LUCID DREAMING FOR WORLD MODELS: LEARNING TO DOUBT IMAGINATION AND DECIDE BY TRUST

Anonymous authors
Paper under double-blind review

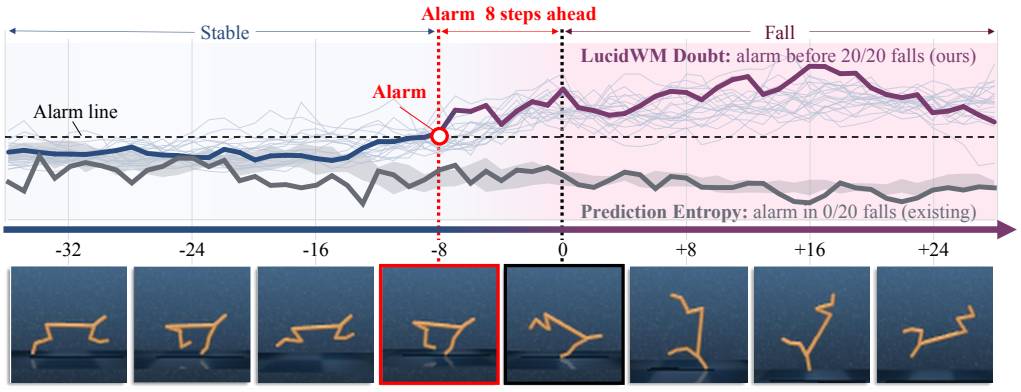

Figure 1: **Our world model with doubt.** The doubt is read from the same head that generates its imagination, with no added parameter. The dark line takes one real fall as an example: the model reports its doubt eight steps before the fall and stays above its alarm line throughout it. We show that it reports all twenty consecutive real falls in the same way, every one before the fall begins. The usual readout, the entropy of the prediction, reports none of them.

ABSTRACT

World models enable agents to learn and plan in imagination, but predictions beyond their experience can become unreliable and mislead decisions. Existing uncertainty estimates derived from predictions can remain overconfident on unfamiliar state-action pairs. We propose the Lucid World Model (LucidWM), which learns doubt from experience and propagates trust through imagination. By integrating Subjective Logic into categorical latent transitions, LucidWM distinguishes predicted outcomes from their evidential support and assigns each transition a degree of doubt. The complement of this doubt defines transition-level trust, which accumulates multiplicatively along imagined trajectories to reweight returns for policy learning and guide action selection. Uncertainty estimation requires no additional parameters or forward passes. Evaluated on four base world models against seventeen uncertainty readouts, LucidWM detects environmental changes and signals uncertainty during action-corrupted rollouts. In a controlled navigation case study, acting on trust reduces the number of steps required to reach the goal from 362 to 190. Fifteen demonstration videos show how LucidWM doubts its dreams and acts on that doubt. Code and videos are available at https://anonymous.4open.science/r/LucidWM.

1 INTRODUCTION

A world model is learned from experience and predicts how the world responds to an agent's actions. With it, the agent can dream: it feeds each predicted state back into the model and imagines future trajectories without further interaction with the environment (Hafner et al., 2025). These trajectories support behaviour learning and planning. Such imagination is especially valuable when real trials are costly (Wu et al., 2022), unsafe (Gao et al., 2024) or impossible to repeat under identical conditions

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.

(Xu et al., 2026). Yet it is often needed to evaluate actions and states beyond the agent's experience, precisely where its world model is least reliable (Yu et al., 2020; Kidambi et al., 2020). Without sufficient experiential support, the model must extrapolate, and prediction errors can compound as each imagined state becomes the input to the next transition (Janner et al., 2019; Asadi et al., 2018).

An unreliable dream is dangerous when believed. The agent may learn behaviours that exploit model errors and plan towards states that are implausible in the real environment (Kurutach et al., 2018). Existing methods use uncertainty estimates to propagate predictive uncertainty or penalise unreliable model-generated transitions (Chua et al., 2018; Yu et al., 2020). Common signals include the entropy or variance of a predictive distribution and disagreement among ensemble members (Sekar et al., 2020). However, a sharp prediction or agreement among models does not, by itself, establish that a queried transition is supported by experience. In our evaluations, these signals can remain insensitive to unfamiliar inputs or even become more confident as prediction reliability deteriorates (Fig. 1, Table 1).

We trace these failures to two distinct limitations. **(i) Doubt concerns support for the input, not just the shape of the output.** A state and an action may each be familiar while their combination is unsupported by observed transitions. The model can still produce a sharp prediction (Hein et al., 2019), so a readout of predictive spread may report confidence despite limited evidence for that transition (Fig. 2). **(ii) Trust depends on the rollout, not just the current step.** Each imagined state inherits its uncertainty from the transitions that produced it. A prediction should therefore be trusted only to the extent that its preceding trajectory is supported, back to the observed starting state. A pointwise uncertainty score alone does not capture this dependence (Berger et al., 2026).

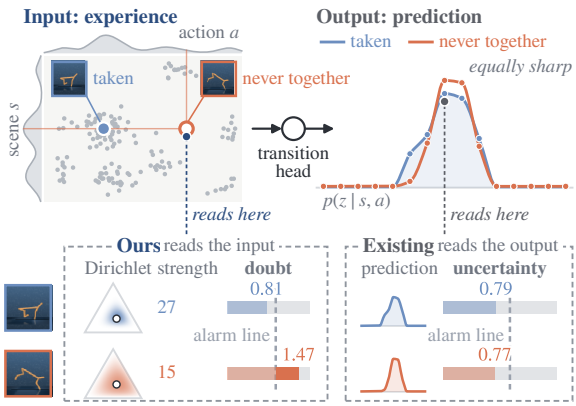

Figure 2: **Same prediction, different experience.** Top: a scene and an action never experienced together are predicted as sharply as a pair that was (real frames: at a fall's onset, and 30 steps before it). Bottom: ours reads the Dirichlet behind the input, whose strength and doubt separate the two; the entropy of the output cannot. Bars: each readout over its own alarm line.

We propose to learn uncertainty from the evidence that experience provides for a transition, rather than infer it solely from the resulting prediction. The world model can then dream lucidly: just as a lucid dreamer knows a dream from waking life, it distinguishes what it predicts from the evidence supporting that prediction. We use the existing transition-head logits to parameterise a Dirichlet distribution over categorical next-state probabilities. Its mean gives the prediction, its concentration above the prior represents learned evidential strength, and the prior's share of the total concentration defines the step's doubt (Jøsang, 2016). Our training objective ties evidence to observed transitions and discourages unsupported evidence. During imagination, we accumulate the complements of these doubts multiplicatively into trajectory trust. This trust governs how strongly the agent relies on imagined continuations when learning and deciding.

We realise this as the Lucid World Model (LucidWM), which integrates doubt and trust into three stages (Fig. 3). **(1) Learning doubt from experience.** We formulate the categorical transition head as a Subjective Logic (SL) opinion, with evidence measured above the Dirichlet prior, and update the posterior by fusing evidence from observation. **(2) Carrying trust through imagination.** We incorporate trust into the standard λ-return, reducing reliance on uncertain imagined continuations. Setting trust to one recovers the standard return. **(3) Deciding by trust.** Before acting, we evaluate the doubt associated with candidate actions and score their imagined futures using trust. LucidWM is designed for world models with categorical latent states and requires no additional learnable parameters or forward passes for uncertainty estimation.

Our contributions are threefold. **(i)** We identify how common predictive readouts can miss unsupported transitions and fail to capture cumulative uncertainty along imagined trajectories. **(ii)** We propose LucidWM, integrating evidential doubt and cumulative trust into world model learn-

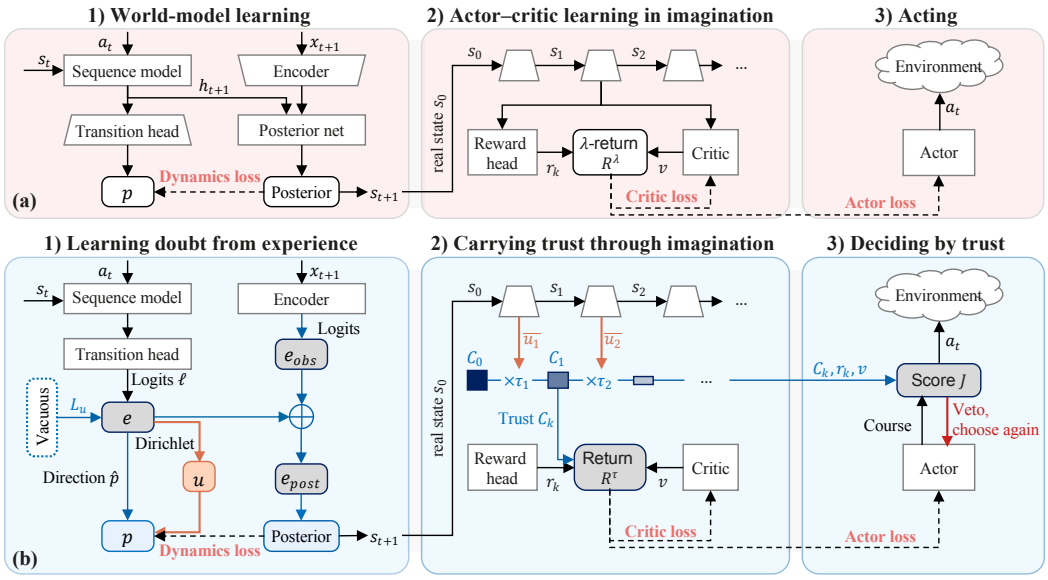

Figure 3: **Doubt is learned from experience and carried as trust.** (a) A standard world model; (b) LucidWM, changes in grey and blue, doubt in orange. 1) The head's evidence e yields a doubt u in place of the fixed ε; the discipline $\mathcal{L}_u$ lets evidence recede toward the vacuous opinion where the dynamics loss does not hold it up. 2) Doubt sets each imagined step's trust τ_k, compounded into C_k; R^τ keeps $\lambda\tau_k$ of each step where R^λ keeps λ. 3) A course whose trust collapses is vetoed.

ing, imagination and decision. **(iii)** Across four world model backbones and seventeen uncertainty readouts, our results demonstrate improved detection of environment changes after observation and rollout drift before decision making. In the reported fall and goal-reaching experiments, LucidWM raises an alarm before each of 20 evaluated falls and reaches the goal in 190 steps, compared with 362 for the base model.

2 Preliminaries and Related Work

World models. A world model (Ha & Schmidhuber, 2018; Hafner et al., 2019) has two parts, an encoder that compresses each observation x_t into a latent state s_t, and a transition that predicts the next state from s_t and the action a_t. Most recent models make the latent discrete, whether the transition is recurrent or a transformer (Hafner et al., 2025; Morihira et al., 2026; Burchi & Timofte, 2025; Zhang et al., 2026). Its stochastic part is G categorical variables of K classes, and the transition predicts each, z_{t+1}, with a categorical head of K logits $\ell(s_t, a_t)$, a softmax with a small uniform mass mixed in so that no class is ever ruled out (Hafner et al., 2025):

$$p(z_{t+1} \mid s_t, a_t) \;=\; (1-\varepsilon)\,\mathrm{softmax}\big(\ell(s_t, a_t)\big) \;+\; \varepsilon/K, \tag{1}$$

where ε is a fixed constant, the same at every state. In training, a second head, the posterior $q(z_{t+1} \mid s_t, a_t, x_{t+1})$, reads the observation that actually arrived, and the transition head is trained to match it. Imagination samples from p for H steps, feeding the sampled state back as the next input, with a reward head supplying r_t; an actor π and a critic v, which estimates the value of a state, are then trained on the imagined trajectory with the λ-return (Hafner et al., 2025; Micheli et al., 2023):

$$R_t^\lambda \;=\; r_t + \gamma\big[(1-\lambda)\,v(s_{t+1}) + \lambda\,R_{t+1}^\lambda\big], \qquad R_H^\lambda = v(s_H), \tag{2}$$

where γ is the discount and the fixed λ sets, at every step, how far the return continues into the imagined future rather than falling back on the critic (we omit DreamerV3's continuation flag).

Subjective Logic. SL (Jøsang, 2016) extends probability by recording how much evidence a belief rests on, so that not knowing is told apart from knowing that an outcome is unlikely. A belief over K outcomes, an *opinion*, is a Dirichlet over the outcome probabilities whose concentration adds

the evidence $e_i \geq 0$ observed for each outcome i to a prior of weight W spread by a base rate, uniform throughout this paper. With $S = \sum_i e_i$ the total evidence and $\hat{p} = e/S$ its direction, the concentration, the uncertainty mass and the mean are:

$$\alpha_i = e_i + W/K, \qquad u = W/(W+S), \qquad P_i = \alpha_i/(W+S) = (1-u)\,\hat{p}_i + u/K. \tag{3}$$

The mean P is the probability the opinion offers to a decision, and the uncertainty mass u, the prior's share of the Dirichlet's strength $W + S$, is one before any evidence and falls as evidence accumulates. We observe that P has exactly the form of Eq. (1), with the fixed ε replaced by u.

Uncertainty in world models. Current world models draw their uncertainty from three sources. The first is the prediction of a single model, read as the entropy or variance of the predicted distribution, its top probability, or the likelihood it gave to the observation that then arrived (Malik et al., 2019; Savov et al., 2025). The second is a head fitted on the model's frozen features, scoring a state by the distance of its features from those seen in training (Lee et al., 2018) or by how poorly a trained network predicts a fixed random network's output on them (Burda et al., 2019); it sees the state, not the action. The third is disagreement among several models trained on the same data (Lakshminarayanan et al., 2017), used as an exploration reward in Plan2Explore (Sekar et al., 2020), as a signal to halt imagination in MOReL (Kidambi et al., 2020), and to weight, mask or truncate imagined rollouts (Buckman et al., 2018; Pan et al., 2020; Frauenknecht et al., 2024). All three are read off what the model produces. We read uncertainty from the evidence behind each imagined step instead. An evidential head reads such evidence for a single prediction (Sensoy et al., 2018; Amini et al., 2020); a world model's prediction is its next input, so its doubt must be carried along the rollout, fused with the observation that follows, and allowed to change what the agent learns.

3 Method

A standard world model pins two constants: its head, Eq. (1), mixes the same ε into every prediction, and its return, Eq. (2), keeps the same share λ of every imagined step. We unfold both into quantities read from the head's own logits, the *doubt* of a transition and the *trust* in a step, and follow them through learning (Sec. 3.1), imagination (Sec. 3.2) and decision (Sec. 3.3), as Fig. 3 summarises.

3.1 Learning doubt from experience

Reading doubt from the transition head. The transition head already computes what we need: its logits, read as the evidence of an opinion. For each categorical variable of the latent, the head's K logits are read as $e = \mathrm{softplus}\,\ell(s_t, a_t)$, a nonnegative amount of evidence per class. By Eq. (3), their total S is the experience the head brings to the scene and action it is asked about, $\hat{p} = e/S$ is its direction, and $u = \max\{W/(W+S),\, \varepsilon\}$ is the *doubt* of the transition, the share of the prediction that no experience supports, floored at the standard head's ε so that the lucid head never trusts a transition more than the standard head does. The prediction is the direction of the evidence; the doubt is its amount. Our head spreads that share uniformly over the classes:

$$\text{STANDARD HEAD} \quad p(z_{t+1} \mid s_t, a_t) = (1-\varepsilon)\,\mathrm{softmax}\big(\ell(s_t, a_t)\big) + \varepsilon/K, \tag{1}$$

$$\text{OUR LUCID HEAD} \quad p(z_{t+1} \mid s_t, a_t) = (1 - u(s_t, a_t))\,\hat{p}(z_{t+1} \mid s_t, a_t) + u(s_t, a_t)/K. \tag{4}$$

Proposition 1 (Standard heads are constant-doubt opinions)**.** Any head of the form Eq. (1) is an instance of Eq. (4) with direction $\mathrm{softmax}(\ell)$ and doubt pinned at $u \equiv \varepsilon$; equivalently, through Eq. (3), it asserts the same evidence total $S \equiv W(1-\varepsilon)/\varepsilon$ for every transition. For example, with DreamerV3's $\varepsilon = 0.01$ and $W = 2$, $S = 198$ (App. C.1).

A standard world model thus assigns every transition, including those it has never experienced, the same amount of evidence. Ours lets the count vary with the transition: prediction and doubt become separate quantities, $\hat{p}$ can be peaked while S is small, and only S sets u. The head predicts as before and reports its doubt at every imagined step from the same forward pass. Read alone, S is only a name for a total; what makes it track experience is how the head is trained, to which we turn next.

Binding evidence to experience. Evidence should grow with experience and with nothing else. The dynamics loss, which trains the transition head toward the posterior, does not enforce this: it decides how the evidence for a transition is allocated across classes, and has zero gradient in u over a

whole interval of doubts (App. C.3), so the total evidence is free. We close the gap with a regulariser, the *evidence discipline*: a pull toward the vacuous opinion, the opinion with no evidence and $u = 1$. Evidential classifiers use such a pull to shed evidence the labels do not support (Sensoy et al., 2018); here its job is different: with the fit blind to the total, the discipline alone sets it:

$$\mathcal{L}_u = \beta_u \sum_g \mathrm{KL}\Big(\mathrm{Dir}\big(e^g(s_t, a_t) + (W/K)\,\mathbf{1}\big) \,\Big\|\, \mathrm{Dir}\big((W/K)\,\mathbf{1}\big)\Big), \tag{5}$$

where g runs over the G categorical variables and β_u weights the term. The dynamics loss now holds evidence up only on the scene–action pairs the agent has taken, and the discipline lets it recede wherever nothing holds it up. On the training distribution the head settles at the most doubtful opinion consistent with the fit (App. C.4); beyond it the doubt rises, as Sec. 4.3 shows.

Fusing observation into prediction. What the model observes should reduce its doubt and never raise it. In a standard world model the posterior is a separate network that reads (s_t, a_t, x_{t+1}) and replaces the prediction outright: what the prediction knew is discarded, and what the observation adds is never counted. Our posterior instead fuses the observation into the prediction with a fusion operator $\oplus$ of SL (Han et al., 2023; Xu et al., 2024); we use cumulative fusion, which adds evidence (Jøsang, 2016) (App. D.1). A linear layer, which replaces the posterior network, reads x_{t+1} alone and reports observation evidence $e_{\mathrm{obs}}(x_{t+1})$; this keeps the two sources independent, so no evidence is counted twice:

$$e_{\mathrm{post}}(s_t, a_t, x_{t+1}) = e(s_t, a_t) \oplus e_{\mathrm{obs}}(x_{t+1}). \tag{6}$$

The posterior is the mean of the fused opinion, Eq. (3) with e_{post}, from which z_{t+1} is drawn in place of q. Fusion never raises the doubt, and brings two more properties: a vacuous observation leaves the state unchanged, and an informative one moves the state toward the observation by exactly the share of doubt it removes. It is also where experience enters the head: the dynamics loss pulls the prediction toward this posterior, whose sharpness bounds the doubt the prediction may keep (App. C.3), so evidence accrues where observations have sharpened the posterior.

3.2 Carrying trust through imagination

Carrying doubt forward as trust. The trust in an imagined step is inherited: each imagined state is the model's own prediction, so a step can be trusted no further than its input. One might carry it in the state, by deducing each imagined opinion from the doubtful one before it; but that recursion settles after one step at a level set by the current step alone, and the lineage is lost (App. D.2). We carry it in the return instead, and show that it then compounds as the trust discounting of SL (App. D.3). The standard return already holds a trust of a kind: at every step it keeps a fixed share λ of the imagined future and hands the rest to the critic. We let that share follow the doubt. Let $\bar{u}_{t+1}$ be the doubt of the transition (s_t, a_t), averaged over the G variables, and $\tau_{t+1} = (1 - \bar{u}_{t+1})/(1 - \varepsilon)$ its trust, equal to one when the doubt sits on its floor ε; the share kept beyond step t becomes $\lambda\tau_{t+1}$:

$$\text{STANDARD RETURN} \quad R_t^{\lambda} = r_t + \gamma\big[(1-\lambda)\, v(s_{t+1}) + \lambda\, R_{t+1}^{\lambda}\big], \tag{2}$$

$$\text{OUR LUCID RETURN} \quad R_t^{\tau} = r_t + \gamma\big[(1-\lambda\tau_{t+1})\, v(s_{t+1}) + \lambda\tau_{t+1}\, R_{t+1}^{\tau}\big], \tag{7}$$

where R_t^{τ} is our trust-weighted return, with the same terminal value $R_H^{\tau} = v(s_H)$. A step with high doubt passes less of the imagined future to the return and more to the critic's value at that step.

Corollary 1. The λ-return Eq. (2) is Eq. (7) with constant trust, $\tau \equiv 1$. With Prop. 1, a standard world model is our lucid world model with both constants pinned.

Compounding trust along the rollout. Trust compounds along the rollout. For a rollout of H steps from a real state s_0, write $C_n = \prod_{k=1}^{n} \tau_k$, with $C_0 = 1$, for the trust carried to its n-th imagined step. Unrolled from s_0, the return shows what this product does:

$$R_0^{\tau} = \sum_{n=1}^{H-1} \big(\lambda^{n-1} C_{n-1} - \lambda^n C_n\big)\, R^{(n)} \;+\; \lambda^{H-1} C_{H-1}\, R^{(H)}, \tag{8}$$

where $R^{(n)} = \sum_{k=0}^{n-1} \gamma^k r_k + \gamma^n v(s_n)$ is the n-step return. The weights are nonnegative and sum to one, and the weight of the n-step return is the share lost between steps $n-1$ and n: one doubtful step lowers the weight of every return beyond it. With $\tau \equiv 1$ the weights are the geometric weights $(1-\lambda)\lambda^{n-1}$ of Eq. (2) (App. D.4). How far imagination counts is thus set by trust, not by a constant: the return reaches as deep as the model's experience reaches, and falls back on the critic beyond it.

3.3 DECIDING BY TRUST

Deciding what to learn from. We train the actor and the critic on R^τ instead of R^λ, so trust decides how much of each imagined step the agent learns from. For this to be safe, reweighting must not change what the critic converges to, and it should make model errors hurt less. Both hold:

Theorem 1 (Contraction and bias bound)**.** For any fixed doubt profile: (i) the critic's update under Eq. (7) is a γ-contraction to $\hat{V}^\pi$, the value of π in the imagined model, as under Eq. (2); (ii) with the critic at the true value V^π and δ_j the model's one-step error at imagined depth j (App. D.4):

$$\left\| \mathbb{E}[R_0^\tau] - V^\pi \right\|_\infty \;\le\; \sum_{j=0}^{H-1} \gamma^j \lambda^j \, C_j \, \delta_j. \tag{9}$$

(i) holds because the weights of Eq. (8) are set by the head, not by the value being learned: trust changes how fast the critic gets there, not where. (ii) weights each depth's model error by $\lambda^j C_j$, the share of the return that reaches that depth, so the bound tightens wherever doubt precedes error.

Deciding what to do. Doubt is a function of (s_t, a), so it is known for any candidate action before the action is taken (App. D.6). We score a candidate future by the trust-weighted sum of its rewards:

$$J(a_{0:H-1}) = \sum_{k=0}^{H-1} \gamma^k \, C_k \, r_k + \gamma^H \, C_H \, v(s_H). \tag{10}$$

J is Eq. (7) with $\lambda = 1$ and the fallback set to zero: untrusted imagined reward contributes nothing, whereas in learning that share went to the critic, so a course the model cannot vouch for cannot win on the critic's word. The simplest use of J is a veto: when trust in the agent's course collapses, every reward beyond that point drops out of J, so the agent withdraws from the course and chooses again (Sec. 4.6). Together, Eqs. (4), (7) and (10) apply one rule three times: keep the share of a prediction that experience supports, and hand the rest to a fallback, the uniform distribution, the critic, or zero.

4 EXPERIMENTS

4.1 EXPERIMENTAL PROTOCOL

Bases. Lucid dreaming applies to any world model with a categorical transition head, and we test it on four bases that span the current designs, in four independent code bases, with the same readout and the same protocol throughout. **DreamerV3** (Hafner et al., 2025) is a *recurrent* state-space model with a *pixel decoder*; **R2-Dreamer** (Morihira et al., 2026) is recurrent but learns its representation *without reconstruction*; **EMERALD** (Burchi & Timofte, 2025) predicts its latent tokens with a *masked transformer*; **OC-STORM** (Zhang et al., 2026) is a *transformer* world model, used here in its visual-only branch. Between them they cover recurrent and transformer dynamics, reconstruction-based and reconstruction-free representations, and flat and spatial latent states.

Environments. We use three environment families to produce situations the model has not experienced. In DeepMind Control (DMC) continuous control (**walker**, **cheetah**, **pendulum**, **finger**, **cartpole**), where the dynamics are smooth, we corrupt the actions fed to imagination; in a first-person **ViZDoom maze**, navigated from 64×64 pixels with discrete actions, we train on one map and then alter it, opening a sealed door or repainting the walls; in **Crafter**, a procedurally generated open world, we corrupt the actions too and let the agent explore and meet caves, lava and terrain it has never seen. Together they cover continuous and discrete actions, proprioceptive and pixel inputs, and situations we inject, build in, or let the agent meet on its own (App. F).

Opponents. We compare against seventeen readouts, eleven in each of the two tests, in the direction their authors define. **Free readouts** of the same pass: entropy, maximum probability, the KL between posterior and prediction, reconstruction and one-step error, and our own formula on the unmodified base (*base* in Table 1); **fitted heads** on frozen features: RND on the latent and on the encoder embedding, an evidential head, Mahalanobis and k-NN distances, a latent ensemble; **multiple forwards**: MC dropout, Laplace, snapshot, self- and deep ensembles. Before decision, every fitted head and ensemble receives the same categorical latent our readout uses, so that what is compared is the readout, not how much the head is allowed to see. Appendix E.3 defines each readout and its cost.

Table 1: (A) After observation: AUROC of the readout, frames of the opened room against familiar ones; 0.5 is chance. (B) Before decision: lift ρ_1/ρ_0, the readout under corrupted actions over the readout under true ones; 1 is no response. Under each column its cost relative to the base ($+c$: heads with c times its parameters; m fwd: m forward passes). Means over five seeds; bold is the best in a row; grey numbers move the wrong way, falling as the model leaves its experience; a dash (–) is a readout the base cannot provide. Each test uses eleven of the seventeen readouts; reconstruction and one-step error need the arriving frame, so appear in (A) only.

(A) After observation				free readouts				fitted heads				multiple forwards	
base	task	**ours** 1×	base 1×	entropy 1×	KL 1×	recon. 1×	1-step 1×	RND +.18	RND-e +.38	evid. +.12	latent +.62	self 3 fwd	deep 3×
DreamerV3	maze	**0.77**	0.24	0.31	0.24	0.06	0.10	0.39	0.27	0.48	0.33	0.21	0.21
R2-Dreamer	maze	**0.83**	0.37	0.36	0.52	–	0.50	0.45	0.66	0.30	0.33	0.41	–
EMERALD	maze	**0.96**	0.72	0.40	0.27	0.05	0.11	0.22	0.20	0.85	0.83	0.26	–
OC-STORM	maze	**0.82**	0.75	0.75	0.37	0.09	0.16	0.53	0.46	0.57	0.59	0.24	–

(B) Before decision				free readouts		fitted heads					multiple forwards		
base	task	**ours** 1×	base 1×	entropy 1×	max p. 1×	Mahal. +.04	kNN +2.4	RND +.13	evid. +.10	latent +.52	MC drop 20 fwd	Laplace 20 fwd	snap. 4 fwd
DreamerV3	walker	**1.30**	0.85	1.03	1.00	1.06	1.07	0.80	1.00	1.04	1.02	1.01	0.93
	cheetah	**2.33**	0.70	1.14	1.14	1.10	0.95	0.93	1.00	1.04	1.04	0.88	0.94
	pendulum	**1.33**	1.00	1.03	1.04	1.08	1.02	0.96	1.00	0.91	1.07	0.84	0.95
	Crafter	**1.49**	0.84	1.02	0.97	0.88	0.93	1.02	1.00	1.01	0.99	0.92	0.95
R2-Dreamer	walker	**3.11**	1.19	1.00	1.01	2.45	1.14	1.36	0.99	1.33	1.31	1.37	–
	cheetah	**2.76**	0.99	1.01	1.02	1.17	1.01	1.18	1.00	1.04	1.21	1.01	–
	pendulum	**1.53**	1.02	0.88	0.87	1.01	0.94	1.11	0.99	0.96	0.96	1.11	–
	Crafter	**1.15**	0.81	1.02	1.00	1.05	1.02	1.05	1.00	1.04	1.00	1.14	1.06
EMERALD	walker	**3.13**	1.10	1.55	1.67	2.08	1.42	1.51	1.00	1.53	0.05	0.06	0.16
	cheetah	**3.67**	1.04	1.42	1.57	2.58	1.56	1.42	1.00	1.72	0.28	0.29	0.25
	pendulum	**2.06**	1.03	1.02	1.03	1.39	1.80	1.87	1.00	1.34	0.62	0.66	0.94
	Crafter	**1.34**	1.01	1.16	1.16	0.89	0.89	1.01	1.00	1.01	0.97	0.97	1.00
OC-STORM	walker	**5.62**	1.16	2.74	2.14	1.08	0.93	2.48	1.00	4.18	4.37	1.08	–
	cheetah	**4.30**	1.06	2.21	2.08	1.12	0.98	1.47	1.03	2.18	2.95	0.69	–
	pendulum	**2.29**	1.03	1.27	1.14	0.86	0.89	0.98	1.00	0.86	1.01	0.80	–
	Crafter	**2.59**	1.04	2.08	1.79	0.82	0.92	0.85	1.01	1.35	0.61	0.55	0.80

4.2 Main results

(A) After observation. *This test asks whether the doubt read after an observation tells that the world has changed.* We open a sealed door in the maze and drive the same routes on the old and the new map; a readout is scored by its AUROC (Table 1). On every base the LucidWM readout recognises the new room. Most readouts of the unmodified base point the wrong way: seen through familiar walls, the new room yields a sharp prediction, so they read it as *more* familiar than the rest of the maze, and the fitted heads and ensembles, at up to three times the cost, are no more reliable.

(B) Before decision. *This test asks whether the doubt read before a decision tells that an imagined future has left the model's experience.* Imagination starts from real states, and its actions are replaced by random ones with growing probability; we report the lift ρ_1/ρ_0, the readout under fully corrupted actions over that under the true ones, so 1 is no response and a readout passes when it rises above it. The LucidWM readout rises and leads on every row, while the same formula on the unmodified base moves little or the wrong way: fed nonsense, a model without doubt barely notices (App. F.5).

4.3 Doubt rises beyond experience

Imagination leaving experience. *This test asks whether doubt rises when imagination leaves what the model has experienced.* Fig. 4 projects the visited states onto a plane, ringed at two standard deviations, and starts 48 random-action rollouts from real states in the cloud, forty steps each, colouring every step by doubt; below, each rollout's doubt and the median are drawn against the 95th percentile on experienced states. The rollouts of both models drift equally far out of the cloud, to twice the

distance of any experienced state from its neighbours. Only LucidWM's doubt rises as they leave: within ten steps its median crosses the 95th percentile, and by the end nine in ten sit above it. The base's doubt never moves: its imagination has left everything it knows, and it is as confident as ever.

Observation leaving experience. *This test asks whether the doubt reads more than the surface of the frames.* We alter the maze in two ways (Fig. 5): the walls of one room are repainted, which changes every pixel and nothing else; and a wall is opened onto a room that was sealed off during training, which changes the layout and shows the agent no texture it has not seen. For the new room we replay the same actions on the training map and read the difference. In the repainted room the lucid doubt rises far above its alarm line again and again; at the new room, it rises each time the agent turns to the opening and falls back as soon as it turns away. Under one alarm rule, set on familiar frames, the LucidWM readout raises an alarm in both rooms and the base's readout in neither. The doubt reads the structure beneath the pixels: every frame at the opening is made of familiar walls, and the doubt rises all the same.

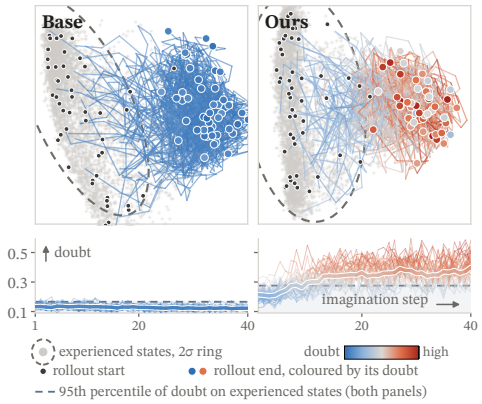

Figure 4: **Ours doubts the departure; base does not.** Top: experienced states and 48 random-action rollouts coloured by doubt. Bottom: doubt along the rollouts.

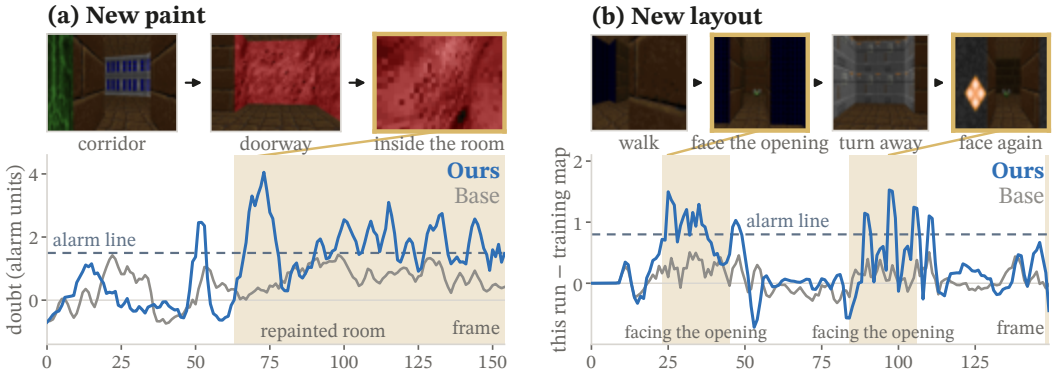

Figure 5: **The alarm fires where the world changed.** (a) A drive into the repainted room. (b) A patrol that twice faces the newly opened room. Doubt in each model's own units; dashed line: alarm line; sand: frames in the changed region, where the gold-outlined pictures were taken.

4.4 Doubt falls as experience grows

Experience accumulating. As the model sees more of the world, its doubt should drain, and drain where its predictions improve. We take one stretch of experience and, at six checkpoints of one run (three in Fig. 6, all in App. F.7), imagine the same future from it and compare it with what happened. At one step the doubt halves, falling at every checkpoint, and the error falls nearly 25-fold, though the two are measured independently: the error by comparing pixels, the doubt from the transition's evidence alone. Step 24 lies past the point where the dynamics turn chaotic, so no amount of training makes the prediction right; there the doubt stays the highest of the steps shown at every checkpoint.

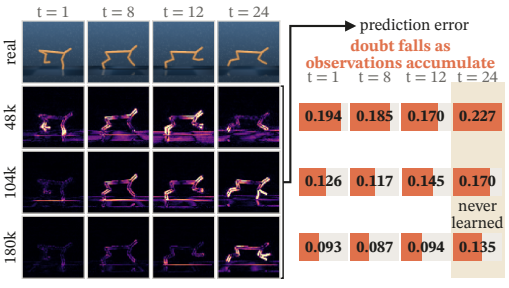

Figure 6: **Learning more, doubting less.** Rows: checkpoints of one run; columns: imagined steps. Tiles: pixel error, dark right, bright wrong. Shaded: chaotic horizon, never learned.

4.5 Ablating trust

Trust as a filter. If trust means what it should, the imagined steps it trusts least should be the ones most wrong. We test this: from 480 real states per run we imagine 15 steps, rank every imagined step by a readout, drop the worst fifth, and measure how far the prediction error of the rest falls, on three DMC tasks. Deep steps are worse on average, so depth alone is a good filter, and over all depths it matches both trust and entropy (Fig. 7a, pale bars). To see what a readout knows beyond depth, we rank within each depth (solid bars): depth alone drops to zero, entropy to almost nothing on two of three tasks, while trust keeps a gain of its own on all three, up to 23%. Over all depths, compounded trust removes 32% of the error against 19% for the doubt of one step; within each depth, the posterior's doubt trails both (Fig. 7b): compounding carries the depth of a dream, the doubt of each transition its drift, so trust reads both how far the dream has gone and how far it has drifted (App. F.8–F.10).

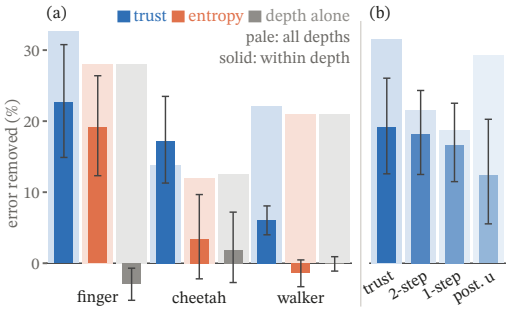

Figure 7: **Trust tells which imagined steps to keep.** Error removed by dropping the worst fifth of steps: (a) three DMC tasks, (b) cheetah from pixels. Pale: over all depths; solid: within each depth; whiskers: bootstrap 95% intervals.

4.6 Acting on trust

A repainted dead end. We run the same weights and seed twice through a dead end whose deep half was repainted after training: every wall is where it was, so a policy that learned the map walks in as before. Once the agent ignores its doubt (*base*); once it acts on trust with the veto of Sec. 3.3 (*ours*). The veto is one rule: when doubt holds above the alarm line for three frames, trust in the course collapses, the agent turns, and its policy resumes. For the first 24 steps the two are one agent. At step 24 the veto fires; ours turns, leaves the dead end and reaches the goal, an armour, at step 190, while base circles the room until step 262 and arrives at step 362 (Fig. 8): the same policy, deciding 238 steps earlier and arriving in half the steps. Over its whole walk the entropy of base crosses its own line for a single frame, so the same rule would never have fired. On both altered maps, 143 of 145 vetoes fired inside a changed room, though the agent is never told where the changes are (App. F.13).

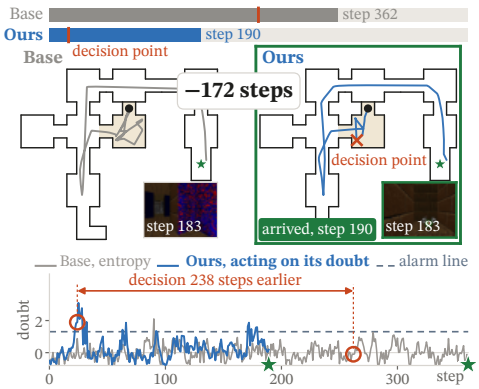

Figure 8: **One veto, half the steps.** Top: steps to the armour. Middle: both routes from the repainted dead end (sand) to the armour (star), and the two views at step 183, seven steps before ours arrives. Bottom: doubt along ours and the entropy of base, each on its own p_{99} alarm line; circles: where ours turns and base leaves.

5 Conclusion

In this work, we identified that existing uncertainty estimates for world models are read from the prediction, and therefore stay confident on exactly what the model has never experienced. To address this, we proposed the Lucid World Model (LucidWM), which learns uncertainty from experience: SL is built into the latent state transition, so that every predicted state carries a doubt, doubt compounds along imagination into trust, and trust weights both what the agent learns from it and what it decides to do. Experiments on four world-model bases and against seventeen readouts show that the doubt recognises a changed world after observation and a drifting rollout before decision, where prior readouts barely move or move the wrong way, and that acting on trust leads to safer choices and faster learning (App. F.11, F.12). A standard world model is the special case with both constants pinned; unfolding them adds no parameter and no forward pass.

Reproducibility Statement

Code and videos are in the anonymous repository linked in the abstract. LucidWM is implemented on four public code bases, one per base; App. E.2 lists the latent, prior weight and imagination horizon of each, App. E.3 defines each of the seventeen readouts and its cost, and App. E.4 gives the protocols of both tests. The formal statements of Sec. 3 are proved in Apps. C and D.

AI Use Statement

We have not used generative AI tools for any task with required disclosure. The theoretical framework, the mathematical claims and their proofs, the method, the experimental design and the interpretation of results are the authors' own; the remaining tasks with required disclosure (synthetic data, translation, dataset cleaning and qualitative analysis) do not arise in this work. We used generative AI tools for tasks with recommended disclosure: editing the paper for readability, formatting references, editing existing code, launching experiment runs and sweeping experimental parameters. Every AI-assisted edit to text, code or parameters was reviewed by the authors, and we take responsibility for the final content of this work, including any text, claims or artifacts produced with the aid of generative AI.

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, 1964.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kavosh Asadi, Dipendra Misra, and Michael L. Littman. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Julia Berger, Bernd Frauenknecht, Sebastian Trimpe, and Bastian Leibe. Biased dreams: Limitations to epistemic uncertainty quantification in latent dynamics models. *Reinforcement Learning Journal*, 7, 2026.

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Maxime Burchi and Radu Timofte. Accurate and efficient world modeling with masked latent transformers. In *International Conference on Machine Learning (ICML)*, 2025.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations (ICLR)*, 2019.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Bernd Frauenknecht, Artur Eisele, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. Trust the model where it trusts itself – model-based actor-critic with uncertainty-aware rollout adaption. In *International Conference on Machine Learning (ICML)*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

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

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR)*, 2017.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Audun Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016.

Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *International Conference on Machine Learning (ICML)*, 2024.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations (ICLR)*, 2023.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-Dreamer: Redundancy-reduced world models without decoders or augmentation. In *International Conference on Learning Representations (ICLR)*, 2026.

Feiyang Pan, Jia He, Dandan Tu, and Qing He. Trust the model when it is confident: Masked model-based actor-critic. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Nedko Savov, Naser Kazemi, Mohammad Mahdi, Danda Pani Paudel, Xi Wang, and Luc Van Gool. Exploration-driven generative interactive environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning (ICML)*, 2020.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. DayDreamer: World models for physical robot learning. In *Conference on Robot Learning (CoRL)*, 2022.

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. medDreamer: Model-based reinforcement learning with latent imagination on complex EHRs for clinical decision support. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Weipu Zhang, Adam Jelley, Trevor McInroe, Amos Storkey, and Gang Wang. Object-centric world models from few-shot annotations for sample-efficient reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

APPENDIX

PAGE

OVERVIEW

A NOTATION . . . 14

B WHAT LUCIDWM CHANGES . . . 14

THEORY

C WHY THE DOUBT READS EXPERIENCE . . . 15
- C.1 The doubt measures the evidence total . . . 15
- C.2 In a table, the doubt counts experience . . . 16
- C.3 The prediction and the fit cannot see the evidence total . . . 16
- C.4 The discipline makes the doubt count experience . . . 17
- C.5 The doubt sets the expected error . . . 18

D WHY TRUST IS CARRIED IN THE RETURN . . . 19
- D.1 Observation never raises the doubt . . . 19
- D.2 In the state, the lineage is lost . . . 20
- D.3 In the return, the doubt compounds . . . 20
- D.4 The lucid return keeps the fixed point and discounts model error . . . 21
- D.5 When trust tightens the bound . . . 22
- D.6 Doubt ranks actions before acting . . . 23

EXPERIMENTS

E EXPERIMENTAL DETAILS . . . 23
- E.1 Every experiment at a glance . . . 23
- E.2 Four bases . . . 23
- E.3 Seventeen readouts and their cost . . . 24
- E.4 Protocols . . . 24

F FURTHER RESULTS . . . 25
- F.1 The alarm fires in the new room . . . 25
- F.2 The alarm fires before the fall . . . 26
- F.3 The alarm fires at nightfall . . . 26
- F.4 The alarm fires on every base . . . 27
- F.5 The worse the actions, the higher the doubt . . . 27
- F.6 The alarm fires at any line, any checkpoint . . . 28
- F.7 The doubt drains as the model learns . . . 28
- F.8 Trust fades faster where the dream drifts . . . 29
- F.9 Trust keeps its edge across ten runs . . . 29
- F.10 Hyperparameter analysis . . . 30
- F.11 Choosing futures by trust cuts deaths . . . 30
- F.12 LucidWM learns Crafter faster than its base . . . 30
- F.13 Vetoes fire where the world changed . . . 31

A NOTATION

Table 2: Symbols of Secs. 2–4. Per variable: defined for each of the G categorical variables, whose index g is dropped when clear. The last column gives the value that recovers the standard world model.

symbol	meaning	defined	standard
World model and imagination			
$x_t,\ s_t,\ a_t$	observation, latent state, action at time t	Sec. 2	
$G,\ K$	number of categorical variables of the state; classes per variable	Sec. 2	
z_{t+1}	one categorical variable of the next state	Eq. (1)	
$\ell(s_t, a_t)$	the transition head's K logits (per variable)	Eq. (1)	
$p(z_{t+1} \mid s_t, a_t)$	the head's prediction	Eqs. (1), (4)	
$q(\cdot \mid s_t, a_t, x_{t+1})$	posterior of a standard world model	Sec. 2	
ε	uniform share mixed into the head (0.01 in DreamerV3); floor of the doubt	Eq. (1)	
$r_t,\ \gamma,\ H$	imagined reward, discount, imagination horizon	Eq. (2)	
$\pi,\ v$	actor, critic	Eq. (2)	
λ	share of the imagined future kept at each step	Eq. (2)	
R_t^λ	λ-return	Eq. (2)	
Opinion of the transition head (per variable)			
$e,\ e_i$	evidence $\mathrm{softplus}\,\ell$; its entry for class i	Eq. (3), Sec. 3.1	
W	prior weight of the opinion	Eq. (3)	
S	total evidence $\sum_i e_i$	Eq. (3)	$W(1-\varepsilon)/\varepsilon$
$\hat{p}$	direction of the evidence, e/S	Eq. (3)	$\mathrm{softmax}(\ell)$
$\alpha_i,\ P$	Dirichlet concentration $e_i + W/K$; mean of the opinion	Eq. (3)	
u	doubt: uncertainty mass $W/(W+S)$, floored at ε	Sec. 3.1	ε
$\mathcal{L}_u,\ \beta_u$	evidence discipline and its weight	Eq. (5)	$\beta_u = 0$
$e_{\mathrm{obs}}(x_{t+1})$	evidence read from the observation alone	Eq. (6)	
$\oplus,\ e_{\mathrm{post}}$	fusion of opinions (here cumulative, which adds evidence); posterior evidence	Eq. (6)	
Trust along imagination and decision			
$\bar{u}_{t+1}$	doubt of the transition (s_t, a_t), averaged over the G variables	Sec. 3.2	ε
τ_{t+1}	trust of that transition, $(1-\bar{u}_{t+1})/(1-\varepsilon)$	Sec. 3.2	1
R_t^τ	trust-weighted return	Eq. (7)	R_t^λ
C_n	trust carried to imagined step n, $\prod_{k=1}^{n} \tau_k$, $C_0 = 1$	Sec. 3.2	1
$R^{(n)}$	n-step return from the real state s_0	Eq. (8)	
$\hat{V}^\pi,\ V^\pi$	value of π in the imagined model; in the environment	Thm. 1	
δ_j	the model's one-step error at imagined depth j	Thm. 1	
$J(a_{0:H-1})$	trust-weighted score of a candidate course of actions	Eq. (10)	$C \equiv 1$
ρ_1/ρ_0	lift: readout under corrupted actions over readout under true ones	Sec. 4.2	

Symbols used inside a single proof in Apps. C–D are defined where they appear.

B WHAT LUCIDWM CHANGES

A standard categorical world model pins two constants: its head mixes the same ε into every prediction, Eq. (1), and its return keeps the same share λ of every imagined step, Eq. (2). LucidWM unfolds both into quantities read from the head's own logits and applies one rule at every stage: keep

the share of a prediction that experience supports, and hand the rest to a fallback. Table 3 lists every change; the statements behind each row are Prop. 1, Prop. 4, Prop. 6, Thm. 1 and Prop. 10. No parameter and no forward pass is added: the posterior is the fused evidence of Eq. (6) in place of a second network, and the readouts u, τ and C are functions of quantities the model already computes.

Table 3: What LucidWM changes in a categorical world model. Every readout is a function of the head's logits; the standard model is recovered by $u \equiv \varepsilon$ and $\tau \equiv 1$.

stage	standard	LucidWM	fallback
prediction, Eq. (4)	$(1-\varepsilon)\,\mathrm{softmax}(\ell) + \varepsilon/K$	$(1-u)\,\hat{p} + u/K$, u from S	uniform $1/K$
training, Eq. (5)	fit only	fit and evidence discipline	the vacuous opinion
posterior, Eq. (6)	a second network q	$e \oplus e_{\mathrm{obs}}$	none: doubt never rises
return, Eq. (7)	share λ per step	share $\lambda\tau_{t+1}$, τ from $\bar{u}$	critic v
decision, Eq. (10)	reward counted in full	reward weighted by C_k	zero

C WHY THE DOUBT READS EXPERIENCE

Section 3.1 reads the total evidence behind a prediction as the experience the head brings to a scene and action. This appendix gives the grounds in five steps (Table 4), with one example pair, in green boxes, followed through all of them. Every result treats the evidence of each scene–action pair on its own and its target as fixed; App. C.4 ends with what carries over to a network with shared parameters.

Table 4: The argument of App. C, one result per step.

step	result	where
measure	the doubt $u = W/(W+S)$ sees the evidence only through its total S	Eq. (3)
ideal	in a table updated by Bayes' rule, S is the joint count; an untaken pair has $u = 1$	Prop. 2
obstacle	neither a readout of the prediction nor the fit can see S	Prop. 3
remedy	the discipline keeps the most doubtful opinion with its mean; $u = 1$ if untaken	Prop. 4
payoff	the doubt sets the expected error of the prediction	Prop. 5

C.1 THE DOUBT MEASURES THE EVIDENCE TOTAL

Notation. An opinion over $K \geq 2$ classes with evidence $e \geq 0$ is the Dirichlet $\boldsymbol{\omega} \sim \mathrm{Dir}(\alpha)$ of Eq. (3), with

$$S = \sum_i e_i, \qquad \alpha = e + \tfrac{W}{K}\mathbf{1}, \qquad \alpha_0 = W + S, \qquad u = \frac{W}{\alpha_0},$$
$$\hat{p} = \frac{e}{S}, \qquad P = \mathbb{E}[\boldsymbol{\omega}] = \frac{\alpha}{\alpha_0} = (1-u)\,\hat{p} + \frac{u}{K}.$$

A *target* q is the distribution the fit pulls the prediction toward (in training, the posterior); the *fit* is $\mathrm{KL}(q \,\|\, P)$; a *readout* is any quantity computed from P. We say *count* for the number N of transitions taken from a pair and *evidence total* for S.

The doubt is the share of the mean P held by the prior: the uncertainty mass of SL (Jøsang, 2016), whose belief masses e_i/α_0 make up the rest. It depends on the evidence only through S, is one at $S = 0$ and falls strictly as S grows. By conjugacy, observed counts add to α, so evidence adds; this is the cumulative fusion of App. D.1.

Proof of Prop. 1. A head of the form Eq. (1) is Eq. (4) with $\hat{p} = \mathrm{softmax}(\ell)$ and $u = \varepsilon$, that is, the opinion with the single total $S_\varepsilon = W(1-\varepsilon)/\varepsilon$ at every input. With the customary $W = 2$ (Jøsang, 2016) and DreamerV3's $\varepsilon = 0.01$, $S_\varepsilon = 198$.

Example. Take $K = 3$ and $W = 2$, and a pair taken $N = 8$ times whose next latent fell in the three classes 6, 2 and 0 times. Then $e = (6, 2, 0)$, $S = 8$ and $u = 0.2$: the prior holds a fifth of the prediction $P \approx (0.67,\ 0.27,\ 0.07)$.

C.2 In a table, the doubt counts experience

Proposition 2 (Doubt counts joint experience)**.** Give each scene–action pair (s, a) its own opinion $\boldsymbol{\omega}(s, a) \sim \mathrm{Dir}((W/K)\mathbf{1})$ over one categorical variable of the next latent, independent across pairs, and update it by Bayes' rule on the transitions experienced from it. After $N(s, a, z')$ transitions $(s, a) \to z'$ for each z',

$$e(s, a) = N(s, a, \cdot), \qquad u(s, a) = \frac{W}{W + N(s, a)}, \qquad N(s, a) = \sum_{z'} N(s, a, z'). \tag{11}$$

The doubt depends on experience only through the joint count $N(s, a)$, and is one for a pair never taken, however often its scene and its action were seen apart.

Proof. The priors are independent across pairs and the likelihood $\prod_{(s,a)} \prod_{z'} \omega_{z'}(s, a)^{N(s,a,z')}$ factorises over pairs. By conjugacy the posterior of (s, a) is $\mathrm{Dir}\big(N(s, a, \cdot) + (W/K)\mathbf{1}\big)$, so $e(s, a) = N(s, a, \cdot)$, $S = N(s, a)$ and $u = W/(W + N(s, a))$; no count of another pair enters.

Example. The evidence $e = (6, 2, 0)$ is exactly the counts; a pair never taken has $e = 0$ and $u = 1$.

In words. A familiar scene and a familiar action that were never taken together, the case of Fig. 2, therefore have doubt one. A network is not a table: it shares one function $e(s_t, a_t)$ across pairs and is trained by a fit, which, as the next step shows, cannot recover the count.

C.3 The prediction and the fit cannot see the evidence total

Many opinions share one prediction (Fig. 9a).

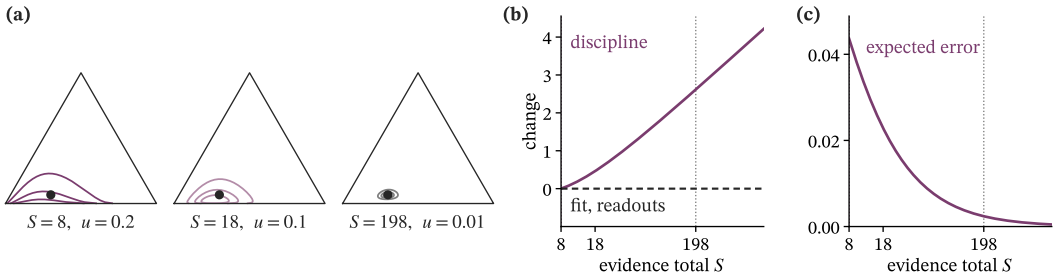

Figure 9: **One prediction, many evidence totals**, for the example pair. (a) Dirichlets with the same mean P (dot) and totals $S = 8$, 18 and 198. (b) Change from $S = 8$ along this family: the fit and every readout stay constant (Prop. 3), while the discipline grows with the total (Lemma 1); among opinions with this mean it prefers the smallest total, $S = 8$, the count. (c) The expected error falls with the total (Prop. 5); at $S = 198$, the total a standard head asserts, it is eighteen times smaller than at the count.

Proposition 3 (The prediction and the fit are blind to the total)**.** Fix a prediction P with full support and let $m = K \min_i P_i$.
(i) For every $u \in (0, m]$, the opinion $\mathrm{Dir}\big((W/u)P\big)$ has mean P and doubt u, and it is the only one; no opinion with mean P has doubt above m.
(ii) Every readout, and the fit to any fixed target q, take the same value on all of these opinions.

Proof. An opinion with doubt u has $\alpha_0 = W/u$, so its mean is P exactly when $\alpha = (W/u)P$, that is, when

$$e = \frac{W}{u}\,P - \frac{W}{K}\,\mathbf{1},$$

which is nonnegative exactly when $u \le m$. Readouts and the fit depend on the opinion only through P.

Example. Here $m = 0.2$, and the opinions with $S = 8$, 18 and 198 all have mean P (Fig. 9a), so every readout and every fit takes one value on them (Fig. 9b).

In words. The readouts of Table 1 that are read off the prediction therefore cannot tell the unexperienced pair of Fig. 2 from an experienced pair with the same sharp prediction. The fit cannot set the total either: its gradient in u vanishes over the whole interval $(0, m]$. This is the non-identifiability of second-order learners whose loss sees only the predictive mean (Jürgens et al., 2024, Thm. 3.2), while losses averaged over the second-order distribution collapse it to a point mass (Bengs et al., 2022, Thm. 1). Something other than the fit must set the total.

C.4 The discipline makes the doubt count experience

The evidence discipline sets the total. We model training in the form of Prop. 2: the discipline takes the place of the prior, charged once on every pair, and the fit the place of the likelihood, charged once per experienced transition. The discipline pulls every pair toward the vacuous opinion; Prior Networks pull toward a flat Dirichlet, and only on inputs out of distribution (Malinin & Gales, 2018). A pair with N experienced transitions and target q is then trained by

$$\mathcal{L}_N(e) = N\,\mathrm{KL}\big(q \,\|\, P\big) + \beta_u\, D(\alpha),$$

$$D(\alpha) = \mathrm{KL}\big(\mathrm{Dir}(\alpha) \,\|\, \mathrm{Dir}(\tfrac{W}{K}\mathbf{1})\big) = \log \frac{\Gamma(\alpha_0)\,\Gamma(\frac{W}{K})^K}{\Gamma(W)\prod_i \Gamma(\alpha_i)} + \sum_i e_i\big[\psi(\alpha_i) - \psi(\alpha_0)\big],$$

with ψ the digamma function. By Prop. 3 an opinion is fixed by its mean P and its strength $\alpha_0 = W + S \ge t_P := W/(K \min_i P_i)$, and the fit depends on P alone. The discipline grows with the total:

Lemma 1 (The discipline grows with the evidence)**.** (i) At a fixed mean P, D is strictly increasing in α_0 on $[t_P, \infty)$.
(ii) $D \ge 0$, with equality only at $e = 0$.

Proof. Idea. At a fixed mean, the derivative of D in the total is a positively weighted sum of $\xi(\alpha_i) - \xi(\alpha_0)$ for a decreasing ξ.

Differentiating gives $\partial D/\partial \alpha_i = e_i\,\psi'(\alpha_i) - S\,\psi'(\alpha_0)$. The function $\xi(y) = y\,\psi'(y)$ is strictly decreasing, since integrating $\psi'(y) = \int_0^\infty t\,e^{-yt}(1 - e^{-t})^{-1} dt$ (Abramowitz & Stegun, 1964, Eq. 6.4.1) by parts gives $\xi(y) = 1 + \int_0^\infty e^{-yt} h'(t)\,dt$ with $h(t) = t/(1 - e^{-t})$ increasing, so also $\xi > 1$; and every $\alpha_i < \alpha_0$. Hence

$$\frac{dD}{d\alpha_0}\Big|_{\alpha = \alpha_0 P} = \frac{1}{\alpha_0} \sum_i e_i\big[\xi(\alpha_i) - \xi(\alpha_0)\big] > 0$$

wherever $e \ne 0$. (ii) is Gibbs' inequality.

Proposition 4 (The discipline sets the evidence total)**.** Let q have full support and $\beta_u > 0$. $\mathcal{L}_N$ has a minimiser, and every minimiser has $u^* = K \min_i P_i^*$: it is the most doubtful opinion with its own mean. Moreover:
(i) without experience, $N = 0$, the doubt is $u^* = 1$;
(ii) with experience, $u^* \ge K \min_i q_i$; for q not uniform, $u^* - K \min_i q_i = \Theta(\beta_u/N)$ as $\beta_u/N \to 0$.

Proof. Idea. The fit sees only the mean and the discipline grows with the total, so the minimiser takes the least total its mean allows; comparing it with the opinion that fits q exactly then bounds the excess. Let $F(P) = \mathrm{KL}(q \,\|\, P)$, $D_{\min}(P) = D(t_P P)$ and $\kappa = \beta_u / N$.

Reduce to the mean. Every opinion with mean P has $\alpha_0 \geq t_P$, so by Lemma 1(i)

$$\mathcal{L}_N(e) \;\geq\; \Phi_N(P) := N\, F(P) + \beta_u\, D_{\min}(P),$$

with equality exactly at $\alpha_0 = t_P$. A minimiser therefore has the smallest total its mean allows: its least likely class holds no evidence, and $u^* = K \min_i P_i^*$. Φ_N is continuous and, as q has full support, grows without bound as $\min_i P_i \to 0$, so a minimiser exists.

(i). At $N = 0$ the loss is $\beta_u D$, which vanishes only at $e = 0$ (Lemma 1(ii)).

(ii), upper bound. The opinion with mean q and doubt $K \min_i q_i$ (Prop. 3) has loss $\beta_u D_{\min}(q)$, so $F(P^*) \leq \kappa\,[D_{\min}(q) - D_{\min}(P^*)] \leq \kappa\, D_{\min}(q)$, and $P^* \to q$ as $\kappa \to 0$. Near q, $D_{\min}$ is Lipschitz and $F \geq \frac{1}{2}\|P^* - q\|_1^2$ (Pinsker), so $\|P^* - q\|_1$ and hence $u^* - K \min_i q_i \leq K\|P^* - q\|_1$ are $O(\kappa)$.

(ii), lower bound. Let $N > 0$ and q be non-uniform (for uniform q, $e^* = 0$ and $u^* = 1 = K \min_i q_i$); then $S^* > 0$. At the class j left without evidence, raising e_j cannot lower the loss, the first inequality below; multiplying it by α_0^*/N gives the second:

$$N\Big(\frac{1}{\alpha_0^*} - \frac{K q_j}{W}\Big) - \beta_u S^* \psi'(\alpha_0^*) \;\geq\; 0 \quad\Longrightarrow\quad 1 - \frac{K q_j}{u^*} \;\geq\; \kappa\, S^*\, \xi(\alpha_0^*) \;>\; \kappa\, S^*,$$

using $\xi > 1$ (proof of Lemma 1). Hence $u^* = K P_j^* > K q_j / (1 - \kappa S^*) \geq K \min_i q_i\,(1 + \kappa S^*)$. By the upper bound $u^* \to K \min_i q_i$, so $K \min_i q_i\, S^* \to W(1 - K \min_i q_i) > 0$ and the excess is at least of order κ.

Example. With $\beta_u = 0.1$ the discipline settles the pair at $u^* \approx 0.24$ against the count's 0.2, and returns $u^* = 1$ for a pair never taken.

In words. When the target is the posterior mean of Prop. 2 after N transitions, $K \min_i q_i \geq W/(W+N)$, the doubt of Eq. (11), with equality when some class was never observed. By (ii), the discipline then never leaves an experienced pair more confident than the count, and errs toward doubt: by an excess of order β_u/N, and by more where every class was observed. Regularised evidential learners take, among equal fits, the one the regulariser prefers (Jürgens et al., 2024, Sec. 3.4), and their uncertainty is then set by the regularisation weight (Bengs et al., 2022; Jürgens et al., 2024); here the weight sets only an excess of order β_u/N above a level fixed by the target.

With a network, the evidence of every pair is one function $e(s_t, a_t)$ of shared parameters. Posterior Networks keep the count of Prop. 2 in such a function with a learned density, setting the evidence of a class to its count times the density at the input (Charpentier et al., 2020); the lucid head needs no density model. The fit rewards a direction that generalises across pairs but never rewards evidence away from the transitions that demand it, so through the shared parameters the discipline lowers the total wherever no transition holds it up. Sec. 4 measures this on four bases: after observation and before decision (Table 1), along a rollout that leaves the experienced states (Fig. 4), and as experience accumulates (Fig. 6).

C.5 The doubt sets the expected error

Proposition 5 (Doubt sets the expected error of the prediction)**.** Let $\boldsymbol{\omega} \sim \mathrm{Dir}(\alpha)$ be the model's opinion on the distribution of a next latent variable, with mean P and doubt u. Then

$$\mathbb{E}\,\|\boldsymbol{\omega} - P\|_2^2 = \big(1 - \|P\|_2^2\big)\,\frac{u}{W+u}, \tag{12}$$

the share $u/(W+u)$ of the largest expected error, $1 - \|P\|_2^2$, that any belief with mean P can carry. For a fixed prediction it rises strictly with the doubt; a standard head gives every transition with that prediction the same share.

Proof. A Dirichlet has $\mathrm{Var}[\omega_i] = P_i(1-P_i)/(\alpha_0+1)$, so $\mathbb{E}\|\boldsymbol{\omega} - P\|_2^2 = (1-\|P\|_2^2)/(\alpha_0+1)$, and $\alpha_0+1 = (W+u)/u$. Any belief with mean P has $\mathbb{E}\|\boldsymbol{\omega}-P\|_2^2 = \mathbb{E}\|\boldsymbol{\omega}\|_2^2 - \|P\|_2^2 \le 1-\|P\|_2^2$, with equality for the belief on the vertices. For fixed P, $1-\|P\|_2^2 > 0$ and $u/(W+u)$ rises with u; a standard head has $\alpha_0 = W/\varepsilon$ everywhere.

Example. Here $1-\|P\|_2^2 = 0.48$: the expected error is 0.044 at the count, $S = 8$, but 0.0024 at $S = 198$, eighteen times smaller for the same prediction (Fig. 9c).

The error is taken under the model's own opinion; Fig. 6 tests whether that opinion is calibrated, and App. D.5 says what calibration buys in the return.

D WHY TRUST IS CARRIED IN THE RETURN

Section 3.2 carries the doubt of each imagined step as trust in the return, and Sec. 3.3 learns and decides by that trust. Carried through the imagined state instead, the doubt of a step would be lost one step later; carried in the return, it compounds, keeps the critic's fixed point and discounts every model error that follows a doubtful step. Table 5 lays out the argument; Apps. D.1–D.5 each give an example in a green box, and Apps. D.2–D.5 follow one imagined rollout.

Table 5: The argument of App. D, step by step.

step	result	where
start	at the real state, fusing in an observation never raises the doubt	Prop. 6
alternative	carried in the state, the doubt of a step is lost one step later	Prop. 7
choice	carried in the return, trust compounds, as SL trust discounting does	Prop. 8
safety	the lucid return keeps the critic's fixed point	Thm. 1
payoff	errors that follow a doubtful step are discounted, where it pays	Prop. 9
action	doubt ranks candidate actions before any is taken	Prop. 10

Notation. Symbols are those of Table 2. A rollout starts at a real state s_0 and takes H imagined steps; step k has the doubt $\bar{u}_k \ge \varepsilon$ of the transition (s_{k-1}, a_{k-1}) and the trust $\tau_k \in [0,1]$, and $C_0 = 1$. We take $\lambda \in (0,1]$ and $\gamma \in (0,1)$, and write $M_n = \lambda^n C_n$ for the share of the return that reaches step n. A doubt profile is *fixed* when $\bar{u}_1, \dots, \bar{u}_H$ are the same on every imagined rollout. Prop. 6 uses the doubt before the floor, $u = W/(W+S)$, as in App. C.

D.1 OBSERVATION NEVER RAISES THE DOUBT

Proposition 6 (Fusion never raises the doubt)**.** Fuse a prediction (evidence e, doubt u) with an observation (evidence e_{obs}, total S_{obs}, doubt $u_{\mathrm{obs}} = W/(W+S_{\mathrm{obs}})$) by cumulative fusion (Jøsang, 2016, Sec. 12.3), $e_{\mathrm{post}} = e + e_{\mathrm{obs}}$, and let u_{post} and P_{post} be the doubt and mean of the result. Then
(i) $u_{\mathrm{post}} \le \min(u, u_{\mathrm{obs}})$, and $1/u_{\mathrm{post}} = 1/u + S_{\mathrm{obs}}/W$;
(ii) $P_{\mathrm{post}} = (1-\theta)\,P + \theta\,\hat{p}_{\mathrm{obs}}$, with $\hat{p}_{\mathrm{obs}} = e_{\mathrm{obs}}/S_{\mathrm{obs}}$ and $\theta = 1 - u_{\mathrm{post}}/u$ the share of the doubt the observation removes; a vacuous observation, $S_{\mathrm{obs}} = 0$, has $\theta = 0$ and leaves the prediction unchanged.

Proof. The total $S + S_{\mathrm{obs}}$ is at least S and at least S_{obs}, and the doubt falls with the total; $W/u_{\mathrm{post}} = W + S + S_{\mathrm{obs}} = W/u + S_{\mathrm{obs}}$. For (ii), $\alpha_{\mathrm{post}} = \alpha + e_{\mathrm{obs}}$, so $P_{\mathrm{post}} = \big((W+S)\,P + S_{\mathrm{obs}}\,\hat{p}_{\mathrm{obs}}\big)/(W+S+S_{\mathrm{obs}})$, and $S_{\mathrm{obs}}/(W+S+S_{\mathrm{obs}}) = 1 - u_{\mathrm{post}}/u$.

Example. Fuse the example pair of App. C, $e = (6,2,0)$ and $u = 0.2$, with an observation $e_{\mathrm{obs}} = (4,0,0)$, $u_{\mathrm{obs}} = 0.33$. Then $e_{\mathrm{post}} = (10,2,0)$ and $u_{\mathrm{post}} = 0.14$, below both. The observation removes $\theta = 0.29$ of the doubt and moves $P \approx (0.67,\ 0.27,\ 0.07)$ 29% of the way to $(1,0,0)$, to $P_{\mathrm{post}} \approx (0.76,\ 0.19,\ 0.05)$.

These identities are specific to cumulative fusion. The reduced Dempster rule of evidential multi-view learning also never raises the uncertainty mass (Han et al., 2023, Prop. 3.3), while averaging fusion, which averages the evidence of the views (Xu et al., 2024), can raise it above the smaller of the two.

D.2 In the state, the lineage is lost

One could carry the doubt in the imagined state, deducing the opinion on each imagined state from the opinion on the state before, as the deduction of SL does (Jøsang, 2016, Ch. 9).

Proposition 7 (Deduction forgets the lineage)**.** Deduce the opinion on each imagined state from the opinion on the one before, which has doubt $\tilde{u}_{k-1}$, with $\tilde{u}_0 = 0$ at the real state. Take the doubt of the step's conditionals as $\bar{u}_k$, the only doubt the head reports. Then

$$\tilde{u}_k = \bar{u}_k + \sigma_k\,\tilde{u}_{k-1} = \sum_{j=1}^{k}\Big(\prod_{l=j+1}^{k}\sigma_l\Big)\,\bar{u}_j, \qquad \sigma_k = u_k^{\circ} - \bar{u}_k,$$

where u_k° is the doubt that deduction assigns to state k when state $k-1$ is vacuous. So $\tilde{u}_1 = \bar{u}_1$, $|\tilde{u}_k - \bar{u}_k| = |\sigma_k|\,\tilde{u}_{k-1} \le |\sigma_k|$ at every depth, and the doubt of an earlier step j reaches depth k only through the factor $\prod_{l=j+1}^{k}\sigma_l$: where $|\sigma|$ is small, a doubtful step is forgotten one step later.

Proof. Deduction gives the deduced opinion the doubt $u_{Y\|X} = u_X\,u_{Y\|\hat{X}} + \sum_i u_{Y|x_i}\,b_X(x_i)$ (Jøsang, 2016, Eq. (9.75)), where u_X and b_X are the doubt and belief masses of the antecedent and $u_{Y\|\hat{X}}$ is the doubt deduced from a vacuous antecedent. With $u_X = \tilde{u}_{k-1}$, $u_{Y\|\hat{X}} = u_k^{\circ}$, $u_{Y|x_i} = \bar{u}_k$ and $\sum_i b_X(x_i) = 1 - \tilde{u}_{k-1}$, this is $\tilde{u}_k = u_k^{\circ}\,\tilde{u}_{k-1} + \bar{u}_k\,(1 - \tilde{u}_{k-1})$, the first form. Unrolling from $\tilde{u}_0 = 0$ gives the second, and $\tilde{u}_{k-1} \le 1$ gives the bound.

Example. Take a rollout of $H = 8$ imagined steps that leaves experience at steps 3 and 4, $\bar{u}_3 = \bar{u}_4 = 0.40$, and is familiar elsewhere, $\bar{u}_k = 0.01$; with $\varepsilon = 0.01$ the trusts are 0.61 at steps 3 and 4 and 1 elsewhere. With $|\sigma_k| \le 0.01$, deduction gives states 3 and 4 the doubt 0.40 and states 5 to 8 the doubt 0.01, each to within 0.005: states 5 to 8 are trusted again (Fig. 10a).

In words. On a trained model on cheetah run, $\sigma = 0.009$ at the median, and exact deduction gives the same doubt at depth fifteen as at depth one, where compounding would have raised it. The recursion settles after one step at a level set by the current step, and the lineage of the rollout is lost.

D.3 In the return, the doubt compounds

Carried in the return instead, trust compounds along the rollout.

Proposition 8 (Carried trust is the trust discounting of SL)**.** Read the k-th imagined step as an advisor in a referral chain that starts at the real state s_0, and its trust τ_k as the projected probability of the binomial opinion that this step is reliable. Then the carried trust $C_n = \prod_{k=1}^{n}\tau_k$ is the projected probability that the transitive trust discounting of SL (Jøsang, 2016, Sec. 14.3.4, Def. 14.7) assigns to the chain of the first n steps, and an opinion with belief masses b reported at depth n is discounted to belief $C_n\,b$ and uncertainty $1 - C_n\sum_i b_i$. With $\tau \equiv 1$ the chain is fully trusted and nothing is discounted.

Proof. Transitive discounting multiplies the projected probabilities of the referral edges along the path (Jøsang, 2016, Eq. (14.13)) and scales the belief of the final opinion by that product (Jøsang, 2016, Def. 14.7). With edge k at τ_k the product is C_n. Only the projected probabilities enter, so the result does not depend on how each τ_k is split into belief, disbelief and uncertainty.

Example. In the return, the rollout carries the trust $C_n = 1$ up to step 2, 0.61 at step 3 and 0.37 from step 4 on: states 5 to 8 inherit the doubt of steps 3 and 4, which deduction dropped (Fig. 10a).

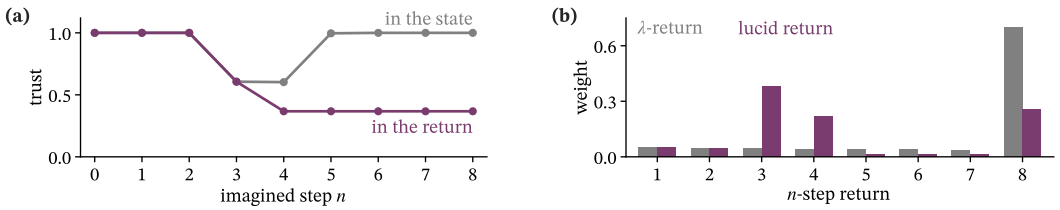

Figure 10: **Trust in the state loses the lineage; trust in the return compounds**, for the example rollout. (a) Trust at each imagined step n: deduced through the state, $(1-\tilde{u}_n)/(1-\varepsilon)$ with $\sigma = 0.009$ (Prop. 7), and carried in the return, C_n (Prop. 8). (b) Weights of the one- to eight-step returns under the λ-return and the lucid return (Lemma 2).

In words. Deduction through the state and discounting along the chain are the two ways SL chains opinions. The first forgets a doubtful step once it is passed, the second carries it to every step built on it (Fig. 10a); this is why LucidWM carries doubt as trust in the return.

D.4 The lucid return keeps the fixed point and discounts model error

The lucid return mixes the n-step returns; Lemma 2 gives the mixture, and Thm. 1 what it does to learning.

Lemma 2 (Unrolled lucid return)**.** For every doubt profile, Eq. (7) unrolls to Eq. (8): the n-step return $R^{(n)}$ has weight $w_n = M_{n-1} - M_n$ for $n < H$ and $w_H = M_{H-1}$. The weights are nonnegative and sum to one, and the returns longer than n steps carry total weight M_n. With $\tau \equiv 1$, $w_n = (1-\lambda)\lambda^{n-1}$, $w_H = \lambda^{H-1}$ and $R^\tau_t = R^\lambda_t$ (Cor. 1).

Proof. Since $R^\tau_H = v(s_H)$, the last step reads $R^\tau_{H-1} = r_{H-1} + \gamma\, v(s_H)$ whatever τ_H. Unrolling Eq. (7) from $t = 0$,

$$R^\tau_0 = \sum_{k=0}^{H-1} \gamma^k M_k\, r_k + \sum_{n=1}^{H-1} \gamma^n M_{n-1}(1-\lambda\tau_n)\, v(s_n) + \gamma^H M_{H-1}\, v(s_H).$$

In $\sum_n w_n R^{(n)}$ the reward r_k collects $\sum_{n>k} w_n = M_k$ and $v(s_n)$ collects w_n, and $M_{n-1}(1-\lambda\tau_n) = M_{n-1} - M_n$, so the two agree. The weights are nonnegative as $\lambda\tau_n \le 1$, telescope to $M_0 = 1$, and those beyond n sum to M_n. With $\tau \equiv 1$, $M_n = \lambda^n$.

Example. With $\lambda = 0.95$, the weight of the eight-step return falls from 0.70 under the λ-return to 0.26, and the three- and four-step returns, which hand over to the critic at steps 3 and 4, gain 0.34 and 0.18 (Fig. 10b).

Theorem 1 (Contraction and bias bound, restated and extended)**.** For any fixed doubt profile: (i) the critic's update under Eq. (7) is a γ-contraction to $\hat{V}^\pi$, the value of π in the imagined model, as under Eq. (2); (ii) with the critic at the true value V^π and $\delta_j = \sup|(\hat{T}-T)V^\pi|$ over the states reached after j imagined steps, where $\hat{T}$ and T are the Bellman operators of π in the model and in the environment,

$$\big\|\, \mathbb{E}[R^\tau_0] - V^\pi \big\|_\infty \;\le\; \sum_{j=0}^{H-1} \gamma^j \lambda^j\, C_j\, \delta_j;$$

(iii) when the trust varies with the state–action pair, (i) holds as it stands, and at each real state s_0, with $\eta = (\hat{T} - T)V^\pi$ the model's one-step error,

$$\big|\, \mathbb{E}[R^\tau_0 \mid s_0] - V^\pi(s_0) \big| \;\le\; \sum_{j=0}^{H-1} \gamma^j \lambda^j\, \mathbb{E}\big[C_j\, |\eta(s_j)| \;\big|\; s_0\big].$$

Proof. Idea. By Lemma 2 the target is a mixture of n-step returns with weights set by the head; each part then follows from one backward recursion.

(i). Let $(\mathcal{R}v)(s_0) = \mathbb{E}[R_0^\tau \mid s_0]$, over imagined rollouts from s_0 under π and the model. By Lemma 2, $R_0^\tau = \sum_{k=0}^{H-1} \gamma^k M_k r_k + \sum_{n=1}^{H} \gamma^n w_n \, v(s_n)$, so

$$|\mathcal{R}v - \mathcal{R}v'|(s_0) \;\le\; \mathbb{E} \sum_{n=1}^{H} \gamma^n w_n \, |v - v'|(s_n) \;\le\; \gamma \, \|v - v'\|_\infty,$$

as $\gamma^n \le \gamma$ and the weights sum to one. For the fixed point, Eq. (7) reads $R_t^\tau = r_t + \gamma v(s_{t+1}) + \gamma\lambda\tau_{t+1}\big(R_{t+1}^\tau - v(s_{t+1})\big)$. Let $v = \hat{V}^\pi$ and $\zeta_t = \mathbb{E}[R_t^\tau \mid s_t] - \hat{V}^\pi(s_t)$, with $\zeta_H = 0$. As $r_t + \gamma\hat{V}^\pi(s_{t+1})$ has mean $\hat{V}^\pi(s_t)$ under the model, $\zeta_t = \gamma\lambda\tau_{t+1}\,\mathbb{E}[\zeta_{t+1} \mid s_t]$, so every ζ_t vanishes: $\hat{V}^\pi$ is a fixed point, and by the contraction the only one.

(ii). Let $v = V^\pi$ and $\zeta_t = \mathbb{E}[R_t^\tau \mid s_t] - V^\pi(s_t)$. Now $r_t + \gamma V^\pi(s_{t+1})$ has mean $(\hat{T}V^\pi)(s_t) = V^\pi(s_t) + \eta(s_t)$, since $TV^\pi = V^\pi$. So $\zeta_t = \eta(s_t) + \gamma\lambda\tau_{t+1}\,\mathbb{E}[\zeta_{t+1} \mid s_t]$, which unrolls to $\zeta_0 = \sum_{j=0}^{H-1} \gamma^j M_j \, \mathbb{E}[\eta(s_j) \mid s_0]$, and $|\mathbb{E}[\eta(s_j) \mid s_0]| \le \delta_j$.

(iii). Both parts use only that the trust of a step is known at the state and action it leaves. The weights of Lemma 2 stay nonnegative and sum to one on every rollout, so (i) holds; with τ_{t+1} inside the expectation, (ii) unrolls to $\zeta_0 = \sum_j \gamma^j \lambda^j \, \mathbb{E}[C_j \, \eta(s_j) \mid s_0]$, which gives the bound.

D.5 When trust tightens the bound

With a critic off the true value, trust also sets how much of the return rests on the critic.

Proposition 9 (When lowering trust tightens the bound)**.** Let the critic err by $\Delta_v = \|v - V^\pi\|_\infty$. For any fixed doubt profile:
(i) the bias is bounded by

$$\big\| \mathbb{E}[R_0^\tau] - V^\pi \big\|_\infty \;\le\; \gamma\Delta_v + \delta_0 + \sum_{j=1}^{H-1} \gamma^j M_j \big(\delta_j - (1-\gamma)\Delta_v\big),$$

and lowering the trust τ_k of one step, $1 \le k \le H-1$, with $C_k > 0$, lowers this bound exactly when the mean of δ_j over $j \ge k$, weighted by $\gamma^j M_j$, exceeds $(1-\gamma)\Delta_v$;
(ii) with the critic at the true value, $\Delta_v = 0$, this is the bound $B(C) = \sum_{j=0}^{H-1} \gamma^j \lambda^j C_j \delta_j$ of Eq. (9), and against the standard return, $\tau \equiv 1$, $B(\mathbf{1}) - B(C) = \sum_{j=0}^{H-1} \gamma^j \lambda^j (1 - C_j)\, \delta_j \ge 0$: the error at each depth is discounted by the trust lost before it.

Proof. (i) By Lemma 2, replacing V^π by v changes R_0^τ by $\sum_{n=1}^{H} \gamma^n w_n \, (v - V^\pi)(s_n)$, at most $\Delta_v \sum_{n=1}^{H} \gamma^n w_n$ in size, and summation by parts gives $\sum_{n=1}^{H} \gamma^n w_n = \gamma - (1-\gamma) \sum_{j=1}^{H-1} \gamma^j M_j$. Adding Thm. 1(ii) gives the bound. Every M_j with $j \ge k$ is proportional to τ_k and no other term depends on it, so the bound is linear in τ_k, whose slope has the sign of $\sum_{j \ge k} \gamma^j M_j \big(\delta_j - (1-\gamma)\Delta_v\big)$. (ii) Subtract the two sums; as $\bar{u} \ge \varepsilon$, $\tau \le 1$, so $C_j \le 1$, and $\delta_j \ge 0$.

Example. In the rollout, errors at depths 0 to 2, up to the first doubtful transition, keep their factors $M_j = 1,\ 0.95,\ 0.90$ in the bound. Errors from depth 3 on, which follow the doubtful steps, enter with $M_3 = 0.52$ down to $M_7 = 0.26$, where the λ-return has λ^j, 0.86 down to 0.70.

In words. Every M_j with $j \ge k$ carries τ_k, so one doubtful step discounts every model error that follows it and none before it; a smaller λ would discount every depth alike. As the doubt rises where the rollout leaves experience (Sec. 4.3), these are the errors made after it has left what the model knows. Prop. 9(i) says where the discount pays when the critic errs: where the model's one-step error beyond a step exceeds $(1-\gamma)\Delta_v$, the critic's error per step. Which imagined steps the trust actually removes is measured in Sec. 4.5.

D.6 DOUBT RANKS ACTIONS BEFORE ACTING

Proposition 10 (Doubt before acting)**.** The doubt $u(s_t, a)$ is computed by the head from (s_t, a) alone, so it is known for every candidate action before any is taken. Above the floor ε, $u(s_t, a) > u(s_t, a')$ exactly when $S(s_t, a) < S(s_t, a')$, for every prior weight $W > 0$.

Proof. The evidence $e = \operatorname{softplus} \ell(s_t, a)$ needs no observation of the outcome, and $u = W/(W + S)$ falls strictly with S for every $W > 0$.

What acting on trust changes in the agent's decisions is measured in Sec. 4.6.

E EXPERIMENTAL DETAILS

LucidWM is compared with seventeen readouts on four bases and three environment families (Table 6). This appendix gives the bases, readouts and protocols; App. F gives the results beyond the main text.

E.1 EVERY EXPERIMENT AT A GLANCE

Table 6: Every experiment in the paper and what it finds. DMC in Table 1(B): walker, cheetah, pendulum.

experiment	bases	environments	result
main text			
after observation, Tab. 1(A)	all four	maze	first on every base
before decision, Tab. 1(B)	all four	DMC, Crafter	first on all 16 rows
leaving experience, Fig. 4	DreamerV3	walker	9 in 10 over the line; base flat
two changed rooms, Fig. 5	DreamerV3	maze	alarm in both; base in neither
learning, Fig. 6	DreamerV3	cheetah	doubt falls with the error
trust as a filter, Fig. 7	DreamerV3	walker, cheetah, finger	gain within depth on all three, up to 23%
acting on trust, Fig. 8	DreamerV3	maze	goal at step 190; base at 362
appendix			
new room, Fig. 11	DreamerV3	maze	25/29 walks; base 0/29
one pass vs. three models, Fig. 12	DreamerV3	maze	8/10 walks; ensembles 0/10
falls, Fig. 13	DreamerV3	cheetah	20/20 before onset; entropy 0/20
nightfall, Fig. 14	DreamerV3	Crafter	peak $2.6\times$ its line; base $< 0.5\times$
every base, Fig. 15	all four	maze	alarm on all four bases
corruption ladder, Fig. 16	OC-STORM, EMERALD	walker, cheetah, Crafter	over fivefold; beats 20-pass readouts
any line, any checkpoint, Fig. 17	DreamerV3	maze	28/28 rules, 38/38 checkpoints
six checkpoints, Fig. 18	DreamerV3	cheetah	doubt halved; error $24.6\times$ lower
dream and reality, Fig. 19	DreamerV3	walker, Crafter	trust 0.16 against 0.42 when calm
trust over ten runs, Fig. 20(a)	DreamerV3	four DMC tasks	11% within depth, first of six
hyperparameters, Fig. 20(b, c)	DreamerV3	cheetah, walker	whole rollout best; return flat
choosing by trust, Fig. 21(a)	DreamerV3	Crafter	deaths/1k steps 3.80 against 4.08
learning Crafter, Fig. 21(b, c)	DreamerV3	Crafter	score 15.6% against 9.0%
where vetoes fire, Fig. 22	DreamerV3	maze	143/145 in a changed room
total: 210 compared cells in Table 1 · 4 bases · 3 environment families, 7 tasks · 17 readouts			

E.2 FOUR BASES

Table 7 lists the four bases.

Table 7: The four bases; $\varepsilon = 0.01$ and $\lambda = 0.95$ on all. $G \times K$: categorical variables and classes of the latent state; W: prior weight; H: imagination horizon.

base	dynamics	$G \times K$	W	H
DreamerV3	recurrent, pixel decoder	32×32	2	15
R2-Dreamer	recurrent, no reconstruction	32×16	1	15
EMERALD	masked transformer, spatial	4×4 cells of 32×32	2	15
OC-STORM	transformer	32×32	2	16

E.3 Seventeen readouts and their cost

Every readout is read on the same frames and imagined futures, and oriented so that higher means less familiar (Table 8). The fitted heads and multiple forwards are RND (Burda et al., 2019), an evidential head (Sensoy et al., 2018), latent disagreement (Sekar et al., 2020), the Mahalanobis distance (Lee et al., 2018), deep nearest neighbours (Sun et al., 2022), deep ensembles (Lakshminarayanan et al., 2017), MC dropout (Gal & Ghahramani, 2016), a Laplace posterior (Daxberger et al., 2021) and snapshot ensembles (Huang et al., 2017); *self*, three samples of one model, is our control for *deep*.

Table 8: The readouts compared in Table 1, as defined on DreamerV3; the other bases read them from their own features. Cost as in Table 1, for tests (A) / (B).

family	readout	what it reads	cost (A / B)
free	base	our doubt, read from the logits of the unmodified base	$1\times$
	entropy	entropy of the latent distribution	$1\times$
	max p.	one minus its largest probability	$1\times$
	KL	divergence of the posterior from the prediction	$1\times$
	recon.	reconstruction error of the arriving frame	$1\times$
	1-step	error of the arriving frame decoded from the prediction	$1\times$
fitted heads	RND	error against a fixed random network, on the model state	$+.18$ / $+.13$
	RND-e	the same, on the encoder embedding	$+.38$
	evid.	vacuity of an evidential head trained on the model state	$+.12$ / $+.10$
	latent	disagreement of five transition heads	$+.62$ / $+.52$
	Mahal.	Mahalanobis distance to the training states	$+.04$
	kNN	distance to the fiftieth nearest training state	$+2.4$
multiple forwards	self	variance of three sampled predictions of one model	3 fwd
	deep	disagreement of three independently trained models	$3\times$
	MC drop	variance over twenty dropout masks on the transition head	20 fwd
	Laplace	variance over twenty weight samples of a Laplace posterior	20 fwd
	snap.	disagreement of four earlier snapshots of the model	4 fwd

E.4 Protocols

Maze. The agent trains on a ViZDoom map with one room sealed off. The *opened* map unseals it; the *repainted* map keeps every wall in place and changes the paint of one room. The repainted dead end of Sec. 4.6 keeps every wall and repaints the deep half of one dead end. A walk is a scripted route that reads the agent's position only, never the model, so that every readout sees the same frames. Learning to doubt leaves the task intact: every trained agent, with LucidWM and without, reaches the goal in all 100 test episodes.

Corrupted actions. From real start states the model imagines forward while each action is replaced, with a set probability, by a random one drawn uniformly from the action space. The lift of a readout is its value under fully corrupted actions over its value under the true ones.

Alarm line. Each readout is set against its own alarm line, taken on familiar frames of the same run; Figs. 13, 11, 15 and 14 show each readout in alarm units, 0 at its familiar level and 1 at its line.

F FURTHER RESULTS

F.1 THE ALARM FIRES IN THE NEW ROOM

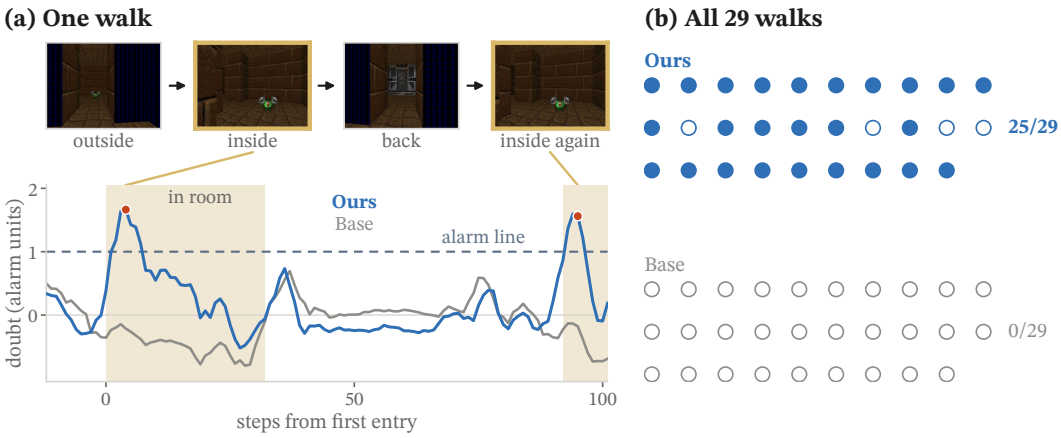

Figure 11: **Into the new room, and into it again.** (a) One walk: views outside, inside (gold), back and inside again, and the doubt of Ours and Base in alarm units; sand: in the room; red: alarms. (b) Walks on which each readout raises its alarm inside the room.

The room sealed in training shows familiar walls in a layout the agent has never experienced. The doubt raises its alarm inside it on 25 of 29 walks and the base's readout on none (Fig. 11); along one walk, it peaks each time the agent enters.

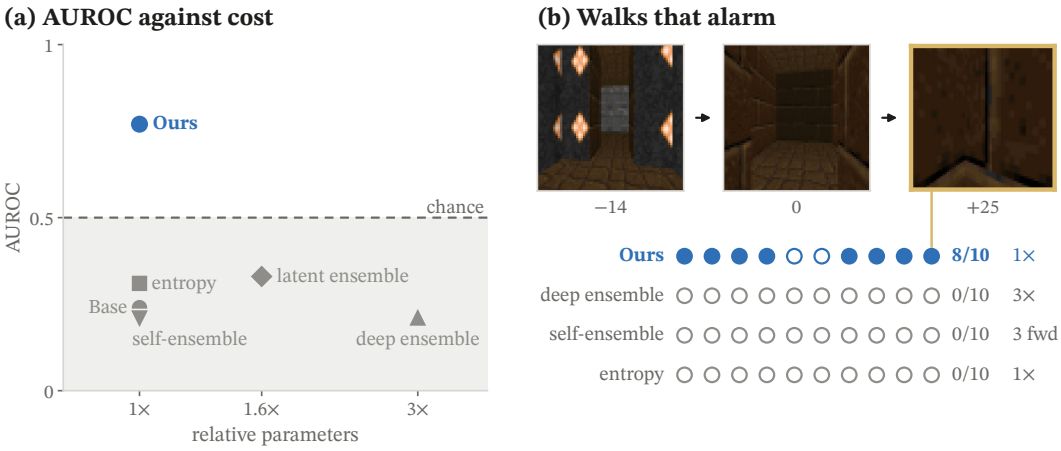

Figure 12: **One pass against three models.** (a) AUROC of the new room against relative parameters, as in Table 1(A); grey band: below chance, where a readout reads the new room as more familiar than the maze. (b) Walks through the opened door on which each readout raises its alarm, with its cost; top: walk 10 at -14, 0 and $+25$ steps from entry, gold where Ours raises its alarm, joined to its dot.

One pass does what three models cannot (Fig. 12): at the cost of one forward pass the doubt ranks the new room above the maze, while the base's latent and deep ensembles, at up to three times the cost, read it as *more* familiar; on ten further walks through the door, the doubt raises its alarm on eight, the deep and self-ensembles and the entropy on none.

F.2 THE ALARM FIRES BEFORE THE FALL

The doubt warns of each of twenty consecutive cheetah falls before it begins (Fig. 13): its alarm fires a median of four steps ahead, while the entropy of the same prediction raises none.

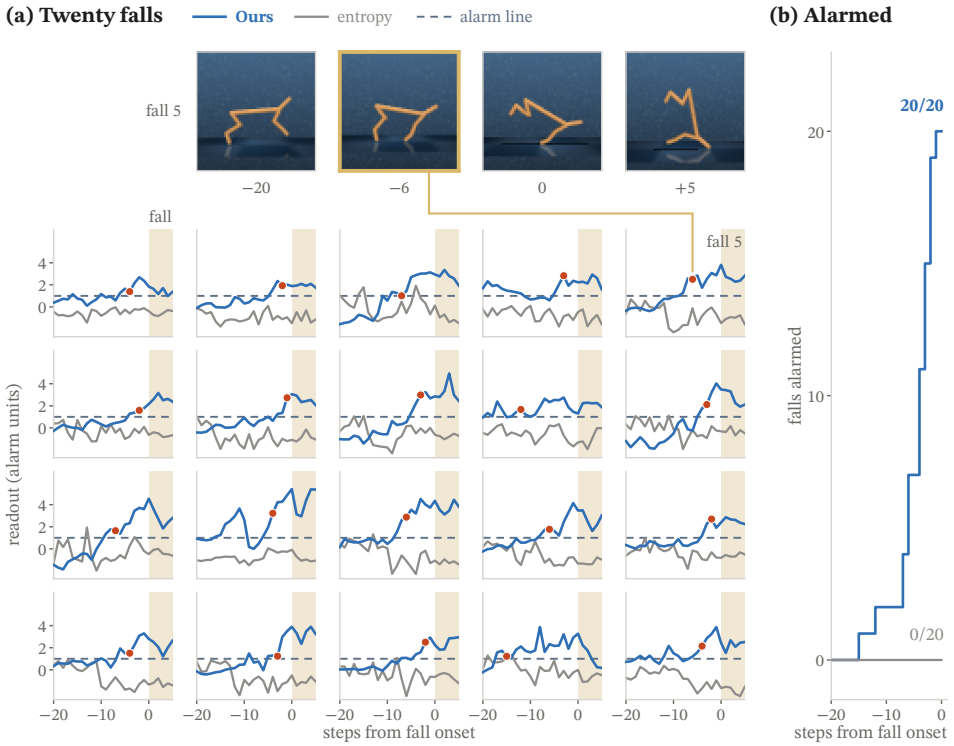

Figure 13: **Twenty falls, twenty alarms before the fall.** (a) Doubt (blue) and the entropy of the same prediction (grey) around the onset of each of twenty consecutive falls, in alarm units; dashed: alarm line; red: onset of the doubt's alarm that holds into the fall; sand: the fall. (b) Falls whose alarm has fired by each step. Top: fall 5, the fall of Fig. 1, at -20, -6, 0 and $+5$ steps; gold: step -6, where its alarm holds; the doubt first crosses the line at -8, which Fig. 1 counts as its lead.

F.3 THE ALARM FIRES AT NIGHTFALL

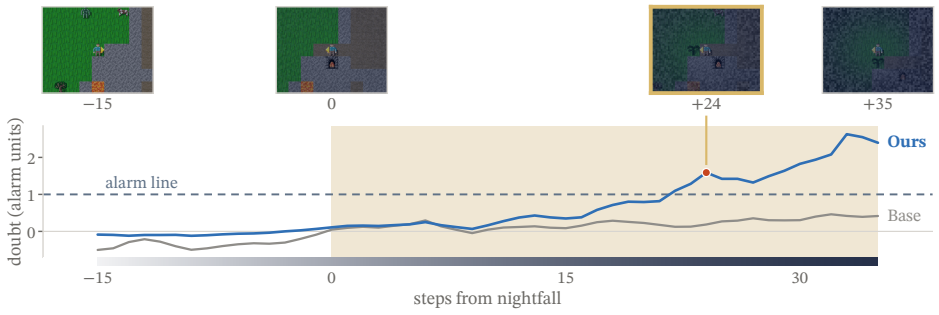

Figure 14: **Nightfall.** Views from fifteen steps before nightfall to thirty-five after (gold: at the alarm), above the doubt of Ours and Base in alarm units; sand: the night; strip: the fading light.

Nightfall darkens Crafter step by step. The doubt climbs over its line within twenty-five steps and peaks above twice it, while the base's readout never reaches half of its own (Fig. 14).

F.4 THE ALARM FIRES ON EVERY BASE

On all four bases the doubt crosses its line at the new room (Fig. 15).

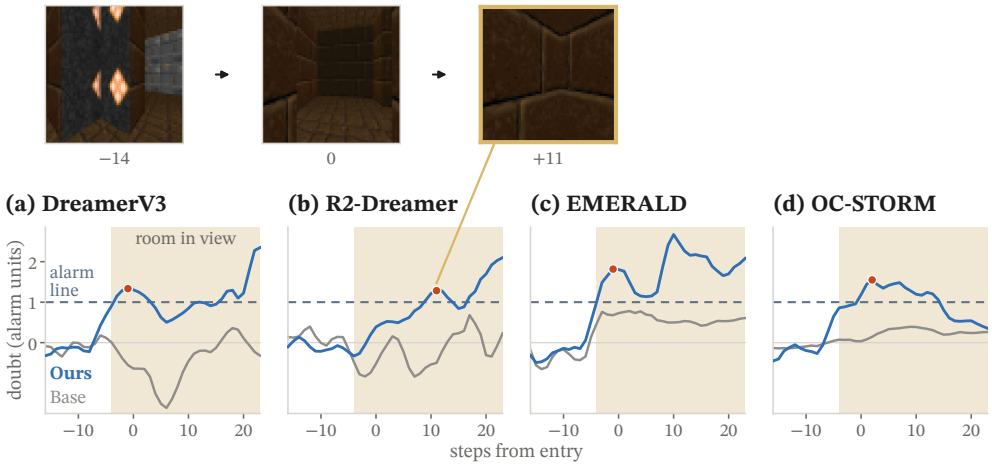

Figure 15: **Four bases, four alarms.** Doubt of Ours and Base on each base as the agent approaches the new room, in alarm units; sand: the room in view; red: the first alarm of Ours. Top: the route plotted in (b) and (d) at −14, 0 and +11 steps from entry, gold at the alarm of (b), joined to its dot; (a) and (c) pass the same door.

F.5 THE WORSE THE ACTIONS, THE HIGHER THE DOUBT

The more actions are corrupted, the higher the doubt climbs (Fig. 16). On OC-STORM its median over starts rises over fivefold on walker and over twofold on Crafter, while the base's readout barely moves. On EMERALD walker and cheetah the doubt rises over threefold from one pass, where MC dropout, Laplace and a snapshot ensemble, at up to twenty passes, fall instead.

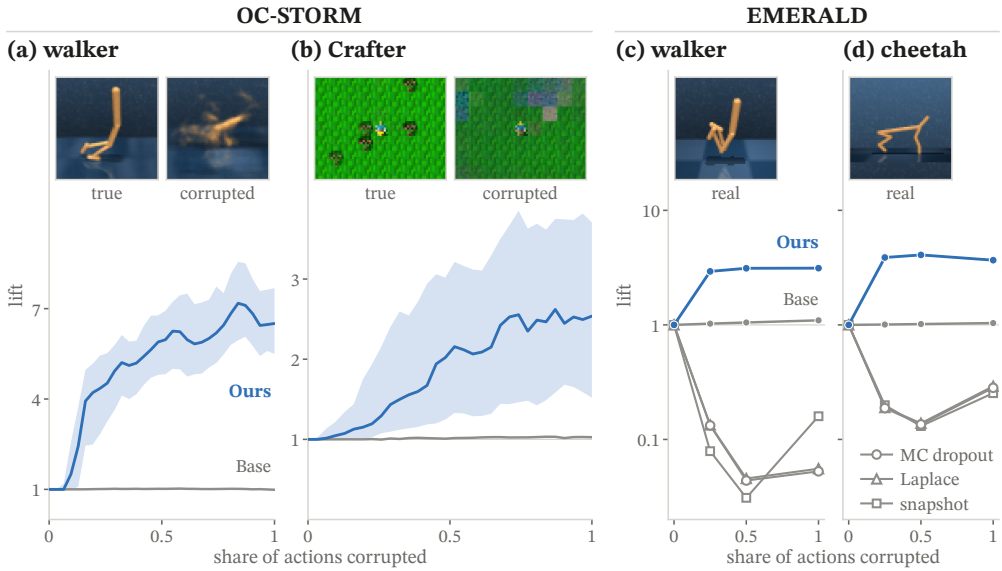

Figure 16: **The worse the actions, the higher the doubt.** Lift against the share of actions corrupted; thin line at one: no response. (a, b) OC-STORM, median over 300 starts (band: interquartile range of Ours); insets: imagination under true and under corrupted actions. (c, d) EMERALD against MC dropout, Laplace and a snapshot ensemble, on a log scale; insets: a real frame of each task.

F.6 THE ALARM FIRES AT ANY LINE, ANY CHECKPOINT

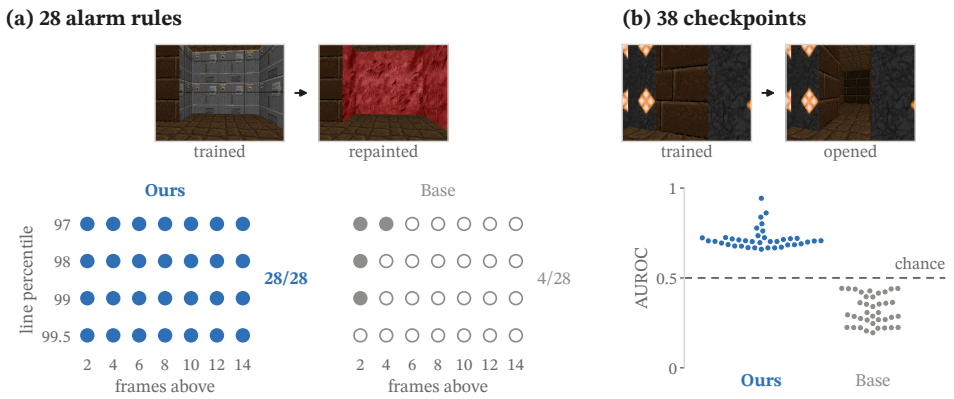

Figure 17: **Any alarm line, any checkpoint.** (a) Rules, placement of the line (rows) by frames required above it (columns), under which each readout raises its alarm in the repainted room. (b) AUROC of the opened room at each of 38 checkpoints of training; dashed: chance. Top: each change seen from one viewpoint, as trained and as changed.

The alarm does not hinge on where its line is drawn or on when the model is read (Fig. 17). In the repainted room the doubt raises its alarm under all 28 rules, four placements of the line by seven required durations, and the base's readout under only 4. Across 38 checkpoints of training, from 0.1M to 3.8M frames, the doubt ranks the frames of the opened room above the familiar ones at every checkpoint, and the base's readout ranks them below chance at every one.

F.7 THE DOUBT DRAINS AS THE MODEL LEARNS

At one step, and averaged over all 24 imagined steps, the doubt falls at every checkpoint of the run of Sec. 4.4 (Fig. 18): from 0.194 at 48k environment steps to 0.093 at 180k, and from 0.181 to 0.096. Over the same span the error falls 24.6-fold at one step and 3.9-fold on average, while the entropy of the same prediction, averaged likewise, rises from 0.79 to 1.01.

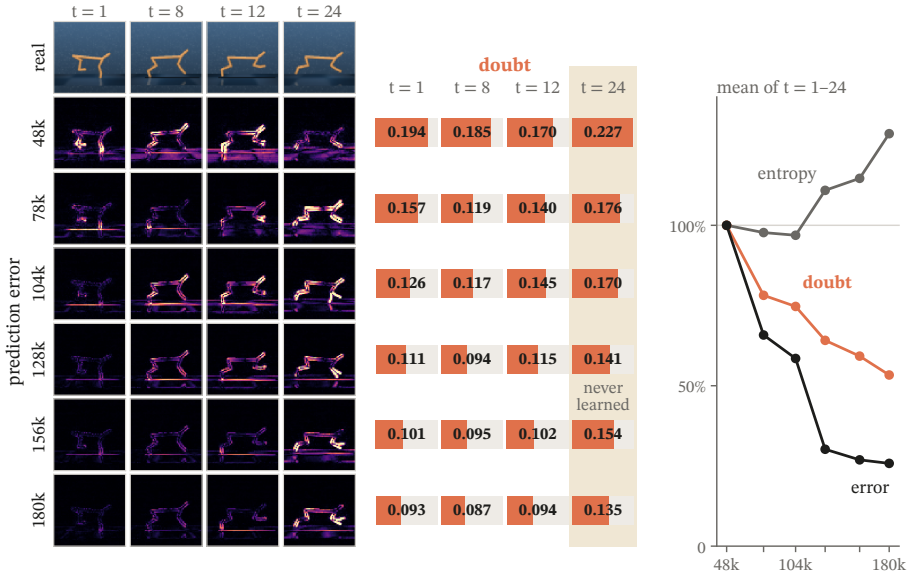

Figure 18: **Six checkpoints, one future.** As Fig. 6, at all six checkpoints of the run. Right: error, doubt and the entropy of the same prediction, each averaged over the 24 steps, relative to the first checkpoint.

F.8 TRUST FADES FASTER WHERE THE DREAM DRIFTS

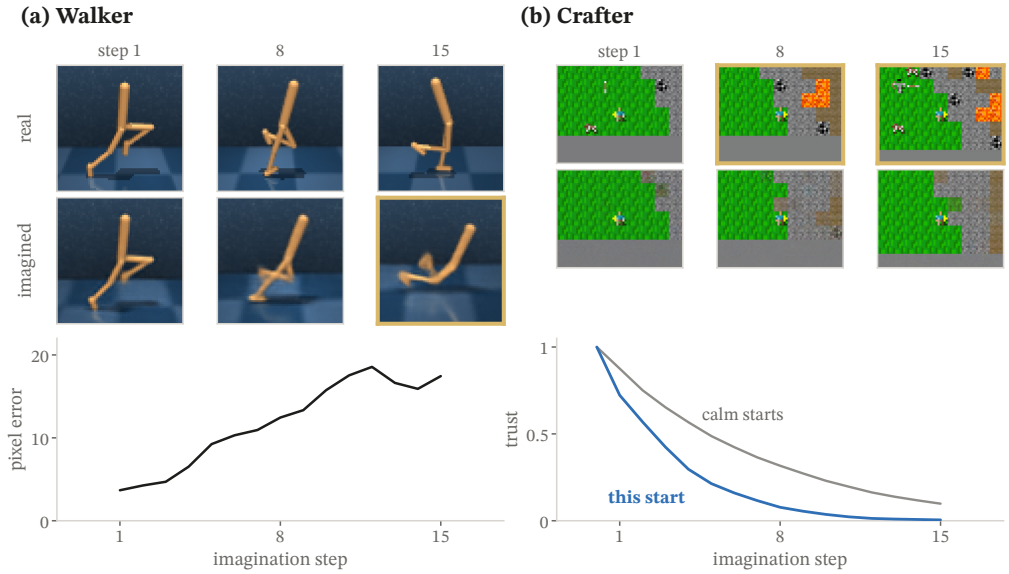

Figure 19: **Trust fades faster where the dream drifts.** Real (top) and imagined (bottom) futures under one action sequence. (a) Walker; gold: the imagined walker has fallen; the pixel error between the two futures. (b) Crafter; gold: lava in the real world; trust of this rollout against the median over thirty calm starts.

Fed one real action sequence, the world and the imagination part ways (Fig. 19). On walker the imagined walker falls while the real one walks on, and the pixel error between the two futures grows almost fivefold. On Crafter lava appears in the real world from the second step and never in the dream, and the trust of this rollout drains far faster than that of rollouts started in calm scenes: 0.16 against 0.42 at step 6.

F.9 TRUST KEEPS ITS EDGE ACROSS TEN RUNS

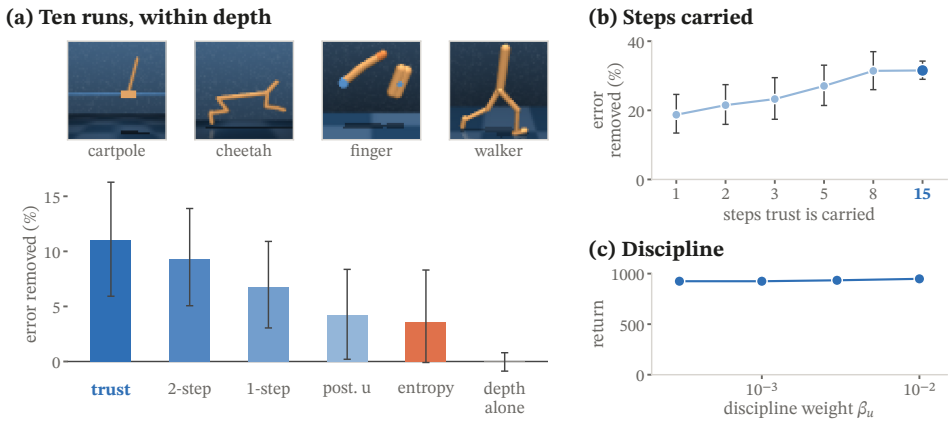

Figure 20: **Trust across ten runs, and two hyperparameters.** (a) Error removed by dropping the worst fifth of imagined steps, ranked within each depth as in the solid bars of Fig. 7, averaged over ten runs across walker, cheetah, finger and cartpole; top: the four tasks. (b) The same, ranked over all depths on cheetah from pixels as in Fig. 7b, with trust compounded over the last w steps; $w = 15$: the whole rollout, as in LucidWM. (c) Return on walker against the discipline weight β_u. Whiskers: bootstrap 95% intervals.

Within each depth, averaged over ten runs across walker, cheetah, finger and cartpole, trust removes 11.0% of the error, more than any other readout of Sec. 4.5 (Fig. 20a): trust compounded over two steps removes 9.3%, the doubt of one step 6.8%, the doubt of the posterior 4.2%, entropy 3.6% and depth alone 0.0%. Trust removes error within depth on each of the ten runs.

F.10 Hyperparameter analysis

LucidWM carries trust through the whole rollout, and no shorter window w does better (Fig. 20b): on cheetah from pixels, the task of Fig. 7b, the error removed grows with every longer window, from 18.7% at $w = 1$, the doubt of one step, to 31.5% at all 15 steps. The weight β_u of the evidence discipline leaves the return in place (Fig. 20c): on walker, from 3×10^{-4} to 10^{-2}, it stays between 924 and 949.

F.11 Choosing futures by trust cuts deaths

Used beyond the veto to choose among whole futures, trust keeps the agent alive longer (Fig. 21a). At every step in Crafter, LucidWM on DreamerV3 imagines eight futures of 15 steps with its own actor and takes the first action of the future that scores highest. With trust, a future is scored by a variant of J (Eq. (10)) that keeps λ, weighting step k by $\lambda^k C_k$, the share of the lucid return that reaches it; without trust, by λ^k alone. Both scores see the same futures, and trust changes the choice at 43% of steps. Over 50 seeds of 1,000 steps, trust cuts deaths from 4.08 to 3.80 per 1,000 steps and lengthens the median life from 236 to 253 steps (paired Wilcoxon, $p = 0.041$ and 0.0016).

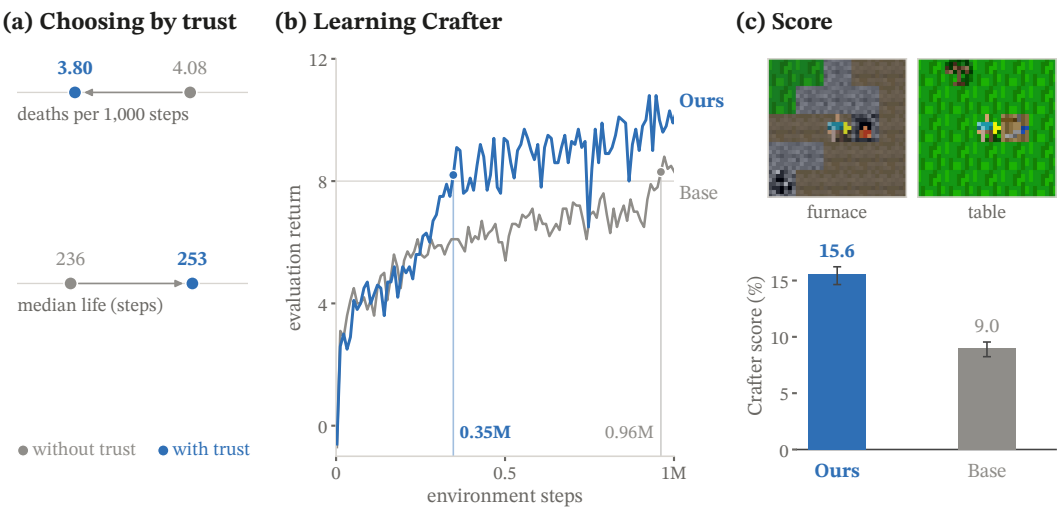

Figure 21: **Fewer deaths, faster learning.** (a) One agent choosing among imagined futures scored with and without trust; median life: per seed, averaged over 50 seeds. (b) Evaluation return over training; dots and lines: where each first reaches 8. (c) The official Crafter score of the 22 achievements over the last 100 evaluation episodes; whiskers: bootstrap 95% intervals over episodes; tiles: from the scored episodes, Ours placing a furnace and Base a table.

F.12 LucidWM learns Crafter faster than its base

The agent of App. F.11 learns Crafter faster than its base and scores higher (Fig. 21b, c). After 0.35M steps it reaches the evaluation return of 8 that its base first reaches after 0.96M, and over the last 100 evaluation episodes of training its Crafter score is 15.6% against 9.0%, with 11.0 achievements unlocked per episode against 8.9.

F.13 VETOES FIRE WHERE THE WORLD CHANGED

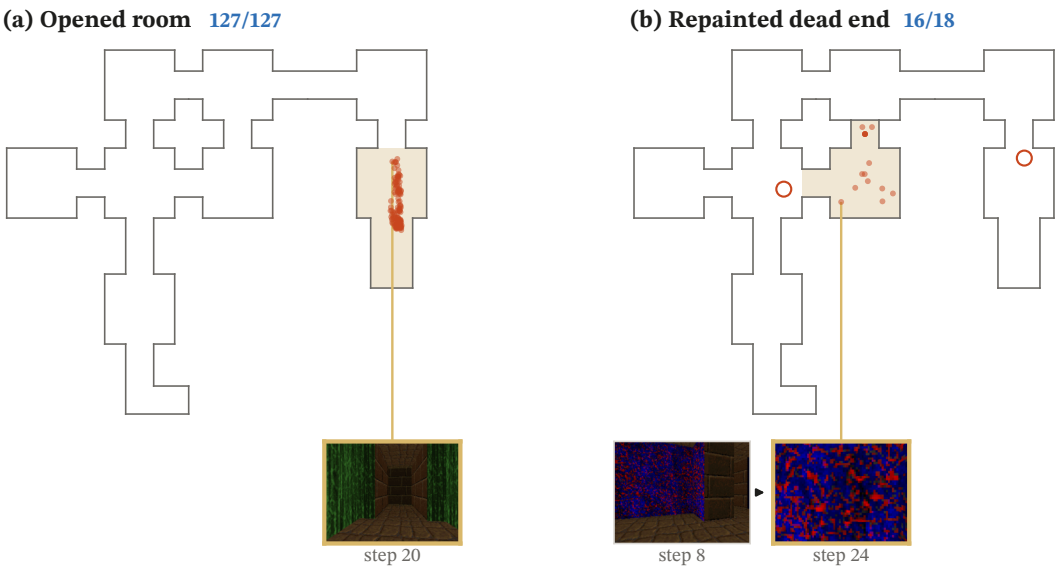

Figure 22: **Vetoes fire where the world changed.** Where every veto fired on (a) the opened map and (b) the repainted dead end; sand: the changed rooms; rings: vetoes outside them. Insets: the agent's view at a veto, gold and joined to its dot; in (b), also the repainted room at step 8. The veto at step 24 in (b) is that of Fig. 8.

The veto of Sec. 4.6 fires where the world has changed (Fig. 22). On the two altered maps, 143 of its 145 vetoes fall inside a changed room: all 127 in the opened room and 16 of 18 in the repainted dead end.