less sparsity. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. arXiv:2505.00509.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/944ecf65a46feb578a43abfd5cddd960-Abstract-Conference.html.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024. URL https://aclanthology.org/2024.acl-long.470/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603, 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1263–1273, 2023a. URL https://proceedings.mlr.press/v216/lippe23a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=itZ6ggvMnzS.

Yang Liu and Yuming Chen. Low-rank dynamics-effective latent carriers for counterfactual rollout in learned world models. *arXiv preprint arXiv:2608.15156*, 2026. URL https://arxiv.org/abs/2608.15156. Version 4, September 2026.

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 553–573, 2023. URL https://proceedings.mlr.press/v213/liu23a.html.

Haoyan Luo, Mateo Espinosa Zarlenga, and Mateja Jamnik. Persistent sparse autoencoders: Learning feature-specific timescales in language model representations. *arXiv preprint arXiv:2607.17117*, 2026. URL https://arxiv.org/abs/2607.17117. Version 2, September 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Jackson, Samuel Coward, and Jakob Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2024.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.11389.

Enrico Parisini, Tapabrata Chakraborti, Chris Harbron, Ben D. MacArthur, and Christopher R. S. Banerji. Leakage and interpretability in concept-based models. *Journal of Machine Learning Research*, 27(174):1–40, 2026. URL https://www.jmlr.org/papers/v27/25-1121.html.

Sai Prasanna, Karim Farid, Raghu Rajan, and André Biedenkapp. Dreaming of many worlds: Learning contextual world models aids zero-shot generalization. In *Reinforcement Learning Conference*, 2024. URL https://arxiv.org/abs/2403.10967.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31504–31520, 2023. URL https://proceedings.mlr.press/v202/shin23a.html.

Nishad Singhi, Jae Myung Kim, Karsten Roth, and Zeynep Akata. Improving intervention efficacy via concept realignment in concept bottleneck models. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/03805.pdf.

Jan Sobotka, Auke Ijspeert, and Guillaume Bellegarda. Reverse-engineering memory in DreamerV3: From sparse representations to functional circuits. In *NeurIPS Workshop on Mechanistic Interpretability*, 2025. URL https://openreview.net/forum?id=JmjqTi4FDF. Workshop spotlight.

Donna Vakalis. The intervention gap in latent world models. *arXiv preprint arXiv:2608.29998*, 2026a. URL https://arxiv.org/abs/2608.29998. Version 1, August 2026.

Donna Vakalis. What a world model represents is three questions. *arXiv preprint arXiv:2607.06640*, 2026b. URL https://arxiv.org/abs/2607.06640. Version 2, August 2026.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23151–23180, 2022. URL https://proceedings.mlr.press/v162/wang22ae.html.

Runze Xia and Richard Jiang. From codes to causes—and where the chain breaks: A causal interpretability audit of quantized delta representations in a transformer world model. *Preprint*, 2026.

Zhuorui Ye, Stephanie Milani, Geoff Gordon, and Fei Fang. LICORICE: Label-efficient concept-based interpretable reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ff0e299b1cd800f006f3516f06177e2a-Abstract-Conference.html.

Jack Young. WriteSAE: Sparse autoencoders for recurrent state. *arXiv preprint arXiv:2605.12770*, 2026. URL https://arxiv.org/abs/2605.12770. Version 4, May 2026.

## A CONCEPT DEFINITIONS

The fourteen trained concepts are *has wood* (inventory $\geq 1$), *wood at least three*, *has sapling*, *has wood pickaxe*, *has wood sword*, *table placed* (achievement flag), four low-vital flags (food, drink, energy, and health at most 3 on the 0–9 scale), *near tree*, *near water*, *near table* (Chebyshev radius 1), and *is dark* (light below 0.5). A3 applies to the first ten, treating them as persistent relative to the four perceptual concepts; this grouping does not imply invariance to actions.

Eight evaluation-only concepts are *has stone*, *has coal*, *stone placed*, *furnace placed*, *near stone*, *near coal*, *near zombie*, and *near cow*. They do not enter the training penalties or the trained-concept aggregates reported in the main text. Concepts can entail one another: *wood at least three*, for example, entails *has wood*. The concept list is not a set of independent latent factors.

## B PER-CONCEPT SCORES

Table 3: Medians across finite checkpoint scores. A-read, A-loc, and A-full show axis AUROC; the final score columns show A-full complement AUROC and its episode-reassignment diagnostic. $n$ gives valid A-full checkpoint counts for axis/complement/reassignment, respectively.

| Concept | A-read | A-loc | A-full | Residual | Permuted | $n$ |
|---|---|---|---|---|---|---|
| has wood | 0.989 | 0.786 | 0.614 | 1.000 | 0.568 | 5/5/5 |
| wood $\geq 3$ | 0.912 | 0.689 | 0.615 | 0.998 | 0.591 | 5/5/5 |
| has sapling | 0.841 | 0.684 | 0.811 | 1.000 | 0.525 | 5/5/5 |
| has wood pickaxe | 0.990 | 0.990 | 0.981 | 1.000 | 0.673 | 5/5/5 |
| has wood sword | 0.964 | 0.960 | 0.978 | 1.000 | 0.751 | 5/5/5 |
| table placed | 0.817 | 0.952 | 0.941 | 0.996 | 0.726 | 5/5/5 |
| food low | 0.699 | 0.603 | 0.647 | 1.000 | 0.672 | 5/5/5 |
| drink low | 0.975 | 0.547 | 0.592 | 1.000 | 0.698 | 5/5/5 |
| energy low | 0.871 | 0.734 | 0.966 | 0.999 | 0.761 | 5/5/4 |
| health low | 0.937 | 0.614 | 0.586 | 0.998 | 0.584 | 5/5/5 |
| near tree | 0.725 | 0.643 | 0.577 | 0.940 | 0.533 | 5/5/5 |
| near water | 0.784 | 0.829 | 0.895 | 0.996 | 0.537 | 5/5/5 |
| near table | 0.695 | 0.595 | 0.709 | 0.933 | 0.537 | 5/5/5 |
| is dark | 0.997 | 0.998 | 0.998 | 0.998 | 0.896 | 5/5/5 |

Undefined one-class comparisons are omitted rather than assigned chance performance. A-full *has wood sword* has five valid checkpoints for each score. The *energy low* reassignment score has four; its axis and complement scores have five. Checkpoint medians use available concepts, so their composition can differ. Medians of checkpoint medians are not interchangeable with medians of this table's concept-level entries.

## C PAIRED READABILITY DIFFERENCES

Let $a_{j,s}$ be the median axis AUROC over available trained concepts for condition $j$ and starting checkpoint $s$. We compute $\Delta_{j,s} = a_{j,s} - a_{\mathrm{none},s}$ before aggregating across checkpoints. Table 4 reports medians and observed ranges of these five paired differences. A difference of medians is generally not the median of paired differences. The calculation is descriptive; pairing fixes starting weights, not subsequent visitation or readout calibration.

## D TRAINING OBJECTIVES AND OPTIMIZATION

**A1 (readout).** Each coordinate has a learned scalar calibration:

$$L_{\mathrm{read}} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{BCE}(\mathrm{sigmoid}(a_k u_{t,k} + b_k), c_{t,k}) \,. \tag{3}$$

A1 supervises a readout. It places no constraint on which coordinates the prediction heads or actor use.

Table 4: Paired differences in checkpoint-level axis AUROC relative to the raw-coordinate continuation control. Positive pairs count starting checkpoints with $\Delta_{j,s} > 0$. Ranges describe the five observations, not confidence intervals.

| Condition | Median $\Delta$ | Observed range | Positive pairs |
|---|---|---|---|
| A-read | 0.360 | $[0.289, 0.456]$ | 5/5 |
| A-loc | 0.341 | $[0.182, 0.379]$ | 5/5 |
| A-full | 0.370 | $[0.073, 0.460]$ | 5/5 |
| B (SAE) | 0.001 | $[-0.002, 0.018]$ | 4/5 |
| S (mask) | 0.017 | $[-0.032, 0.056]$ | 3/5 |

**A2 (residual cross-covariance).** The implemented statistic, averaged over concepts, is

$$L_{\text{res}} = \frac{1}{K} \sum_{k=1}^{K} \left[ \frac{\|\widehat{\text{Cov}}(r, c_k)\|_2^2}{\widehat{\text{Var}}(c_k)\, \widehat{\widehat{\text{Var}}}(r)} - \frac{d}{n-1} \right]_+, \tag{4}$$

with numerical stabilizers omitted, $[x]_+ = \max(x, 0)$, and $n = 1024$ flattened batch positions. The subtracted term comes from an independent-sample approximation; given temporal dependence, estimated normalization, and clipping, a value of zero is not a calibrated certificate of erasure for these replay batches. In the population, zero cross-covariance characterizes linear guardedness for affine predictors under the convex-loss conditions of LEACE (Belrose et al., 2023). Equation 4 is a surrogate inspired by that condition, not the LEACE transformation. Because it excludes the whole designated block, sibling concept coordinates can still predict or entail one another.

**A3 (readout after ablation).** The implementation zeros every coordinate in $D$, keeps $z_t$, and takes one imagination step with the action from the training batch. The next-state axis readout is encouraged to match a running base rate for the concept:

$$L_{\text{abl}} = \frac{1}{|P|} \sum_{k \in P} \text{KL}\Big(\text{Bern}(\bar{p}_k) \,\|\, \text{Bern}\Big(\text{sigmoid}(a_k \tilde{h}_{t+1,k} + b_k)\Big)\Big), \tag{5}$$

where $P$ indexes the ten concepts the training objective treats as persistent. A3 thus constrains the model's own readout after an artificial ablation. It does not supervise changes in predicted images, rewards, or actor outputs, and by itself it cannot show causal necessity.

**Controller and predictive-loss brake.** The multiplicative update uses an exponential moving average of relative constraint violation, with learning rate $0.05$, update period 100 gradient steps, EMA half-life 2000 gradient steps, relative-violation clipping at $\pm 100$, and log-update clipping at $\pm 0.5$. Multipliers lie in $[10^{-6}, 10^{3}]$. A2 and A3 are released when A1 reaches its working target. Targets are A1 $0.25$, A2 $0.02$, A3 $0.02$, and B $0.07$. These settings determine the controller updates; they do not certify any AUROC level or hard erasure constraint.

The brake compares predictive image, reward, and continuation losses with $1.25$ times a baseline maximum-loss curve. At a triggered update, it halves positive multipliers, bounded below by $\mu_{\min}$, and skips ascent. Penalty gradients remain part of the optimization. The recorded final A2 means are $2.63$ for A-loc and $1.93$ for A-full; A3 is $0.0252$ for A-full, and A1 is $0.214$ for A-read. Thus the experimental conditions are defined by the objectives applied, rather than by an assumption that all working targets were attained.

**Normalization and interpretation.** The A2 subtraction is $d/(n-1) \approx 1.99$ under the independent-position approximation. Replay sequences have temporal dependence, and the statistic also depends on variance heterogeneity, denominator estimation, and positive-part clipping. Increasing variance in unrelated coordinates can change the loss normalization without a proportional change in probe accuracy. We therefore interpret complement AUROC directly as the reported residual-predictability measure, rather than converting the training surrogate into an information estimate.

The logged `model_loss` includes weighted penalties. Training return is a separate quantity, summarized by averaging checkpoint-wise ratios of the final 30 logged return values. Overlapping windows and policy-dependent visitation limit interpretation of this telemetry as out-of-sample task performance.

## E EVALUATION AND IMPLEMENTATION DETAILS

All six continuation conditions start from the same five baseline weight checkpoints. They use fresh replay, reset world-model optimizer state, and a launcher adapter for additional parameters. Continuation runs disable TF32. Hardware and code fingerprints are recorded per run segment. The runs use V100, L40, and L40S hardware; checkpoint pairing does not assert numerical equivalence across devices.

The evaluation split is by episode, with 70% used to fit probes and 30% to score both probes and axis readouts. No frames are treated as independent checkpoint replicates. AUROC assigns average ranks to ties; a constant predictor has AUROC $0.5$ when both classes are present. The implementation agrees with a reference implementation on heavily tied data to a maximum absolute difference of $2 \times 10^{-16}$. Raw held-out predictions and labels are stored with the evaluation outputs; the paper's aggregates use evaluator-version-2 records.

Episode reassignment operates on complete label sequences, followed by repetition or truncation to the destination episode length. The resulting probes and original-label probes use the same representation partition. Section 5 gives the interpretation of this transformation. No independence claim or chance threshold is inferred from one reassignment.

Compute accounting is approximately 50 GPU-days for the continuations and 45 for the separate controls. The 35 checkpoint-evaluation jobs use about 6.3 GPU-hours in total. Checkpoints and rollout arrays are stored separately from the manuscript assets; aggregate records, training summaries, and asset-generation scripts accompany the paper.