ADAPTIVE PRECISION ALLOCATION CAN STARVE ITSELF: CONSEQUENCE-AWARE ALLOCATION FOR BOUNDED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A recurrent world model operating under a hard per-transition representation budget must decide *where* to spend its available precision. We show that adaptive precision allocation can *starve itself*: when source-referred importance is repeatedly measured through the precision assignment it is used to update, low precision attenuates the very signal needed for a group to recover rate. We formalize this reference-point failure through *multi-step predictive consequence*—how strongly current representation error propagates into future latent predictions under the same action continuation—and show that the resulting feedback admits locally absorbing starvation fixed points. Measuring all groups from a common allocation-independent reference removes the assignment-dependent gain. Building on this analysis, we introduce **CAP-WM**, which predicts state-dependent multi-step consequence and converts it into an exact recurrent precision allocation by reverse water filling at a no-bypass interface. A linear–Gaussian characterization recovers the same groupwise allocation rule under a standardized group-block-isotropic specialization. In trained models, adaptive source-referred refresh reaches ≤ 4 active groups from ordinary initializations with probability 0.74–0.91, versus 0.06–0.11 for memory-referred refresh. Under a directly countable serialized interface, CAP-WM reaches 90% retention of the no-channel DreamerV3 reference at 4 bits/coordinate, while the tested adaptive controls that reach this criterion require 6; at the tight nominal rate, CAP-WM improves DMC-GB2 IQM by $+88\,[55, 119]$ over Uniform and $+24\,[8, 40]$ over One-step consequence. The additional multi-step benefit is conditional rather than universal, resolving when precision is restrictive and predictive dependence extends beyond one latent transition. More broadly, the results expose a design hazard for adaptive resource allocation: an assignment can become self-reinforcing when it attenuates the importance signal used to revise itself.

1 INTRODUCTION

World models compress interaction histories into latent states that support prediction, imagination, and control (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025). Recent work has improved *what* these states represent through stronger dynamics, alternative sequence models, pretrained visual features, and redundancy reduction (Hansen et al., 2024; Zhang et al., 2023; Alonso et al., 2024; Zhou et al., 2025; Morihira et al., 2026). We study a complementary resource-allocation question: *when recurrent state transmission is constrained by a hard per-transition precision budget, where should the available precision be spent?*

A natural strategy is to protect components that are surprising or locally sensitive. Neither criterion, however, directly captures how current representation error propagates through future predictions. Nuisance variation can be difficult to predict yet dynamically inconsequential, while a predictable component may deserve high precision because a small current error changes many future latent states. We therefore allocate precision according to *multi-step predictive consequence*: how strongly corrupting one current state group changes future latent predictions under the same action continuation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

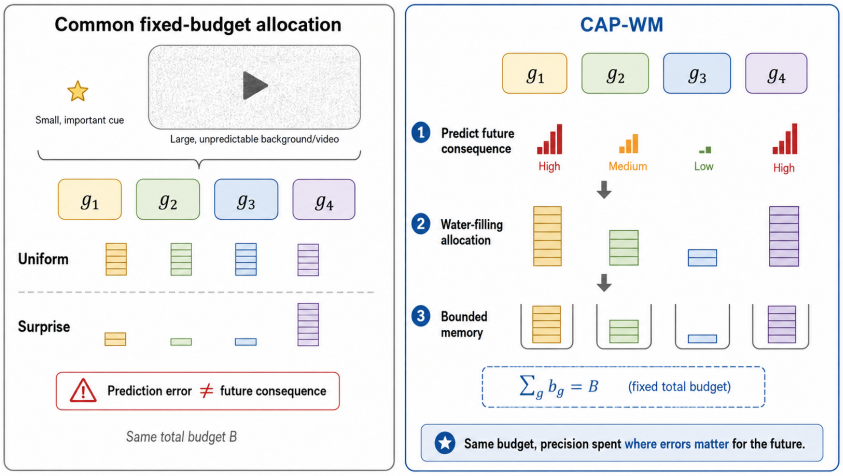


Figure 1: **Which information deserves scarce recurrent precision?** Uniform allocation ignores state-dependent consequence, while Surprise can prioritize unpredictable but dynamically inconsequential variation. CAP-WM allocates the same hard recurrent budget according to how strongly current representation error propagates into future latent predictions.

A reference-point failure. The allocation problem contains a less obvious difficulty: importance must itself be measured at an operating point. In our recurrent channel, source-referred sensitivity measured through the precision assignment being updated takes the form

$$S_g(b) = \kappa(b_g)^2 c_g, \qquad \kappa(b) = 1 - e^{-2b/d}, \tag{1}$$

where c_g denotes intrinsic predictive consequence. Reduced precision can therefore attenuate the same importance signal required for a group to recover precision. For a fixed intrinsic profile, we show that the resulting adaptive measurement–allocation map admits locally absorbing fixed points that can starve arbitrary groups, including the maximally consequential one (Proposition 1). Measuring all groups from a common allocation-independent reference removes this assignment-dependent gain (Proposition 2). The obstruction is therefore not adaptive allocation itself, but *assignment-attenuated refreshed measurement*: a broader design hazard whenever the current resource assignment attenuates the importance signal used to revise that assignment.

CAP-WM. We introduce **CAP-WM**, *Consequence-Aware Precision Allocation for World Models*. CAP-WM separates consequence measurement from the allocation it subsequently controls. A replay-derived teacher measures multi-step predictive consequence from a common reference; a lightweight state-conditioned predictor estimates this quantity online; and reverse water filling converts the predicted scores into a hard per-transition allocation at a no-bypass recurrent interface. The key contribution is not reverse water filling or local sensitivity in isolation, but their use with an allocation-independent *multi-step predictive* criterion together with the identification and removal of the self-attenuating reference failure above.

The resulting rule also has a tractable normative characterization. For a linear–Gaussian system, predictive rate–distortion allocates rate along the eigenmodes of $\Sigma^{1/2} G_{\mathcal{H}} \Sigma^{1/2}$. Under the standardized group-block-isotropic specialization of Corollary 2, the unrestricted predictive optimum reduces exactly to CAP-WM's groupwise reverse-water-filling rule (Proposition 4). We use this correspondence to characterize the groupwise allocator, not as a global optimality claim for the nonlinear world model.

Evidence and boundary. We evaluate the resource claim, the multi-step contribution, and the reference-point mechanism separately. Under a directly countable serialized recurrent-feature interface, CAP-WM is the only tested allocator to reach 90% retention of the no-channel DreamerV3 reference at 4 bits per coordinate; Surprise, a parameter-matched End-to-end gate, One-step consequence, and Multi-step value sensitivity first reach the criterion at 6 bits, while Uniform does not

reach it on the evaluated grid. Accounting explicitly for the allocation map, the corresponding retention crossing uses 33.1% less effective recurrent payload. Under the nominal channel at the tight rate $\bar{B} = 0.20$, CAP-WM improves DMC-GB2 IQM by $+88\,[55, 119]$ over Uniform and $+24\,[8, 40]$ over One-step consequence.

The proposed mechanism is independently measurable. Replacing the common reference by adaptive source-referred refresh contracts the mean active set from $8.1\,[7.6, 8.6]$ to $3.4\,[3.0, 3.9]$ groups, while finite interventions validate both score–damage and allocation–damage associations. A ResNet-50 mixed-precision experiment exhibits the same structural feedback under a distinct attenuation mechanism, supporting the scope of assignment-dependent refreshed measurement beyond the recurrent Gaussian channel.

The benefit specific to multi-step consequence is conditional rather than universal. Prospective action-repeat interventions vary the amount of predictive sensitivity lying beyond one latent transition, and a joint resource–horizon experiment yields a resolved interaction of $+31\,[8, 54]$: the additional CAP-WM–One-step benefit is largest when precision is restrictive and predictive dependence extends beyond one latent transition, but is unresolved when either condition weakens. Rate sweeps, the prospective intervention, and transfer across three world-model backbones consistently support this boundary.

Contributions.

- **Self-reinforcing measurement failure in adaptive allocation.** We identify assignment-attenuated refreshed sensitivity as a source of precision starvation, characterize locally absorbing fixed points that can suppress even highly consequential groups, and connect the theoretical obstruction to reachable allocation dynamics in trained models.
- **Consequence-aware exact-budget allocation.** CAP-WM measures multi-step predictive consequence from an allocation-independent reference, predicts it online, and converts it into hard per-transition allocations at a no-bypass state interface; a linear–Gaussian specialization recovers the same groupwise reverse-water-filling form.
- **Mechanism- and boundary-focused validation.** Exact serialized-payload controls, finite interventions, active-set diagnostics, prospective horizon manipulation, and three-backbone transfer establish both the operational resource benefit and the regime in which multi-step consequence adds beyond one-step sensitivity.

2 Related Work

World models and bounded representation. World models learn latent states for prediction, imagination, and control (Ha & Schmidhuber, 2018; Hafner et al., 2020; Schrittwieser et al., 2020), with recent work improving dynamics, planning, generative modeling, and representation quality (Hafner et al., 2025; Hansen et al., 2024; Micheli et al., 2023; Zhang et al., 2023; Alonso et al., 2024; Zhou et al., 2025; Morihira et al., 2026; Zhang et al., 2025; Peng et al., 2026). These methods primarily improve *what* the latent state represents or *how* it is learned. CAP-WM instead asks how a fixed per-transition recurrent resource should be distributed across the current state. This connects to rate-limited control, rate–distortion, information bottlenecks, and learned compression (Tatikonda & Mitter, 2004; Nair & Evans, 2004; Cover & Thomas, 2006; Still, 2014; Alemi et al., 2017; Ballé et al., 2018; Minnen et al., 2018; Choi et al., 2019; Cui et al., 2021), but differs in using future predictive effect as the allocation criterion under a hard no-bypass recurrent budget.

Sensitivity-aware precision allocation. Mixed-precision methods allocate unequal precision using local sensitivity, including second-order, Hessian-trace, and Fisher-style criteria (Dong et al., 2019; 2020; Feng et al., 2025). Recent rate–distortion allocation methods similarly use calibrated component-wise distortion models and reverse water filling for mixed-precision KV-cache quantization (Zuo et al., 2026). CAP-WM differs in allocating a recurrent state at every transition according to multi-step predictive consequence, while additionally analyzing how the allocation can distort the importance signal used to update itself. Damage-guided resource allocation has also been explored in efficient generation: MOSAIK predicts the fidelity loss induced by regional coarsening and assigns heterogeneous patch sizes under a token budget (Hami et al., 2026). CAP-WM instead allo-

cates recurrent precision under a hard per-transition budget and defines damage through propagation into future latent predictions.

With $\mathcal{H} = \{1\}$, CAP-WM reduces to one-step predictive Jacobian sensitivity, which we include as the *One-step consequence* control. Its distinction is twofold: it measures how current representation error propagates through multiple future latent predictions, and it measures this consequence from a common allocation-independent reference. The latter removes the assignment-dependent attenuation that can make refreshed source-referred importance self-reinforcingly starve a group. Our claim is therefore about *assignment-attenuated refreshed measurement*, not adaptive precision allocation in general.

Predictive and resource-rational coding. Efficient-coding and resource-rational perspectives study how limited representational resources should be distributed (Attneave, 1954; Barlow, 1961; Sims, 2003; Griffiths et al., 2015), while predictive-coding frameworks emphasize prediction error (Rao & Ballard, 1999; Friston, 2005). CAP-WM instead allocates according to *predictive consequence*: a component can be difficult to predict yet cheap to corrupt, or predictable yet important because its error propagates strongly into future latent states. The contribution is an operational hard-budget allocation rule together with a reference-point analysis of when the importance measurement itself can be distorted by the assignment it updates.

3 Consequence-Aware Precision Allocation

CAP-WM treats recurrent state transmission as a hard allocation problem: every transition receives the same total resource, but that resource may be distributed differently across the current state. The construction separates four questions that are often conflated—*what is bounded*, *what information is consequential*, *where consequence should be measured*, and *how a fixed budget should be allocated*. Its key distinction is to allocate by multi-step predictive consequence measured at a common allocation-independent reference, avoiding assignment-attenuated self-starvation while retaining an exact reverse-water-filling solution. We then relate this fixed-group rule to predictive Gaussian rate–distortion under explicit structural conditions.

3.1 Bounded Recurrent Interface

DreamerV3 uses recurrent state $s_t = (h_t, z_t)$. After the backbone's native categorical sampling, we expose the parameter-free numerical state

$$u_t = \phi(s_t) = [h_t; \mathrm{vec}(z_t)] \in \mathbb{R}^{1536}, \tag{2}$$

partitioned into $G = 16$ equal-width groups of width $d = 96$. After standardization using running moments estimated from training replay, group g passes through

$$D_{t,g} = e^{-2b_{t,g}/d}, \qquad m_{t,g} = (1 - D_{t,g})\bar{u}_{t,g} + \sqrt{D_{t,g}(1 - D_{t,g})}\,\varepsilon_{t,g}, \quad \varepsilon_{t,g} \sim \mathcal{N}(0, I_d). \tag{3}$$

Definition 1 (Bounded recurrent interface)**.** *At every transition,*

$$b_{t,g} \geq 0, \qquad \sum_{g=1}^{G} b_{t,g} = B, \tag{4}$$

and all subsequent recurrent dynamics and actor–critic computation consume only $m_t = \mathrm{concat}_g(m_{t,g})$*. The untransmitted* u_t*, predicted consequence* $\widehat{C}_t$*, and allocation* b_t *are not exposed as additional recurrent features.*

The allocator may inspect the current candidate state to configure the channel, but its only downstream effect is through the transmitted state m_t. The same no-bypass interface is used for posterior candidates during interaction/replay and prior candidates during imagination; in particular, $b_{t,g} = 0$ implies $m_{t,g} \equiv 0$. For an ideal standardized Gaussian source, $b_{t,g}$ equals mutual information in nats. For learned features, $\bar{B} = B/(Gd)$ is only a controlled *nominal Gaussian rate*, not empirical mutual information or hardware bitrate. Appendix A.1 gives the exact interface, standardization, channel identities, and categorical-state convention.

3.2 MULTI-STEP PREDICTIVE CONSEQUENCE

Scarce precision should protect errors that propagate, not simply components that are surprising at the current step. Let $F_{\theta,h}(m, a_{t:t+h-1})$ denote the differentiable h-step prior-prediction map from bounded memory m under a fixed replay action continuation. With $\mathcal{H} = \{1, 2, 4, 8\}$ and $\gamma_h = 1/4$, define the intrinsic, memory-referred consequence

$$c_{t,g}(m) = \sum_{h\in\mathcal{H}} \gamma_h \left\| \frac{\partial F_{\theta,h}(m, a_{t:t+h-1})}{\partial m_g} \right\|_F^2 . \tag{5}$$

The teacher differentiates the same surrogate computation graph used by the executed backbone for categorical rollouts, conditional on the realized stochastic trace, while treating the replay action sequence as fixed. Appendix A.3 makes this convention explicit.

CAP-WM does not measure this quantity through the deployed adaptive allocation. Instead it uses a common reference channel $b_g^{\mathrm{ref}} = B/G$. The corresponding source-referred pathwise teacher is

$$C_{t,g}^{\mathrm{path}} = \sum_{h\in\mathcal{H}} \gamma_h \left\| \frac{\partial F_{\theta,h}(m_t^{\mathrm{ref}}, a_{t:t+h-1})}{\partial \bar{u}_{t,g}} \right\|_F^2 = \kappa(b_g^{\mathrm{ref}})^2 c_{t,g}(m_t^{\mathrm{ref}}), \qquad \kappa(\beta) = 1 - e^{-2\beta/d}. \tag{6}$$

Thus, at the same operating point and linearization, source- and memory-referred consequence differ only by the channel gain. A group-dependent reference therefore changes the relative scores presented to the allocator even if the intrinsic consequence profile is unchanged.

The score also has a local distortion interpretation. Consider an isotropic perturbation family $\delta_{t,g} = \sqrt{D_{t,g}}\,\xi_{t,g}$ with $\mathbb{E}[\xi_{t,g}] = 0$, $\mathbb{E}[\xi_{t,g}\xi_{t,g}^\top] = I_d$, and finite fourth moment. Its covariance matches the channel-error covariance for the ideal standardized Gaussian source and serves as a local second-moment surrogate for learned features. Let $\Delta_{t,g}^{\mathrm{pred}}$ denote the resulting weighted squared multi-step predictive deviation.

Lemma 1 (Local predictive damage)**.** *At the reference memory, as $D_{t,g} \downarrow 0$,*

$$\mathbb{E}[\Delta_{t,g}^{\mathrm{pred}}] = c_{t,g}(m_t^{\mathrm{ref}}) D_{t,g} + o(D_{t,g}). \tag{7}$$

Hence $c_{t,g}$ is the leading-order local cost of corrupting group g, while $D_{t,g} = e^{-2b_{t,g}/d}$ is the corresponding nominal distortion scale. With $\mathcal{H} = \{1\}$, the construction reduces to one-step predictive Jacobian sensitivity, yielding the primary One-step consequence control. Appendix A.4 gives the perturbation definition and proof; Appendix A.3 gives replay conditioning and stochastic Jacobian estimation.

3.3 A COMMON REFERENCE PREVENTS SELF-REINFORCING STARVATION

Lemma 1 identifies intrinsic consequence, but a source-referred measurement taken through the allocation being updated returns $\kappa(b_g)^2 c_g$. The measurement therefore mixes consequence with the current fidelity assignment. To isolate this feedback, fix $c_{1:G} > 0$ and define

$$S_g(b) = \kappa(b_g)^2 c_g, \qquad T(b) = \mathrm{WF}\big(\kappa(b)^2 \odot c; B\big)\,. \tag{8}$$

Because $\kappa(0) = 0$, reducing a group's precision suppresses the same importance signal required for that group to recover precision.

Proposition 1 (Starvation under an adaptive source-referred reference)**.** *Assume $B > 0$ and $c_{1:G} > 0$ is independent of the operating allocation. For every nonempty proper subset $\mathcal{S} \subsetneq [G]$, T has a fixed point assigning zero rate to all $g \in \mathcal{S}$, including a fixed point that starves a maximally consequential group. Moreover, starvation is locally absorbing: there exists $\beta > 0$ such that, for any group g, $b_g \le \beta$ implies $T(b)_g = 0$.*

This proposition is an existence and local-absorption statement, not a global-convergence theorem. The fixed-profile assumption is exact in the linear–Gaussian model of Section 3.5; in the nonlinear model, $c_g(m(b))$ can also move with the reference. Appendix A.2 proves the result, while Appendix A.2.1 separates this nonlinear re-linearization pathway from the assignment-dependent source gain.

CAP-WM removes the assignment-dependent gain by measuring all groups at the same allocation-independent operating point, $b_g^{\mathrm{ref}} = B/G$.

Proposition 2 (Common-reference invariance). *For $B > 0$, under the common reference,*

$$C_g^{\mathrm{path}} = \kappa(B/G)^2 c_g(m^{\mathrm{ref}}), \qquad \mathrm{WF}(C^{\mathrm{path}}; B) = \mathrm{WF}(c(m^{\mathrm{ref}}); B). \tag{9}$$

The common positive gain is absorbed into the water level, so source- and memory-referred scores induce the same allocation at this operating point. This removes the *assignment-induced* starvation mechanism above; ordinary water filling may still assign zero rate to intrinsically low-consequence groups. Under a non-uniform reference, source- and memory-referred scores need not agree. Section 5.3 tests whether the theoretically isolated mechanism is reachable in trained models.

3.4 Consequence-Weighted Exact-Budget Allocation

The pathwise teacher uses future replay actions and is training-only. We distill it into a state-conditioned predictor

$$\widehat{C}_{t,1:G} = \mathrm{softplus}(g_\omega(\mathrm{sg}[\bar{u}_t])) + \epsilon. \tag{10}$$

At deployment, $\widehat{C}_t$ depends only on the current candidate state. The predictor output is used only to configure the channel; neither $\widehat{C}_t$ nor b_t is concatenated to the transmitted memory or provided separately to the dynamics, actor, or critic. Because the common reference rescales intrinsic consequence by the same positive factor for every group, replacing c_t by its predicted pathwise counterpart preserves the allocation semantics. Combining this equivalence with the local damage surrogate gives

$$\min_{b_{t,1:G}} \sum_{g=1}^{G} \widehat{C}_{t,g} e^{-2b_{t,g}/d} \quad \text{s.t.} \quad b_{t,g} \geq 0,\ \sum_g b_{t,g} = B. \tag{11}$$

Proposition 3 (Consequence-weighted reverse water filling). *For $B > 0$, because $\widehat{C}_{t,g} > 0$, Equation 11 is strictly convex and has the unique solution*

$$b^\star_{t,g} = \frac{d}{2}\left[\log \frac{\widehat{C}_{t,g}}{\lambda_t}\right]_+, \qquad \sum_g b^\star_{t,g} = B, \tag{12}$$

with unique water level $\lambda_t > 0$.

Reverse water filling and local Jacobian sensitivity are classical ingredients. CAP-WM's contribution is the allocation criterion and reference construction: multi-step predictive consequence is measured at a common allocation-independent operating point to avoid self-attenuation, distilled into a deployable state-conditioned predictor, and enforced through a hard no-bypass recurrent interface. Predictor supervision, stop-gradient separation, the KKT derivation, bisection cost, and the complete recurrent algorithm are in Appendices A.3–A.4.

3.5 Linear–Gaussian Characterization

A linear–Gaussian system provides both an exact fixed-profile setting for Proposition 1 and a normative predictive rate–distortion comparison. Consider

$$x_{t+1} = Ax_t + \mathsf{U}a_t + w_t, \qquad y_t = Cx_t. \tag{13}$$

For two trajectories with identical open-loop actions and process noise that differ only by current error e_t,

$$\sum_{h\in\mathcal{H}} \gamma_h \|CA^h e_t\|_2^2 = e_t^\top G_{\mathcal{H}} e_t, \qquad G_{\mathcal{H}} = \sum_{h\in\mathcal{H}} \gamma_h (A^h)^\top C^\top C A^h, \tag{14}$$

and the consequence of coordinate group g is

$$c_g = \mathrm{tr}\big(P_g^\top G_{\mathcal{H}} P_g\big), \tag{15}$$

which is independent of the rate allocation.

Proposition 4 (Gaussian predictive rate allocation). *Let $x_t \sim \mathcal{N}(0, \Sigma)$, let $B > 0$, and let $\eta_1, \ldots, \eta_r > 0$ be the nonzero eigenvalues of $\Sigma^{1/2} G_{\mathcal{H}} \Sigma^{1/2}$. Under total Gaussian rate B, the minimum expected predictive distortion is attained by*

$$R_i^\star = \frac{1}{2}\left[\log \frac{\eta_i}{\lambda}\right]_+, \qquad \sum_{i=1}^{r} R_i^\star = B. \tag{16}$$

The unrestricted Gaussian optimum depends jointly on source uncertainty and predictive geometry; CAP-WM is therefore not the unrestricted transform optimum. Under the standardized group-block-isotropic specialization of Corollary 2—$\Sigma = I$, zero cross-group predictive coupling, and within-group isotropy—the d modes in group g share eigenvalue c_g/d, and Proposition 4 reduces exactly to Equation 12. These are sufficient structural conditions, not assumptions imposed on the neural model. Appendix A.5 gives the proof, and Appendix A.5.1 measures how far the learned local geometry departs from this specialization.

Scope. The theory is normative rather than a global optimality result for the nonlinear world model. CAP-WM uses a local predictive linearization, a nominal Gaussian distortion surrogate, fixed coordinate groups, and open-loop predictive consequence. Nonlinear reference drift, replay–policy mismatch, grouping restrictions, measurement conventions, and the requirements for transferring the bounded-interface principle beyond the primary RSSM setting are detailed in Appendix A.6.

4 Experiments

We evaluate three claims: **(Q1)** whether consequence-aware allocation improves performance under the same hard recurrent resource; **(Q2)** whether *multi-step predictive consequence* adds value beyond one-step sensitivity and alternative adaptive allocation signals; and **(Q3)** whether the learned score and reference-point analysis correspond to independently measurable damage and allocation dynamics rather than only return. Sections 5.1–5.3 address these questions in turn; Section 5.4 examines transfer and boundary conditions.

Protocol and controls. We evaluate DreamerV3 (Hafner et al., 2025) on 20-task DMC (Tassa et al., 2018), DMC-Subtle (Morihira et al., 2026), six DMC-GB2 tasks with held-out visual shifts (Almuzairee et al., 2024), and all 26 Atari 100K games (Kaiser et al., 2020). Unless stated otherwise, DMC-style experiments use five training seeds and 10 evaluation episodes per seed. Matched variants share the same 1536-dimensional no-bypass recurrent interface, $G = 16$ groups, interaction budget, actor–critic objective, channel, and hard per-transition resource. Uniform tests state-dependent allocation; One-step consequence changes only $\mathcal{H} = \{1, 2, 4, 8\}$ to $\{1\}$ and is the primary multi-step control. Surprise, Static task sensitivity, a parameter-matched End-to-end gate, Multi-step value sensitivity, and VIB-WM provide additional controls. We evaluate both the nominal Gaussian budget $\bar{B} = B/1536$ and a separately trained exact serialized recurrent-feature interface carrying 2, 4, or 6 bits/coordinate. We report stratified-bootstrap 95% intervals and paired differences over matched task/game–seed units (Agarwal et al., 2021); the central Subtle/Video CAP-WM–One-step comparison is independently repeated with 10 seeds. Full protocols, estimands, resource accounting, and run-level outcomes are in Appendices B.1 and B.16; the additional architecture-transfer and joint resource–horizon protocols are in Appendix C.

5 Results

5.1 CAP-WM Preserves More Performance per Transmitted Bit

We first evaluate a directly countable recurrent resource. Relative to the no-channel DreamerV3 reference $S_{\mathrm{DV3}} = 820$, let K_{90} denote the smallest tested payload retaining at least 90% of that score. Table 1 shows that CAP-WM reaches this threshold at 4 bits/coordinate, whereas Surprise, End-to-end gate, One-step consequence, and Multi-step value sensitivity require 6, and Uniform does not reach it on the tested grid. At the tightest payload, CAP-WM reaches IQM 619.

The separation is strongest when precision is scarce. CAP-WM–One-step is $+58\,[28, 87]$ and $+40\,[13, 66]$ at 2 and 4 bits/coordinate, but $-2\,[-22, 18]$ at 6 bits. The nominal channel shows the same boundary: at $\bar{B} = 0.20$, CAP-WM improves Uniform by $+88\,[55, 119]$ and One-step by $+24\,[8, 40]$, whereas CAP-WM–One-step is unresolved at $\bar{B} = 0.80$ and 1.60. Including allocation-map metadata, the K_{90} crossings become 4.0306 versus 6.0254 effective bits/coordinate, corresponding to a 33.1% reduction in effective recurrent payload. CAP-WM adds 2.7% parameters and 6.2% action-selection latency (Appendix B.2). Appendix B.8 evaluates a pre-branch rule based on resource pressure.

Method	2 bits	4 bits	6 bits	K_{90}
Uniform	493 [468, 517]	648 [623, 672]	735 [713, 756]	> 6
Surprise	522 [497, 546]	671 [646, 695]	749 [727, 770]	6
End-to-end gate	548 [523, 572]	690 [666, 713]	776 [754, 797]	6
One-step consequence	561 [536, 585]	699 [675, 722]	785 [764, 805]	6
Multi-step value	586 [561, 610]	721 [697, 744]	**786** [765, 806]	6
CAP-WM	**619** [594, 643]	**739** [715, 762]	783 [762, 802]	**4**

Table 1: **Exact serialized recurrent-feature payload on DMC-GB2.** IQM with stratified-bootstrap 95% intervals. K_{90} is the smallest tested feature payload retaining at least 90% of the no-channel DreamerV3 reference; > 6 means the criterion is not reached on the tested grid.

5.2 Multi-Step Consequence Helps When Predictive Dependence Extends Beyond One Step

We next isolate what is specific to *multi-step predictive consequence*. At $\bar{B} = 0.20$ on DMC-GB2, CAP-WM reaches IQM 666, improving the parameter-matched End-to-end gate by $+30\,[12, 47]$, One-step consequence by $+24\,[8, 40]$, and Multi-step value sensitivity by $+16\,[1, 31]$. An independent 10-seed Subtle/Video experiment confirms the primary horizon contrast: CAP-WM reaches 501 versus 397 for One-step, with paired difference $+104\,[63, 143]$ and $P(\text{improvement}) = 0.989$. A within-method horizon ablation yields $534 \to 565 \to 594 \to 608 \to 603$ as $\mathcal{H}$ expands over 1, 2, 4, 8, and 16 steps, showing saturation rather than a mechanical benefit from additional Jacobian terms.

The same hypothesis predicts when the additional horizon should matter. Before allocator branching, action repeats $r \in \{2, 4, 8\}$ yield $L_{\text{long}} = 0.69, 0.55, 0.37$ and subsequent CAP-WM–One-step differences $+45\,[22, 68]$, $+24\,[8, 40]$, and $+6\,[-13, 25]$. At the permissive rate $\bar{B} = 0.80$, the corresponding differences shrink to $+7\,[-8, 22]$, $-3\,[-16, 10]$, and $-1\,[-14, 12]$, all unresolved. The resulting resource–action-repeat interaction is $+31\,[8, 54]$. Thus the additional multi-step benefit is resolved when restrictive precision coincides with substantial dependence beyond one transition (Appendix C.2). Because action repeat also changes temporal discretization, this is an applicability test rather than a causal mediation analysis.

5.3 Measured Damage and Allocation Dynamics Support the Proposed Mechanism

Return alone does not establish the intended predictive semantics. Under finite interventions, CAP-WM's deployable score correlates with measured multi-step predictive damage at $\rho = 0.66$, versus 0.49 for One-step and 0.29 for Surprise; the pathwise teacher reaches 0.82. A histogram-preserving shuffle of CAP-WM's per-state rates reduces allocation–damage correlation from 0.63 to 0.04 and Subtle/Video return from 504 to 287. Closed-loop return loss likewise rises from 2.1% in the lowest consequence quartile to 12.4% in the highest. Together, these interventions validate the score–damage and allocation–damage links (Figure 2; Appendix B.5).

The reference-point experiments test the second mechanism directly. Replacing the common reference by adaptive source-referred refresh reduces DMC-Subtle return from 608 to 575 and contracts the active set from $8.1\,[7.6, 8.6]$ to $3.4\,[3.0, 3.9]$ groups. Isolating the two pathways, memory-referred re-linearization retains $7.7\,[7.2, 8.2]$ groups, whereas the assignment-dependent source gain alone contracts the set to $3.9\,[3.5, 4.3]$. Across ordinary initializations, source-referred refresh reaches ≤ 4 active groups with probability 0.74–0.91, versus only 0.06–0.11 for memory-referred refresh. Together with robustness to fixed allocation-independent reference magnitudes, these results connect Proposition 1 to reachable allocation dynamics and identify assignment-dependent attenuation, rather than nonlinear re-linearization or reference tuning, as the relevant collapse mechanism (Appendices A.2.1 and B.12).

5.4 Transfer and Boundary Conditions

World-model transfer. The tight-budget multi-step effect transfers beyond the primary DreamerV3 backbone. At $\bar{B} = 0.20$, CAP-WM–One-step is $+27\,[8, 46]$ on R2-Dreamer and $+25\,[8, 42]$

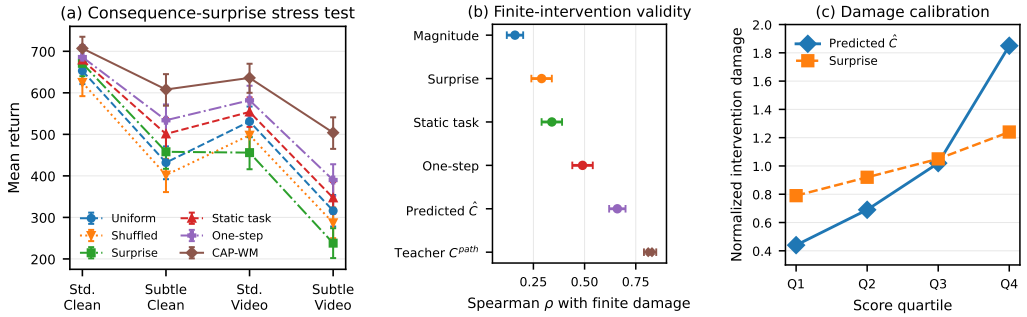


Figure 2: **Mechanism validation.** (a) Stress test for selected allocation signals. (b) Association between candidate scores and finite predictive damage. (c) Damage across score quartiles.

on TD-MPC2; at $\bar{B} = 0.80$, the corresponding differences contract to $+6\,[-7, 19]$ and $+3\,[-10, 16]$, respectively, and are unresolved. On Atari 100K at $\bar{B} = 0.80$, CAP-WM reaches human-normalized IQM 0.630 versus 0.530 for Uniform, while its difference from One-step consequence is $+0.041\,[-0.008, 0.088]$ and unresolved. Together, these results support transfer across world-model parameterizations and control domains, while the additional multi-step benefit remains concentrated in the restrictive-precision regime (Appendix C.1).

Sensitivity convention	Active / 16	Top-1 (%)
Frozen full-precision source	$15.4\,[14.8, 16.0]$	75.68
Adaptive source-referred	$10.8\,[10.0, 11.6]$	72.91
Adaptive quantized-value	$15.1\,[14.4, 15.8]$	75.47
Pilot source-referred fixed point	11	—

Table 2: **Reference-point starvation outside world models.** ResNet-50 mixed-precision post-training quantization at average branch precision 4 bits. Conditions differ in how sensitivity is referenced and refreshed; model, budget, allocator, and calibration protocol are fixed.

Structural analogue outside world models. Table 2 shows the same qualitative failure under a distinct attenuation mechanism: source-referred refresh contracts the active set and degrades accuracy, whereas removing assignment-dependent attenuation largely restores both. Because attenuation here arises from quantization-bin absorption rather than the Gaussian-channel gain, this experiment is a structural rather than literal analogue of Proposition 1 (Appendix B.13).

6 Conclusion

CAP-WM allocates a hard recurrent precision budget according to multi-step predictive consequence measured from a common allocation-independent reference. We show that measuring source-referred importance through the adaptive assignment itself can create self-reinforcing starvation: reduced precision attenuates the same importance signal needed for a group to recover rate. CAP-WM removes this assignment-dependent feedback and converts the resulting consequence signal into an exact per-transition allocation through reverse water filling. Exact serialized-payload experiments, finite predictive and closed-loop interventions, and transfer across three world-model backbones support the resulting allocation principle. A joint resource–horizon test further identifies its empirical boundary: the additional benefit of multi-step over one-step consequence is resolved when precision is restrictive and predictive dependence extends beyond one latent transition but becomes unresolved when either condition weakens. More broadly, these results suggest a design principle for adaptive resource allocation: when the current assignment can attenuate its own importance measurement, the consequence should be measured from an allocation-independent reference before resources are reassigned.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and feedback on technical presentation. The authors made all final scientific decisions and independently reviewed and verified the theoretical arguments, experimental methodology, reported results, and final manuscript. The authors take full responsibility for all claims and content.

REPRODUCIBILITY STATEMENT

Appendices A – C provide the complete methodological details, proofs, experimental protocols, ablations, and supporting numerical results.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Abdulaziz Almuzairee, Nicklas Hansen, and Henrik I. Christensen. A recipe for unbounded data augmentation in visual reinforcement learning. *Reinforcement Learning Journal*, 1:130–157, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper26.html.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6bdde0373d53d4a501249547084bed43-Abstract-Conference.html.

Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3): 183–193, 1954. doi: 10.1037/h0054663.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Horace B. Barlow. Possible principles underlying the transformations of sensory messages. In Walter A. Rosenblith (ed.), *Sensory Communication*, pp. 217–234. MIT Press, 1961.

Maxime Burchi and Radu Timofte. Learning transformer-based world models with contrastive predictive coding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YK9G4Htdew.

Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3146–3154, 2019. doi: 10.1109/ICCV.2019.00324.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai. Asymmetric gained deep image compression with continuous rate adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10532–10541, 2021. doi: 10.1109/CVPR46437.2021.01039.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18518–18529, 2020.

Chao Feng, Guoxu Zhou, Xun Wu, and Qibin Zhao. Plmq: Piecewise linear mixed-precision quantization for deep neural networks. *Neurocomputing*, 651:131000, 2025. doi: 10.1016/j.neucom.2025.131000.

Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005. doi: 10.1098/rstb.2005.1622.

Thomas L. Griffiths, Falk Lieder, and Noah D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7 (2):217–229, 2015. doi: 10.1111/tops.12142.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

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

Mohammadreza Hami, Mohammadreza Samadi, Chao Gao, and Negar Hassanpour. MOSAIK: Multi-patch content-aware spatial allocation of image tokens for efficient generation. *arXiv preprint arXiv:2608.05450*, 2026.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xCPJHtDB.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

David Minnen, Johannes Ballé, and George D. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10771–10780, 2018.

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*, 2026. URL https://openreview.net/forum?id=Je2QqXrcQq.

Girish N. Nair and Robin J. Evans. Stabilizability of stochastic linear systems with finite feedback data rates. *SIAM Journal on Control and Optimization*, 43(2):413–436, 2004. doi: 10.1137/S0363012902402116.

Zhao-Han Peng, Shaohui Li, Zhi Li, Shulan Ruan, Yu Liu, and You He. From observations to events: Event-aware world models for reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fc09b26b85ab3abb2832bd555a2e4215-Abstract-Conference.html.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003. doi: 10.1016/S0304-3932(03)00029-1.

Susanne Still. Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989, 2014. doi: 10.3390/e16020968. URL https://doi.org/10.3390/e16020968.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. URL https://arxiv.org/abs/1801.00690.

Sekhar Tatikonda and Sanjoy K. Mitter. Control under communication constraints. *IEEE Transactions on Automatic Control*, 49(7):1056–1068, 2004. doi: 10.1109/TAC.2004.831187.

Boxuan Zhang, Runqing Wang, Wei Xiao, Weipu Zhang, Jian Sun, Gao Huang, Jie Chen, and Gang Wang. DyMoDreamer: World modeling with dynamic modulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c56069f6be512aec8c6e445b98b21d5a-Abstract-Conference.html.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. STORM: Efficient stochastic transformer based world models for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27147–27166, 2023. URL https://openreview.net/forum?id=WxnrX42rnS.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Fei Zuo, Zikang Zhou, Hao Cong, Xiaoyan Xi, and Ho Fai Leung. Ratequant: Optimal mixed-precision kv cache quantization via rate-distortion theory. *arXiv preprint arXiv:2605.06675*, 2026.

Appendix

A Method Details and Proofs

This appendix makes the implementation and theoretical scope of CAP-WM explicit. We first specify the bounded recurrent interface and the reference-point mechanism, then detail the replay-derived consequence teacher and exact-budget allocator, prove the linear–Gaussian characterization, and finally quantify the gap between the idealized group specialization and the learned neural geometry.

A.1 Bounded Interface, Standardization, and Channel Identities

Native state and transmission boundary. The backbone uses a 512-dimensional deterministic RSSM state and 32 categorical stochastic variables with 32 classes each,

$$s_t = (h_t, z_t), \qquad h_t \in \mathbb{R}^{512}, \qquad z_{t,i} \in \{e_1, \dots, e_{32}\} \subset \mathbb{R}^{32}, \quad i = 1, \dots, 32, \tag{17}$$

so $\mathrm{vec}(z_t) \in \mathbb{R}^{1024}$ and the numerical recurrent feature has dimension $D_u = 512 + 1024 = 1536$.

Definition 1 inserts the bounded interface at the recurrent state boundary. Writing

$$(\widetilde{h}_{t-1}, \widetilde{z}_{t-1}) = \mathrm{Split}(m_{t-1}), \tag{18}$$

where Split is a fixed bookkeeping map assigning the first 512 transmitted coordinates to the deterministic branch and the remaining 1024 to the stochastic-feature branch, the next deterministic state is computed only from bounded memory,

$$h_t = f_\theta(\widetilde{h}_{t-1}, \widetilde{z}_{t-1}, a_{t-1}). \tag{19}$$

The native prior and observation-conditioned posterior then produce

$$z_t^- \sim p_\theta(z_t \mid h_t), \qquad z_t^+ \sim q_\theta(z_t \mid h_t, e_t), \qquad e_t = \mathrm{Enc}_\theta(o_t), \tag{20}$$

and CAP-WM forms the parameter-free candidates $u_t^\pm = \phi(h_t, z_t^\pm) = [h_t; \mathrm{vec}(z_t^\pm)] \in \mathbb{R}^{1536}$. The interface acts on u_t^+ during interaction and replay reconstruction and on u_t^- during imagination.

The Gaussian channel is applied *after* the backbone's native categorical sampling; CAP-WM neither replaces p_θ or q_θ by a Gaussian latent model nor assumes that z_t is Gaussian. After transmission, coordinates originating from $\mathrm{vec}(z_t)$ remain bounded numerical features and are not re-projected to one-hot vectors, because such a projection would restore information removed by the channel. The consequence head and allocator may inspect the current candidate to choose the channel parameters, but u_t, $\widehat{C}_t$, and b_t are not concatenated to the transmitted memory or supplied separately to downstream dynamics, actor, or critic. Their only downstream effect is through m_t. This is the no-bypass condition used throughout the paper.

Quantity	Symbol	Value
Deterministic RSSM state	D_h	512
Categorical stochastic state	(N_z, K_z)	$(32, 32)$
Transmission feature	$\phi(s_t)$	$[h_t; \mathrm{vec}(z_t)]$
Transmitted dimension	$D_u = D_h + N_z K_z$	1536
Groups	G	16
Group width	$d = D_u / G$	96
Consequence head	g_ω	$1536 \to 256 \to 256 \to 16$
Head parameters	—	463,376

Table 3: **Bounded recurrent-interface specification.** CAP-WM operates on the ordinary numerical DreamerV3 state after native categorical sampling and inserts no trainable projection before the bounded interface.

Grouping and the deterministic/stochastic boundary. The 1536 coordinates are partitioned into $G = 16$ contiguous groups of equal width $d = 96$. Because $512/96$ is not an integer, the partition is intentionally not aligned with the deterministic/stochastic boundary: groups 1–5 contain only deterministic coordinates, group 6 contains the last 32 deterministic and first 64 stochastic coordinates, and groups 7–16 contain only stochastic coordinates. Equal widths keep the channel law and allocation objective comparable across groups rather than confounding content with group size. The fixed-coordinate-permutation ablation reported in Appendix B.12 tests sensitivity to this arbitrary coordinate partition.

Standardization. CAP-WM maintains training-only exponential-moving-average first and second moments of posterior candidate features drawn from replay, initialized from the first learner minibatch. At learner update k,

$$\mu_g^{(k)} = \beta_{\mathrm{std}}\mu_g^{(k-1)} + (1-\beta_{\mathrm{std}})\widehat{\mu}_g^{(k)}, \tag{21}$$

$$\nu_g^{(k)} = \beta_{\mathrm{std}}\nu_g^{(k-1)} + (1-\beta_{\mathrm{std}})\widehat{\nu}_g^{(k)}, \tag{22}$$

with minibatch moments taken jointly over batch and replay-time dimensions, $\beta_{\mathrm{std}} = 0.99$, and $\epsilon_{\mathrm{std}} = 10^{-5}$. The moments and the operations below are coordinatewise within each group. With $\sigma_g^{2(k)} = \max(\nu_g^{(k)} - (\mu_g^{(k)})^2, 0)$,

$$\bar{u}_{t,g} = \frac{u_{t,g} - \mathrm{sg}[\mu_g]}{\sqrt{\mathrm{sg}[\sigma_g^2] + \epsilon_{\mathrm{std}}}}. \tag{23}$$

The same statistics are applied to prior and posterior candidates, updated only from posterior replay features, and frozen at evaluation. Imagined rollouts never update normalization. Standardization controls marginal scale only; it does not assert Gaussianity, whiten cross-group covariance, or align the fixed groups with predictive eigenmodes.

Channel identities. For an ideal standardized Gaussian group $\bar{u}_{t,g} \sim \mathcal{N}(0, I_d)$, let $D_{t,g} = \exp(-2b_{t,g}/d)$ and let $m_{t,g}$ be given by Equation 3. With $\delta_{t,g} = m_{t,g} - \bar{u}_{t,g}$,

$$\mathbb{E}[\delta_{t,g}\delta_{t,g}^\top] = D_{t,g}I_d, \qquad \tfrac{1}{d}\,\mathbb{E}\|\delta_{t,g}\|_2^2 = D_{t,g}. \tag{24}$$

Moreover $\mathrm{Cov}(m_{t,g}) = (1 - D_{t,g})I_d$ and the conditional noise covariance is $D_{t,g}(1 - D_{t,g})I_d$, so

$$I(\bar{u}_{t,g}; m_{t,g}) = -\tfrac{d}{2}\log D_{t,g} = b_{t,g} \text{ nats}. \tag{25}$$

These two identities justify the nominal Gaussian rate interpretation only for the ideal reference source. The channel derivative with the operating rate held fixed is, in contrast, exact for any source law:

$$\left.\frac{\partial m_{t,g}}{\partial \bar{u}_{t,g}}\right|_{b_{t,g}} = (1 - D_{t,g})I_d = \kappa(b_{t,g})I_d, \qquad \kappa(\beta) = 1 - e^{-2\beta/d}. \tag{26}$$

At $b_{t,g} = 0$, $D_{t,g} = 1$, both channel coefficients vanish, and $m_{t,g} \equiv 0$. Hence a starved group transmits neither its source value nor fresh noise, and $\kappa(0) = 0$ exactly.

Memory- and source-referred consequence. Equation 5 defines the intrinsic memory-referred quantity $c_{t,g}(m)$. Treating the operating allocation b as fixed during differentiation, Equation 26 and the chain rule give the corresponding source-referred sensitivity

$$S_g(b) = \sum_h \gamma_h \left\| \frac{\partial F_{\theta,h}}{\partial \bar{u}_{t,g}} \right\|_F^2 = \kappa(b_g)^2\, c_{t,g}(m(b)). \tag{27}$$

Equation 6 is the special case $b = b^{\mathrm{ref}}$. Under a uniform reference, C^{path} and c differ only by the common constant $\kappa_{\mathrm{ref}}^2 := \kappa(B/G)^2$; under an adaptive non-uniform reference, the multiplicative gain is group dependent. This distinction is the mechanism isolated in Section 3.3 and proved below.

A.2 Proofs for the Reference-Point Results

Throughout this subsection $B > 0$ and $c_1, \dots, c_G > 0$ are held fixed, as assumed in Proposition 1. This assumption is exact in the linear–Gaussian specialization because Equation 15 is independent of the operating allocation. Let

$$\Delta = \left\{ b \in \mathbb{R}^G_{\geq 0} : \sum_g b_g = B \right\}, \qquad S_g(b) = \kappa(b_g)^2 c_g, \qquad T(b) = \mathrm{WF}(S(b); B). \tag{28}$$

Observation 1. *For every $b \in \Delta$, some g satisfies $b_g \geq B/G$. Therefore*

$$\max_g S_g(b) \geq \kappa(B/G)^2 c_{\min} > 0,$$

so the water-filling problem defining $T(b)$ is well posed.

Lemma 2 (Uniform lower bound on the water level)**.** *For every $b \in \Delta$, the water level $\lambda(b)$ of $T(b)$ satisfies*

$$\lambda(b) \geq \lambda_{\min} := \kappa(B/G)^2 c_{\min} e^{-2B/d} > 0. \tag{29}$$

Proof. Let $S_{\max} = \max_g S_g(b)$, attained at $g^\star$. Because the total allocated rate is B, $T(b)_{g^\star} \leq B$. Equation 12 therefore gives $\frac{d}{2} \log(S_{\max}/\lambda) \leq B$, or $\lambda \geq S_{\max} e^{-2B/d}$. Observation 1 lower-bounds $S_{\max}$. □

Lemma 3 (T is a continuous self-map of Δ)**.** *$T : \Delta \to \Delta$ is continuous.*

Proof. For weights S with $\max_g S_g > 0$, define

$$R_S(\lambda) = \frac{d}{2} \sum_g r(S_g, \lambda), \qquad r(s, \lambda) = \begin{cases} [\log(s/\lambda)]_+, & s > 0, \\ 0, & s = 0. \end{cases}$$

For every $\lambda > 0$, this extension is continuous in $s \geq 0$; moreover, R_S is continuous and strictly decreasing wherever positive, with $R_S(\lambda) \to \infty$ as $\lambda \downarrow 0$ and $R_S(\max_g S_g) = 0$. Hence the root satisfying $R_S(\lambda) = B$ is unique. Because each $S_g(b)$ is continuous in b and Lemma 2 keeps the root uniformly away from zero, the root and the resulting water-filled rate vector vary continuously with b, including at points where individual weights vanish. Thus T is a continuous self-map of Δ. □

Proof of Proposition 1. Existence. Fix any nonempty proper $\mathcal{S} \subsetneq [G]$ and define

$$\Delta_{\mathcal{S}} = \{ b \in \Delta : b_g = 0 \; \forall g \in \mathcal{S} \}. \tag{30}$$

$\Delta_{\mathcal{S}}$ is nonempty, compact, and convex. It is also T-invariant: if $g \in \mathcal{S}$, then $b_g = 0$, so $\kappa(b_g) = 0$ and $S_g(b) = 0$; because the water level is positive, Equation 12 gives $T(b)_g = 0$. Thus $T(\Delta_{\mathcal{S}}) \subseteq \Delta_{\mathcal{S}}$. Lemma 3 and Brouwer's fixed-point theorem yield $b^\star \in \Delta_{\mathcal{S}}$ with $T(b^\star) = b^\star$. Since the argument requires only $c_g > 0$, it applies in particular to $\mathcal{S} = \{\arg\max_g c_g\}$.

Local absorption. By Lemma 2, group g is inactive whenever $\kappa(b_g)^2 c_g \leq \lambda_{\min}$. If $c_g \leq \lambda_{\min}$, this holds for every feasible b_g and we may set $\beta_g = B$. Otherwise, since κ is strictly increasing and $\kappa(0) = 0$, define

$$\beta_g = -\frac{d}{2} \log\left(1 - \sqrt{\lambda_{\min}/c_g}\right) > 0. \tag{31}$$

Then $b_g \leq \beta_g$ implies $T(b)_g = 0$. Setting $\beta = \min_g \beta_g > 0$ proves the stated one-step threshold. Once a group reaches zero, its source-referred weight remains zero on subsequent iterations, so the starved set is absorbing. □

Proof of Proposition 2. At $b_g^{\mathrm{ref}} = B/G$, Equation 27 gives $C^{\mathrm{path}} = \alpha c$ with the common positive factor $\alpha = \kappa(B/G)^2$. Water filling is invariant to a positive common rescaling of all weights: if λ solves

$$\frac{d}{2} \sum_g \left[\log \frac{\alpha c_g}{\lambda} \right]_+ = B,$$

then $\lambda' = \lambda/\alpha$ solves the corresponding equation for weights c, and the allocated rates are identical. Hence

$$\mathrm{WF}(C^{\mathrm{path}}; B) = \mathrm{WF}(c; B).$$

The resulting allocation is a direct function of the fixed consequence profile rather than a fixed point of an assignment-dependent measurement map. Ordinary water filling can still assign zero rate to groups below its water level; what disappears is the self-referential mechanism that can keep an otherwise consequential group suppressed because its current allocation attenuates its own score. □

Proposition 5 (Convention independence at a common reference)**.** *At $b_g^{\mathrm{ref}} = B/G$, source-referred and memory-referred consequence induce the same allocation,*

$$\mathrm{WF}\big(C^{\mathrm{path}}; B\big) = \mathrm{WF}\big(c(m^{\mathrm{ref}}); B\big)\,. \tag{32}$$

Under a non-uniform reference the two conventions need not agree.

Proof. The positive half follows immediately from Equation 6 and the common-rescaling invariance above. For a non-uniform reference, the source-referred weights become $\kappa(b_g^{\mathrm{ref}})^2 c_g$. The multiplier is then group dependent, so it cannot in general be absorbed into a single water level. For example, two groups with equal intrinsic consequence but unequal reference rates receive equal rates under the memory-referred convention and generally unequal rates under the source-referred convention. □

Remark 1 (Nonlinear reference drift). Proposition 1 deliberately isolates assignment-dependent source attenuation by holding c_g fixed. In the nonlinear model, changing the operating reference also changes the memory at which the rollout Jacobian is evaluated, so $c_g(m(b))$ can drift with b. This second pathway is neither assumed away empirically nor assigned a fixed sign. Appendix A.2.1 measures it directly by comparing memory-referred and source-referred adaptive refresh under matched states, action continuations, channel noise, and JVP probes. The common reference fixes a single operating point for all groups, preventing this operating-point movement from entering the teacher comparison itself.

Remark 2 (Common rescalings do not change the allocation). Several expressions differ only by positive constants shared across groups, including the common source gain κ_{ref}^2 and the group-width factor that appears in Corollary 2. Such constants are absorbed into the water level and do not change the realized allocation. This invariance is precisely what fails under an adaptive source-referred reference, where the gain varies with g.

The theorem does not quantify how often the starved fixed points are reached in a trained nonlinear model. We therefore treat reachability as an empirical question rather than constructing a favorable synthetic consequence profile; the basin analysis in Appendix B.12 evaluates the adaptive source-referred, memory-referred, and common-reference maps on held-out trained states from multiple ordinary initializations.

A.2.1 Empirical Nonlinear Reference Drift

Proposition 1 isolates assignment-dependent attenuation under a fixed intrinsic consequence profile. In the nonlinear world model, however, moving the reference also changes the memory state at which the rollout Jacobian is evaluated, so $c_g(m(b))$ can itself vary with b. We therefore quantify this additional pathway directly on trained models, without retraining.

For the drift statistics below, we use DMC-Subtle CAP-WM checkpoints at $\bar{B} = 0.20$ and 512 held-out replay states. Consequence is measured strictly *memory-referred*,

$$c_{t,g}(m) = \sum_{h \in \mathcal{H}} \gamma_h \left\| \frac{\partial F_{\theta,h}(m, a_{t:t+h-1})}{\partial m_g} \right\|_F^2, \tag{33}$$

so this diagnostic does not contain the source-channel factor $\kappa(b_g)^2$. We use $K = 16$ Rademacher probes because this is a one-time offline diagnostic rather than a training-time target.

Let

$$b^{\mathrm{ref}} = \frac{B}{G}\mathbf{1}, \qquad c_t^{\mathrm{ref}} = c_t(m(b^{\mathrm{ref}})).$$

Operating reference	$\rho_{\mathrm{drift}} \uparrow$	$A_{\mathrm{drift}} \downarrow$
Common uniform	1.00	0.00
Deployed CAP-WM	$0.89\,[0.87, 0.91]$	$0.08\,[0.07, 0.09]$
Random simplex #1	$0.84\,[0.81, 0.87]$	$0.12\,[0.10, 0.14]$
Random simplex #2	$0.80\,[0.77, 0.83]$	$0.15\,[0.13, 0.17]$
Starvation stress	$0.75\,[0.71, 0.78]$	$0.19\,[0.17, 0.22]$

Table 4: **Nonlinear reference drift on DMC-Subtle at $\bar{B} = 0.20$.** Both statistics use memory-referred consequence and therefore exclude the assignment-dependent source gain $\kappa(b_g)^2$. Brackets denote 95% bootstrap intervals over the 512 sampled held-out replay states.

In addition to the common uniform reference, we evaluate four operating points: CAP-WM's deployed allocation $b_t^\star$; two fixed random simplex allocations

$$b^{\mathrm{rnd},j} = Br^{(j)}, \qquad r^{(j)} \sim \mathrm{Dirichlet}(\mathbf{1});$$

and a starvation-stress allocation obtained by first freezing c_t^{ref} and iterating

$$b_t^{(k+1)} = \mathrm{WF}\left(\kappa(b_t^{(k)})^2 \odot c_t^{\mathrm{ref}}; B\right), \qquad b_t^{(0)} = b^{\mathrm{ref}}, \tag{34}$$

to convergence. Only after this stress allocation is obtained do we re-evaluate the true nonlinear memory-referred profile at the resulting operating point.

For each replay state, all operating points use the same future action continuation, channel-noise realization, and Rademacher probes. We measure rank stability relative to the common reference as

$$\rho_{\mathrm{drift}}(b) = \rho_{\mathrm{Spearman}}\left(c_t(m(b)), c_t^{\mathrm{ref}}\right), \tag{35}$$

and the corresponding change in the allocation decision as

$$A_{\mathrm{drift}}(b) = \frac{1}{2B}\left\|\mathrm{WF}(c_t(m(b)); B) - \mathrm{WF}(c_t^{\mathrm{ref}}; B)\right\|_1. \tag{36}$$

Because both allocations contain total mass B, $A_{\mathrm{drift}} \in [0, 1]$ measures normalized budget mass that must be redistributed after changing the nonlinear measurement reference. This uses the same $1/(2B)$ normalization as the allocation-displacement diagnostic in Appendix B.11.

Interpretation criterion. If intrinsic drift largely preserves consequence ordering and produces only small A_{drift}, while memory-referred adaptive refresh retains an active set close to the common-reference regime, then nonlinear reference drift is present but is insufficient by itself to reproduce the starvation observed under source-referred refresh. Conversely, if memory-referred refresh contracts the active set to a degree comparable to source-referred refresh, then nonlinear reference drift itself explains a substantial fraction of the collapse, and Proposition 1 should be interpreted as only one contributing mechanism.

The deployed CAP-WM allocation induces measurable but moderate nonlinear reference drift. The intrinsic consequence profile retains a Spearman rank correlation of $\rho_{\mathrm{drift}} = 0.89\,[0.87, 0.91]$ with the common-reference profile, while the resulting reverse-water-filling allocation changes by only $A_{\mathrm{drift}} = 0.08\,[0.07, 0.09]$ in normalized budget displacement.

For scale, Appendix B.11 uses the same $1/(2B)$ normalization and reports allocation displacements of 0.11 under the nuisance shift and 0.35 under the dynamics shift. These quantities have different estimands: the shift statistics compare pre/post environmental allocations, whereas A_{drift} is a within-state reference perturbation. We therefore do not treat them as a direct statistical comparison. They nevertheless provide a common numerical scale on which the deployed-reference drift, $0.08\,[0.07, 0.09]$, is modest.

Both random references produce larger deviations than the deployed operating point. Their rank agreements are $0.84\,[0.81, 0.87]$ and $0.80\,[0.77, 0.83]$, with corresponding normalized allocation displacements $0.12\,[0.10, 0.14]$ and $0.15\,[0.13, 0.17]$. The deliberately severe starvation-stress reference produces the largest change, with rank agreement $0.75\,[0.71, 0.78]$ and allocation displacement $0.19\,[0.17, 0.22]$.

These measurements confirm that $c_g(m(b))$ is not globally allocation independent in the trained nonlinear world model. The fixed-profile hypothesis in Proposition 1 is therefore an analytical isolation of the assignment-gain pathway rather than an exact description of all nonlinear reference dynamics.

Does nonlinear drift alone reproduce active-set starvation? Rank and displacement statistics show that drift exists; they do not show whether it is strong enough to cause starvation. We therefore isolate the two pathways on the outcome that matters. The memory-referred map

$$T_{\mathrm{mem}}(b) = \mathrm{WF}(c(m(b)); B) \tag{37}$$

re-evaluates the nonlinear consequence profile at every iterate, so it carries nonlinear reference drift but no assignment-dependent gain. The fixed-profile source map of Proposition 1,

$$T_{\mathrm{src}}(b) = \mathrm{WF}\big(\kappa(b)^2 \odot c^{\mathrm{ref}}; B\big)\,, \tag{38}$$

is the map already used to construct the starvation-stress point in Equation 34; it carries the assignment-dependent gain but no drift. If drift alone were sufficient for starvation, T_{mem} would collapse; if the gain is the operative mechanism, T_{src} collapses while T_{mem} does not.

Both maps are evaluated on the 256-state sample of the basin-of-attraction analysis in Appendix B.12, from four initialization families, with shared future action continuations, matched channel-noise realizations and JVP probes, and a group counted as active when $b_{t,g} > 10^{-4}B/G$, the threshold used for the trained-system diagnostic. We quote the uniform-initialization values here rather than duplicating that table.

From uniform initialization, T_{mem} retains $7.7\,[7.2, 8.2]$ active groups, whereas T_{src} contracts the active set to $3.9\,[3.5, 4.3]$. Nonlinear re-linearization by itself therefore does not reproduce the collapse, whereas the assignment-dependent gain does, even with the intrinsic profile held fixed. The separation holds across all four initialization families: T_{src} reaches at most four active groups with probability 0.74–0.91, T_{mem} with probability only 0.06–0.11.

Both isolated values overlap the corresponding trained-system diagnostics reported in Appendix B.12: T_{mem} overlaps the common-reference value $8.1\,[7.6, 8.6]$, and T_{src} overlaps the adaptive-reference value $3.4\,[3.0, 3.9]$. The trained adaptive system contains both pathways at once; the isolated maps show that the collapse it exhibits is reproduced by the gain alone and not by drift alone.

Overall, nonlinear reference drift contributes to operating-point dependence, but the isolated-map comparison shows that drift alone does not reproduce the observed active-set collapse, whereas the assignment-dependent source attenuation does. This supports the mechanism isolated by Proposition 1 without extending the proposition to a global nonlinear fixed-point theorem.

A.3 Teacher Target, Replay Conditioning, and Estimation

Target space. The teacher and finite-intervention analysis use the same standardized feature space as the bounded interface. Let u^+_{t+h} be the posterior candidate from replay observation o_{t+h} and define the detached future target

$$v^+_{t+h} = \mathrm{sg}\Big[\mathrm{concat}\Big(\mathcal{N}_1(u^+_{t+h,1}), \ldots, \mathcal{N}_G(u^+_{t+h,G})\Big)\Big]\,. \tag{39}$$

$F_{\theta,h}(m, a_{t:t+h-1})$ rolls the prior forward h steps from m under the supplied action prefix and predicts in this standardized state-feature space. Reward, value, policy gradients, task labels, and privileged simulator state do not enter the teacher target; the score is defined by predictive geometry alone.

Reference channel. Teacher Jacobians are evaluated at the common operating point

$$b^{\mathrm{ref}}_g = \frac{B}{G}, \qquad D^{\mathrm{ref}} = \exp\!\left(-\frac{2B}{Gd}\right). \tag{40}$$

A reference memory m^{ref}_t is sampled from Equation 3 at this allocation. The reference specifies where importance is *measured*; it is not the deployed allocation, which is $b^\star_t$ from Equation 12.

Differentiation through categorical prior rollouts. The teacher differentiates the surrogate computation graph used by the executed backbone while retaining discrete categorical samples in the forward rollout. For each teacher construction, the forward categorical samples and reference-channel noise define one realized stochastic rollout trace, and the action continuation is fixed to replay. Jacobian–vector products are taken through that surrogate graph without differentiating through replay actions. Thus $F_{\theta,h}$ denotes the differentiable surrogate prior-prediction map conditional on the realized stochastic trace; the resulting Jacobian is a surrogate-gradient object, not a classical derivative of the discrete sampling operation. Across replay states and teacher draws, the training target averages over the stochastic traces encountered by this construction.

Pathwise teacher. For a realized teacher trace, write

$$J_{t,g,h} = \frac{\partial F_{\theta,h}(m_t^{\mathrm{ref}}, a_{t:t+h-1})}{\partial \bar{u}_{t,g}} = \kappa_{\mathrm{ref}} \left.\frac{\partial F_{\theta,h}}{\partial m_{t,g}}\right|_{m_t^{\mathrm{ref}}}, \tag{41}$$

where $\kappa_{\mathrm{ref}} = \kappa(B/G)$. Hence

$$C_{t,g}^{\mathrm{path}} = \sum_{h\in\mathcal{H}} \gamma_h \|J_{t,g,h}\|_F^2 = \kappa_{\mathrm{ref}}^2 c_{t,g}(m_t^{\mathrm{ref}}).$$

The implementation computes the source-referred left-hand side by differentiating the composition of reference channel and prior rollout. Proposition 5 explains why this induces the same allocation as memory-referred consequence at the common reference.

Remark 3 (Relation to one-step sensitivity). For $\mathcal{H} = \{1\}$, $C_{t,g}^{\mathrm{path}} = \|J_{t,g,1}\|_F^2 = \mathrm{tr}(J_{t,g,1}^\top J_{t,g,1})$. Under squared prediction error this is the corresponding local Gauss–Newton/Hessian-trace-style sensitivity. The One-step consequence baseline therefore changes only the predictive horizon, not the reference, grouping, channel, predictor capacity, or allocation rule.

Replay-conditional population target. Let $H_{\max} = \max \mathcal{H}$, $\mathbf{a}_t = a_{t:t+H_{\max}-1}$, let ξ_t^{ref} denote reference-channel randomness, and let ζ_t denote the categorical prior-rollout trace used by the straight-through teacher graph. Write

$$\rho_{\mathcal{D}}(\mathbf{a}_t, \xi_t^{\mathrm{ref}}, \zeta_t \mid \bar{u}_t)$$

for the replay-conditional distribution induced by the buffer and teacher construction. Because g_ω is trained by squared error in *log consequence*, the population-optimal prediction is the conditional log-mean

$$\ell_{t,g}^{\mathcal{D}}(\bar{u}_t) = \mathbb{E}_{\rho_{\mathcal{D}}}\left[\log(C_{t,g}^{\mathrm{path}} + \epsilon) \,\middle|\, \bar{u}_t\right]. \tag{42}$$

Its positive-space image $C_{t,g}^{\mathcal{D}} = \exp(\ell_{t,g}^{\mathcal{D}})$ is the conditional geometric mean of $C_{t,g}^{\mathrm{path}} + \epsilon$, not the arithmetic conditional mean $\mathbb{E}[C_{t,g}^{\mathrm{path}} \mid \bar{u}_t]$.

Stochastic Jacobian estimator. Materializing $J_{t,g,h}$ is unnecessary. For independent Rademacher probes $v^{(k)} \in \{-1,+1\}^d$ with $\mathbb{E}[v^{(k)}v^{(k)\top}] = I_d$, $\|J_{t,g,h}\|_F^2 = \mathbb{E}_v\|J_{t,g,h}v\|_2^2$. With $K = 2$ probes,

$$\widehat{Q}_{t,g,h} = \frac{1}{K}\sum_{k=1}^{K} \|J_{t,g,h}v^{(k)}\|_2^2, \qquad \widetilde{C}_{t,g}^{\mathrm{path}} = \sum_{h\in\mathcal{H}} \gamma_h \widehat{Q}_{t,g,h}. \tag{43}$$

Conditioned on the realized reference memory, rollout trace, and action continuation, $\widehat{Q}_{t,g,h}$ is unbiased for $\|J_{t,g,h}\|_F^2$, so $\widetilde{C}_{t,g}^{\mathrm{path}}$ is unbiased before the logarithm. The implemented predictor loss is

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{G}\sum_{g=1}^{G}\left[\log \widehat{C}_{t,g} - \log\Big(\mathrm{sg}[\widetilde{C}_{t,g}^{\mathrm{path}}] + \epsilon\Big)\right]^2. \tag{44}$$

Finite-K unbiasedness in positive space does not make the sampled log target unbiased; by Jensen, the finite-probe log target has a downward Monte Carlo bias. We therefore treat Equation 44 as a Monte Carlo approximation to Equation 42 and report sensitivity to $K \in \{1,2,4\}$. A fresh teacher target is built every four learner updates; intermediate updates execute the ordinary backbone and actor–critic losses without reconstructing Jacobians.

Replay–policy mismatch. The teacher is explicitly replay conditional. If behavior or dynamics change rapidly, replay continuations can differ from those induced by the current policy. CAP-WM has no theorem eliminating this distribution mismatch. The deployable head receives no future action sequence, and the current-policy comparison in the experiments measures the practical size of the mismatch rather than turning the teacher into a policy-conditioned causal quantity.

A.4 Allocation: Damage Surrogate, KKT, Cost, and Gradient Separation

Predictive-deviation definition. Let $P_g \in \mathbb{R}^{D_u \times d}$ inject group g. For the local asymptotic in Lemma 1, write the perturbation family as

$$\delta_{t,g}(D_{t,g}) = \sqrt{D_{t,g}}\,\xi_{t,g}, \qquad \mathbb{E}[\xi_{t,g}] = 0, \qquad \mathbb{E}[\xi_{t,g}\xi_{t,g}^\top] = I_d,$$

with finite fourth moment. Hence $\mathbb{E}[\delta_{t,g}\delta_{t,g}^\top] = D_{t,g} I_d$. The weighted multi-step predictive deviation is

$$\Delta_{t,g}^{\mathrm{pred}} = \sum_{h\in\mathcal{H}} \gamma_h \left\| F_{\theta,h}(m + P_g\delta_{t,g}, a_{t:t+h-1}) - F_{\theta,h}(m, a_{t:t+h-1}) \right\|_2^2 . \tag{45}$$

The covariance model is exact for the ideal standardized Gaussian reference source and is used as a local second-moment surrogate for learned features. The scaled-family formulation makes explicit that the statement in Lemma 1 is a local $D_{t,g} \downarrow 0$ result rather than a claim about arbitrary error distributions sharing the same covariance.

Proof of Lemma 1. Fix the replay action continuation and abbreviate

$$J_h^{\mathrm{mem}} = \left.\frac{\partial F_{\theta,h}}{\partial m_g}\right|_{m=m_t^{\mathrm{ref}}}.$$

Assume $F_{\theta,h}$ is twice continuously differentiable in a neighborhood of m_t^{ref}. Taylor expansion gives

$$F_{\theta,h}\big(m_t^{\mathrm{ref}} + P_g\delta_{t,g}, a_{t:t+h-1}\big) - F_{\theta,h}\big(m_t^{\mathrm{ref}}, a_{t:t+h-1}\big) = J_h^{\mathrm{mem}}\delta_{t,g} + r_h(\delta_{t,g}), \tag{46}$$

where locally $\|r_h(\delta)\| \le c_h\|\delta\|^2$ for some finite c_h. Using $\mathbb{E}[\delta_{t,g}\delta_{t,g}^\top] = D_{t,g}I_d$,

$$\mathbb{E}\big\|J_h^{\mathrm{mem}}\delta_{t,g}\big\|_2^2 = \mathrm{tr}\big(J_h^{\mathrm{mem}}\mathbb{E}[\delta_{t,g}\delta_{t,g}^\top]J_h^{\mathrm{mem}\top}\big) = D_{t,g}\big\|J_h^{\mathrm{mem}}\big\|_F^2. \tag{47}$$

Because $\delta_{t,g} = \sqrt{D_{t,g}}\xi_{t,g}$ and $\mathbb{E}\|\xi_{t,g}\|^4 < \infty$,

$$\mathbb{E}\|r_h(\delta_{t,g})\|^2 = O(D_{t,g}^2),$$

and Cauchy–Schwarz gives

$$\begin{aligned} 2\,\mathbb{E}\,|\langle J_h^{\mathrm{mem}}\delta_{t,g}, r_h(\delta_{t,g})\rangle| &\le 2\left(\mathbb{E}\|J_h^{\mathrm{mem}}\delta_{t,g}\|^2\right)^{1/2}\left(\mathbb{E}\|r_h(\delta_{t,g})\|^2\right)^{1/2} \\ &= O(D_{t,g}^{1/2})\,O(D_{t,g}) = O(D_{t,g}^{3/2}) = o(D_{t,g}). \end{aligned}$$

Therefore

$$\mathbb{E}\left\|F_{\theta,h}(m_t^{\mathrm{ref}} + P_g\delta_{t,g}) - F_{\theta,h}(m_t^{\mathrm{ref}})\right\|_2^2 = D_{t,g}\big\|J_h^{\mathrm{mem}}\big\|_F^2 + o(D_{t,g}).$$

Summing over $h \in \mathcal{H}$ with weights γ_h and applying Equation 5 yields Equation 7. □

Remark 4 (Deviation versus error increase). Lemma 1 concerns the predictive *deviation* in Equation 45, which is the local quantity underlying Equation 11. The finite-intervention experiment instead measures the increase in predictive *error* relative to the replay target v_{t+h}^+. Let

$$e_h = F_{\theta,h}(m_t^{\mathrm{ref}}, a_{t:t+h-1}) - v_{t+h}^+$$

denote the unperturbed prediction residual. Using Equation 46 and expanding the squared error gives

$$\mathbb{E}\big[\Delta_{t,g}^{\mathrm{err}}\big] = \mathbb{E}\big[\Delta_{t,g}^{\mathrm{pred}}\big] + 2\sum_{h\in\mathcal{H}} \gamma_h \left\langle e_h,\, \mathbb{E}[r_h(\delta_{t,g})]\right\rangle, \tag{48}$$

because the first-order residual cross term vanishes:

$$\langle e_h, J_h^{\mathrm{mem}} \mathbb{E}[\delta_{t,g}] \rangle = 0.$$

If $F_{\theta,h}$ is affine over the perturbation region, then $r_h \equiv 0$ and the two quantities coincide exactly in expectation. For a twice differentiable predictor, $\mathbb{E}[r_h(\delta_{t,g})] = O(D_{t,g})$, so the curvature–residual term in Equation 48 is also $O(D_{t,g})$. It can therefore be of the same asymptotic order as the leading predictive deviation; its sign and magnitude depend on the interaction between local model curvature and the current prediction residual.

Accordingly, the finite-intervention correlations are empirical tests of whether the learned score predicts realized finite predictive damage, not corollaries of Lemma 1. This distinction also explains why $\rho(C^{\mathrm{path}}, \Delta^{\mathrm{err}})$ can be high without being expected to equal one.

The perturbation model $\mathbb{E}[\delta\delta^\top] = D_{t,g} I_d$ is exact for the ideal Gaussian reference source through Equation 24 and is used as a local second-moment surrogate otherwise. It is not a claim about the exact conditional error distribution of an arbitrary learned feature.

Proof of Proposition 3. Write

$$f(b) = \sum_g \widehat{C}_{t,g} e^{-2b_g/d}.$$

Since $\widehat{C}_{t,g} > 0$ by Equation 10,

$$\frac{\partial^2 f}{\partial b_g^2} = \frac{4}{d^2} \widehat{C}_{t,g} e^{-2b_g/d} > 0.$$

Thus f is strictly convex. The feasible set $\{b \geq 0 : \sum_g b_g = B\}$ is nonempty, compact, and convex, so the minimizer is unique.

With multiplier ν for the equality constraint and $\mu_g \geq 0$ for $-b_g \leq 0$,

$$\mathcal{L} = \sum_g \widehat{C}_{t,g} e^{-2b_g/d} + \nu \left(\sum_g b_g - B \right) - \sum_g \mu_g b_g. \tag{49}$$

Stationarity gives

$$-\frac{2}{d} \widehat{C}_{t,g} e^{-2b_g/d} + \nu - \mu_g = 0.$$

For an active group $b_g > 0$, complementary slackness gives $\mu_g = 0$. Defining $\lambda_t = \nu d/2$ yields

$$b_g = \frac{d}{2} \log \frac{\widehat{C}_{t,g}}{\lambda_t}.$$

For an inactive group $b_g = 0$, dual feasibility gives $\widehat{C}_{t,g} \leq \lambda_t$. Combining the two cases yields Equation 12.

Finally, define

$$R(\lambda) = \frac{d}{2} \sum_g \left[\log \frac{\widehat{C}_{t,g}}{\lambda} \right]_+ .$$

For $\lambda \in (0, \max_g \widehat{C}_{t,g}]$, R is continuous and strictly decreasing wherever $R > 0$, with $R(\lambda) \to \infty$ as $\lambda \downarrow 0$ and $R(\max_g \widehat{C}_{t,g}) = 0$. Hence for every $B > 0$ there is a unique λ_t satisfying $R(\lambda_t) = B$, completing the proof. □

Proposition 6 (Allocator cost)**.** *Bisection in $\ell = \log \lambda$ produces an allocation satisfying*

$$\left| \sum_g b_{t,g} - B \right| \leq \varepsilon B$$

after at most

$$\left\lceil \log_2 \frac{G}{\varepsilon} \right\rceil$$

bisection steps from the bracket below. Each step requires $O(G)$ scalar operations. The total allocation cost is therefore

$$O\left(G \log \frac{G}{\varepsilon}\right)$$

per recurrent transition and requires no automatic differentiation at deployment. For the fixed $G = 16$ used in our experiments, this reduces to $O(G \log(1/\varepsilon))$ with respect to numerical tolerance.

Proof of Proposition 6. Let

$$\widetilde{R}(\ell) = R(e^{\ell}).$$

On any interval with fixed active set $A(\ell) = \{g : \widehat{C}_{t,g} > e^{\ell}\}$,

$$\widetilde{R}'(\ell) = -\frac{d}{2}|A(\ell)|.$$

Hence $\widetilde{R}$ is continuous, non-increasing, piecewise linear, and globally Lipschitz with constant $dG/2$.

Bracket. Let $\widehat{C}_{\max} = \max_g \widehat{C}_{t,g}$. At

$$\ell_{\mathrm{hi}} = \log \widehat{C}_{\max},$$

we have $\widetilde{R}(\ell_{\mathrm{hi}}) = 0 \le B$. At

$$\ell_{\mathrm{lo}} = \log \widehat{C}_{\max} - \frac{2B}{d},$$

the maximally weighted group alone contributes B, so $\widetilde{R}(\ell_{\mathrm{lo}}) \ge B$. The unique root is therefore bracketed in an interval of width $2B/d$.

Rate. After k bisection steps, the bracket width is

$$\frac{2B}{d\,2^k}.$$

Using the Lipschitz constant $dG/2$, any point in the current bracket differs from the root in total allocated rate by at most

$$\frac{dG}{2}\frac{2B}{d\,2^k} = \frac{GB}{2^k}.$$

Thus

$$\frac{GB}{2^k} \le \varepsilon B$$

whenever

$$k \ge \log_2 \frac{G}{\varepsilon}.$$

Each iteration evaluates G clipped logarithms, so the total scalar cost is

$$O\left(G \log \frac{G}{\varepsilon}\right).$$

The group width d cancels between the initial bracket width and the Lipschitz constant. No automatic differentiation is required at deployment: Equation 12 is evaluated in the forward pass from detached scores. □

Gradient separation. The allocator consumes detached scores,

$$b_t^{\star} = \mathrm{WF}\Big(\mathrm{sg}[\widehat{C}_t]; B\Big), \tag{50}$$

so world-model, actor, and critic losses cannot train g_ω through the choice of rate allocation. Conversely, the predictor input $\mathrm{sg}[\bar{u}_t]$ and the teacher target in Equation 44 are detached from the backbone. Writing Θ for backbone parameters,

$$\mathcal{L}_{\mathrm{WM}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}, \qquad \nabla_{\Theta}\mathcal{L}_{\mathrm{con}} = 0, \qquad \nabla_{\omega}\mathcal{L}_{\mathrm{base}} = 0 \quad \text{through allocation.} \tag{51}$$

Thus the consequence head is trained to approximate the specified predictive quantity rather than being free to redefine its semantics through the control objective. By contrast, an end-to-end gate optimized directly for return is not trained to predict finite predictive damage; any correlation it exhibits with that quantity is therefore diagnostic rather than an identified target. This separation is what makes the intervention analysis informative about CAP-WM's intended allocation criterion.

Exact budget and bounded control. The budget is enforced by construction rather than through a KL or other soft penalty. At every bounded transition,

$$\sum_g b^\star_{t,g} = B$$

up to the numerical tolerance of Proposition 6. The actor, critic, and subsequent dynamics consume m_t, never the untransmitted u_t. During interaction and replay reconstruction the interface is applied to the posterior candidate; during imagination the same predictor–allocator–channel sequence is applied to the prior candidate. Thus imagined rollouts train behavior through the same bounded interface used at control time.

The consequence head is supervised from replay-derived posterior candidates and shared with the prior branch. This relies on the prior–posterior alignment already learned by the backbone; CAP-WM introduces no oracle future target inside imagined rollouts.

Complete bounded transition and teacher update. The deployment transition and the training-only teacher are separated below to make execution order and information flow explicit.

Algorithm 1 CAP-WM bounded recurrent transition

Require: previous bounded memory m_{t-1}, action a_{t-1}, observation o_t or $\varnothing$, total budget B

$(\widetilde{h}_{t-1}, \widetilde{z}_{t-1}) \leftarrow \mathrm{Split}(m_{t-1})$
$h_t \leftarrow f_\theta(\widetilde{h}_{t-1}, \widetilde{z}_{t-1}, a_{t-1})$
if $o_t \neq \varnothing$ **then**
  $z_t \sim q_\theta(\cdot \mid h_t, \mathrm{Enc}_\theta(o_t))$
else
  $z_t \sim p_\theta(\cdot \mid h_t)$
end if
$\bar{u}_t \leftarrow \mathcal{N}([h_t; \mathrm{vec}(z_t)])$
$\widehat{C}_t \leftarrow \mathrm{softplus}(g_\omega(\mathrm{sg}[\bar{u}_t])) + \epsilon$
$b^\star_t \leftarrow \mathrm{WF}(\mathrm{sg}[\widehat{C}_t]; B)$
for $g = 1, \ldots, G$ **do**
  $D_{t,g} \leftarrow \exp(-2b^\star_{t,g}/d)$
  $m_{t,g} \leftarrow (1 - D_{t,g})\bar{u}_{t,g} + \sqrt{D_{t,g}(1 - D_{t,g})}\,\varepsilon_{t,g}$
end for
return m_t

Algorithm 2 Training-only consequence-teacher update

Require: replay candidate $\bar{u}_t$, replay action continuation $a_{t:t+H_{\max}-1}$, total budget B

$\widehat{C}_t \leftarrow \mathrm{softplus}(g_\omega(\mathrm{sg}[\bar{u}_t])) + \epsilon$
Sample m^{ref}_t from the channel with $b^{\mathrm{ref}}_g = B/G$
Roll the prior forward under replay actions using the executed categorical surrogate; hold the realized stochastic trace fixed for this teacher graph
$\widetilde{C}^{\mathrm{path}}_{t,g} \leftarrow \sum_{h\in\mathcal{H}} \gamma_h \frac{1}{K} \sum_{k=1}^{K} \|J_{t,g,h} v^{(k)}\|_2^2$, $v^{(k)} \sim \mathrm{Rad}$, $K = 2$
$\omega \leftarrow \omega - \eta \nabla_\omega \frac{1}{G} \sum_g \left[\log \widehat{C}_{t,g} - \log(\mathrm{sg}[\widetilde{C}^{\mathrm{path}}_{t,g}] + \epsilon)\right]^2$

A.5 Full Linear–Gaussian Characterization

Consider Equation 13 and compare two trajectories under the same open-loop actions and process-noise realization, differing only in the current error $e_t = x_t - \hat{x}_t$. The common action and noise terms cancel, so

$$e_{t+h} = A^h e_t,$$

and the horizon-h readout error is $CA^h e_t$. Therefore Equation 14 holds with

$$G_{\mathcal{H}} = \sum_{h\in\mathcal{H}} \gamma_h (A^h)^\top C^\top C A^h \succeq 0.$$

Restricting the current error to group g gives

$$c_g = \mathrm{tr}\big(P_g^\top G_{\mathcal{H}} P_g\big)\,,$$

which is Equation 15 and is exactly the linear specialization of the intrinsic consequence in Equation 5. Because $CA^h P_g$ does not depend on the operating rate allocation, each c_g is allocation independent. Hence the fixed-profile hypothesis of Proposition 1 is exact in this linear system. Importantly, this conclusion requires neither group block-diagonality nor within-group isotropy; those stronger assumptions enter only in the groupwise specialization below.

Proof of Proposition 4. Let $B > 0$ and

$$x_t \sim \mathcal{N}(0, \Sigma)$$

and define

$$d_{\mathcal{H}}(x_t, \hat{x}_t) = \mathbb{E}\left[(x_t - \hat{x}_t)^\top G_{\mathcal{H}}(x_t - \hat{x}_t)\right].$$

Since $G_{\mathcal{H}} \succeq 0$, let

$$L = G_{\mathcal{H}}^{1/2}, \qquad q_t = L x_t.$$

Then

$$d_{\mathcal{H}}(x_t, \hat{x}_t) = \mathbb{E}\,\|q_t - L\hat{x}_t\|_2^2\,.$$

For any test channel $p(\hat{x}_t \mid x_t)$, define $\hat{q}_t = L\hat{x}_t$. This preserves distortion, and data processing gives

$$I(q_t; \hat{q}_t) \le I(x_t; \hat{x}_t).$$

Thus any feasible code for the original weighted problem induces a feasible code for q_t with no larger rate and the same distortion.

Conversely, q_t lies in $\mathrm{range}(L)$. For squared-error distortion, any reconstruction $\hat{q}_t$ may be projected onto $\mathrm{range}(L)$ without increasing distortion or mutual information, so we may restrict attention to $\hat{q}_t \in \mathrm{range}(L)$. Given any such test channel $p(\hat{q}_t \mid q_t)$, define

$$\hat{x}_t = L^\dagger \hat{q}_t,$$

where $L^\dagger$ is the Moore–Penrose pseudoinverse. Then $L\hat{x}_t = \hat{q}_t$, so the original weighted distortion equals $\mathbb{E}\|q_t - \hat{q}_t\|_2^2$. Moreover, $x_t \to q_t \to \hat{q}_t \to \hat{x}_t$ is a Markov chain, hence

$$I(x_t; \hat{x}_t) \le I(q_t; \hat{q}_t).$$

Therefore the weighted Gaussian rate–distortion problem is equivalent to ordinary squared-error Gaussian rate–distortion coding of q_t on its nonzero subspace.

The covariance of q_t is

$$\Sigma_q = G_{\mathcal{H}}^{1/2} \Sigma G_{\mathcal{H}}^{1/2}.$$

Its nonzero eigenvalues coincide with those of

$$\Sigma^{1/2} G_{\mathcal{H}} \Sigma^{1/2},$$

because these two matrices have the nonzero spectra of $AA^\top$ and $A^\top A$ for $A = G_{\mathcal{H}}^{1/2}\Sigma^{1/2}$. Denote these eigenvalues by $\eta_1, \ldots, \eta_r > 0$. Classical Gaussian rate–distortion theory diagonalizes the corresponding predictive modes and allocates

$$R_i^\star = \frac{1}{2}\left[\log \frac{\eta_i}{\lambda}\right]_+, \qquad \sum_{i=1}^{r} R_i^\star = B,$$

which is Equation 16. Directions with $\eta_i = 0$ incur no predictive distortion and optimally receive zero rate. □

Corollary 1 (Shared eigenbasis)**.** *If Σ and $G_{\mathcal{H}}$ share orthonormal eigenvectors v_i, with*

$$s_i^2 = v_i^\top \Sigma v_i, \qquad \zeta_i = v_i^\top G_{\mathcal{H}} v_i,$$

then

$$\eta_i = s_i^2 \zeta_i,$$

and Proposition 4 gives

$$R_i^\star = \frac{1}{2}\left[\log \frac{s_i^2 \zeta_i}{\lambda}\right]_+ .$$

If the source is additionally standardized, $\Sigma = I$, then

$$R_i^\star = \frac{1}{2}\left[\log \frac{\zeta_i}{\lambda}\right]_+ .$$

Thus, in this shared-eigenbasis standardized case, relative rate is determined by predictive consequence rather than by mode-dependent source variance.

Corollary 2 (Standardized group-block-isotropic case)**.** *Assume $\Sigma = I$ and that the coordinate groups form an orthogonal partition, with injections $P_g \in \mathbb{R}^{D_u \times d}$ satisfying*

$$P_g^\top P_{g'} = \delta_{gg'} I_d, \qquad \sum_{g=1}^{G} P_g P_g^\top = I.$$

Assume further that the predictive Gramian has no cross-group coupling and is isotropic within each group:

$$P_g^\top G_{\mathcal{H}} P_{g'} = \begin{cases} \frac{c_g}{d} I_d, & g = g', \\ 0, & g \neq g'. \end{cases} \tag{52}$$

The diagonal condition is consistent with Equation 15, since

$$\operatorname{tr}\big(P_g^\top G_{\mathcal{H}} P_g\big) = c_g.$$

Then the d coordinates spanning group g form an invariant predictive eigenspace with eigenvalue c_g/d. Consequently, Proposition 4 reduces to the groupwise allocation

$$b_g^\star = \frac{d}{2}\left[\log \frac{c_g}{\lambda'}\right]_+, \qquad \lambda' = d\lambda, \qquad \sum_g b_g^\star = B, \tag{53}$$

with zero-consequence groups receiving zero rate by continuity. This is exactly the groupwise reverse-water-filling form of Equation 12 when the allocator is supplied the true linear consequence c_g.

Proof. For any group g,

$$\begin{aligned} G_{\mathcal{H}} P_g &= \left(\sum_{g'=1}^{G} P_{g'} P_{g'}^\top\right) G_{\mathcal{H}} P_g \\ &= \sum_{g'=1}^{G} P_{g'} \left(P_{g'}^\top G_{\mathcal{H}} P_g\right) \\ &= P_g \frac{c_g}{d} I_d. \end{aligned}$$

Hence $\operatorname{range}(P_g)$ is an invariant eigenspace of $G_{\mathcal{H}}$ and all d modes in that group have eigenvalue c_g/d. Because $\Sigma = I$, these are also the predictive rate–distortion eigenmodes in Proposition 4. Each active mode therefore receives

$$\frac{1}{2}\left[\log \frac{c_g/d}{\lambda}\right]_+ = \frac{1}{2}\left[\log \frac{c_g}{d\lambda}\right]_+ .$$

Summing over the d modes in group g and defining $\lambda' = d\lambda$ yields Equation 53. □

What the characterization does and does not establish. Proposition 4 is a characterization of the full Gaussian predictive rate–distortion problem. It requires neither a shared eigenbasis nor any predefined grouping: the optimal transform directions are the predictive eigenmodes of

$$\Sigma^{1/2} G_{\mathcal{H}} \Sigma^{1/2},$$

so source uncertainty and predictive geometry jointly determine rate.

Corollary 1 separates these two factors when Σ and $G_{\mathcal{H}}$ share an eigenbasis. Corollary 2 is a stricter specialization: the source is standardized, the predefined group subspaces have no cross-group predictive coupling, and $G_{\mathcal{H}}$ is isotropic within each group. Only under these additional structural conditions does the unrestricted Gaussian optimum reduce exactly to CAP-WM's fixed-group consequence-weighted water-filling form.

CAP-WM does not assume that these conditions hold in the learned nonlinear world model. Group-wise standardization removes marginal scale but does not whiten cross-group correlations, force $P_g^\top G_{\mathcal{H}} P_{g'} = 0$ for $g \neq g'$, impose within-group isotropy, or align the fixed groups with the globally optimal predictive eigenbasis. The nonlinear method additionally replaces the global quadratic geometry by a local predictive Jacobian evaluated at a common reference channel and uses $D = \exp(-2b/d)$ as a nominal Gaussian distortion surrogate.

The scoped theoretical connection is therefore the following: the full linear–Gaussian problem allocates rate according to the spectrum of $\Sigma^{1/2} G_{\mathcal{H}} \Sigma^{1/2}$; under the stated standardized group-block-isotropic specialization, this optimum reduces exactly to the same groupwise reverse-water-filling form used by CAP-WM. This correspondence motivates the allocator but is not a global optimality theorem for the nonlinear neural model. The remaining theory-to-neural gaps are summarized in Appendix A.6.

A.5.1 Learned Geometry Relative to the Group Specialization

Corollary 2 gives sufficient structural conditions under which the unrestricted Gaussian predictive-rate solution reduces exactly to CAP-WM's fixed-group rule. CAP-WM does not assume that these conditions hold exactly in the nonlinear world model. We therefore measure both the degree to which the learned predictive geometry approaches this specialization and the predictive-distortion cost of retaining the fixed groups.

Using 16 held-out DMC-Subtle states at $\bar{B} = 0.20$, we materialize the local predictive Gramian at the common reference,

$$J_{t,h} = \frac{\partial F_{\theta,h}(m_t^{\mathrm{ref}}, a_{t:t+h-1})}{\partial m_t}, \qquad \widehat{G}_{t,\mathcal{H}} = \sum_{h \in \mathcal{H}} \gamma_h J_{t,h}^\top J_{t,h}. \tag{54}$$

Because this is a one-time offline geometric diagnostic, we materialize the Jacobian with basis-vector JVPs rather than the stochastic $K = 2$ estimator used during training.

Cross-group concentration. We measure the fraction of squared Frobenius energy contained inside the predefined group blocks:

$$\eta_{\mathrm{block}} = \frac{\sum_{g=1}^{G} \left\| P_g^\top \widehat{G}_{t,\mathcal{H}} P_g \right\|_F^2}{\left\| \widehat{G}_{t,\mathcal{H}} \right\|_F^2}. \tag{55}$$

For an entry-exchangeable squared-energy reference with equal-width groups, the expected within-block fraction is simply the fraction of matrix entries lying inside the G diagonal blocks,

$$\eta_{\mathrm{null}} = \frac{G d^2}{D_u^2} = \frac{1}{G} = 0.0625. \tag{56}$$

We use this only as a coordinate-level reference scale; it is not a generative model for a random positive-semidefinite Gramian.

Within-group isotropy. Let

$$c_{t,g} = \mathrm{tr}\left(P_g^\top \widehat{G}_{t,\mathcal{H}} P_g \right).$$

We define

$$\epsilon_{\mathrm{iso}} = \frac{\sum_g \left\| P_g^\top \widehat{G}_{t,\mathcal{H}} P_g - \frac{c_{t,g}}{d} I_d \right\|_F^2}{\sum_g \left\| P_g^\top \widehat{G}_{t,\mathcal{H}} P_g \right\|_F^2}. \tag{57}$$

Condition	$\eta_{\text{block}} \uparrow$	$\epsilon_{\text{iso}} \downarrow$	Excess distortion $\downarrow$
Original $G = 16$	$0.72\,[0.68, 0.76]$	$0.42\,[0.38, 0.47]$	$8.6\%\,[6.3, 11.5]$
Fixed random permutation	$0.70\,[0.66, 0.74]$	$0.44\,[0.40, 0.49]$	$9.4\%\,[6.9, 12.6]$
Uniform allocation	—	—	$48.9\%\,[41.0, 58.2]$

Table 5: **Neural predictive geometry relative to the exact group-block-isotropic specialization.** Excess distortion is measured relative to the unrestricted predictive-eigenmode solution under the same standardized local Gaussian surrogate. Brackets denote 95% bootstrap intervals over the sampled held-out states. The Uniform row provides a scale reference for the cost of discarding consequence-adaptive rate allocation.

Because $(c_{t,g}/d)I_d$ is the Frobenius projection of each block onto the isotropic subspace, $1 - \epsilon_{\text{iso}}$ is the fraction of within-block squared Frobenius energy captured by the isotropic component.

Predictive-distortion price. Under the standardized $\Sigma = I$ local Gaussian surrogate, let $\nu_{t,1}, \ldots, \nu_{t,r}$ denote the nonzero eigenvalues of $\widehat{G}_{t,\mathcal{H}}$. The unrestricted predictive-eigenmode allocation is

$$R_{t,i}^{\text{eig}} = \frac{1}{2}\left[\log \frac{\nu_{t,i}}{\lambda_t}\right]_+, \qquad \sum_i R_{t,i}^{\text{eig}} = B, \tag{58}$$

with predictive distortion

$$D_t^{\text{eig}} = \sum_i \nu_{t,i} e^{-2R_{t,i}^{\text{eig}}}. \tag{59}$$

For CAP-WM's fixed groups,

$$b_t^{\text{grp}} = \text{WF}(c_t; B), \qquad D_t^{\text{grp}} = \sum_g c_{t,g} e^{-2b_{t,g}^{\text{grp}}/d}. \tag{60}$$

Although $\widehat{G}_{t,\mathcal{H}}$ can contain nonzero cross-group blocks, Equation 60 is the expected distortion of the groupwise Gaussian surrogate used here: the reconstruction error is zero-mean with block-diagonal isotropic covariance, so cross-group quadratic terms vanish in expectation and only the diagonal-block traces $c_{t,g}$ contribute. The off-block geometry is nevertheless retained by the unrestricted eigenmode solution through the spectrum of $\widehat{G}_{t,\mathcal{H}}$ and therefore affects the gap below.

The corresponding uniform allocation has

$$D_t^{\text{unif}} = e^{-2B/(Gd)} \sum_g c_{t,g}. \tag{61}$$

We report excess distortion relative to the unrestricted predictive-eigenmode solution:

$$\Delta_{\text{grp}} = \frac{D_t^{\text{grp}}}{D_t^{\text{eig}}} - 1, \qquad \Delta_{\text{unif}} = \frac{D_t^{\text{unif}}}{D_t^{\text{eig}}} - 1. \tag{62}$$

Interpretation criterion. Large η_{block} together with a small Δ_{grp} would indicate that the exact group specialization is a useful local approximation. Low block concentration but a modest distortion gap would instead indicate that the structural conditions are not literally satisfied while the practical cost of the fixed-group restriction remains limited. A large Δ_{grp} would show that fixed grouping is materially suboptimal even within the local Gaussian surrogate and would restrict Corollary 2 primarily to qualitative motivation.

The learned predictive geometry is substantially more block concentrated than the unstructured-coordinate reference. Under the original $G = 16$ grouping, $0.72\,[0.68, 0.76]$ of the squared Gramian energy lies inside the predefined group blocks, approximately $11.5\times$ the entry-exchangeable reference level $1/G = 0.0625$. The geometry is therefore strongly block concentrated, although it is not exactly block diagonal.

Within the blocks,

$$\epsilon_{\text{iso}} = 0.42\,[0.38, 0.47],$$

so the isotropic projection captures approximately 58% of the within-block squared Frobenius energy. The exact within-group isotropy condition of Corollary 2 is therefore only partially realized, and the corollary should not be interpreted as an exact empirical description of the learned neural geometry.

The more operational comparison is the predictive-distortion gap. Consequence-weighted fixed-group allocation incurs $8.6\%\,[6.3, 11.5]$ excess distortion relative to the unrestricted predictive-eigenmode solution, whereas Uniform incurs $48.9\%\,[41.0, 58.2]$. At the point estimate, the fixed-group allocator therefore recovers approximately

$$1 - \frac{8.6}{48.9} = 82.4\%$$

of the distortion gap between Uniform and the unrestricted predictive-eigenmode solution. Thus most of the gain available from nonuniform predictive allocation is retained despite imperfect block structure and residual within-group anisotropy.

Under the fixed coordinate permutation, $\eta_{\mathrm{block}} = 0.70\,[0.66, 0.74]$, $\epsilon_{\mathrm{iso}} = 0.44\,[0.40, 0.49]$, and the excess fixed-group distortion is $9.4\%\,[6.9, 12.6]$. These values are close to those obtained with the original grouping and are consistent with the broader grouping robustness result in Appendix B.12, where a fixed coordinate permutation changes mean return only modestly.

This diagnostic is deliberately conditioned on the standardized $\Sigma = I$ surrogate used in Corollary 2. Accordingly, Δ_{grp} quantifies the combined restriction imposed by predefined coordinate groups and a single scalar rate per group. It does *not* include additional distortion arising from residual non-whitened source covariance. Quantifying that separate effect would require a covariance-aware Gaussian test channel and lies outside the scope of this diagnostic.

A.6 Scope and Limitations

Nominal rate and exact-bit validation. The variables B and $\bar{B}$ are nominal Gaussian rate parameters. The identities $D = \exp(-2b/d)$ and $I(\bar{u}_g; m_g) = b_g$ hold for an ideal standardized Gaussian source, not for arbitrary learned neural features. The separately trained serialized experiment therefore serves a different purpose: it validates the allocation principle under a directly countable recurrent-feature payload. State-dependent allocation metadata is not hidden inside the feature count; its explicit overhead is analyzed separately in Appendix B.2.

Locality. Lemma 1 is a local first-order statement for an isotropic perturbation family whose second moment matches the ideal channel-error scale. CAP-WM evaluates sensitivities once at the common reference and does not re-linearize the nonlinear world model for every candidate rate vector. It therefore does not solve a globally non-convex nonlinear rate–distortion problem. The lemma concerns predictive deviation; the finite-intervention experiment measures predictive-error increase relative to a replay target, whose additional curvature–residual term is made explicit in Remark 4.

Categorical rollout derivatives. The teacher differentiates the surrogate computation graph used by the executed backbone while conditioning on the realized stochastic rollout trace. The resulting Jacobian is therefore a surrogate-gradient object associated with that implemented stochastic computation, not a classical derivative of the discrete sampling map itself. Appendix A.3 states the differentiation convention used throughout the paper.

Measurement convention. Proposition 1 isolates one specific failure mechanism: source-referred importance is attenuated by the assignment through which it is measured. A memory-referred score lacks this multiplicative gain, although its nonlinear operating point can still move with b. Proposition 5 shows why the common reference makes source- and memory-referred conventions allocation equivalent at the same linearization point. We therefore claim a structural obstruction for assignment-attenuated refreshed measurement, not for adaptive allocation in general.

Score semantics. C^{path} is an open-loop predictive sensitivity under a fixed future action continuation. It is not a causal measure of reward, value, or closed-loop policy importance. The behavioral intervention is complementary evidence that predictive ranking has downstream relevance; it does not change the semantics of the teacher.

Replay conditioning. The deployable head predicts a replay-conditional consequence target while observing only the current candidate state. If the behavior policy or dynamics change quickly, replay can lag the current policy. CAP-WM provides no formal distribution-shift bound. The current-policy diagnostic quantifies this mismatch empirically at matched states but does not eliminate it.

Grouping and source covariance. The partition is fixed and equal width. Standardization removes marginal scale but does not whiten cross-group covariance, block-diagonalize $G_{\mathcal{H}}$, or learn the globally optimal predictive eigenbasis. The grouping and local-Gramian diagnostics quantify these departures. The fixed-group Gaussian distortion comparison is conditioned on the standardized $\Sigma = I$ surrogate and therefore does not account for residual non-whitened source covariance.

Theory-to-neural gap. The theory isolates tractable components of the allocation problem rather than establishing global optimality of the complete neural system. Proposition 1 is exact for a fixed intrinsic consequence profile; Appendix A.2.1 shows that the profile can move in the nonlinear model and compares that movement with the additional source-gain pathway. Proposition 4 gives the unrestricted linear–Gaussian predictive optimum; Corollary 2 recovers CAP-WM's groupwise rule only under a standardized group-block-isotropic specialization. Appendix A.5.1 measures how closely the learned local predictive geometry approximates this structure. These diagnostics quantify the theory-to-neural gap rather than close it with a uniform nonlinear guarantee.

Applicability beyond the primary RSSM setting. Definition 1 requires an identifiable latent-state boundary through which subsequent prediction and control must pass. Appendix C.1 evaluates this principle on TD-MPC2, where the bounded interface is inserted at a distinct deterministic latent-dynamics boundary. This supports transfer beyond the primary RSSM-style implementation, but not architecture independence: models without an identifiable state or token boundary require a separate definition of the no-bypass interface and of how allocation metadata configures it.

B Additional Experiments and Results

This appendix provides the complete experimental evidence underlying the main claims. We first specify the matched bounded-interface protocol and evaluate resource efficiency under both nominal and exactly serialized recurrent budgets. We then examine alternative allocation signals, finite predictive and closed-loop interventions, effective-horizon and prospective boundary tests, nonstationarity, reference-point reachability and robustness, and the mixed-precision structural analogue. Learning curves and unsorted run-level outcomes provide the corresponding transparency record.

B.1 Full Experimental Setup

Benchmarks. We use four complementary visual-control evaluations. *DeepMind Control Suite* (DMC) (Tassa et al., 2018) provides continuous control from pixels; we follow the 20-task protocol and train for 1M environment interactions per task. *DMC-Subtle* (Morihira et al., 2026) contains Ball in Cup Catch, Cartpole Swingup, Finger Turn, Point Mass, and Reacher variants whose task-critical objects are reduced to only a few pixels. *DMC-GB2* (Almuzairee et al., 2024) contains Walker Walk, Walker Stand, Cheetah Run, Finger Spin, Cartpole Swingup, and Cup Catch, with held-out photometric and geometric test distributions; agents train only on clean observations and are evaluated zero-shot under the shifts. Finally, *Atari 100K* (Kaiser et al., 2020) evaluates sample-efficient discrete-action control on all 26 games.

Common bounded interface. All matched DreamerV3 bounded-interface comparisons (Hafner et al., 2025) use the same numerical recurrent feature $u_t \in \mathbb{R}^{1536}$, $G = 16$ groups of width $d = 96$, interaction budget, actor–critic objective, and hard per-transition resource. The nominal experiments use the stochastic Gaussian channel of the main method, whereas the serialized experiments replace only this transmission channel by the fixed-width quantization boundary described below. The main CAP-WM configuration uses

$$\mathcal{H} = \{1, 2, 4, 8\}, \qquad \gamma_h = \frac{1}{4}, \qquad K_{\mathrm{JVP}} = 2, \tag{63}$$

and constructs one consequence-teacher target every four learner updates. The configuration is shared across tasks; the rate, horizon set, grouping, and teacher frequency are never tuned per environment.

Consequence target and reference point. For each horizon h, $F_{\theta,h}$ predicts the detached standardized posterior target v^+_{t+h} obtained from replay observation o_{t+h}. Teacher sensitivities are evaluated around the same uniform reference channel for every group and under the replayed open-loop action continuation, yielding $C^{\mathrm{path}}_{t,g}$. The common reference prevents importance from being defined around the adaptive allocation it is itself intended to choose. The deployable predictor sees only the current standardized candidate state; future replay actions are used to construct supervision and are not available to the online allocator.

Matched control taxonomy. **Uniform** assigns

$$b_{t,g} = \frac{B}{G}. \tag{64}$$

Shuffled consequence is evaluated at the tight nominal operating point $\bar{B} = 0.20$ on the stress-test conditions, where the correspondence between predicted consequence and allocated precision has the most leverage; it is not run across the dense rate sweep or the serialized payload grid. It first computes CAP-WM's complete state-dependent rate vector and then independently permutes its entries over latent groups before transmission. It preserves the consequence-predictor architecture, parameter count, total rate, and complete per-state rate histogram while breaking systematic rate–consequence alignment.

Surprise uses the same water-filling allocator with

$$S^{\mathrm{surp}}_{t,g} = \frac{1}{d}\left\|\bar{u}_{t,g} - \bar{u}^{\mathrm{prior}}_{t,g}\right\|_2^2 + \epsilon. \tag{65}$$

One-step consequence restricts CAP-WM to $\mathcal{H} = \{1\}$:

$$S^{\mathrm{1step}}_{t,g} = \|J_{t,g,1}\|_F^2 = \mathrm{tr}\left(J^\top_{t,g,1} J_{t,g,1}\right). \tag{66}$$

For squared prediction error this controls directly for local Gauss–Newton/Hessian-trace-style sensitivity.

Static task sensitivity uses a fixed training-only calibration set $\mathcal{D}_{\mathrm{cal}}$ to compute

$$S^{\mathrm{static}}_{g} = \mathbb{E}_{\mathcal{D}_{\mathrm{cal}}}\left[\left\|\nabla_{m_{t,g}} \mathcal{L}_{\mathrm{AC}}\right\|_2^2\right] + \epsilon, \tag{67}$$

then freezes this profile and supplies it to the same water-filling allocator. It is task aware but neither state conditioned nor multi-step predictive.

Multi-step value sensitivity is the closest dynamic control-oriented alternative. Using the same replay continuation, common reference, grouping, and horizons as CAP-WM, its pathwise teacher is

$$C^{V,\mathrm{path}}_{t,g} = \sum_{h\in\mathcal{H}} \gamma_h \left\| \frac{\partial V_\psi\left(F_{\theta,h}(m^{\mathrm{ref}}_t, a_{t:t+h-1})\right)}{\partial m^{\mathrm{ref}}_{t,g}} \right\|_2^2. \tag{68}$$

A deployable predictor with the same $1536 \to 256 \to 256 \to 16$ architecture and stop-gradient structure as CAP-WM regresses the replay-conditional log value-sensitivity target and supplies its positive scores to the same water-filling allocator. Hence the comparison changes the *semantic target*—future value rather than future latent prediction—while matching the recurrent budget, state conditioning, multi-step horizon, score-predictor capacity, and allocation rule.

End-to-end gate tests whether an unrestricted learned rate policy is sufficient. A parameter-matched head

$$q_t = g_\phi(\bar{u}_t), \qquad b_{t,g} = B\frac{\exp(q_{t,g})}{\sum_j \exp(q_{t,j})} \tag{69}$$

uses the same $1536 \to 256 \to 256 \to 16$ architecture as the consequence head. Unlike CAP-WM, gradients from the ordinary world-model and actor–critic objectives flow through the channel and into g_ϕ; there is no consequence teacher and no semantic stop-gradient constraint. The softmax enforces non-negative rates whose sum is exactly B at every transition.

Exact serialized extensions for the strongest controls. For Multi-step value sensitivity, the deployable predictor supplies positive scores $\widehat{C}^V_{t,g}$ to the same exact integer allocator used by CAP-WM, $q^\star_t = \arg\min_{q_g\in\{0,\dots,8\},\,\sum_g q_g=Q_t} \sum_g \widehat{C}^V_{t,g} 2^{-2q_g}$, whose greedy one-bit-at-a-time solution is exact because marginal gains decrease monotonically with q_g. For the End-to-end gate, logits $\ell_{t,1:G}$ define $p_{t,g} = \exp(\ell_{t,g})/\sum_j \exp(\ell_{t,j})$ and real-valued bit targets $r_{t,g} = Q_t p_{t,g}$; these are mapped to integer widths by a bounded largest-remainder projection $q_t = \Pi^{\mathrm{BLR}}_{Q_t}(r_t)$ satisfying $q_{t,g} \in \{0,\dots,8\}$ and $\sum_g q_{t,g} = Q_t$. The projection assigns $q_{t,g} = \min\{8, \lfloor r_{t,g}\rfloor\}$ and distributes remaining bits in descending order of fractional remainder among groups with $q_{t,g} < 8$ until the exact total is reached. During training, gradients through the integer projection use a straight-through estimator with $Q_t p_{t,g}$ as the backward surrogate. Thus both added controls obey exactly the same discrete recurrent-feature payload as CAP-WM while retaining their original allocation semantics.

VIB-WM (Alemi et al., 2017) is a generic compression comparator. Its KL coefficient is selected by a fixed development protocol to obtain an average KL magnitude near the numerical operating point. Because VIB imposes a soft expected KL while CAP-WM imposes a hard per-transition nominal Gaussian budget, it is not treated as an exact rate match.

Unconstrained DreamerV3 and R2-Dreamer are contextual references rather than matched controls.

Offline pathwise-teacher diagnostic. The replay-computed $C^{\mathrm{path}}_{t,g}$ is additionally retained as a non-deployable held-out diagnostic. Because it conditions on the realized replay action continuation, it is never used to choose online actions or to report behavioral return. It is used only to quantify predictor–teacher agreement, intervention headroom, and the gap between the deployable score and the pathwise quantity it is trained to predict.

Nominal rate budgets. We report

$$\bar{B} = \frac{B}{Gd} = \frac{B}{1536} \tag{70}$$

in nominal Gaussian nats per transmitted coordinate. The dense sweep

$$\bar{B} \in \{0.05, 0.10, 0.20, 0.40, 0.80, 1.60\}$$

corresponds under Uniform to nominal distortions

$$\exp(-2\bar{B}) = \{0.905, 0.819, 0.670, 0.449, 0.202, 0.041\}.$$

The sweep is evaluated on all six DMC-GB2 training tasks. The 20-task DMC and DMC-Subtle suites use $\bar{B} \in \{0.20, 0.80\}$. No rate is tuned per environment. Outside the exactly Gaussian reference model, $\bar{B}$ is a controlled nominal rate parameter rather than a measurement of empirical mutual information.

Operational fixed-payload channel. To validate the allocation principle under a resource that is directly countable in bits, we separately replace the Gaussian channel by a serialized fixed-width quantization boundary. For each group, $q_{t,g} \in \{0,\dots,8\}$ denotes transmitted bits per coordinate, giving the exact recurrent payload

$$K_t = d\sum_{g=1}^{G} q_{t,g} \qquad \text{bits/transition.} \tag{71}$$

We externally fix

$$\frac{K_t}{D_u} \in \{2, 4, 6\} \qquad \text{bits/coordinate} \tag{72}$$

for every transition. Standardized coordinates are clipped to $[-4, 4]$ and uniformly quantized with $2^{q_{t,g}}$ reconstruction levels for $q_{t,g} \geq 1$; $q_{t,g} = 0$ transmits the group as zero. All methods are trained from scratch with the identical quantizer and straight-through gradient estimator. Uniform distributes the integer bit widths as evenly as possible. Adaptive methods greedily assign each remaining group-bit to the largest marginal reduction in the separable surrogate

$$\sum_g S_{t,g} 2^{-2q_{t,g}}, \tag{73}$$

where $S_{t,g}$ is Surprise, One-step consequence, or CAP-WM consequence. This greedy assignment is not an approximation. The objective is separable across groups and the marginal reduction obtained by granting group g its $(q+1)$-th bit,

$$S_{t,g}\left(2^{-2q}-2^{-2(q+1)}\right)=\tfrac{3}{4}\,S_{t,g}\,2^{-2q}, \tag{74}$$

is strictly decreasing in q. For a separable objective with decreasing marginal returns under a cardinality constraint, greedy marginal assignment attains the integer optimum, so the discrete allocator solves Equation 73 exactly under the prescribed payload. The continuous and discrete channels are therefore both exact-budget allocators, differing only in whether rate is real valued or integral, which is what makes the two experiments comparable. The payload count excludes packet headers and is exactly the packed recurrent feature payload. This validation is separate from the nominal Gaussian channel and does not equate B with physical bits.

Evaluation and uncertainty. Unless explicitly extended, DMC-style experiments use five independent training seeds and 10 evaluation episodes per seed. The confirmatory CAP-WM–One-step Subtle/Video experiment uses 10 training seeds. DMC-style aggregates report mean, median, and IQM as appropriate; Atari uses human-normalized IQM, optimality gap, median, and mean. Aggregate 95% intervals use stratified bootstrap resampling (Agarwal et al., 2021). Offline diagnostics on held-out replay states draw from a single held-out DMC-Subtle pool at $\bar{B}=0.20$: the Gramian analysis uses 16 states (Appendix A.5.1), the basin and isolated-pathway analysis 256 (Appendix B.12), and the drift statistics 512 (Appendix A.2.1).

For paired comparisons, let T denote the aggregate statistic being reported and let X and Y contain matched task/game–seed runs. The estimand is

$$\widehat{\Delta}=T(X)-T(Y). \tag{75}$$

A bootstrap replicate jointly resamples the matched units and computes

$$\Delta^{(b)}=T(X^{(b)})-T(Y^{(b)}). \tag{76}$$

Pairing therefore affects the resampling distribution and interval, while the point estimate remains the difference of the two reported aggregates. This is distinct from computing $T(X-Y)$. Probability of improvement is the fraction of paired replicates for which $\Delta^{(b)}>0$.

Two paired comparisons are primary and were fixed before the corresponding runs: CAP-WM versus Uniform, which tests whether state-dependent allocation helps at all, and CAP-WM versus One-step consequence, which tests the multi-step contribution that distinguishes the method. Every other paired comparison reported in this appendix — against Surprise, Static task sensitivity, the End-to-end gate, and Multi-step value sensitivity — is exploratory and characterizes alternative allocation semantics rather than testing a pre-specified hypothesis. All intervals are reported at the nominal 95% level without multiplicity adjustment. Because several exploratory comparisons are drawn at the same operating point, individual exploratory intervals should be read as descriptive; conclusions that depend on resolution rest on the two primary comparisons, on the confirmatory 10-seed Subtle/Video experiment, and on the within-method horizon analysis, none of which are selected post hoc.

B.2 Exact Serialized-Payload Validation

Relative to the no-channel DreamerV3 reference $S_{\mathrm{DV3}}=820$, define the operational retention threshold

$$K_{90}=\min\{K: S(K)\geq 0.9S_{\mathrm{DV3}}\}, \tag{77}$$

where K is the tested serialized payload in bits per transmitted coordinate. All six variants in Table 6 use the same quantized no-bypass recurrent interface and obey the same exact recurrent-feature payload at every transition. CAP-WM is the first allocator on the tested grid to reach the 90%-retention criterion, doing so at 4 bits/coordinate. Surprise, End-to-end gate, One-step consequence, and Multi-step value sensitivity require 6 bits/coordinate, while Uniform remains below the criterion over the tested range. At the tightest payload, CAP-WM retains 75.5% of the no-channel reference compared with 60.1% for Uniform, 66.8% for End-to-end gate, 68.4% for One-step consequence, and 71.5% for Multi-step value sensitivity.

Method	2 bits/coord.		4 bits/coord.		6 bits/coord.	
	IQM	Ret.	IQM	Ret.	IQM	Ret.
Uniform	493 [468, 517]	60.1%	648 [623, 672]	79.0%	735 [713, 756]	89.6%
Surprise	522 [497, 546]	63.7%	671 [646, 695]	81.8%	749 [727, 770]	91.3%
End-to-end gate	548 [523, 572]	66.8%	690 [666, 713]	84.1%	776 [754, 797]	94.6%
One-step consequence	561 [536, 585]	68.4%	699 [675, 722]	85.2%	785 [764, 805]	95.7%
Multi-step value sensitivity	586 [561, 610]	71.5%	721 [697, 744]	87.9%	**786** [765, 806]	**95.9%**
CAP-WM	**619** [594, 643]	**75.5%**	**739** [715, 762]	**90.1%**	783 [762, 802]	95.5%

Table 6: **DMC-GB2 IQM under an exact serialized recurrent-feature payload with the strongest allocation controls.** Brackets are stratified-bootstrap 95% intervals. "Ret." denotes retention relative to the unconstrained DreamerV3 reference $S_{\mathrm{DV3}} = 820$; Equation 77 defines the operational threshold. Bold marks the best point estimate per payload.

Payload	CAP-WM − One-step	95% paired CI	$P(\text{improvement})$
2 bits/coord.	$+58$	$[28, 87]$	0.99
4 bits/coord.	$+40$	$[13, 66]$	0.97
6 bits/coord.	-2	$[-22, 18]$	0.44

Table 7: **Primary paired CAP-WM–One-step comparisons under the exact serialized payload.**

The CAP-WM–One-step separation remains strongly rate dependent. At 2 and 4 bits/coordinate, the paired differences are $+58\,[28, 87]$ and $+40\,[13, 66]$, respectively. At 6 bits/coordinate, the difference is $-2\,[-22, 18]$ and is unresolved. The stronger exact-payload controls sharpen the same boundary. Relative to the End-to-end gate, CAP-WM improves IQM by $+71\,[40, 101]$ at 2 bits and $+49\,[21, 77]$ at 4 bits, while the 6-bit difference $+7\,[-14, 28]$ is unresolved. Relative to Multi-step value sensitivity, CAP-WM improves by $+33\,[5, 61]$ at 2 bits, while the 4- and 6-bit differences, $+18\,[-6, 42]$ and $-3\,[-24, 18]$, include zero. Thus CAP-WM remains separated from the strongest learned and control-oriented comparators under the most restrictive payload, whereas the methods become statistically difficult to distinguish as the interface becomes permissive.

Allocation-map side information. The feature-payload count above excludes packet framing. If the state-dependent integer allocation vector $q_{t,1:G}$ must also be communicated explicitly, a naive four-bit code per group costs $16\times4 = 64$ bits/transition, or 0.0417 bits/coordinate. A tighter enumerative code exploits the known total $Q = \sum_g q_g$. The numbers of admissible vectors are $N_Q(32) = 513{,}374{,}088{,}549$, $N_Q(64) = 70{,}886{,}845{,}397{,}481$, and by symmetry $N_Q(96) = 513{,}374{,}088{,}549$, requiring 39, 47, and 39 bits, respectively. These correspond to effective rates 2.0254, 4.0306, and 6.0254 bits/coordinate; the conservative naive code gives 2.0417, 4.0417, and 6.0417. Hence, explicit allocation-map accounting does not change the operational K_{90} separation: using enumerative coding, CAP-WM reaches the retention criterion at 4.0306 effective bits/coordinate, whereas the tested adaptive controls that cross the criterion require 6.0254 bits/coordinate, corresponding to a 33.1% reduction in effective recurrent payload. This comparison covers the serialized recurrent feature and its allocation map; it does not claim an end-to-end reduction in hardware energy or total system traffic.

The exact-payload experiment therefore reproduces and strengthens the principal resource-dependent pattern observed with the nominal channel: CAP-WM separates most strongly from local sensitivity and stronger adaptive alternatives when precision is scarce, while those differences become unresolved as the interface becomes permissive. Because this experiment uses a distinct discrete channel family, we use it to validate the operational resource claim rather than to calibrate the nominal Gaussian rate.

B.3 Nominal Rate Frontier and Matched Controls

Paired comparisons support the budget-dependent interpretation. At $\bar{B} = 0.20$, CAP-WM improves Uniform by $+88$ IQM points with paired 95% CI $[55, 119]$ and $P(\text{improvement}) = 0.98$. Against One-step consequence, the paired difference at the same rate is $+24\,[8, 40]$ with $P(\text{improvement}) =$

Comparison	2 bits/coord.	4 bits/coord.	6 bits/coord.
CAP-WM − End-to-end gate	$+71\,[40, 101]$	$+49\,[21, 77]$	$+7\,[-14, 28]$
CAP-WM − Multi-step value sensitivity	$+33\,[5, 61]$	$+18\,[-6, 42]$	$-3\,[-24, 18]$

Table 8: **Exploratory paired comparisons against the additional strong exact-payload controls.** Intervals use the same paired stratified-bootstrap analysis as the completed exact-payload evaluation.

Feature rate	Feature bits	Enum. map	Naive map	Effective rate (enum.)	Effective rate (naive)
2 bits/coord.	3072	39	64	2.0254	2.0417
4 bits/coord.	6144	47	64	4.0306	4.0417
6 bits/coord.	9216	39	64	6.0254	6.0417

Table 9: **Explicit allocation-map overhead for the serialized interface.** The original experiment exactly controls recurrent-feature payload; the last two columns additionally encode the adaptive allocation map.

0.96. At $\bar{B} = 0.80$, CAP-WM–Uniform remains positive at $+21\,[2, 40]$, whereas CAP-WM–One-step is $-3\,[-16, 10]$. At $\bar{B} = 1.60$, CAP-WM–One-step is $-1\,[-12, 11]$. The additional multi-step benefit is therefore concentrated where the recurrent channel is restrictive; small point-estimate reversals at permissive rates are unresolved and rule out a universal-dominance interpretation.

A complementary summary is the smallest tested nominal rate at which a method retains 90% of the unconstrained reference score. Let $S_{\mathrm{DV3}} = 820$ denote the measured DMC-GB2 training IQM of unconstrained DreamerV3 and define

$$\bar{B}_{90} = \min\{\bar{B} : S(\bar{B}) \geq 0.9 S_{\mathrm{DV3}}\}. \tag{78}$$

On the evaluated grid, CAP-WM and One-step consequence first cross this criterion at $\bar{B} = 0.40$, whereas Uniform and Surprise first cross at $\bar{B} = 0.80$. This statistic shows that state-dependent predictive sensitivity is already valuable for rate efficiency. The contribution specific to multi-step propagation is identified instead by the low-rate paired comparisons, the independent Subtle/Video confirmation, and the horizon analyses below.

At $\bar{B} = 0.20$, CAP-WM improves DMC IQM from 568 to 654 relative to the matched Uniform control and improves DMC-Subtle IQM from 421 to 602. At $\bar{B} = 0.80$, the standard-DMC gain narrows, whereas the DMC-Subtle difference remains substantial. VIB-WM also remains below CAP-WM at similar average KL magnitudes. Because VIB-WM uses a soft expected KL rather than the same hard recurrent channel, this comparison is contextual and is not treated as an exact rate match.

What does the bound buy? The unconstrained references address a different question from the matched allocation controls: whether imposing the recurrent bottleneck can itself be useful. On standard DMC, the bounded CAP-WM model is nearly cost-neutral at $\bar{B} = 0.80$, reaching 793 versus 812 for unconstrained DreamerV3 (-2.3%). Under nuisance, the bounded models can instead improve over that reference. CAP-WM reaches 663 versus 505 on DMC-Subtle ($+31\%$) and 586 versus 472 on the zero-shot Photo-Hard shift ($+24\%$).

Importantly, this effect is not unique to CAP-WM. Uniform under the same bounded interface also improves over unconstrained DreamerV3, reaching 548 versus 505 on DMC-Subtle and 498 versus 472 on Photo-Hard, but its gains are substantially smaller. Thus compression, stochastic-channel regularization, or both can contribute to the constrained–unconstrained improvement. The controlled evidence for CAP-WM's allocation principle therefore comes from comparisons among methods sharing the same bounded interface, which isolate the additional effect of *where* the available precision is spent.

We consequently treat the constrained–unconstrained comparisons as contextual rather than as causal estimates of the value of consequence-aware allocation. They show that the imposed bottleneck need not be purely detrimental under nuisance, while the matched bounded-interface controls determine which allocation signal is preferable.

Method	0.05	0.10	0.20	0.40	0.80	1.60
Uniform	306 [282, 329]	438 [413, 462]	578 [552, 603]	696 [672, 719]	777 [756, 797]	811 [793, 828]
Surprise	333 [309, 357]	472 [448, 496]	612 [587, 636]	721 [698, 743]	785 [765, 804]	813 [795, 830]
One-step consequence	352 [327, 377]	509 [484, 533]	642 [617, 666]	**760** [738, 781]	**801** [781, 820]	**817** [800, 833]
CAP-WM	**381** [356, 405]	**531** [505, 556]	**666** [641, 690]	758 [736, 779]	798 [779, 816]	816 [799, 832]

Table 10: **Full rate–performance frontier on the six DMC-GB2 training tasks.** Aggregate IQM with 95% stratified-bootstrap confidence intervals. Bold marks the best point estimate in each column. Beyond the tight-rate regime, One-step consequence has slightly higher point estimates at several rates, but the paired CAP-WM–One-step differences are unresolved.

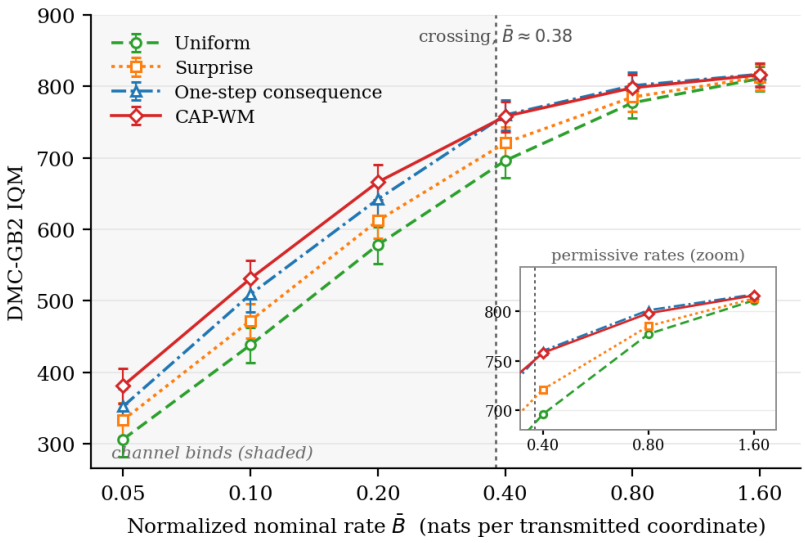


Figure 3: **Fixed-budget performance–rate frontier on DMC-GB2.** All methods share the same recurrent channel, grouping, and training budget. CAP-WM separates most strongly at low rate; from $\bar{B} = 0.40$ onward, its difference from One-step consequence is unresolved by paired intervals. Table 10 reports the complete numerical frontier.

Strong adaptive comparators. The tight DMC-GB2 operating point additionally compares alternative state-dependent allocation semantics under the same recurrent channel. Table 12 includes the frozen Static task-sensitivity profile, the parameter-matched End-to-end gate, One-step consequence, Multi-step value sensitivity, and CAP-WM.

At $\bar{B} = 0.20$, the paired CAP-WM differences are $+30\,[12, 47]$ against the End-to-end gate, $+24\,[8, 40]$ against One-step consequence, and $+16\,[1, 31]$ against Multi-step value sensitivity, with probabilities of improvement 0.97, 0.96, and 0.93, respectively. These controls distinguish several alternative explanations: unrestricted return-trained gating, local predictive sensitivity, and multi-step value sensitivity all improve over equal allocation, yet CAP-WM has the highest point estimate at the tight operating point and its paired differences against these three controls are resolved.

At $\bar{B} = 0.80$, the adaptive methods are tightly clustered and the paired intervals for adjacent comparisons include zero; One-step consequence is slightly above CAP-WM in point estimate. This is consistent with the central boundary condition rather than with universal superiority: the contribution specific to multi-step predictive consequence is clearest when precision is scarce.

B.4 Stress Tests, Representation Selectivity, and Zero-Shot Robustness

Full consequence–surprise stress test. The stress test compares allocation signals with deliberately different semantics rather than defining a universal ranking among them. Static task sensitivity improves over Uniform in the subtle conditions but cannot adapt its profile to the current state. The End-to-end gate reaches 414 on Subtle/Video versus 390 for One-step consequence, whereas their

Method	$\bar{B}$	DMC (20 tasks) IQM ↑	DMC (20 tasks) Mean ↑	DMC-Subtle (5 tasks) IQM ↑	DMC-Subtle (5 tasks) Mean ↑
Uniform	0.20	568 [545, 590]	552	421 [378, 462]	432
Surprise	0.20	602 [579, 624]	586	439 [397, 480]	458
VIB-WM	≈ 0.21	590 [566, 613]	575	449 [408, 489]	463
CAP-WM	0.20	**654** [631, 676]	**637**	**602** [561, 641]	**608**
Uniform	0.80	763 [742, 782]	747	548 [507, 588]	559
Surprise	0.80	774 [754, 793]	756	562 [522, 601]	574
VIB-WM	≈ 0.79	771 [751, 790]	754	568 [528, 607]	579
CAP-WM	0.80	**793** [774, 811]	**777**	**663** [625, 699]	**667**
DreamerV3	unconstr.	812	795	505	514
R2-Dreamer	unconstr.	818	802	690	684

Table 11: **Complete DMC and DMC-Subtle results.** VIB-WM is approximately aligned by average KL magnitude and is not an exact same-channel control. Unconstrained rows are contextual references rather than matched comparisons.

Method	$\bar{B} = 0.20$	$\bar{B} = 0.80$
Uniform	578 [552, 603]	777 [756, 797]
Static task sensitivity	629 [604, 653]	786 [766, 805]
End-to-end gate	636 [611, 660]	788 [768, 807]
One-step consequence	642 [617, 666]	**801** [781, 820]
Multi-step value sensitivity	650 [625, 674]	795 [775, 813]
CAP-WM	**666** [641, 690]	798 [779, 816]

Table 12: **Strong allocation controls on the six DMC-GB2 training tasks.** Aggregate IQM under identical within-column recurrent budgets.

ordering reverses on DMC-GB2 at $\bar{B} = 0.20$ (636 versus 642). Multi-step value sensitivity reaches 452 on Subtle/Video, compared with 504 for CAP-WM. Thus alternative adaptive signals can be competitive in individual regimes, but CAP-WM has the strongest point estimate across all four tight-budget stress conditions while retaining the predictive-consequence semantics tested independently by intervention.

Ten-seed confirmation of CAP-WM versus One-step. Because multi-step propagation is the central distinction from local sensitivity, we double the seed count only for the pre-specified CAP-WM–One-step comparison on Subtle/Video. This confirmatory run uses an independent set of ten training seeds rather than extending the five-seed stress-test runs of Table 13; the two are therefore separate samples of the same estimand (504 and 501 for CAP-WM, 390 and 397 for One-step consequence) and neither is a subset of the other.

The completed matched seed-level outcomes are reported in Appendix B.16; they are printed in run order, include a local reversal, and reproduce the aggregate means reported here.

Representation probes. We freeze representations from Subtle/Video and fit probes on held-out transitions. Simulator state and target position measure task-relevant accessibility, whereas nuisance-video identity measures linearly decodable irrelevant appearance information.

Surprise retains 68% background-identity accuracy, far above the 10% chance level, whereas CAP-WM reduces nuisance decodability to 24% while preserving substantially more task-relevant information. These probes are consistent with selective representation but are not used as causal evidence for the allocation mechanism.

DMC-GB2 zero-shot robustness. CAP-WM gives its largest gains over the bounded controls on photometric shifts, where appearance changes while simulator dynamics are unchanged. The smaller

Method	Standard/Clean	Subtle/Clean	Standard/Video	Subtle/Video
Uniform	653 [621, 684]	432 [392, 471]	531 [493, 567]	316 [278, 354]
Shuffled consequence	625 [592, 656]	402 [361, 441]	498 [459, 535]	287 [250, 324]
Surprise	671 [640, 701]	458 [417, 497]	456 [416, 494]	238 [202, 274]
Static task sensitivity	679 [648, 708]	501 [461, 539]	554 [518, 589]	347 [309, 385]
End-to-end gate	690 [659, 719]	549 [509, 587]	591 [554, 627]	414 [375, 452]
One-step consequence	686 [655, 715]	534 [494, 572]	582 [545, 617]	390 [351, 428]
Multi-step value sensitivity	700 [670, 729]	574 [534, 611]	614 [576, 648]	452 [410, 489]
CAP-WM	**707** [677, 735]	**608** [569, 645]	**636** [600, 670]	**504** [465, 541]

Table 13: **Full consequence–surprise stress test at $\bar{B} = 0.20$.** Mean return across five tasks with uncertainty intervals.

Method	Mean return	95% CI	Training seeds
One-step consequence	397	[370, 416]	10
CAP-WM	**501**	[483, 529]	10

Table 14: **Confirmatory Subtle/Video comparison.** The paired CAP-WM–One-step difference is $+104\,[63, 143]$ with $P(\text{improvement}) = 0.989$.

geometric-shift gains are consistent with, but do not by themselves prove, selective suppression of irrelevant appearance variation.

B.5 Finite-Intervention and Allocation Diagnostics

Return alone does not establish that the learned CAP-WM score has the predictive semantics claimed by the method. We therefore test the score directly by intervening on the transmitted state. For each held-out state, we perturb one group, $m_{t,g} \leftarrow m_{t,g} + \delta_{t,g}$ with $\delta_{t,g} \sim \mathcal{N}(0, \sigma^2 I)$, hold the replay action continuation fixed, and measure

$$\Delta^{\text{err}}_{t,g} = \sum_{h \in \mathcal{H}} \gamma_h \left[d\Big(F_{\theta,h}(m^{(g)}_t, a_{t:t+h-1}), v^+_{t+h} \Big) - d\big(F_{\theta,h}(m_t, a_{t:t+h-1}), v^+_{t+h} \big) \right]. \tag{79}$$

This quantity is the *open-loop predictive damage* caused by a finite representation intervention. It is deliberately aligned with CAP-WM's predictive definition of consequence and is not interpreted as a closed-loop causal measure of reward, value, or policy importance.

Equation 79 measures the increase in predictive error relative to the replay target and is therefore distinct from the predictive deviation Δ^{pred} characterized by Lemma 1. Appendix A.4 derives their local relationship for a scaled perturbation family. For the analogous finite-intervention scale used here, the curvature–residual correction is $O(\sigma^2)$, paralleling the $O(D)$ correction in the local analysis. It can therefore be of the same asymptotic order as the leading predictive-deviation term; its sign and magnitude depend on local model curvature and the unperturbed prediction residual.

The finite-intervention analysis is therefore an *empirical test* of whether the learned score predicts realized predictive damage, not a corollary of Lemma 1. This distinction also explains why the pathwise teacher can correlate strongly with Δ^{err} without being expected to achieve unit correlation.

The ordering remains qualitatively stable when the perturbation magnitude is doubled, although every correlation decreases as the intervention moves farther from the local linearization underlying the teacher. The deployable predictor reaches

$$\rho(\widehat{C}, C^{\text{path}}) = 0.78\,[0.75, 0.81],$$

while the non-deployable teacher reaches

$$\rho(C^{\text{path}}, \Delta^{\text{err}}) = 0.82\,[0.79, 0.85].$$

The first quantity measures how well the online score approximates its replay-derived teacher; the second measures how well that teacher predicts finite realized damage. Their separation exposes

Method	State NRMSE ↓	Target error (px) ↓	Background acc. ↓
Uniform	0.281 ± 0.018	8.2 ± 0.6	$29 \pm 3\%$
Shuffled consequence	0.302 ± 0.020	9.0 ± 0.7	$32 \pm 3\%$
Surprise	0.327 ± 0.021	10.1 ± 0.8	$68 \pm 4\%$
One-step consequence	0.231 ± 0.017	6.3 ± 0.5	$37 \pm 3\%$
CAP-WM	$\mathbf{0.176} \pm 0.014$	$\mathbf{4.1} \pm 0.4$	$\mathbf{24} \pm 3\%$

Table 15: **Representation probes on Subtle/Video.** Lower is better for state and target errors; lower background accuracy indicates less linearly decodable nuisance information. These probes measure linear accessibility and do not estimate mutual information or allocated rate.

Method	Clean	Photo-Easy	Photo-Hard	Geom-Easy	Geom-Hard
DreamerV3	762	622	472	610	470
Uniform	719	635	498	602	451
Surprise	731	610	432	606	459
CAP-WM	**752**	**688**	**586**	**637**	**501**

Table 16: **DMC-GB2 zero-shot return at $\bar{B} = 0.80$.** Entries are arithmetic means across the six held-out task evaluations. The unconstrained DreamerV3 row is contextual. These means are not directly comparable with the DMC-GB2 training IQM $S_{\mathrm{DV3}} = 820$ used in Equation 78.

the two approximation steps between the theoretical local quantity and the deployable allocation signal. Because Static task sensitivity is calibrated from the actor–critic objective rather than from multi-step predictive error, its correlation with Δ^{err} is reported as a cross-objective diagnostic rather than as a matched-objective test.

The quartile analysis provides a complementary non-rank-based view. After normalizing mean intervention damage to one, quartiles of $\widehat{C}$ have mean damage

$$\{0.44, 0.69, 1.02, 1.85\},$$

whereas quartiles formed by Surprise have

$$\{0.79, 0.92, 1.05, 1.24\}.$$

Thus predicted consequence separates low- from high-damage interventions far more sharply than Surprise. We additionally evaluate a matched short-horizon closed-loop intervention to test whether this predictive ranking is associated with downstream behavioral sensitivity without redefining consequence as a reward-causal quantity. For each held-out state, groups are ranked by the deployable score $\widehat{C}_{t,g}$ and divided into quartiles Q_1–Q_4. The same simulator state is cloned into clean and perturbed rollouts; the group perturbation magnitude is held fixed across quartiles and matches the finite-intervention scale above, environment stochasticity is synchronized, and both copies are rolled out for $H_{\mathrm{beh}} = 20$ environment steps under the deployed actor. We measure short-horizon percentage return loss $L_R = 100(R_{\mathrm{clean}} - R_{\mathrm{pert}})/(|R_{\mathrm{clean}}| + \epsilon)$ and normalized mean action deviation $\Delta_a = H_{\mathrm{beh}}^{-1} \sum_{\tau=0}^{H_{\mathrm{beh}}-1} \|a_{t+\tau}^{\mathrm{pert}} - a_{t+\tau}^{\mathrm{clean}}\|_2/(2\sqrt{d_a})$ for DMC actions in $[-1, 1]^{d_a}$. Return loss increases monotonically from $2.1\,[1.5, 2.8]\%$ in Q_1 to $4.7\,[3.8, 5.7]\%$, $8.2\,[7.1, 9.4]\%$, and $12.4\,[10.8, 14.0]\%$ in Q_2–Q_4, while normalized action deviation increases from $0.028\,[0.022, 0.034]$ to $0.047\,[0.040, 0.054]$, $0.076\,[0.067, 0.085]$, and $0.112\,[0.101, 0.123]$. Across all state–group interventions, $\rho(\widehat{C}, L_R) = 0.58\,[0.52, 0.64]$ and $\rho(\widehat{C}, \Delta_a) = 0.55\,[0.49, 0.61]$. These results provide an independent downstream association: groups ranked as more predictively consequential tend to produce greater behavioral disruption under the matched intervention, while we do not interpret $\widehat{C}$ as an exact causal control-importance measure.

The intervention analysis tests whether the score identifies damaging representation perturbations; the allocation analysis tests whether the deployed allocator actually assigns precision accordingly. CAP-WM reaches allocation–damage correlation $\rho = 0.63$, compared with 0.47 for One-step consequence, 0.43 for Multi-step value sensitivity, and 0.36 for the End-to-end gate. Independently shuffling CAP-WM's per-state rate vector reduces the correlation to 0.04 while preserving the same total rate and per-state rate histogram.

Signal	Small perturbation		$2\times$ perturbation	
	$\rho \uparrow$	Top-Q $\uparrow$	$\rho \uparrow$	Top-Q $\uparrow$
Latent magnitude	$0.16\,[0.12, 0.20]$	0.28	$0.13\,[0.09, 0.17]$	0.27
Surprise	$0.29\,[0.24, 0.34]$	0.35	$0.25\,[0.20, 0.30]$	0.33
Static task sensitivity	$0.34\,[0.29, 0.39]$	0.38	$0.29\,[0.24, 0.34]$	0.35
End-to-end gate score	$0.36\,[0.31, 0.41]$	0.39	$0.31\,[0.26, 0.36]$	0.36
Multi-step value sensitivity	$0.44\,[0.39, 0.49]$	0.45	$0.38\,[0.33, 0.43]$	0.41
One-step consequence	$0.49\,[0.44, 0.54]$	0.47	$0.42\,[0.37, 0.47]$	0.43
Predicted $\widehat{C}$	$\mathbf{0.66}\,[0.62, 0.70]$	**0.61**	$\mathbf{0.55}\,[0.50, 0.59]$	**0.54**
Pathwise teacher C^{path}	$0.82\,[0.79, 0.85]$	0.68	$0.68\,[0.64, 0.72]$	0.60

Table 17: **Full finite-intervention analysis.** Spearman correlation and top-quartile overlap with measured predictive damage. Random top-quartile overlap is 0.25. Bold denotes the best deployable score; the pathwise teacher is a non-deployable offline reference.

Metric	Q_1	Q_2	Q_3	Q_4
Return loss (%)	$2.1\,[1.5, 2.8]$	$4.7\,[3.8, 5.7]$	$8.2\,[7.1, 9.4]$	$12.4\,[10.8, 14.0]$
Normalized action deviation	$0.028\,[0.022, 0.034]$	$0.047\,[0.040, 0.054]$	$0.076\,[0.067, 0.085]$	$0.112\,[0.101, 0.123]$

Table 18: **Closed-loop behavioral damage by CAP-WM consequence quartile.** All entries use matched 20-step rollouts from cloned simulator states with the same finite representation-intervention scale across quartiles.

The shuffled control is important because it separates the existence of a state-dependent rate profile from its alignment with predictive consequence. Together, the score–damage and allocation–damage diagnostics support both links required by the proposed mechanism: the learned score predicts which representation perturbations are damaging, and the allocator assigns more precision to those groups. These measurements support CAP-WM's intended predictive semantics without claiming that the local teacher is an exact nonlinear, closed-loop, or reward-causal importance measure.

B.6 Multi-Step Horizon Ablation and Effective-Horizon Diagnostic

Incremental horizon evidence. The existing horizon ablation changes only the set of future transitions included in the same consequence construction. Mean DMC-Subtle return at $\bar{B} = 0.20$ is 534, 565, 594, 608, and 603 for $\{1\}$, $\{1, 2\}$, $\{1, 2, 4\}$, $\{1, 2, 4, 8\}$, and $\{1, 2, 4, 8, 16\}$, respectively. Paired analysis of successive horizon extensions is shown below.

Effective predictive horizon. For each held-out state/group pair we define the fraction of main-teacher sensitivity lying beyond one step,

$$L_{\mathrm{long}} = \frac{\sum_{h\in\{2,4,8\}} \gamma_h \|J_{t,g,h}\|_F^2}{\sum_{h\in\{1,2,4,8\}} \gamma_h \|J_{t,g,h}\|_F^2}. \tag{80}$$

Under the main configuration the weights are uniform, so γ_h cancels and L_{long} is the plain share of teacher sensitivity carried by horizons beyond one. We retain γ_h so that the diagnostic remains the correct quantity under the non-uniform weightings examined in the horizon ablation.

One calibration point helps interpret the diagnostic. Three of the four uniformly weighted horizons lie in the numerator, so an exactly even distribution of sensitivity across $\mathcal{H}$ would yield $L_{\mathrm{long}} = 0.75$. All three suites fall below this reference, indicating that the one-step term carries more than its 25% equal-share baseline. The claim therefore rests on the relative amount of sensitivity that extends beyond one transition, and particularly on the ordering across suites, rather than on interpreting the absolute value as a universal long-horizon threshold.

The diagnostic provides a data-based boundary indicator. DMC-Subtle and DMC-GB2 place more teacher sensitivity beyond one transition than Atari, and within DMC-GB2 and Atari, environments with larger long-horizon mass tend to show larger CAP-WM–One-step gains. This pattern is consistent with the weaker aggregate multi-step separation on Atari, where the predictive tail available to distinguish multi-step from one-step allocation is smaller on average.

Variant	Subtle/Clean	Subtle/Video	$\rho(b, \Delta^{\mathrm{err}})$	Top-Q overlap
Uniform	432	316	n/a	n/a
Shuffled consequence	402	287	0.04	0.26
Surprise	458	238	0.27	0.34
Static task sensitivity	501	347	0.32	0.37
End-to-end gate	549	414	0.36	0.39
Multi-step value sensitivity	574	452	0.43	0.44
One-step consequence	534	390	0.47	0.47
CAP-WM	**608**	**504**	**0.63**	**0.59**

Table 19: **Targeted allocation diagnostics at $\bar{B} = 0.20$.** The Uniform correlation is undefined because every group receives the same rate.

Change	Δ mean	95% paired CI	$P(\text{improvement})$
$\{1\} \to \{1, 2\}$	$+31$	$[11, 50]$	0.97
$\{1, 2\} \to \{1, 2, 4\}$	$+29$	$[10, 48]$	0.96
$\{1, 2, 4\} \to \{1, 2, 4, 8\}$	$+14$	$[2, 27]$	0.92
$\{1, 2, 4, 8\} \to \{1, 2, 4, 8, 16\}$	-5	$[-18, 8]$	0.29

Table 20: **Paired incremental horizon analysis on DMC-Subtle.** The benefit saturates after horizon 8 rather than increasing merely because more Jacobian terms are added.

B.7 Prospective Validation of the Effective-Horizon Principle

The cross-suite analysis suggests that CAP-WM is most useful when predictive dependence extends beyond one latent transition. Because that observation is correlational, we test the stronger prospective prediction by intervening on temporal discretization within the *same* six DMC-GB2 environments.

Controlled intervention and prospective prediction. We vary action repeat

$$r \in \{2, 4, 8\} \tag{81}$$

while keeping the environment, observation space, reward, recurrent interface, $\bar{B} = 0.20$, $G = 16$, optimizer, network sizes, and total simulator-frame budget fixed. Learner updates are normalized by simulator frames so that larger action repeat does not receive more physical interaction.

For each task, seed, and action-repeat condition, we first train a common Uniform bounded model to the pre-specified 200K-frame checkpoint. On a fixed held-out replay slice from this checkpoint we measure

$$L_{\mathrm{long}} = \frac{\sum_{h \in \{2,4,8\}} \gamma_h \|J_{t,g,h}\|_F^2}{\sum_{h \in \{1,2,4,8\}} \gamma_h \|J_{t,g,h}\|_F^2}. \tag{82}$$

Only after this measurement is recorded do we fork the matched checkpoint into One-step consequence and CAP-WM and continue both to the common final simulator-frame budget. Thus L_{long} is measured before observing the CAP-WM–One-step outcome.

The intervention yields the pre-outcome ordering

$$L_{\mathrm{long}}(r = 2) > L_{\mathrm{long}}(r = 4) > L_{\mathrm{long}}(r = 8),$$

which predicts, before the paired return comparison is evaluated,

$$\Delta_{\mathrm{CAP-1step}}(r = 2) > \Delta_{\mathrm{CAP-1step}}(r = 4) > \Delta_{\mathrm{CAP-1step}}(r = 8).$$

The $r = 4$ row is not an independent measurement. DMC's standard action repeat is four, so that condition coincides with the primary DMC-GB2 configuration and reproduces its entries exactly ($642, 666, +24\,[8, 40]$, $P(\text{improvement}) = 0.96$, $L_{\mathrm{long}} = 0.55$). The intervention therefore extends an established operating point in both directions rather than constituting a separate experiment with its own calibration.

Suite	Mean L_{long}	Relation to CAP-WM–One-step gain
DMC-Subtle	$0.63\,[0.58, 0.68]$	–
DMC-GB2	$0.55\,[0.51, 0.59]$	$\rho = 0.71\,[0.19, 0.94]$
Atari 100K	$0.34\,[0.30, 0.38]$	$\rho = 0.41\,[0.08, 0.67]$

Table 21: **Effective-horizon diagnostic.** The second column averages long-horizon teacher mass over held-out state/group pairs. The last column is the Spearman correlation between per-environment L_{long} and the paired CAP-WM–One-step performance gain within the corresponding suite.

Action repeat	L_{long}	One-step IQM	CAP-WM IQM	CAP-WM − One-step
2	$0.69\,[0.64, 0.74]$	$613\,[588, 637]$	$658\,[633, 682]$	$+45\,[22, 68]$
4	$0.55\,[0.51, 0.59]$	$642\,[617, 666]$	$666\,[641, 690]$	$+24\,[8, 40]$
8	$0.37\,[0.33, 0.41]$	$603\,[577, 628]$	$609\,[583, 634]$	$+6\,[-13, 25]$

Table 22: **Prospective action-repeat intervention on DMC-GB2 at $\bar{B} = 0.20$.** L_{long} is measured from the common pre-branch checkpoint before the CAP-WM–One-step comparison is evaluated. Brackets denote 95% bootstrap intervals over matched task–seed units.

The measured ordering follows the prediction fixed before branching: finer temporal discretization ($r = 2$) places more predictive sensitivity beyond one latent step and produces the largest multi-step gain, whereas at $r = 8$ most useful dependence is compressed into fewer latent transitions and the CAP-WM–One-step difference is unresolved. Absolute performance is not monotone in r: both $r = 2$ and $r = 8$ fall below the standard $r = 4$ for every method, so the effect is carried by the paired difference rather than by a level shift that would move the controls equally. The probabilities of improvement are 0.99, 0.96, and 0.70 for $r = 2, 4, 8$, respectively.

As a secondary quantitative test, we fit a task-fixed-effect regression over the 18 task–repeat cells,

$$\Delta_{j,r} = \alpha_j + \beta L_{\text{long},j,r} + \varepsilon_{j,r}, \tag{83}$$

where $\Delta_{j,r}$ is the paired CAP-WM–One-step return difference for task j. The fitted slope is

$$\beta = 121\ [49, 193] \qquad \text{return points per unit } L_{\text{long}}, \tag{84}$$

equivalently $+12.1\,[4.9, 19.3]$ return points for a 0.1 increase in long-horizon mass.

This intervention provides prospective evidence for the effective-horizon boundary. Temporal discretization changes L_{long} in the pre-specified direction before allocator outcomes are evaluated, and the subsequent CAP-WM–One-step difference changes in the predicted order. Because action repeat also changes temporal discretization and potentially control difficulty, the experiment supports L_{long} as a prospective applicability diagnostic rather than establishing causal mediation through L_{long}.

B.8 An Operational Pre-Branch Rule for Choosing the Horizon Set

The experiments establish a conditional result: multi-step consequence provides its clearest additional benefit when the recurrent resource is restrictive, whereas One-step consequence becomes statistically indistinguishable, and occasionally slightly better in point estimate, as the interface becomes permissive. A practitioner therefore needs an observable condition for choosing the horizon set rather than an oracle that selects whichever method attains the larger return. We operationalize the resource-pressure component of this boundary using a statistic computed from a common checkpoint before branching into the competing allocators.

A binding indicator available before branching. Reverse water filling assigns zero rate to groups whose consequence falls below the water level. We therefore measure the fraction of groups that would be inactive under the consequence profile at a common Uniform checkpoint. Specifically, on held-out replay states, we first construct the pathwise multi-step consequence profile $C^{\text{path}}_{t,1:G}$ using the same common reference and replay-conditioned teacher as in the main method, and compute the probe allocation

$$b^{\text{probe}}_t = \mathrm{WF}\Big(C^{\text{path}}_t; B\Big)\,. \tag{85}$$

We then define

$$\zeta = \mathbb{E}_t \left[\frac{1}{G} \left| \left\{ g : b_{t,g}^{\text{probe}} \leq 10^{-4} \frac{B}{G} \right\} \right| \right]. \tag{86}$$

Thus, ζ is the mean inactive-group fraction induced by the consequence profile and the candidate budget, using the same activity threshold as Equation 93. Computing it requires neither return measurements nor training either competing allocator: both candidate branches share the same Uniform checkpoint, following the before-branching logic of the L_{long} experiment in Appendix B.7. Importantly, the selector uses the offline pathwise consequence profile rather than the subsequently trained CAP-WM predictor, preventing it from inheriting information from the branch whose performance it is intended to predict.

The statistic is also independently anchored. At $\bar{B} = 0.20$ on DMC-GB2 the measured $\zeta = 0.49$ corresponds to 8.2 active groups of 16, which agrees with the common-reference active-set size $8.1\,[7.6, 8.6]$ measured separately in Appendix B.12 from the trained recurrent allocations. The binding statistic is therefore not a new quantity introduced for this diagnostic but a rescaling of one the paper already reports.

Selection rule and the status of its threshold. We use DMC-Subtle only as the calibration suite and freeze the threshold thereafter:

$$\mathcal{H}_{\text{sel}}(\zeta) = \begin{cases} \{1, 2, 4, 8\}, & \zeta \geq \tau, \\ \{1\}, & \zeta < \tau, \end{cases} \qquad \tau = 0.25. \tag{87}$$

The same τ is applied without rate-specific adjustment to the DMC-GB2 rate sweep and the serialized-payload experiment evaluated below. The value has a direct interpretation: multi-step consequence is selected only when at least approximately one quarter of the latent groups become inactive under the pre-branch consequence-weighted allocation.

We state plainly what the calibration does and does not determine. On the two DMC-Subtle operating points the measured inactive fractions are $\zeta = 0.53$ at $\bar{B} = 0.20$ and $\zeta = 0.10$ at $\bar{B} = 0.80$, so calibration only establishes that τ must lie in $(0.10, 0.53]$. That interval is wider than the range over which the transfer decisions are actually determined, and it therefore does *not* pin down $\tau = 0.25$: within the admissible interval the specific value is our choice, not a quantity the calibration data forces. What the calibration does supply is an exclusion of the two degenerate regimes, one in which the rule always selects multi-step consequence and one in which it never does. Whether the chosen value is defensible is settled after the fact by the robustness intervals reported below, not by the calibration itself.

Because the allocator-return experiments were completed before this selector analysis was introduced, we treat the selector as an exploratory operationalization of the empirical boundary established by the main experiments rather than as an additional confirmatory hypothesis. The rule, threshold, and interpretation were fixed before evaluating ζ on the transfer settings and were not retuned after observing those measurements.

Decision criterion. The diagnostic tests whether resource pressure alone provides a useful pre-branch signal. A successful statistic should separate restrictive from permissive operating points without access to allocator returns. Values near τ are naturally interpreted as an uncertain transition regime, particularly where CAP-WM and One-step consequence are statistically indistinguishable. Conversely, a persistently large ζ at operating points where One-step consequence is reliably preferable would falsify inactive fraction as an adequate operationalization of this boundary rather than motivate post-hoc threshold adjustment.

Observed selector behavior and threshold robustness. The measured binding statistic decreases sharply as recurrent capacity increases: ζ falls from 0.78 at $\bar{B} = 0.05$ to 0.49 at $\bar{B} = 0.20$, drops below the frozen threshold to 0.22 at $\bar{B} = 0.40$, and approaches zero at the two most permissive rates. Accordingly, the selector chooses multi-step consequence at $\bar{B} \in \{0.05, 0.10, 0.20\}$ and One-step consequence at $\bar{B} \in \{0.40, 0.80, 1.60\}$. These decisions select the CAP-WM runs with IQM 381, 531, and 666 at the three restrictive rates and the One-step runs with IQM 760, 801, and 817 thereafter. Thus, without access to DMC-GB2 returns, the frozen rule's selected branch coincides with the higher completed IQM at all six tested rates.

	Normalized nominal rate $\bar{B}$					
	0.05	0.10	0.20	0.40	0.80	1.60
Binding fraction ζ	0.78	0.65	0.49	0.22	0.08	0.02
Selected $\mathcal{H}$	$\{1,2,4,8\}$	$\{1,2,4,8\}$	$\{1,2,4,8\}$	$\{1\}$	$\{1\}$	$\{1\}$
One-step consequence	352	509	642	760	801	817
CAP-WM ($\mathcal{H} = \{1,2,4,8\}$)	381	531	666	758	798	816
Selector	381	531	666	760	801	817

Table 23: **Pre-branch horizon selection on the DMC-GB2 rate sweep.** The binding fraction ζ is measured from the pathwise consequence profile of a common Uniform checkpoint before training either competing allocator. The selected horizon set follows Equation 87 with the single threshold $\tau = 0.25$ calibrated on DMC-Subtle and frozen thereafter. The One-step and CAP-WM rows are the completed runs reported in Table 10; the Selector row reports the corresponding completed branch chosen by the profile-only rule and does not represent an additional trained model or an oracle selection based on DMC-GB2 returns.

The decisions are also insensitive to the precise placement of the threshold. Among the operating points assigned to One-step consequence, the largest measured inactive fraction is 0.22, whereas among those assigned to multi-step consequence, the smallest is 0.49. Hence every

$$0.22 < \tau \leq 0.49 \tag{88}$$

produces exactly the same six DMC-GB2 decisions. The observed behavior is therefore supported by a substantial separation in the binding statistic rather than by a narrowly tuned value of τ.

Serialized-payload check. For the discrete recurrent-feature channel, we use the same pre-branch consequence profile but apply the exact greedy integer allocator of Appendix B.1. The corresponding binding statistic is

$$\zeta_{\mathrm{bit}} = \mathbb{E}_t \left[\frac{1}{G} \left| \left\{ g : q_{t,g}^{\mathrm{probe}} = 0 \right\} \right| \right], \tag{89}$$

where q_t^{probe} obeys the prescribed integer payload. The measured values at 2, 4, and 6 bits/coordinate are

$$\zeta_{\mathrm{bit}} = (0.68,\ 0.37,\ 0.10).$$

With the same frozen $\tau = 0.25$, the selector chooses multi-step consequence at 2 and 4 bits/coordinate and One-step consequence at 6 bits/coordinate. This agrees with the observed performance regime: CAP-WM is separated from One-step consequence at 2 and 4 bits, whereas their difference at 6 bits is unresolved. Every threshold satisfying $0.10 < \tau \leq 0.37$ yields the same three serialized-payload decisions. Intersecting this interval with the DMC-GB2 robustness interval gives

$$0.22 < \tau \leq 0.37, \tag{90}$$

a range that contains $\tau = 0.25$ and preserves every decision across both channel families. This intersection provides a post hoc robustness check on the frozen choice: $\tau = 0.25$ lies inside a nontrivial interval over which all nine observed decisions remain unchanged. Thus, the observed decisions are not knife-edge in τ, although this analysis does not make the exploratory selector confirmatory.

What this adds, and what it does not. The selector introduces neither a new allocation mechanism nor a new training objective. Instead, it converts one component of the empirical boundary—whether the recurrent resource is sufficiently binding—into an observable pre-branch decision variable derived from the same consequence geometry used by the allocator. The selector does not choose between methods using their observed returns, and after calibration its threshold is not adjusted across DMC-GB2 rates or between the two evaluated channel families.

Three limitations remain. First, ζ captures resource pressure only. The second condition identified in the main paper—the amount of predictive sensitivity lying beyond one latent transition—is summarized separately by L_{long}; a task may therefore have a restrictive budget yet contain too little long-horizon predictive structure to benefit from the multi-step teacher. Second, the invariance

Method	$\bar{B}$	IQM ↑	Opt. gap ↓	Median ↑	Mean ↑
DreamerV3 (Hafner et al., 2025)	unconstr.	0.504	0.500	0.503	1.134
TWISTER (Burchi & Timofte, 2025)	unconstr.	–	–	0.770	1.620
DIAMOND (Alonso et al., 2024)	unconstr.	0.641	0.480	0.373	1.459
DyMoDreamer (Zhang et al., 2025)	unconstr.	–	–	0.713	1.566
EAWM (Peng et al., 2026)	unconstr.	0.593	0.474	0.651	1.290
Uniform	0.80	$0.530\,[0.48, 0.58]$	$0.500\,[0.47, 0.53]$	0.50	1.17
Surprise	0.80	$0.556\,[0.51, 0.61]$	$0.491\,[0.46, 0.52]$	0.53	1.21
One-step consequence	0.80	$0.589\,[0.54, 0.64]$	$0.476\,[0.45, 0.50]$	0.57	1.28
CAP-WM	0.80	$\mathbf{0.630}\,[0.58, 0.68]$	$\mathbf{0.460}\,[0.43, 0.49]$	**0.61**	**1.35**

Table 24: **Atari 100K aggregate performance at $\bar{B} = 0.80$.** Unconstrained rows are contextual only and are not used in matched-rate claims. Brackets are marginal bootstrap intervals.

Comparison	Paired ΔIQM	95% paired CI	$P(\text{improvement})$
CAP-WM − Uniform	$+0.100$	$[0.047, 0.151]$	0.99
CAP-WM − Surprise	$+0.074$	$[0.023, 0.123]$	0.97
CAP-WM − One-step consequence	$+0.041$	$[-0.008, 0.088]$	0.88

Table 25: **Paired Atari 100K IQM comparisons.** Paired bootstrap analysis over matched game–seed runs.

interval of Equation 90 is established over nine operating points from two channel families at a single backbone, so it characterizes the settings tested here rather than a threshold transferable by default to arbitrary architectures. Third, ζ cannot determine whether a small One-step advantage at permissive rates reflects a genuine absence of useful long-horizon structure or the higher estimation variance of the multi-step teacher. The selector should therefore be interpreted as an empirical operationalization of the observed resource-pressure boundary rather than as a universal guarantee.

B.9 Backbone Transfer

The original recurrent-backbone transfer experiment is subsumed by the three-backbone comparison in Appendix C.1, which reports DreamerV3, R2-Dreamer, and TD-MPC2 under the same tight and permissive operating points. We therefore avoid duplicating the same DreamerV3 and R2-Dreamer results here.

B.10 Atari 100K Transfer

All matched-rate variants use the same DreamerV3 backbone, latent grouping, bounded channel, interaction budget, and $\bar{B} = 0.80$. Surprise, One-step consequence, and CAP-WM use the same water-filling procedure and differ in the score supplied to it; Uniform is the non-adaptive equal-budget reference.

Point estimates follow the same ordering across IQM, optimality gap, median, and mean. However, the marginal intervals of adjacent variants overlap, so these intervals alone do not establish that every adjacent comparison is resolved. We therefore complement them with a paired analysis over the matched runs.

Paired Atari analysis. The matched variants share game and seed identities. The paired bootstrap preserves these matched game–seed units when comparing methods, thereby removing between-game and between-seed variation shared by the pair.

The paired analysis resolves CAP-WM's improvement over Uniform and Surprise: both 95% intervals lie above zero. The CAP-WM–One-step point estimate is also positive, but its interval includes zero, so we do not claim that this pairwise difference is individually resolved. Atari therefore supports transfer of the qualitative allocation-signal ordering and provides statistically resolved gains over Uniform and Surprise, while the stronger separation between one-step and multi-step consequence is established by the DMC stress and horizon analyses.

Shift	Method	Pre	Immediate	Post mean	Recovery	A_{shift}
Nuisance	Uniform	640	421	548	190K	0.00
	Surprise	666	352	522	228K	0.31
	One-step consequence	682	476	596	145K	0.19
	CAP-WM	**705**	**545**	**636**	**103K**	0.11
Dynamics	Uniform	640	305	444	$>$ 250K	0.00
	Surprise	666	336	493	212K	0.27
	One-step consequence	682	351	522	184K	0.29
	CAP-WM	**705**	**374**	**570**	**142K**	0.35

Table 26: **Online behavior under nuisance and dynamics shifts.** "Post mean" covers the 250K post-shift interactions. The experiment measures selective reallocation and recovery under the tested shifts; it is not a general continual-learning benchmark.

Shift	Pre	Immediate	50K	100K	200K
Nuisance	$0.78\,[0.74, 0.81]$	$0.70\,[0.65, 0.75]$	$0.75\,[0.71, 0.79]$	$0.77\,[0.73, 0.80]$	$0.78\,[0.74, 0.81]$
Dynamics	$0.78\,[0.74, 0.81]$	$0.54\,[0.48, 0.60]$	$0.64\,[0.59, 0.69]$	$0.71\,[0.66, 0.75]$	$0.76\,[0.72, 0.79]$

Table 27: **Post-shift consequence-target tracking.** Spearman correlation between $\widehat{C}$ and C^{path} on held-out replay states.

B.11 Controlled Non-Stationarity

We run the shift experiment on the six DMC-GB2 training tasks at the tight operating point $\bar{B} = 0.20$, using the same five training seeds as the main DMC-GB2 study. Each agent is first trained on the clean environment to the common pre-shift checkpoint and is then updated online for a further 250K interactions after a single controlled change. CAP-WM receives no task identifier or change-point signal.

In the *nuisance shift*, the clean background is replaced by an independent video while the simulator state, transition law, action space, and reward function are unchanged. In the *dynamics shift*, rendering and reward are held fixed while the pre-specified task-specific physical parameter used by the shift protocol is changed and then kept fixed for the entire 250K post-shift window. The same shift instance is used for every allocation method within a task/seed pair.

"Pre" denotes the evaluation immediately before the change, "Immediate" the first evaluation after it, and "Post mean" the mean return over the full 250K post-shift window. "Recovery" is the number of post-shift interactions required to regain 90% of the method's final post-shift score; values exceeding the observation window are reported as $>$ 250K.

We measure the normalized allocation displacement

$$A_{\text{shift}} = \frac{1}{2B} \left\| \mathbb{E}[b_t \mid \text{post}] - \mathbb{E}[b_t \mid \text{pre}] \right\|_1 \tag{91}$$

and define

$$R_{\text{sel}} = \frac{A_{\text{dynamics}}}{A_{\text{nuisance}} + \epsilon}. \tag{92}$$

A_{shift} measures how much the average allocation changes, whereas R_{sel} asks whether a method reallocates more strongly for a change in predictive dynamics than for an appearance-only nuisance.

The selectivity ratios are 0.87 for Surprise, 1.53 for One-step consequence, and 3.18 for CAP-WM. Surprise therefore reallocates slightly more under the nuisance shift than under the dynamics shift, whereas CAP-WM exhibits the opposite and much stronger selectivity. We interpret this as evidence that consequence-aware allocation responds more strongly when the predictive effect of state information changes. It is not a guarantee against catastrophic forgetting.

After a dynamics shift, predictor–teacher agreement falls from 0.78 to 0.54 and progressively recovers to 0.76 by 200K interactions. After a nuisance shift, the drop is substantially smaller and

Training checkpoint	Spearman ρ	Top-Q overlap	Allocation displacement
100K	$0.62\,[0.57, 0.67]$	$0.68\,[0.64, 0.72]$	$0.18\,[0.16, 0.20]$
500K	$0.78\,[0.74, 0.82]$	$0.79\,[0.75, 0.83]$	$0.10\,[0.09, 0.12]$
1M	$0.88\,[0.85, 0.91]$	$0.87\,[0.84, 0.90]$	$0.06\,[0.05, 0.07]$

Table 28: **Replay–current-policy consequence mismatch during training.** All quantities are evaluated offline on the same held-out states; no additional agent training is required.

$\bar{B}$	MARE $\downarrow$	$R^2(D_{\text{pred}}, D_{\text{emp}}) \uparrow$	$\rho(b, -\log D_{\text{emp}}) \uparrow$
0.10	3.8%	0.93	0.94
0.20	3.1%	0.95	0.96
0.40	2.7%	0.96	0.97
0.80	2.4%	0.97	0.97

Table 29: **Channel-distortion calibration.** The implemented stochastic channel closely realizes its prescribed distortion schedule after standardization. This validates the operational channel law used by the allocator; it does not estimate the mutual information of the non-Gaussian learned representation.

agreement returns to its pre-shift level. This analysis measures whether the deployable predictor reestablishes agreement with newly observed replay-derived targets as fresh experience accumulates; it does not estimate the full discrepancy between replay and the current-policy action distribution.

Replay-conditioned versus current-policy consequence. We therefore measure that remaining mismatch directly at 100K, 500K, and 1M interactions without retraining the agent. At each checkpoint, the same held-out states are used to construct two open-loop teachers over $\mathcal{H} = \{1, 2, 4, 8\}$: C^{replay} uses the recorded replay continuation, whereas C^{π} uses a deterministic continuation generated once from the current actor mean starting at the same unperturbed reference memory. The action sequences are held fixed during Jacobian evaluation, so the comparison differs only in the continuation and does not differentiate through future policy responses. Spearman agreement between the two consequence profiles rises from $0.62\,[0.57, 0.67]$ at 100K to $0.78\,[0.74, 0.82]$ at 500K and $0.88\,[0.85, 0.91]$ at 1M; top-quartile overlap rises from $0.68\,[0.64, 0.72]$ to $0.79\,[0.75, 0.83]$ and $0.87\,[0.84, 0.90]$. The normalized displacement between the water-filled allocations, $\|WF(C^{\text{replay}}; B) - WF(C^{\pi}; B)\|_1/(2B)$, correspondingly decreases from $0.18\,[0.16, 0.20]$ to $0.10\,[0.09, 0.12]$ and $0.06\,[0.05, 0.07]$. Thus replay–policy mismatch is largest early in training and becomes modest at the final checkpoint; this quantifies, rather than eliminates, the distribution-mismatch limitation of the replay-conditioned teacher.

B.12 Implementation Robustness, Channel Calibration, and Computational Cost

Channel-distortion calibration. For held-out transitions, we compare the prescribed per-group distortion

$$D_{t,g} = \exp(-2b_{t,g}/d)$$

with the realized distortion

$$D_{t,g}^{\text{emp}} = d^{-1} \left\| m_{t,g} - \bar{u}_{t,g} \right\|_2^2 .$$

Across the tested operating points, mean absolute relative error remains below 4%, R^2 increases from 0.93 to 0.97, and allocated rate is strongly rank-correlated with negative log realized distortion. Thus the implemented channel closely follows the distortion schedule assumed by reverse water filling. This is an implementation check rather than an empirical rate–distortion fit of the learned latent distribution: the nominal Gaussian rate $\bar{B}$ should not be interpreted as the true mutual information of an arbitrary neural representation.

Grouping and estimator robustness. We next vary the grouping granularity and the two principal estimator-cost parameters while holding the environment, budget, backbone, and remaining training configuration fixed.

Component	Setting	Mean return ↑
Groups G	4	574
	8	598
	16	**608**
	16 + fixed random coordinate permutation	603
	32	601
JVP probes	1	598
	2	**608**
	4	611
Teacher update interval	every update	612
	every 4 updates	**608**
	every 8 updates	594

Table 30: **Implementation sensitivity on DMC-Subtle at $\bar{B} = 0.20$.** Bold denotes the default configuration used throughout the main experiments, not necessarily the largest point estimate in each block.

Teacher operating point	Mean return	Active groups / 16
Current adaptive allocation	575	$3.4\,[3.0, 3.9]$
Common uniform reference	**608**	$8.1\,[7.6, 8.6]$

Table 31: **Reference-point mechanism diagnostic on DMC-Subtle at $\bar{B} = 0.20$.** The active-set statistic is computed from the same recurrent allocations as the return ablation using Equation 93.

Performance improves substantially from $G = 4$ to $G = 16$ and changes only modestly at $G = 32$ (608 versus 601). A fixed random coordinate permutation at $G = 16$ gives 603, providing no evidence that the result depends strongly on the original coordinate ordering. Appendix A.5.1 provides the complementary geometric diagnostic: despite imperfect group-block structure, the fixed grouping retains most of the predictive distortion improvement available to unrestricted eigenmode allocation. Together, these results indicate that the main effect is not tied to a privileged coordinate partition.

Two JVP probes give nearly the same return as four (608 versus 611), whereas one probe gives 598. Constructing a teacher target every four learner updates similarly gives 608, close to the 612 obtained by constructing a target every update, while reducing teacher cost. We therefore use two probes and a four-update interval as the default compute–accuracy trade-off. Horizon sensitivity is analyzed separately in Appendix B.6, where the benefit saturates at $\mathcal{H} = \{1, 2, 4, 8\}$.

Reference-point mechanism in trained models. The reference-channel ablation directly tests the allocation-independent teacher argument of Section 3.3. Measuring consequence around the current adaptive assignment reduces mean return from 608 to 575 and contracts the mean active set from $8.1\,[7.6, 8.6]$ to $3.4\,[3.0, 3.9]$ groups.

We define group g as active when

$$b_{t,g} > 10^{-4}\frac{B}{G}. \tag{93}$$

This pattern is the trained-system counterpart of Proposition 1: source-referred refresh can suppress the same importance signal required for a low-rate group to recover. Under a common allocation-independent reference, Proposition 2 removes this assignment-dependent gain; ordinary water filling may still assign zero rate to genuinely low-consequence groups. The discriminating prediction is therefore not sparsity itself, but assignment-dependent contraction that can exclude intrinsically consequential groups. Appendix A.2.1 separately isolates nonlinear re-linearization from this source-gain pathway.

Reachability of self-reinforcing starvation. Proposition 1 establishes existence and local absorption for a fixed intrinsic consequence profile; it does not assert global convergence. We therefore test

whether the corresponding failure is reachable from ordinary allocations. On 256 held-out DMC-Subtle states at $\bar{B} = 0.20$, we iterate three maps: the fixed-profile source map T_{src} of Equation 38, which contains the assignment-dependent gain without profile drift; the memory-referred map T_{mem} of Equation 37, which re-linearizes without that gain; and the allocation-independent map

$$T_{\mathrm{common}} = \mathrm{WF}\big(c(m^{\mathrm{ref}}); B\big) .$$

We evaluate uniform, noisy-uniform, random-feasible, and trained-allocation initializations. Noisy-uniform and random-feasible conditions use 16 draws per state, with bootstrap resampling grouped by state. Iteration stops when $\|b^{(k+1)} - b^{(k)}\|_1/(2B) < 10^{-4}$ or after 50 iterations. Collapse is defined as convergence to at most four active groups, while high-c starvation denotes zero allocation to at least one group in the top quartile of the common-reference consequence ranking.

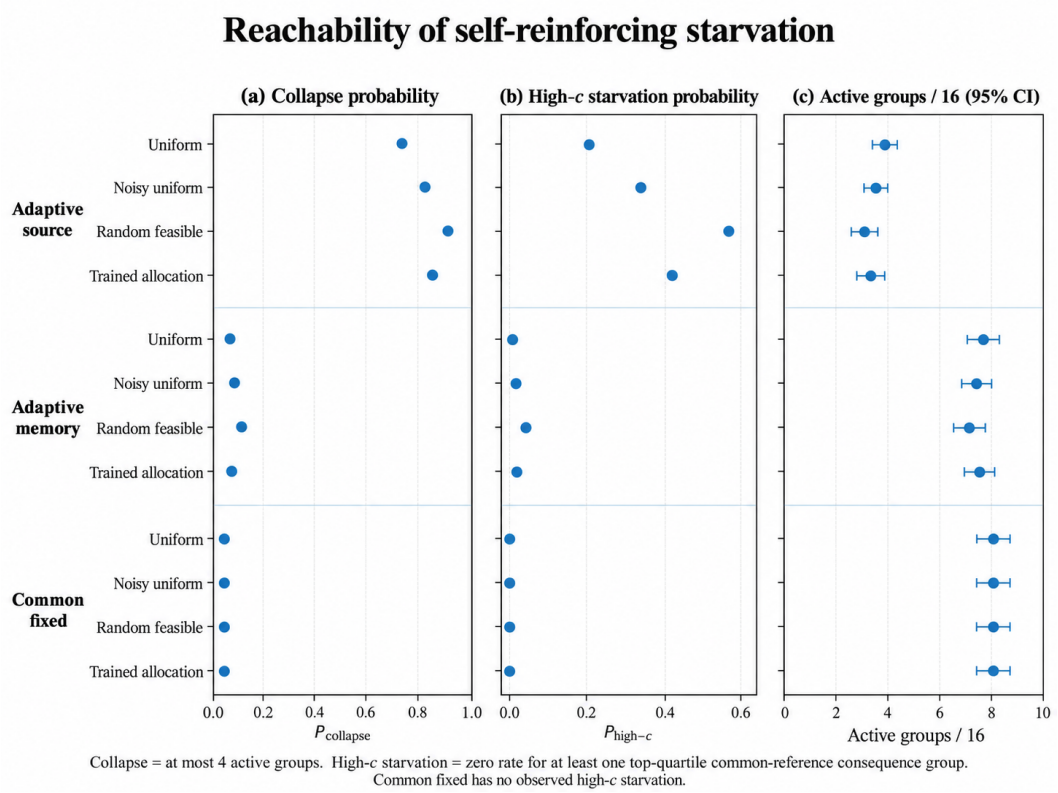


Figure 4: **Reachability of self-reinforcing starvation.** Basin-of-attraction diagnostic on 256 held-out DMC-Subtle states at $\bar{B} = 0.20$. **(a)** Probability of convergence to at most four active groups. **(b)** Probability of assigning zero rate to at least one top-quartile common-reference consequence group. **(c)** Mean active-set size with 95% bootstrap intervals. Adaptive source refresh exhibits substantially larger collapse and high-c starvation across initialization families, whereas adaptive memory refresh largely avoids the failure. The common fixed map is initialization independent; its 0.03 collapse rate reflects genuinely sparse fixed-reference water-filling solutions and is accompanied by no observed top-quartile starvation.

Figure 4 shows a sharp separation between assignment-dependent source refresh and the two controls that remove that gain. Adaptive source refresh reaches collapsed solutions with probability 0.74–0.91 across the four initialization families and exhibits high-c starvation with probability 0.21–0.57, while retaining only 3.2–3.9 active groups on average. In contrast, adaptive memory refresh has collapse probabilities of only 0.06–0.11, high-c starvation of 0.01–0.04, and retains 7.2–7.7 active groups. The common fixed map is initialization independent, has $P_{\mathrm{collapse}} = 0.03$, and shows no observed top-quartile starvation. Its nonzero collapse rate therefore reflects states whose genuine fixed-reference water-filled solution is itself sparse, rather than assignment-induced starvation.

Teacher operating point	Mean return	Active / 16	ρ_C	Δ_b
Fixed $\alpha = 0.5$	600	$7.7\,[7.2, 8.2]$	$0.93\,[0.91, 0.95]$	$0.06\,[0.05, 0.07]$
Fixed $\alpha = 1$	**608**	$8.1\,[7.6, 8.6]$	1.00	0.00
Fixed $\alpha = 2$	604	$8.3\,[7.8, 8.8]$	$0.95\,[0.93, 0.97]$	$0.05\,[0.04, 0.06]$
Fixed $\alpha = 4$	598	$8.5\,[8.0, 9.0]$	$0.90\,[0.87, 0.93]$	$0.09\,[0.07, 0.11]$
Current adaptive	575	$3.4\,[3.0, 3.9]$	$0.61\,[0.56, 0.66]$	$0.31\,[0.28, 0.34]$

Table 32: **Sensitivity to the allocation-independent teacher reference on DMC-Subtle at $\bar{B} = 0.20$.** ρ_C is Spearman agreement with the default $\alpha = 1$ consequence profile and Δ_b is normalized allocation displacement from the default. The $\alpha = 0.5$ active-set value coincides numerically with an independent memory-referred measurement in Figure 4; the two quantities are obtained from different evaluations and neither is derived from the other.

Method	Params	Train time	Updates/s	Peak memory	Action latency
DreamerV3	18.3M	9.6 h (1.00×)	100.0	8.4 GB	2.84 ms
Uniform	18.3M	10.0 h (1.04×)	96.2	8.6 GB	2.92 ms
CAP-WM	18.8M	11.6 h (1.21×)	82.7	9.2 GB	3.10 ms

Table 33: **Measured computational cost under a fixed DMC training budget.** All measurements use the same hardware and interaction budget.

These results establish practical reachability of the proposed failure mode without strengthening Proposition 1 into a global-convergence claim.

Robustness to the allocation-independent reference magnitude. To distinguish allocation independence from the particular default magnitude $b_g^{\mathrm{ref}} = B/G$, we evaluate fixed references

$$b_g^{\mathrm{ref}}(\alpha) = \alpha B/G, \qquad \alpha \in \{0.5, 1, 2, 4\},$$

while keeping the deployed budget fixed at $\sum_g b_{t,g} = B$. Each non-default fixed-reference condition is trained on DMC-Subtle at $\bar{B} = 0.20$ with five seeds, and we compare its held-out consequence profile and induced allocation with the default $\alpha = 1$ teacher.

All fixed references remain close to the default: mean return varies from 598 to 608, consequence-profile agreement remains 0.90–0.95 for non-default references, and normalized allocation displacement is only 0.05–0.09. The current adaptive reference differs substantially more, with return 575, profile agreement $0.61\,[0.56, 0.66]$, allocation displacement $0.31\,[0.28, 0.34]$, and only $3.4\,[3.0, 3.9]$ active groups. Under a fixed intrinsic profile, changing the common reference magnitude applies the same scalar source gain to all groups and therefore does not alter their water-filled ordering. The remaining empirical variation across fixed references is consequently consistent with nonlinear movement of the linearization point rather than the assignment-dependent feedback isolated by Proposition 1. Thus the reference-point result is not specific to the default choice $\alpha = 1$.

Computational cost. CAP-WM adds a $463{,}376$-parameter consequence predictor at deployment. Jacobian–vector products are required only to construct teacher targets during training and are absent from action selection.

Relative to the matched Uniform interface, CAP-WM adds 2.7% parameters, 7.0% peak memory, and 6.2% action-selection latency; training is approximately 16% slower. Consequence-teacher construction accounts for approximately 14.8% of learner time. This separation is consistent with the algorithm: deployment adds one consequence-head forward pass and a one-dimensional water-level solve, whereas teacher JVPs are training-only.

The parameter increase does not by itself explain the performance gain. Shuffled consequence uses the same predictor architecture and parameter count, the same total rate, and the same per-state rate histogram, but assigns those rates to different groups and loses most of CAP-WM's advantage. The measured gain therefore depends on the alignment between allocated precision and predictive consequence rather than on additional model capacity alone.

B.13 The Obstruction Outside World Models: Mixed-Precision Quantization

Proposition 1 identifies a failure mode caused by *assignment-attenuated refreshed measurement*: the precision currently assigned to a component attenuates the importance signal subsequently used to revise that same assignment. The proposition is proved for the recurrent channel of Section 3.3; it does not by itself imply the same functional map for a hard quantizer. We therefore test a *structural analogue* in mixed-precision post-training quantization, where second-order sensitivity is a standard signal for assigning precision (Dong et al., 2020). In this setting, attenuation arises through quantization-bin absorption rather than through the Gaussian-channel gain $\kappa(b_g)$. The experiment asks whether this distinct precision-dependent measurement mechanism produces the same qualitative feedback, and whether removing the assignment dependence removes the collapse.

Experimental setting. We quantized an ImageNet-1K pretrained ResNet-50. Its $3{+}4{+}6{+}3 = 16$ residual bottleneck branches define $G = 16$ allocation groups. A single bit width is shared by all convolutional weights in a residual branch; the stem, classifier, and down-sampling projection shortcuts remain fixed at 8 bits, so a zero-bit residual branch is implemented as a well-defined bypass rather than a dimensionality change. Activations remain in FP32. Each group g receives $q_g \in \{0, 1, \dots, 8\}$ bits per weight under the fixed *branch-level precision budget*

$$\frac{1}{G}\sum_{g=1}^{G} q_g = \bar{q}, \qquad \bar{q} \in \{3, 4, 5\} \text{ bits.} \tag{94}$$

Because residual branches contain different numbers of parameters, $\bar{q}$ is an average branch precision rather than a parameter-weighted model bitrate. This experiment therefore tests the reference-dependent allocation mechanism, not compression efficiency. The primary operating point is $\bar{q} = 4$ bits. We use symmetric per-output-channel hard quantization for weights; $q_g = 0$ sets the residual-branch contribution to zero. No retraining, distillation, or quantization-aware fine-tuning is performed.

Bit widths are assigned using the same greedy marginal rule as in the serialized-payload experiment of Section B.2. Because the surrogate is separable and its marginal distortion reduction decreases with bit width, the greedy procedure attains the integer optimum under the prescribed total budget.

Sensitivity estimation and calibration. We use five disjoint calibration splits, each containing 1,024 ImageNet training images, and evaluate every final assignment on the full ImageNet validation set. Group importance is measured with a Hessian-trace-style directional-curvature proxy. To isolate the effect of hard quantization without differentiating through rounding with a straight-through estimator, we use the same symmetric finite-difference Hutchinson estimator for the two source-referred conditions:

$$\widehat{S}_g^{\mathrm{src}}(q; \delta_g) = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathcal{L}(Q_q(w_g + \delta_g v^{(k)})) - 2\mathcal{L}(Q_q(w_g)) + \mathcal{L}(Q_q(w_g - \delta_g v^{(k)}))}{\delta_g^2}, \tag{95}$$

where $v^{(k)}$ is Rademacher and $K = 32$. The primary finite-difference scale is $\delta_g = \sigma_g/4$, where σ_g denotes the full-precision within-group weight standard deviation. We additionally evaluate

$$\delta_g \in \{\sigma_g/8, \sigma_g/4, \sigma_g/2\} \tag{96}$$

to test sensitivity to the finite-difference scale.

Sensitivity conventions. We compare three sensitivity conventions while keeping the model, precision budget, allocator, calibration protocol, and, for the adaptive conditions, refresh schedule fixed.

Frozen source reference. Sensitivity is measured once at full precision, $Q_\infty(w) = w$, and the resulting profile is held fixed throughout allocation. This is the mixed-precision analogue of measuring importance from an allocation-independent reference.

Adaptive source reference. Sensitivity is re-estimated after every assignment round at the bit widths currently in force, after which the allocator is rerun from the refreshed scores. We terminate when the starved set remains unchanged for three consecutive refreshes or after 12 rounds. This defines

Condition	Active groups / 16	Top-1 (%)	Rounds to absorption
FP32 reference	16/16	76.13	—
Frozen source reference	15.4 [14.8, 16.0]	75.68 [75.57, 75.79]	—
Adaptive source reference	10.8 [10.0, 11.6]	72.91 [72.42, 73.37]	4.6 [4, 6]
Adaptive quantized-value reference	15.1 [14.4, 15.8]	75.47 [75.31, 75.62]	—
Pilot source-referred fixed point	11	—	—

Table 34: **Observed reference-dependent starvation outside world models.** Mixed-precision post-training quantization of ResNet-50 on ImageNet-1K at $\bar{q} = 4$ bits with $\delta_g = \sigma_g/4$ and no retraining. FP32 is a contextual accuracy reference rather than a rate-matched condition. The adaptive-source condition refreshes sensitivity through the quantizer at the currently assigned precision, whereas the adaptive quantized-value condition applies the same perturbation after quantization. The final row reports the fixed point obtained from the disjoint pilot analysis before final validation. Brackets are five-split bootstrap intervals.

a discrete self-referential allocation map, denoted T_q, that is structurally analogous to the update pattern in Equation 28, but is not the same map as the Gaussian-channel $T(b) = \mathrm{WF}(\kappa(b)^2 \odot c; B)$.

The attenuation mechanism is correspondingly different. Under hard quantization, if δ_g is small relative to the current quantization step, perturbations of the underlying full-precision weights increasingly remain inside the same quantization bin:

$$Q_q(w_g + \delta_g v) \approx Q_q(w_g - \delta_g v) \approx Q_q(w_g).$$

The measured source-referred curvature then decreases after the group has already been assigned low precision. Thus low precision can suppress the importance signal used by the next allocation round, paralleling the assignment-dependent feedback of Proposition 1 without sharing its specific gain function κ.

Adaptive quantized-value reference. We instead perturb the already-quantized value:

$$\widehat{S}_g^{\mathrm{qval}}(q; \delta_g) = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathcal{L}(Q_q(w_g) + \delta_g v^{(k)}) - 2\mathcal{L}(Q_q(w_g)) + \mathcal{L}(Q_q(w_g) - \delta_g v^{(k)})}{\delta_g^2}. \tag{97}$$

Because the perturbation is applied downstream of quantization, it cannot be absorbed by the current quantization bin. The allocator, stopping criterion, and refresh schedule are otherwise identical to those of the adaptive source-reference condition. This control therefore retains adaptive re-estimation while removing the specific bin-absorption pathway.

Pilot fixed-point analysis. On a disjoint pilot split, we measure the source-referred attenuation curve across bit widths and iterate the induced discrete allocation map T_q from the frozen-reference sensitivity profile. At $\bar{q} = 4$ bits, this pilot analysis converges to a fixed point with 11 active groups, where a group is defined as active when $q_g \geq 1$. The 11-group pilot fixed point is recorded before the final validation evaluation and is subsequently compared with the fixed point reached by the adaptive source-reference runs.

Primary 4-bit result. At the primary 4-bit operating point, the frozen source-reference allocator retains $15.4/16$ active residual branches and achieves 75.68% top-1 accuracy. When source-referred sensitivity is refreshed through the currently assigned quantizer, the active set contracts to $10.8/16$ groups after 4.6 assignment rounds on average and top-1 accuracy falls to 72.91%. Relative to the frozen reference, adaptive source-referred refresh therefore reduces the active set substantially and lowers top-1 accuracy by 2.77 percentage points ($75.68 \to 72.91$).

The observed adaptive fixed point closely matches the independently obtained 11-active-group pilot fixed point. In the median calibration split, five branches are starved; two belong to the top quartile of the frozen-reference sensitivity ranking, including the third-most-sensitive branch. None re-enters the active set before the allocation reaches absorption. This trajectory is consistent with the proposed feedback: low assigned precision attenuates the refreshed source-referred measurement, which makes subsequent allocation less favorable to the same group.

δ_g	Frozen source Active / 16	Adaptive source Active / 16	Adaptive quantized value Active / 16
$\sigma_g/8$	15.2 [14.6, 15.8]	9.6 [8.8, 10.4]	14.9 [14.2, 15.6]
$\sigma_g/4$	15.4 [14.8, 16.0]	10.8 [10.0, 11.6]	15.1 [14.4, 15.8]
$\sigma_g/2$	15.6 [15.0, 16.0]	12.6 [11.8, 13.4]	15.0 [14.4, 15.7]

Table 35: **Finite-difference-scale robustness at $\bar{q} = 4$.** Adaptive source-referred refresh contracts the active set at all three finite-difference scales, most strongly when perturbations are smallest relative to the current quantization bin. The adaptive quantized-value condition remains close to the frozen reference because its perturbations are applied downstream of quantization. Brackets are five-split bootstrap intervals.

By contrast, adaptive refresh with the quantized-value convention retains $15.1/16$ active groups and achieves 75.47% top-1 accuracy, remaining close to the frozen source-reference result and far from the adaptive source-reference condition. The controlled difference is where the finite-difference perturbation is applied relative to the quantizer. Applying it downstream of quantization removes the bin-absorption pathway while keeping the model, budget, allocator, adaptive refresh, and stopping rule fixed. This is a structural analogue of the convention dependence characterized by Proposition 5, not an application of the proposition's Gaussian-channel identity.

Finite-difference robustness. At $\bar{q} = 4$, the median ratio of perturbation magnitude to the 4-bit quantization step is 0.32, 0.63, and 1.26 for $\delta_g = \sigma_g/8, \sigma_g/4, \sigma_g/2$, respectively. Thus the primary $\sigma_g/4$ condition does not make 4-bit groups automatically invisible to the estimator; the strongest attenuation appears only after groups have moved to lower precision. Across the complete pre-specified scale sweep, adaptive source-referred refresh contracts the active set at every tested scale, with weaker contraction as the perturbation becomes large relative to the quantization bin.

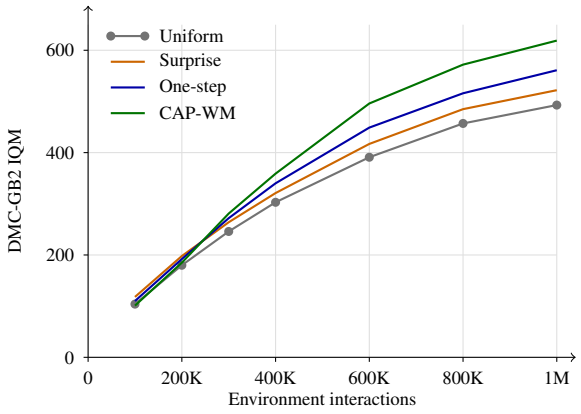
The corresponding adaptive source-reference top-1 accuracies are 71.88, 72.91, and 74.08 for $\delta_g = \sigma_g/8, \sigma_g/4, \sigma_g/2$, respectively. Frozen source-reference accuracies are 75.62, 75.68, and 75.64, while the adaptive quantized-value convention reaches 75.36, 75.47, and 75.42. Thus the finite-difference scale changes the magnitude of the source-referred collapse but not its qualitative presence: smaller perturbations produce stronger bin absorption and a larger accuracy loss, whereas the post-quantization convention remains comparatively stable.

Budget dependence. The same pattern weakens as the precision budget becomes more permissive. At $\bar{q} = 3, 4, 5$ bits, frozen/adaptive-source active-set sizes are $14.2/8.6$, $15.4/10.8$, and $16.0/13.4$, respectively; the corresponding top-1 accuracies are $74.86/68.97$, $75.68/72.91$, and $75.91/74.76$. Across these completed runs, both the active-set gap and accuracy gap narrow as more branches can retain nonzero precision.

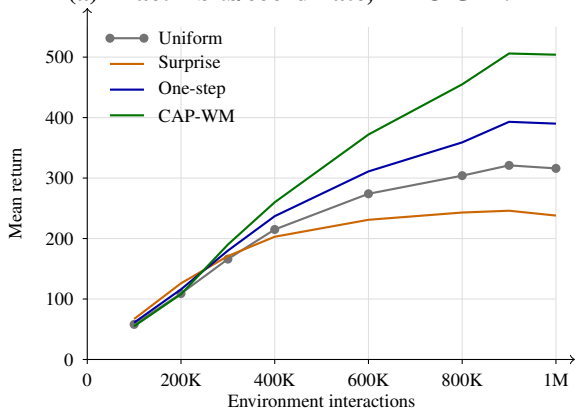
What the experiment establishes. The ResNet experiment changes the architecture, task, objective, channel, and resource from the world-model setting. It therefore does not test the Gaussian-channel map of Proposition 1 literally. What it preserves is the structural dependency at issue: an importance score is refreshed through the precision assignment that it is subsequently used to update.

Under the adaptive source-referred convention, this dependency produces a self-reinforcing contraction of the active set, including starvation of branches that rank highly under the allocation-independent reference. Two controls identify the relevant attenuation pathway. First, increasing the finite-difference scale weakens bin absorption and monotonically reduces the active-set collapse. Second, moving the perturbation downstream of quantization removes bin absorption while preserving adaptive refresh, the budget, allocator, and architecture; the active set and top-1 accuracy then remain close to the frozen-reference condition.

These results therefore support the *structural scope* of the reference-point analysis: assignment-dependent attenuation of a refreshed importance measurement can create self-reinforcing starvation outside the specific recurrent Gaussian channel in which Proposition 1 is proved. They do not imply that adaptive recalibration generally fails, nor that every mixed-precision quantization method exhibits this behavior. Measuring all components from a common reference, or using a measurement



(a) Exact 2 bits/coordinate, DMC-GB2.



(b) DMC-Subtle + video, $\bar{B} = 0.20$.

Figure 5: **Learning dynamics in the two tight-resource settings that carry the central claims.** (a) DMC-GB2 aggregate IQM under an exact serialized payload of 2 bits per recurrent coordinate. (b) Mean return on the Subtle/Video stress condition under the nominal channel. Curves begin at the first evaluated checkpoint (100K interactions); CAP-WM trails Uniform there in both panels (101 versus 104 and 55 versus 58) and separates only after the consequence predictor has accumulated replay supervision, so the advantage is not an initialization effect. Final checkpoints match the corresponding aggregate tables exactly. All plotted values are listed in Table 36.

convention that removes the assignment-dependent attenuation, eliminates the observed feedback in this experiment.

B.14 Learning Curves

Figure 5 reports training-time performance for the two settings that carry the central empirical claims: the exact 2-bit recurrent payload on DMC-GB2 and the Subtle/Video stress test at $\bar{B} = 0.20$. The curves are evaluated at fixed interaction checkpoints without smoothing. In both settings CAP-WM is not ahead at the first reported checkpoint; its advantage emerges only later, after the consequence predictor has received substantial replay supervision. This pattern is consistent with a learned teacher–predictor effect rather than an advantage already present at the first evaluation checkpoint.

For reproducibility, the checkpoint values plotted above are listed explicitly in Table 36.

B.15 Raw-Result Reporting Policy

Task-, game-, and seed-level tables are exported directly from completed run logs and are never reconstructed from aggregate IQM, means, confidence intervals, or paired differences. They are

Setting	Method	100K	200K	300K	400K	600K	800K	900K	1M
2-bit DMC-GB2	Uniform	104	180	246	303	391	457	–	493
	Surprise	118	198	264	321	417	485	–	522
	One-step	110	192	272	340	449	516	–	561
	CAP-WM	101	185	281	359	496	572	–	619
Subtle/Video	Uniform	58	109	166	215	274	304	321	316
	Surprise	67	126	171	203	231	243	246	238
	One-step	61	116	180	237	311	359	393	390
	CAP-WM	55	108	190	260	372	455	506	504

Table 36: **Learning-curve checkpoints.** All checkpoints use the same evaluation protocol as the corresponding final aggregate. Small late-stage regressions are retained rather than forcing monotone curves.

printed in run order rather than sorted by performance, so that per-unit reversals remain visible. The headline aggregates corresponding to the printed tables are recomputed directly from those entries: the 6×5 DMC-GB2 table below reproduces the rate-sweep IQMs of Table 10 at $\bar{B} = 0.20$, the matched ten-seed table reproduces the confirmatory Subtle/Video means, and the Atari game table reproduces both the IQM and mean columns of Table 24. For experiments whose raw logs are not attached, the aggregate tables in this appendix are the complete numerical record.

B.16 Task-, Game-, and Seed-Level Results

The aggregate claims are complemented by unsorted task/game/seed-level reporting at the critical operating points. The outcomes are heterogeneous, as expected for matched RL evaluations: CAP-WM does not dominate every environment or every matched seed. Aggregate conclusions continue to use the pre-specified IQM/mean estimands and paired bootstrap analyses.

DMC-GB2 task–seed results. Table 37 reports all 6×5 task–seed returns at $\bar{B} = 0.20$. Recomputing IQM over the 30 entries gives 578, 612, 642, and 666 for Uniform, Surprise, One-step consequence, and CAP-WM, respectively, matching the corresponding column of Table 10. CAP-WM is below One-step consequence on 5 of the 30 task–seed units, and the four methods follow the aggregate ordering on only 18 of 30; the aggregate claim is rate efficiency, not per-unit dominance.

Ten-seed Subtle/Video comparison. The matched seed means in Table 38 average to 397 for One-step consequence and 501 for CAP-WM. The paired effect is heterogeneous across seeds: per-seed differences range from -18 to $+176$ with standard deviation 57, including one reversal (seed 3), while the aggregate difference remains $+104$.

Seed	One-step	CAP-WM	CAP-WM − One-step
1	352	444	+92
2	431	576	+145
3	389	371	-18
4	455	614	+159
5	318	494	+176
6	402	479	+77
7	377	511	+134
8	468	526	+58
9	341	467	+126
10	437	528	+91
Mean	**397**	**501**	**+104**

Table 38: **Matched 10-seed Subtle/Video mean returns.** Seed identities are kept in run order rather than sorting methods independently.

Atari game-level scores. Table 39 reports human-normalized game means over the five training seeds. CAP-WM is below One-step consequence on 8 of the 26 games, consistent with the un-

Task	Seed	Uniform	Surprise	One-step	CAP-WM
Walker Walk	1	647	688	724	776
Walker Walk	2	622	676	677	698
Walker Walk	3	635	674	701	742
Walker Walk	4	689	725	791	811
Walker Walk	5	633	694	670	769
Walker Stand	1	611	662	700	690
Walker Stand	2	600	582	643	665
Walker Stand	3	569	580	629	659
Walker Stand	4	657	693	703	691
Walker Stand	5	622	648	652	722
Cheetah Run	1	532	560	633	647
Cheetah Run	2	549	570	665	667
Cheetah Run	3	587	632	660	688
Cheetah Run	4	609	636	659	729
Cheetah Run	5	532	601	600	534
Finger Spin	1	617	632	666	615
Finger Spin	2	692	728	759	777
Finger Spin	3	594	612	631	637
Finger Spin	4	683	701	708	774
Finger Spin	5	648	655	698	741
Cartpole Swingup	1	466	524	517	564
Cartpole Swingup	2	500	577	564	640
Cartpole Swingup	3	417	467	463	496
Cartpole Swingup	4	533	577	617	653
Cartpole Swingup	5	444	507	519	573
Cup Catch	1	503	566	560	620
Cup Catch	2	526	581	608	612
Cup Catch	3	461	505	554	503
Cup Catch	4	511	573	544	635
Cup Catch	5	482	511	531	579

Table 37: **Unsorted DMC-GB2 task–seed returns at $\bar{B} = 0.20$.** The table contains several local reversals; the paper's claim is aggregate rate efficiency, not universal task/seed dominance.

resolved aggregate CAP-WM–One-step paired comparison; resampling these game means alone reproduces the point estimate $+0.041$ with an interval that includes zero, and the narrower interval reported in Table 25 follows from resampling matched game–seed units rather than game means. Recomputing the middle-50% IQM from these game means gives 0.530, 0.556, 0.589, and 0.630; arithmetic means are 1.170, 1.210, 1.280, and 1.350.

Reporting scope. The tables above are sufficient for the manuscript-level transparency check: the primary DMC result is exposed at task–seed resolution, the central multi-step comparison at matched-seed resolution, and Atari at full game resolution.

Game	Uniform	Surprise	One-step	CAP-WM
Alien	0.309	0.047	-0.088	0.143
Amidar	-0.051	0.090	-0.264	-0.093
Assault	0.481	0.520	0.430	0.575
Asterix	1.013	0.900	0.381	0.820
Bank Heist	0.220	0.255	-0.247	0.382
Battle Zone	0.579	0.301	0.866	0.554
Boxing	1.042	1.214	1.019	1.216
Breakout	0.023	0.259	0.585	0.143
Chopper Command	0.165	0.530	-0.007	0.249
Crazy Climber	4.187	4.763	4.904	5.399
Demon Attack	2.660	2.651	1.940	2.861
Freeway	0.995	0.721	1.227	1.305
Frostbite	-0.167	0.396	-0.015	0.187
Gopher	0.681	0.781	1.228	1.017
Hero	0.442	0.290	0.179	0.288
Jamesbond	0.224	0.350	0.604	0.536
Kangaroo	0.232	0.705	0.784	0.632
Krull	0.403	0.273	0.337	0.137
Kung Fu Master	0.812	1.096	0.916	1.090
Ms Pacman	0.415	0.345	0.113	0.115
Pong	0.818	0.781	0.617	1.209
Private Eye	-0.362	-0.042	-0.122	0.023
Qbert	0.808	0.875	1.215	0.722
Road Runner	6.030	4.903	6.657	4.668
Seaquest	-0.207	0.289	0.200	0.559
Up N Down	8.665	8.169	9.822	10.361
IQM	0.530	0.556	0.589	0.630
Mean	1.170	1.210	1.280	1.350

Table 39: **Atari 100K game-level human-normalized scores at $\bar{B} = 0.80$.** Game means are shown in the canonical benchmark order rather than sorted by performance. The heterogeneous game-wise ordering is consistent with the unresolved aggregate CAP-WM–One-step comparison.

C Additional Generality and Joint-Boundary Validation

C.1 Transfer Beyond RSSM-Style World Models

The DreamerV3–R2-Dreamer comparison changes the recurrent representation learner while retaining an RSSM-style world-model structure. To test whether the observed allocation behavior depends on this family, we apply the same bounded-interface principle to TD-MPC2 (Hansen et al., 2024), an architecturally distinct deterministic latent-dynamics world model whose learned latent state directly supports subsequent prediction and control. This experiment asks whether the tight-budget benefit specific to multi-step consequence survives a change in latent parameterization and world-model training objective.

Bounded latent interface. For the TD-MPC2 transfer experiment, we use its 512-dimensional latent state, $\boldsymbol{z}_t \in \mathbb{R}^{512}$, produced by the encoder at observation steps or by the latent dynamics during model rollout. We retain the same allocation granularity as in the primary experiments, $G = 16$, giving equal-width groups of $d_z = 512/16 = 32$ coordinates. Each group is standardized using the same running-statistics protocol as the recurrent experiments. For a total per-transition budget B, the transmitted state is

$$\boldsymbol{m}_t = \mathcal{Q}(\boldsymbol{z}_t; \boldsymbol{b}_t)\,, \qquad \sum_{g=1}^{16} b_{t,g} = B, \qquad \bar{B} = \frac{B}{512},$$

where $\mathcal{Q}$ is the same stochastic precision-channel family used in the main experiments. All subsequent latent dynamics, value prediction, policy computation, and model-based control consume $\boldsymbol{m}_t$ rather than the untransmitted $\boldsymbol{z}_t$. Predicted latent states inside model rollouts are passed through the same bounded interface before the next latent transition. The consequence scores and allocation vector configure only the transmission channel and are never supplied as auxiliary downstream features. Thus the transfer preserves the no-bypass constraint and the $G = 16$ allocation granularity while changing the underlying world-model architecture and native latent dimensionality.

Consequence construction. For a fixed action continuation, we define the groupwise predictive consequence

$$c_{t,g}(\boldsymbol{m}) = \sum_{h \in \mathcal{H}} \gamma_h \left\| \frac{\partial F_h(\boldsymbol{m}, \boldsymbol{a}_{t:t+h-1})}{\partial \boldsymbol{m}_g} \right\|_F^2, \qquad \mathcal{H} = \{1, 2, 4, 8\},$$

where F_h denotes the h-step latent prediction induced by TD-MPC2. Following the main protocol, the training-only pathwise teacher is evaluated from a common allocation-independent reference and distilled into a state-conditioned consequence predictor. One-step consequence changes only the horizon set to $\mathcal{H} = \{1\}$, whereas Uniform assigns the same rate to every group. All variants otherwise share the same backbone, objectives, latent grouping, interaction budget, channel, and normalized total rate.

Protocol. We evaluate the six DMC-GB2 training tasks with five matched training seeds at $\bar{B} \in \{0.20, 0.80\}$. The tight-budget condition tests whether the additional multi-step benefit transfers beyond RSSM-style world models; the permissive condition tests whether the loss of separation observed on the recurrent backbones transfers as well. Aggregate IQM and paired stratified-bootstrap intervals follow the primary evaluation protocol and preserve matched task–seed units. Comparisons are made within backbone; equal $\bar{B}$ denotes equal normalized latent-state rate rather than equal absolute latent dimensionality across architectures.

Results. Figure 6 reveals the same tight-versus-permissive pattern across all three backbones. At $\bar{B} = 0.20$, the paired CAP-WM–One-step differences are $+24\,[8, 40]$ on DreamerV3, $+27\,[8, 46]$ on R2-Dreamer, and $+25\,[8, 42]$ on TD-MPC2. The additional multi-step effect is therefore resolved at the restrictive operating point under all three tested parameterizations. On TD-MPC2, CAP-WM also exceeds Uniform by $+77\,[47, 107]$, separating the broader benefit of consequence-aware allocation from the incremental contribution of extending the predictive horizon.

At $\bar{B} = 0.80$, the CAP-WM–One-step differences contract to $-3\,[-16, 10]$, $+6\,[-7, 19]$, and $+3\,[-10, 16]$ on DreamerV3, R2-Dreamer, and TD-MPC2, respectively, and all are unresolved.

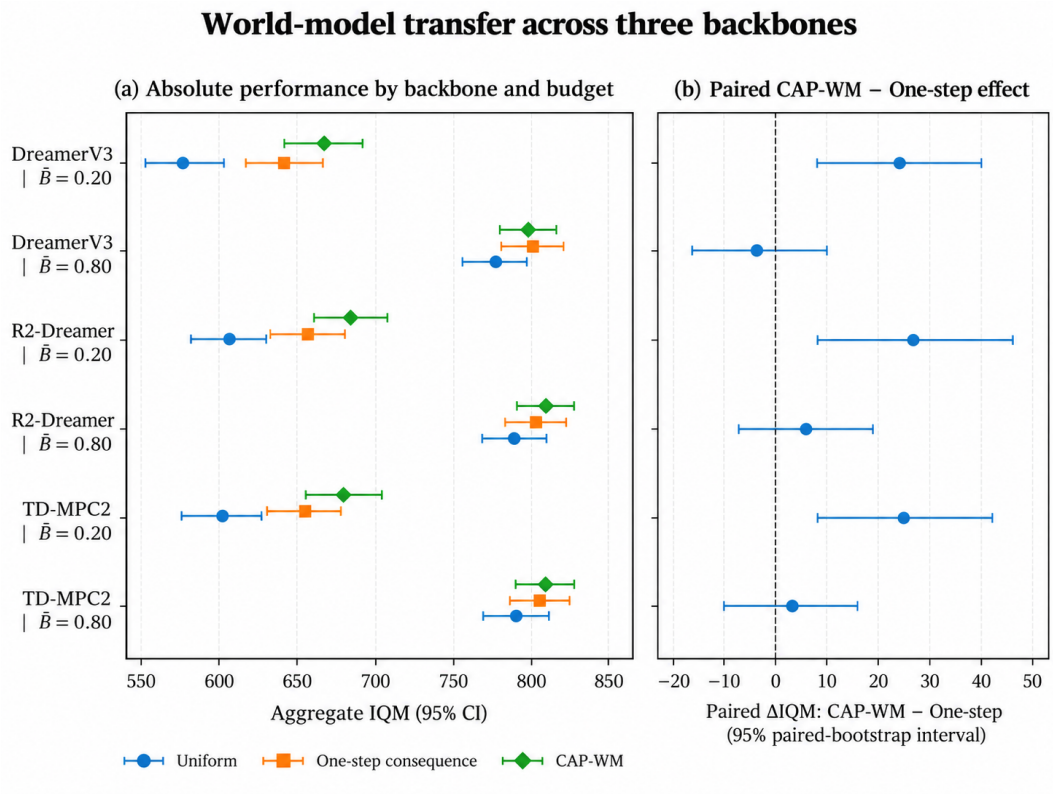


Figure 6: **World-model transfer across three backbones. (a)** Aggregate IQM with 95% stratified-bootstrap intervals for Uniform, One-step consequence, and CAP-WM at restrictive ($\bar{B} = 0.20$) and permissive ($\bar{B} = 0.80$) rates. **(b)** The primary paired comparison, $\Delta\text{IQM} = \text{CAP-WM} - \text{One-step}$, with 95% paired-bootstrap intervals. Positive intervals excluding zero indicate resolved additional benefit from multi-step consequence.

CAP-WM still exceeds Uniform on TD-MPC2 by $+17$ $[2, 32]$, indicating that adaptive consequence-aware allocation can remain useful even when the additional multi-step term no longer separates from one-step consequence. The cross-backbone result therefore supports the proposed allocation principle while preserving the central boundary: the benefit specific to the multi-step criterion is concentrated in the restrictive-rate regime.

Scope. The three-backbone comparison supports CAP-WM as an allocation principle rather than a modification tied to the primary DreamerV3 implementation. It does not establish architecture independence or a universal transition rate. All three tested architectures expose an identifiable latent-state boundary through which subsequent predictive and control computation can be constrained. Architectures without such a boundary require a separate definition of the bounded representation and remain outside the scope of these experiments.

C.2 Joint Resource–Horizon Boundary Test

The preceding experiments identify two empirical conditions associated with the additional benefit of multi-step consequence over One-step consequence: the recurrent representation budget must be sufficiently restrictive, and predictive sensitivity must extend appreciably beyond the first latent transition. The rate sweep and prospective action-repeat intervention examine these conditions separately. We therefore cross resource level and action repeat in a compact factorial experiment to test whether the CAP-WM–One-step difference varies jointly with resource pressure and long-horizon predictive dependence.

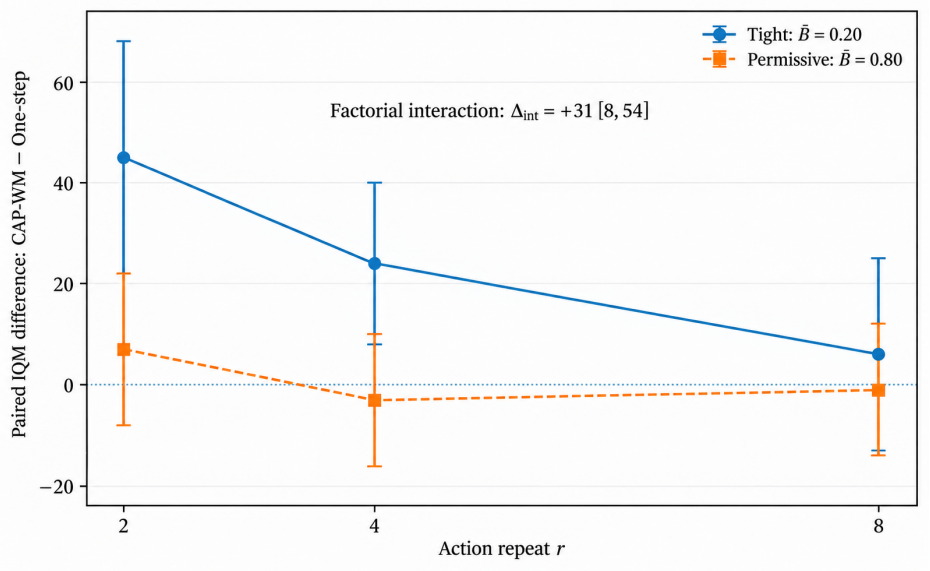


Figure 7: **Joint resource–horizon boundary test on DMC-GB2.** Each point reports the paired CAP-WM–One-step IQM difference with its 95% paired-bootstrap interval at action repeats $r \in \{2, 4, 8\}$. Under restrictive precision ($\bar{B} = 0.20$), the additional multi-step effect contracts as action repeat increases; under permissive precision ($\bar{B} = 0.80$), all three differences are unresolved. The resulting factorial interaction is $\Delta_{\text{int}} = +31\,[8, 54]$. The $\bar{B} = 0.20$ conditions and the $r = 4$ comparison reuse completed experiments; the remaining permissive-rate conditions complete the factorial design.

Design. We evaluate

$$\bar{B} \in \{0.20, 0.80\}, \qquad r \in \{2, 4, 8\},$$

where r denotes the environment action repeat. The complete $\bar{B} = 0.20$ column reuses the prospective action-repeat experiment, while the $r = 4$ row reuses the matched rate-sweep experiment. Completing the factorial design therefore requires only the $\bar{B} = 0.80$, $r \in \{2, 8\}$ conditions for CAP-WM and One-step consequence. Horizon sets, evaluation rules, and allocation hyperparameters are held fixed across all conditions and are not modified after observing these runs.

For each operating point, define

$$\Delta(\bar{B}, r) = \text{IQM}_{\text{CAP}}(\bar{B}, r) - \text{IQM}_{\text{1step}}(\bar{B}, r).$$

The joint-boundary hypothesis predicts that Δ should be largest when the recurrent resource is restrictive and substantial predictive sensitivity lies beyond one transition. We quantify the interaction as

$$\Delta_{\text{int}} = [\Delta(0.20, 2) - \Delta(0.20, 8)] - [\Delta(0.80, 2) - \Delta(0.80, 8)]\,.$$

Uncertainty is estimated by paired stratified bootstrap over matched task–seed units.

Results. Figure 7 sharpens the proposed applicability boundary. At $\bar{B} = 0.20$, increasing action repeat from $r = 2$ to $r = 8$ reduces the independently measured L_{long} from 0.69 to 0.55 and 0.37, while the CAP-WM–One-step difference contracts from $+45\,[22, 68]$ to $+24\,[8, 40]$ and $+6\,[-13, 25]$. The separation is resolved at $r \in \{2, 4\}$ but not at $r = 8$.

At $\bar{B} = 0.80$, the corresponding differences are $+7\,[-8, 22]$, $-3\,[-16, 10]$, and $-1\,[-14, 12]$, all unresolved. The difference-in-differences interaction is

$$\Delta_{\text{int}} = +31\,[8, 54],$$

whose paired-bootstrap interval excludes zero. Thus changing temporal discretization produces a substantially larger change in the CAP-WM–One-step effect when recurrent precision is restrictive than when it is permissive. Together with the prospectively measured ordering in L_{long}, this supports the proposed boundary: the additional benefit of multi-step consequence is resolved when

restrictive precision coincides with appreciable predictive dependence beyond one latent transition, but becomes unresolved when either condition weakens.

Scope of the interaction. The experiment directly manipulates resource level and temporal discretization; it does not manipulate L_{long} itself. Because changing action repeat can alter both predictive horizon and control difficulty, the resolved interaction should not be interpreted as causal mediation through L_{long}. Instead, the prospective L_{long} measurements establish that the temporal intervention moves long-horizon predictive dependence in the predicted direction, while the factorial comparison shows that the CAP-WM–One-step advantage depends jointly on temporal regime and resource pressure. The result is therefore an applicability test of the proposed boundary rather than a universal causal characterization.