What Should a Finite World Model Preserve? Recomposable Consequence–Constraint Factorization for Finite Representations

Anonymous authors
Paper under double-blind review

Abstract

Finite world models must decide which distinctions deserve limited representational resolution. We identify a *recomposable consequence–constraint factorization*: learn how preserving each latent distinction at different resolutions affects downstream quality, then combine that reusable consequence model with current decision relevance and capacity only at allocation time. Uniform consequence error bounds allocation regret and yields exact recovery below the oracle margin. Under a unified calibrated-quantizer protocol, StaticImportance+DP recovers only 1.67% of held-out oracle allocations, ConditionalParametric+DP reaches $93.37\% \pm 1.32\%$, the more flexible FullConditional+DP reaches $85.74\% \pm 4.38\%$, and Direct-MLP+Slack reaches $53.56\% \pm 6.63\%$. Thus static importance is insufficient, consequence–constraint factorization transfers substantially better than the tested direct allocator, and additional function-space flexibility need not improve allocation. An independently reconstructed learned-latent diagnostic further shows that measurable consequence-shape variation can persist without robustly changing marginal allocation priority once near-tie comparisons are removed. The resulting principle is to preserve *decision-relevant rate structure*, not every measurable variation in the consequence function.

1 Introduction

World models compress interaction histories into internal variables for prediction, memory, planning, and control. Under finite latent capacity, they must decide *which distinctions deserve representational resolution*, a problem arising in compact latent dynamics, discrete world models, compressed memories, and token-limited agents.

Predictive/world-model methods preserve forecasting-relevant distinctions (Littman et al., 2002; Singh et al., 2004; Boots et al., 2011; Ha and Schmidhuber, 2018; Hafner et al., 2019; 2020; Hansen et al., 2024; Hafner et al., 2025); information bottlenecks and learned compression trade information or distortion against rate (Tishby et al., 1999; Alemi et al., 2017; Achille and Soatto, 2018; Agustsson et al., 2017; Ballé et al., 2018); decision-focused learning incorporates downstream optimization (Donti et al., 2017; Wilder et al., 2019; Elmachtoub and Grigas, 2022). We instead study finite latent allocation when *decision relevance, observational support, and capacity disagree*.

We study that allocation problem directly. For representational direction i, let d_i denote downstream decision relevance, o_i observational support, r_i allocated representational capacity, and

$$D_i(r_i; o_i, h) \tag{1}$$

the distortion remaining under context h. The central object is

$$\boxed{\boldsymbol{r}^\star = \arg\min_{\boldsymbol{r} \in \mathbb{Z}^m_{\geq 0}} \sum_{i=1}^{m} d_i D_i(r_i; o_i, h) \quad \text{s.t.} \quad 0 \leq r_i \leq K_i,\ \forall i, \qquad \sum_i r_i \leq B.} \tag{2}$$

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.

Controlled falsification ruled out fixed observation/Fisher weighting, decision sensitivity, and hard observable projection: none captures how marginal representational value changes with allocation and budget.

Oracle experiments show that optimal allocation is budget-conditioned,

$$\boldsymbol{r}^\star = f(\boldsymbol{d}, \boldsymbol{o}, B), \tag{3}$$

with observability and relevance changing roles with scarcity. Measured direction-specific distortion–rate curves nearly reconstruct the brute-force oracle, suggesting:

$$\boxed{\underbrace{\text{learn allocation consequences}}_{\widehat{D}_\theta} + \underbrace{\text{optimize the current finite budget}}_{\text{explicit constrained solver}}.} \tag{4}$$

We call this *Learn–then–Allocate*. It is a structured predict–then–optimize (PTO) construction whose decision variables distribute finite internal resolution across latent distinctions.

Why this is a world-model problem. The budget in eq. (2) need not be interpreted literally as scalar bits. In a modern world model, finite internal capacity may appear as a limited number of object tokens, latent groups, memory slots, codebook entries, retained feature channels, or task-relevant subspaces. Across these realizations, the same question recurs: when not every distinction can be preserved at full resolution, which distinctions deserve the next unit of representational capacity? A static saliency score answers only which component appears important in isolation. It does not specify whether the *next* unit of capacity remains valuable after that component has already received some resolution. This distinction motivates a rate-conditioned consequence profile rather than a single importance value.

Our contribution is not a new generic PTO architecture, nor the observation that marginal value can depend on rate. We use PTO to answer a representation-specific interface question: *what should be predicted so that finite latent resolution can be recomposed under changing sensing, decision relevance, and capacity?* We propose a *recomposable consequence–constraint factorization*. A consequence model $\widehat{D}_i(k; o_i, h)$ describes the downstream effect of retaining latent direction i at level k under the current observation context; transient relevance d_i and capacity B enter only when the optimizer forms the current allocation. The consequence model may be parametric or flexible—its complexity should match the observed variation of the rate–consequence law. Conceptually, allocation is therefore a context-conditioned decision rather than a fixed property of the latent space. Theoretically and empirically, we connect consequence error to allocation regret and test oracle reconstruction, mechanism structure, model parsimony, factorized-versus-direct transfer, and quantizer, distributional, temporal, and coupling boundaries.

2 Problem Formulation

Let x denote an underlying task state and $y \sim p(y \mid x, h)$ an observation under context h. A finite representation exposes m allocatable directions or factors. For direction i, $d_i \geq 0$ is decision relevance, o_i denotes available observational support, and $r_i \in \mathbb{Z}_{\geq 0}$ is the amount of representational capacity assigned to it. Each direction has a maximum feasible rate $K_i \in \mathbb{Z}_{\geq 0}$, so $0 \leq r_i \leq K_i$. The residual distortion is $D_i(r_i; o_i, h)$.

In the separable setting,

$$D_{\text{task}}(\boldsymbol{r}) = \sum_{i=1}^{m} d_i D_i(r_i; o_i, h), \tag{5}$$

with budget

$$0 \leq r_i \leq K_i, \ \forall i, \qquad \sum_i r_i \leq B. \tag{6}$$

Observability and relevance play distinct roles. A direction with high d_i but weak o_i may be important yet expensive to represent reliably. A direction with high o_i but small d_i may be

Table 1: One interpretation of the abstract allocation variables in finite world models. The formulation is agnostic to the physical unit of capacity.

Quantity	World-model interpretation
i	latent group, object/token family, memory unit, codebook block, or task-relevant subspace
k or r_i	retained resolution/capacity assigned to component i
$D_i(k; o_i, h)$	downstream loss remaining after component i is represented at level k under current sensing/context
d_i	current decision relevance of component i
B	total internal representation budget available to the allocator

easy to encode yet unimportant downstream. What matters for allocation is therefore not a static importance score, but the capacity-dependent reduction

$$D_i(r_i; o_i, h) - D_i(r_i + 1; o_i, h), \qquad 0 \le r_i < K_i. \tag{7}$$

The separation in Table 1 is deliberate: (o_i, h) describe what can be reliably preserved, d_i describes whether preserving it matters now, and B describes what is jointly feasible. The learned consequence model is intended to remain reusable when d_i and B change, instead of absorbing them into one task-specific allocation policy.

3 Decision-Weighted Distortion Allocation

3.1 General discrete allocation

Theorem 1 (Exact discrete allocation)**.** *For arbitrary direction-specific distortion functions $D_i : \{0, \dots, K_i\} \to \mathbb{R}$, nonnegative d_i, per-direction caps K_i, and integer budget B, the global optimum of eq. (2) is obtained by*

$$F(j, b) = \min_{0 \le k \le \min\{b, K_j\}} \{F(j-1, b-k) + d_j D_j(k; o_j, h)\}, \tag{8}$$

with $F(0, b) = 0$. A direct implementation has complexity $O(mB^2)$.

The proof and low-rate unused-budget discussion are deferred to Appendix A.

3.2 Marginal and continuous special cases

Define the k-th marginal reduction

$$\Delta_i(k; o_i, h) = D_i(k-1; o_i, h) - D_i(k; o_i, h), \tag{9}$$

and decision-weighted marginal gain $g_i(k) = d_i \Delta_i(k; o_i, h)$.

The quantity $g_i(k)$ makes the representation-specific mechanism explicit. If each direction is collapsed to a budget-independent scalar s_i, then its effective priority cannot express how the value of additional resolution changes after previous units have been assigned. By contrast, a rate-conditioned profile supplies the sequence

$$g_i(1),\ g_i(2),\ \dots,\ g_i(K_i), \tag{10}$$

so two directions can exchange priority as capacity is consumed. Rate dependence by itself is not the claimed novelty; classical allocation models already exploit diminishing returns. The representation-specific formulation allows the consequence law itself to depend on observational support and latent context, so the marginal value of rate need not follow one fixed global shape. The full discrete table $D_i(k; o_i, h)$ is the most general realization considered here, while parametric rate curves remain admissible when their structure is known and adequate.

Corollary 1 (Discrete marginal water filling)**.** *If every $\Delta_i(k)$ is nonnegative and satisfies diminishing returns, $\Delta_i(k+1) \le \Delta_i(k)$, then an optimal allocation is obtained by repeatedly purchasing the largest currently available $d_i \Delta_i(r_i + 1)$.*

The condition is not automatic. Our low-bit experiments contain $D_i(1) > D_i(0)$, so the general discrete problem is genuinely required.

For differentiable convex decreasing curves, the continuous optimum equalizes decision-weighted marginal reduction, $d_i[-D_i'(r_i)] = \mu$. The Gaussian high-rate form and derivation are given in Appendix B; neither is assumed by the discrete theory.

4 Recomposable Consequence–Constraint Factorization

The architecture separates representation consequence modeling from constrained allocation. We refer to the resulting implementation as *Learn–then–Allocate*. A consequence model estimates

$$(h, o_i, k) \mapsto \widehat{D}_\theta(i, k \mid h, o_i), \qquad k = 0, \dots, K_i. \tag{11}$$

If relevance is unknown, an auxiliary predictor may estimate $\widehat{d}_i = d_\psi(h, i)$. Our controlled experiments provide d_i in order to isolate the representation-allocation mechanism.

At inference time,

$$\boxed{\widehat{\boldsymbol{r}}^\star = \operatorname*{arg\,min}_{\substack{\boldsymbol{r} \in \mathbb{Z}^m_{\geq 0},\ 0 \leq r_i \leq K_i \\ \sum_i r_i \leq B}} \sum_i \widehat{d}_i \widehat{D}_\theta(i, r_i \mid h, o_i).} \tag{12}$$

Budget B is not a predictor input: changing it only alters the optimizer's feasible set, enabling *structural reuse across budgets* without extrapolating a budget-specific policy.

Representation-specific factorization. This is a PTO factorization, not an alternative to PTO. Its representation-level hypothesis is that the reusable statistical object is a *consequence model separated from transient relevance and capacity*, not allocation identity and not a single importance score. We learn $(h, o_i, k) \mapsto \widehat{D}_i(k; o_i, h)$ and deliberately exclude the current budget from that predictor; current relevance and capacity are recomposed explicitly through the objective and feasible set. The consequence family itself need not be fully flexible: a low-dimensional parametric law is preferable when it captures the observed rate structure, while a richer model is justified only when normalized consequence shape varies materially with context. A direct policy instead learns the allocation map jointly and therefore entangles statistical prediction with a particular relevance–constraint configuration. Consequence targets can be obtained without oracle allocation labels by masking, coarsening, pruning, merging tokens, restricting memory blocks, or otherwise rate-limiting latent components and measuring induced downstream loss.

4.1 Consequence prediction error and allocation regret

Let

$$\mathcal{R}_B = \left\{ \boldsymbol{r} \in \mathbb{Z}^m_{\geq 0} : 0 \leq r_i \leq K_i,\ \forall i,\ \sum_i r_i \leq B \right\} \tag{13}$$

denote the feasible allocation set, and define

$$F(\boldsymbol{r}) = \sum_{i=1}^m d_i D_i(r_i; o_i, h), \qquad \widehat{F}(\boldsymbol{r}) = \sum_{i=1}^m d_i \widehat{D}_i(r_i; o_i, h). \tag{14}$$

Let

$$\boldsymbol{r}^\star \in \operatorname*{arg\,min}_{\boldsymbol{r} \in \mathcal{R}_B} F(\boldsymbol{r}), \qquad \widehat{\boldsymbol{r}} \in \operatorname*{arg\,min}_{\boldsymbol{r} \in \mathcal{R}_B} \widehat{F}(\boldsymbol{r}). \tag{15}$$

Theorem 2 (Consequence error controls allocation regret)**.** *Assume that, for every direction i and every rate level k appearing in $\mathcal{R}_B$,*

$$\left| \widehat{D}_i(k; o_i, h) - D_i(k; o_i, h) \right| \leq \epsilon_i. \tag{16}$$

Define $E_d = \sum_i d_i \epsilon_i$. Then

$$\boxed{F(\widehat{\boldsymbol{r}}) - F(\boldsymbol{r}^\star) \leq 2E_d = 2 \sum_i d_i \epsilon_i.} \tag{17}$$

If $\epsilon_i \le \epsilon$ for all i, then

$$F(\widehat{\boldsymbol{r}}) - F(\boldsymbol{r}^\star) \le 2\epsilon \sum_i d_i. \tag{18}$$

The proof is in Appendix A; the bound shows that the relevant learned error is objective perturbation, not allocation-label error.

Corollary 2 (Margin-based exact allocation recovery)**.** *Assume the oracle allocation $\boldsymbol{r}^\star$ is unique, and define its allocation margin*

$$\gamma_B = \min_{\boldsymbol{r} \in \mathcal{R}_B,\ \boldsymbol{r} \neq \boldsymbol{r}^\star} \left[F(\boldsymbol{r}) - F(\boldsymbol{r}^\star)\right]. \tag{19}$$

If

$$\boxed{2 \sum_i d_i \epsilon_i < \gamma_B,} \tag{20}$$

then

$$\boxed{\widehat{\boldsymbol{r}} = \boldsymbol{r}^\star.} \tag{21}$$

If the uniform error condition holds at every feasible rate $0 \le k \le K_i$ for every direction i, the same consequence model applies without retraining to any total budget B up to the maximum useful capacity $\sum_i K_i$; changing B only changes the feasible set. This does not imply distributional robustness of $\widehat{D}_\theta$ (Appendix A).

This structure also clarifies a boundary. The solver can generalize across constraints it receives explicitly, but it cannot by itself make $\widehat{D}_\theta$ distributionally robust. Hence

$$\boxed{\underbrace{\text{constraint OOD}}_{\text{handled by optimizer}} \neq \underbrace{\text{environment/context OOD}}_{\widehat{D}_\theta \text{ must generalize}}.} \tag{22}$$

We test both regimes below.

Why consequence prediction is the reusable target. The regret bound and margin condition formalize the advantage of predicting consequences rather than allocation labels. The learner need not reproduce the discrete optimizer's output directly. It only needs to keep the objective perturbation induced by consequence error sufficiently small; if that perturbation is below the oracle allocation margin, the exact optimizer output is recovered automatically. Thus the statistical task and the constraint-solving task remain separated: $\widehat{D}_\theta$ models what is lost at each feasible representation level, while the solver composes those predictions with the current d_i and B. This separation is especially useful when the feasible set changes more often than the consequence model should be retrained.

5 Results

Full protocols are in Appendix C.

5.1 RQ1: oracle allocation changes with scarcity

Across $d_i \in \{0.5, 1.5, 4.0\}$, $o_i \in \{0.15, 0.8, 4.0\}$, and $B \in \{2, 4, 6, 8\}$, the oracle exhibits a regime shift: at $B = 2$, $\rho(r^\star, o) = 0.770$ versus $\rho(r^\star, d) = 0.192$, whereas at $B = 8$ the ordering reverses (0.379 versus 0.785). Thus the relative roles of observability and relevance change with representational scarcity.

5.2 RQ2: measured distortion curves reconstruct the oracle

We measure

$$D_i(r; o_i) = \mathbb{E}\left[(x_i - q_r(\mathbb{E}[x_i \mid y_i]))^2\right] \tag{23}$$

and insert those curves into theorem 1. The resulting allocator reproduces 98.61% of brute-force oracle allocations, with 98.61% active-set match, 0.056 bit mean ℓ_1 error, and

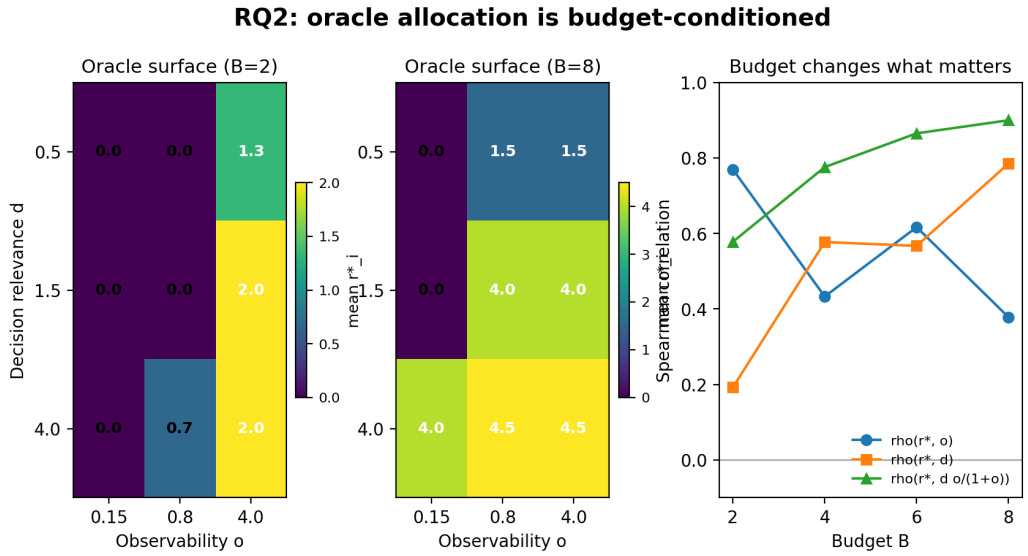


Figure 1: Oracle allocation surfaces and budget-dependent regime shift. At low budget, allocation tracks what is both useful and reliably observable; as budget increases, decision relevance becomes more dominant.

approximately 10^{-4} mean regret. For $B \in \{4, 6, 8\}$, all tested allocations are recovered exactly.

Under the fixed-range uniform quantizer used in the controlled stress test, the high-rate closed form reaches only 15.3% exact match because the empirical low-bit distortion curve can be nonmonotone. This pathology is quantizer-specific rather than a pillar of the method. A five-seed robustness rerun with MSE-calibrated uniform and Lloyd–Max quantizers restores essentially monotone curves and still gives Learned-DP lower mean joint-OOD regret than Direct-MLP+Slack (0.0270 vs. 0.0442, and 0.0276 vs. 0.0379, respectively); full results are in Appendix D.

5.3 Mechanism ablation: minimum structure required by the consequence model

We hold the DP solver fixed and change only the consequence representation. **StaticImportance** uses one budget-independent score $d_i[D_i(0) - D_i(8)]$; **ScalarLinear-DP** uses the same endpoint amplitude with constant marginal value; **GenericHighRate-DP** imposes a fixed diminishing-return shape; **FullCurve-DP** uses the measured $D_i(k; o_i, h)$ table.

In the original RQ1 pipeline, pairwise allocation preference changes from $B = 2$ to $B = 8$ for 55.6% of FullCurve pairs, 11.1% of GenericHighRate pairs, and 0% of StaticImportance/ScalarLinear pairs. Under MSE-calibrated quantization, where the severe low-bit pathology is removed, static/linear regret rises from 0.160 to 1.070 as B grows from 2 to 8, whereas GenericHighRate remains close to the full-curve oracle $(0.065, 0.029, 0.019, 0.0003)$. The robust conclusion is therefore structural: scarcity-sensitive allocation cannot generally be represented by a budget-independent importance scalar. A reusable consequence model must retain how marginal loss changes with rate; in our formulation that rate–consequence relationship is additionally conditioned on observation and context through $D_i(k; o_i, h)$.

The GenericHighRate result is informative rather than a failed baseline. Once a surrogate preserves a plausible diminishing-return profile, it recovers substantially more allocation structure than static or constant-marginal alternatives. The conclusion is therefore a minimum-structure statement, not a preference for maximal flexibility: the consequence model must retain rate-dependent marginal effects, but its functional complexity should be determined empirically.

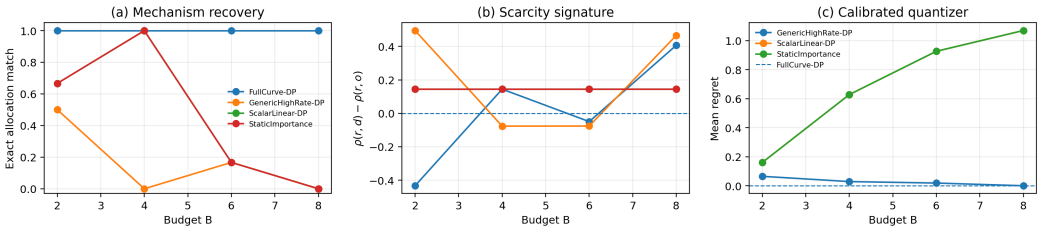


Figure 2: Representation-specific mechanism ablation. Holding the optimizer fixed, static or constant-marginal surrogates cannot reproduce scarcity-sensitive allocation; rate-dependent models recover substantially more structure. The learned formulation generalizes this idea by conditioning the consequence profile on observation and context rather than assuming one fixed rate law.

Table 2: Unified calibrated-quantizer benchmark. Values are mean ± SD over five seeds; StaticImportance is deterministic. Curve MAE/RMSE are reported only for learned consequence predictors.

Method	Exact	Active	Mean ℓ_1	Regret	MAE/RMSE
StaticImportance+DP	1.67%	1.67%	4.800	0.63905	—
ConditionalParametric+DP	**93.37 ± 1.32%**	**96.33%**	**0.133**	**0.000397 ± 0.000169**	0.01042/0.01238
FullConditional+DP	85.74 ± 4.38%	95.19%	0.285	0.002074 ± 0.001535	**0.01033/0.01204**
Direct-MLP+Slack	53.56 ± 6.63%	83.33%	0.931	0.03336 ± 0.00539	—

5.4 Unified factorization benchmark

We place the principal structured and direct baselines under one calibrated-uniform protocol: the same consequence-curve table, observability split, held-out relevance values, unseen budgets, five neural seeds, true objective, and evaluation set of 540 joint held-out (d, o, B) conditions. StaticImportance, ConditionalParametric, and FullConditional all use the *same* discrete DP; Direct-MLP+Slack uses the same explicit slack decoder as the fairness rerun. This removes the numerical ambiguity caused by comparing independently regenerated robustness experiments.

Three conclusions follow. First, StaticImportance+DP is a deliberately strong structural diagnostic—it receives the true held-out endpoint amplitude yet fails because constant marginal value cannot represent rate-dependent allocation. Second, the three-parameter ConditionalParametric model is sufficient in this calibrated Gaussian regime and generalizes better at the *allocation* level than FullConditional, even though FullConditional has slightly lower held-out curve MAE/RMSE. Prediction-space accuracy therefore need not order allocation quality. Third, factorized consequence prediction plus explicit optimization transfers substantially better than the tested direct allocation learner; the latter still fits approximately 96% of seen training allocations, so the held-out gap is not explained by gross training underfit.

The earlier 97.22% RQ4 result and 60.44% calibrated-quantizer robustness result come from independently regenerated protocols and are retained only as separate robustness checks in the appendix. The unified comparison in table 2 is the primary cross-method result.

5.5 Decision-relevance boundary in a reconstructed learned latent

We next ask whether observable variation in a consequence function is sufficient to justify a more flexible consequence model. Using the stated RQ5 architecture, protocol, and ten seeds, but independently reconstructing the models because the original checkpoints were not retained, we measure per-coordinate consequence curves over four goal contexts, three sensing-noise levels, and $k = 0, \ldots, 4$. This diagnostic is therefore a boundary study, not an RQ5 replication; its numerical results are not pooled with the original RQ5 control evaluation.

Naive endpoint normalization is unstable when $D_i(0) - D_i(4)$ is nearly zero, so we apply a measurement-validity gate before interpreting shape differences. Context-dependent normalized-shape variation remains substantial as the minimum retained endpoint effect increases from 0 to 0.02:

$$S: \ 3.108 \to 0.724 \to 0.471 \to 0.265 \to 0.214.$$

Thus the shape variation is not purely a denominator artifact. However, the decision-relevant marginal-order reversal rate decreases after excluding near-tie comparisons:

$$16.82\% \to 12.34\% \to 9.63\% \to 8.40\%$$

for marginal-separation thresholds $\epsilon = 0, 0.001, 0.0025, 0.005$. Before inspecting this learned-latent diagnostic, we specified 10% as a practical stopping criterion for whether effect-size-filtered marginal-order reversal would justify a more flexible consequence model; the value is a diagnostic go/no-go threshold, not a theoretically privileged constant. More importantly, the reversal rate falls systematically from 16.82% to 8.40% as near-tie comparisons are removed. We therefore classify this as a *Case-2* boundary: function-space consequence variation is real, but it does not robustly translate into marginal-priority changes large enough to justify additional consequence-model flexibility.

This observation is the empirical analogue of theorem 2 and corollary 2: consequence functions need not be pointwise identical for the optimizer to return the same allocation. The relevant model complexity is determined by whether approximation error changes the decision-relevant ordering or crosses an allocation margin, not by whether every measurable feature of the consequence curve is reproduced.

5.6 RQ5: learned-latent control and horizon-dependent loss under frozen allocation

In a nonlinear partially observed 2-D navigation model with four learned latent dimensions, ten independently trained models are evaluated. Learned-DP approximately halves one-step representation regret relative to Uniform (0.0228 ± 0.0090 vs. 0.0456 ± 0.0208). When the initial allocation is frozen across a rollout, regret normalized to Uniform rises $0.50, 0.67, 0.82, 1.19$ for $H = 1, 10, 25, 50$. This supports viewing allocation as context-conditioned rather than permanent, but does not itself validate dynamic re-allocation; trajectory divergence and variance may also contribute. Under joint context+sensing shift, the one-step advantage disappears (0.0478 ± 0.0235 vs. 0.0447 ± 0.0218).

The natural sequential extension is therefore not to treat one allocation as part of the learned representation forever, but to index the allocation by the current decision context:

$$\boldsymbol{r}_t^\star = \underset{\sum_i r_{i,t} \le B_t}{\arg\min} \sum_i d_{i,t} D_i(r_{i,t}; o_{i,t}, h_t). \tag{24}$$

RQ5 does not evaluate this receding re-allocation policy; it deliberately freezes $\boldsymbol{r}_0$ to probe how long a locally useful allocation remains competitive as the closed-loop trajectory evolves. The observed crossover therefore motivates, rather than proves the benefit of, state-conditioned re-solving.

5.7 Failure boundary: controlled nonseparability

We add an unmodeled pairwise coupling term to the separable objective (Appendix C.5). As coupling grows from $\lambda = 0.05$ to 0.50, exact match falls from 95.83% to 67.36%, while normalized regret rises only from 0.072% to 1.51%.

Thus separability can fail structurally before it fails behaviorally; the figure and protocol are in Appendix C.5.

6 Related Work

Classical rate–distortion and bit-allocation theory already establish that marginal value can vary with rate (Shannon, 1948; Berger, 1971; Cover and Thomas, 2006; Gersho and

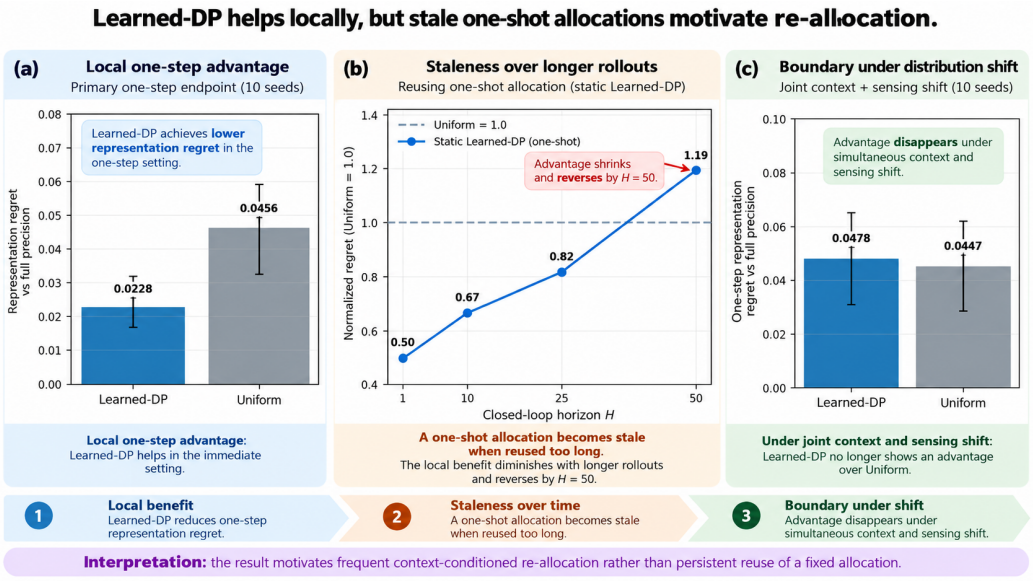


Figure 3: Relative value of a frozen allocation decays with rollout horizon. Uniform aggregate regret is normalized to 1.0 independently at each horizon. Static Learned-DP rises from 0.50 at $H = 1$ to 1.19 at $H = 50$, indicating that the local advantage of the initial context-specific allocation progressively disappears, with the aggregate mean crossing above Uniform at $H = 50$ under persistent reuse. This diagnostic does not by itself identify the cause of the decay.

Gray, 1992; Gray and Neuhoff, 1998). Learned compression extends rate allocation to trainable codecs (Agustsson et al., 2017; Ballé et al., 2018). Our claim is narrower and representation-specific: the rate–consequence law itself is conditioned on what the current observation supports and on latent context, then recomposed with current decision relevance and finite internal capacity. Information bottlenecks, task-aware abstractions, and world models study which distinctions should survive compression (Tishby et al., 1999; Alemi et al., 2017; Ferns et al., 2004; Gelada et al., 2019; Zhang et al., 2021; Hafner et al., 2019; Hansen et al., 2024); we study how much resolution each surviving distinction should receive when observation support, relevance, and scarcity vary independently.

Dynamic-computation and token-selection methods allocate compute or token count (Graves, 2016; Rao et al., 2021; Bolya et al., 2023). Learn–then–Allocate instead treats representational resolution itself as the scarce resource. Structurally, the method is predict–then–optimize / decision-focused learning (Donti et al., 2017; Wilder et al., 2019; Elmachtoub and Grigas, 2022; Mandi et al., 2022; Amos and Kolter, 2017; Agrawal et al., 2019). The contribution is therefore not a new generic optimizer. PTO supplies the general predict–then–optimize template; our representation-specific contribution is the *prediction interface*: model reusable representation consequence separately from transient relevance and capacity, then let the optimizer recompose them. The new parsimony audit further shows that this interface need not be nonparametric when a low-dimensional rate law is adequate.

7 Discussion

The scientific claim is not that consequence prediction lies outside PTO, nor that rate-dependent marginal utility is itself new. The representation-specific claim is the *consequence–constraint factorization*: expose a reusable model of representation consequence, separate it from transient relevance and capacity, and recompose them only when an allocation is needed. The unified benchmark establishes three distinct facts under one protocol: static importance is too weak, a simple conditional rate family can be sufficient, and explicit factorization transfers better than the tested direct allocator. The learned-latent boundary diagnostic then adds the missing qualification: function-space consequence variation alone is

not evidence that a richer model is allocation-relevant. Model complexity should increase only when those variations alter decision-relevant marginal ordering or exceed the allocation margin.

The experiments also delimit the scope. Explicit optimization structurally adapts to the supplied budget but does not make $\widehat{D}_\theta$ robust to context shift. A one-shot allocation can lose its advantage as a trajectory evolves, motivating but not proving state-conditioned re-allocation. Coupled codebooks, attention, or latent factors may require structured consequence models beyond the separable objective. Relevance d_i is supplied in primary experiments, large-scale consequence supervision remains open, and RQ5 is synthetic.

Operational implication for adaptive world models. The formulation suggests an interface between representation learning and resource management. Instead of exposing a single saliency value or a precomputed allocation, the representation module exposes a consequence model over feasible resolution levels, and the allocator recomposes that model with current task relevance and capacity. The experiments refine the model-selection rule: use the simplest consequence family that preserves *decision-relevant rate structure*. Observable curve-shape variation by itself is insufficient; what matters is whether approximation changes marginal priority or crosses the optimizer's decision boundary. Such a model may therefore be parametric in stable regimes and more flexible only when context variation is demonstrably allocation-relevant.

What remains to establish at scale. The controlled experiments isolate mechanism and failure boundaries, but do not establish that the same gains will persist in large pretrained world models. At that scale, components may interact strongly through attention, shared codebooks, or recurrent memory, violating separability; measuring $D_i(k; o_i, h)$ may also be expensive. A direct next test is therefore to group modern latent tokens or memory units into allocatable blocks, estimate their consequence profiles through structured ablations, and compare static saliency, parametric rate models, and learned consequence profiles under a fixed token or memory budget. This would test whether the mechanism identified here survives the coupling and dimensionality of contemporary world-model architectures.

8 Conclusion

Finite representation creates scarcity, but the appropriate consequence model is not determined by function-space expressivity alone. Under one unified calibrated protocol, static importance fails, a three-parameter conditional rate family recovers 93.37% of held-out oracle allocations, the more flexible FullConditional predictor is lower at 85.74%, and factorized consequence prediction transfers substantially better than the tested direct allocator. An independently reconstructed learned-latent diagnostic further shows that genuine context-dependent consequence-shape variation need not survive as decision-relevant marginal-priority reversal once near-tie comparisons are removed. Together with the regret and margin results, these experiments support a narrower principle:

factorize consequence from constraint;
preserve decision-relevant rate structure.

The simplest adequate consequence family is therefore the one that preserves the optimizer-relevant structure, not necessarily the one that best reproduces every measurable feature of the consequence function.

AI Use Disclosure

AI-assisted tools were used for limited manuscript editing, formatting assistance, and code-related support during development. The authors reviewed and verified all generated content, analyses, experimental results, and final manuscript text.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 136–145, 2017.

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.

Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Byron Boots, Sajid M. Siddiqi, and Geoffrey J. Gordon. Closing the learning-planning loop with predictive state representations. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 120–128, 2011.

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

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 162–169, 2004.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2170–2179, 2019.

Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998. doi: 10.1109/18.720541.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

Jayanta Mandi, V
'ictor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14935–14947. PMLR, 2022.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pages 13937–13949, 2021.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3–4):379–423, 623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 512–519, 2004.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

A Deferred Theory Details

A.1 Proof of Theorem 1 and unused-budget remark

Proof. Condition on assigning a feasible $k \in \{0, \dots, \min\{b, K_j\}\}$ units to direction j. The remaining first $j-1$ directions must solve the same problem under residual budget $b-k$; otherwise that sub-allocation could be replaced by a better one. Optimal substructure yields eq. (8), and backtracking returns a global optimum. □

Remark 1 (Useful versus nominal capacity)**.** *The inequality $\sum_i r_i \le B$ is intentional. In low-rate regimes a nominal increment can increase distortion, so forcing exact budget use can be suboptimal. The general dynamic program may leave capacity unused.*

A.2 Proof of Theorem 2

Proof. By eq. (16), for every feasible $\boldsymbol{r}$, $|F(\boldsymbol{r}) - \widehat{F}(\boldsymbol{r})| \le E_d$. Therefore,

$$\begin{aligned} F(\widehat{\boldsymbol{r}}) &\le \widehat{F}(\widehat{\boldsymbol{r}}) + E_d \\ &\le \widehat{F}(\boldsymbol{r}^\star) + E_d \\ &\le F(\boldsymbol{r}^\star) + 2E_d, \end{aligned} \tag{25}$$

where the middle inequality follows from optimality of $\widehat{\boldsymbol{r}}$ under $\widehat{F}$. □

A.3 Proof of Corollary 2

Proof. For any $\boldsymbol{r} \neq \boldsymbol{r}^\star$,

$$\widehat{F}(\boldsymbol{r}) \ge F(\boldsymbol{r}) - E_d \ge F(\boldsymbol{r}^\star) + \gamma_B - E_d, \tag{26}$$

while

$$\widehat{F}(\boldsymbol{r}^\star) \le F(\boldsymbol{r}^\star) + E_d. \tag{27}$$

If $\gamma_B > 2E_d$, every alternative allocation has strictly larger learned objective than $\boldsymbol{r}^\star$. □

A.4 Budget reuse and scope of the guarantee

Remark 2 (Structural reuse across budgets)**.** *If eq. (16) holds for every feasible rate level $0 \le k \le K_i$ of every direction i, then theorem 2 applies without retraining to any total budget B up to the maximum useful capacity $\sum_i K_i$. Changing B changes only the feasible set $\mathcal{R}_B$. This is the theoretical counterpart of the unseen-budget evaluations in RQ3 and RQ4. In RQ5, $K_i = 4$ for all four latent coordinates, so $\sum_i K_i = 16$ and the evaluated budgets $B \in \{6, 10, 14\}$ are all feasible.*

Remark 3 (What the guarantee does not imply)**.** *The bound is deterministic and worst-case in the consequence error. It does not imply that $\widehat{D}_\theta$ generalizes under a new observation or task distribution. In particular, the empirical MAE/RMSE reported in RQ3 are average predictive diagnostics and are not substituted directly into eq. (17), which assumes a uniform error bound.*

B Gaussian High-Rate Special Case

For the Gaussian observation model

$$x_i \sim \mathcal{N}(0,1), \qquad y_i = \sqrt{o_i}\, x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1),$$

the total mean-squared distortion decomposes orthogonally as

$$D_i(r_i; o_i) = \underbrace{\frac{1}{1+o_i}}_{\text{irreducible observation error}} + \underbrace{D_i^{\text{repr}}(r_i; o_i)}_{\text{representation error}}. \tag{28}$$

The first term is independent of r_i and therefore does not affect the allocation optimum. The high-rate approximation is applied only to the representation-dependent term,

$$D_i^{\mathrm{repr}}(r_i; o_i) \approx c\, \frac{o_i}{1+o_i}\, 2^{-2r_i}, \tag{29}$$

which, when substituted into the KKT condition for active coordinates, yields the continuous-relaxation closed form

$$r_i^\star = \frac{1}{2}\left[\log_2 \frac{d_i\, o_i/(1+o_i)}{\nu}\right]_+, \tag{30}$$

where ν is chosen to satisfy the total capacity constraint and absorbs constants independent of i (including the Lagrange multiplier, c, and $2\ln 2$).

The fixed-range uniform quantizer used in the controlled stress test violates the high-rate approximation. At $o = 0.8$,

$$D(0) = 0.983, \qquad D(1) = 6.780, \qquad D(2) = 0.915.$$

This nonmonotonic point is quantizer-specific rather than a universal low-rate phenomenon; the general discrete formulation remains valid because it does not assume monotonicity or diminishing returns.

C Experimental Protocols and Reproducibility

C.1 Controlled oracle and learned-consequence benchmarks

The controlled state directions are independent with $x_i \sim \mathcal{N}(0,1)$ and observations

$$y_i = \sqrt{o_i} x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1).$$

For each condition, we estimate distortion by Monte Carlo sampling. The initial consequence-generalization experiment (RQ3) uses 16,000 samples per scalar distortion curve; the independent direct-policy rerun (RQ4) uses 12,000. The scalar posterior mean is uniformly quantized after clipping to $[-3, 3]$. For $r > 0$, the quantizer has 2^r uniformly spaced reconstruction levels; $r = 0$ maps the posterior mean to zero. We use $r \in \{0, \ldots, 8\}$.

The RQ3 consequence predictor receives four features

$$[\log o,\ r/8,\ \mathbb{I}(r=1),\ \mathbb{I}(r=0)]$$

and uses two hidden layers of width 48 with tanh activations, followed by a scalar output. It is trained with Adam (3×10^{-3} learning rate, 10^{-5} weight decay) for 3500 optimization steps using a relative squared-error objective

$$\mathcal{L}_D = \frac{1}{N}\sum_n \left(\frac{\widehat{D}_n - D_n}{|D_n| + 0.25}\right)^2.$$

Training observability values are

$$\{0.12, 0.20, 0.35, 0.55, 0.90, 1.50, 2.50, 4.50\},$$

and held-out observability values are

$$\{0.16, 0.28, 0.70, 1.20, 3.20\}.$$

Held-out relevance values are $\{0.8, 2.4, 5.0\}$ and held-out budgets are $\{3, 5, 7\}$. The test set contains 540 allocation conditions formed from six relevance permutations, 30 observation-order permutations, and three budgets. Budget B is never provided to the consequence predictor.

C.2 RQ4 fairness rerun: Direct-MLP with slack

RQ4 is a new fairness rerun designed specifically to remove the decoder asymmetry in the original direct-policy baseline. The direct allocator receives

$$[\log d_1, \log d_2, \log d_3, \log o_1, \log o_2, \log o_3, B/8]$$

and uses three ReLU hidden layers of width 96 followed by a four-way Softmax output corresponding to the three representation directions plus an explicit slack channel. For a requested budget B, all B units are distributed over these four channels; units assigned to slack are discarded before constructing the representation. Hence the decoded representation always satisfies $\sum_i r_i \le B$ and can leave capacity unused.

Training uses the seen budgets $\{2, 4, 6, 8\}$, the observability grid

$$\{0.12, 0.20, 0.35, 0.55, 0.90, 1.50, 2.50, 4.50\},$$

and relevance permutations of $[0.5, 1.5, 4.0]$. Evaluation uses held-out observability

$$\{0.16, 0.28, 0.70, 1.20, 3.20\},$$

held-out relevance permutations of $[0.8, 2.4, 5.0]$, and unseen budgets $\{3, 5, 7\}$. Five independent neural seeds are used: $\{11, 22, 33, 44, 55\}$. Learned-DP is retrained/evaluated on the same rerun protocol, and both methods are scored using the same true distortion curves and oracle allocator.

Across the joint held-out (d, o, B) test, Learned-DP obtains

$$97.22\% \pm 1.75\% \text{ exact match}, \qquad 0.000871 \pm 0.000834 \text{ regret},$$

whereas Direct-MLP+Slack obtains

$$56.22\% \pm 1.75\% \text{ exact match}, \qquad 1.782 \pm 0.424 \text{ regret}.$$

The corresponding active-set matches are 98.33% and 79.52%, with mean ℓ_1 allocation errors 0.116 and 1.563.

To separate budget shift from the other held-out variables, we also evaluate held-out d/o with *seen* budgets. Across the same five seeds, Direct-MLP+Slack achieves $63.44\% \pm 3.51\%$ exact match and 1.631 ± 0.349 regret, while Learned-DP achieves $93.31\% \pm 1.97\%$ and 0.00196 ± 0.00052. Thus the remaining direct-policy gap is not attributable to unseen budget alone.

Finally, the oracle leaves capacity unused in 1.56% of training conditions but in none of the held-out d/o conditions, either at seen or unseen budgets. The slack channel is therefore necessary for a fair formulation, but test-time unused capacity is not the main source of the observed OOD gap.

Table 3: Five-seed RQ4 fairness rerun under jointly held-out relevance, observability, and budgets.

Method	Exact	Active	Mean ℓ_1	Regret
Learned-DP	**97.22±1.75%**	**98.33%**	**0.116**	**0.000871±0.000834**
Direct-MLP+Slack	56.22±1.75%	79.52%	1.563	1.782±0.424

Table 4: Budget-wise regret in the RQ4 fairness rerun (mean ± SD over five seeds).

Method	$B = 3$	$B = 5$	$B = 7$
Learned-DP	0.00012	0.00188±0.00256	0.00062±0.00085
Direct-MLP+Slack	0.442±0.269	3.004±0.862	1.899±0.649

C.3 Learned-latent dynamics experiment

We use a compact partially observed 2-D navigation process with state

$$s_t = (x_t, y_t, v_{x,t}, v_{y,t}),$$

goal $g \in \mathbb{R}^2$, and nuisance $n \in \mathbb{R}^2$. Positions are sampled uniformly from $[-1.5, 1.5]^2$, velocities from $\mathcal{N}(0, 0.35^2 I)$, and nuisance variables from $\mathcal{N}(0, I)$. The action space contains four accelerations $\{\pm x, \pm y\}$ with magnitude 0.18. Training targets use

$$v_{t+1} = 0.88 v_t + 0.18 e_{a_t} + \xi_t, \qquad x_{t+1} = x_t + v_{t+1},$$

with $\xi_t \sim \mathcal{N}(0, 0.025^2 I)$.

The action target is generated by an explicit one-step greedy expert,

$$a_t^\star = \arg \min_{a \in \{+x, -x, +y, -y\}} \left[\|p_t + v_{t+1}(a) - g\|_2^2 + 0.05 \|v_{t+1}(a)\|_2^2 \right], \tag{31}$$

where $v_{t+1}(a) = 0.88 v_t + 0.18 e_a$. The first five predetermined runs use seeds $\{11, 22, 33, 44, 55\}$; five additional seeds are sampled independently as described in the randomness and aggregation subsection, yielding ten trained models in total. Action-head training accuracy is approximately 94–95% for every seed.

The 14-dimensional observation concatenates position, velocity, goal, two sinusoidal nuisance/state terms, state–goal products, velocity–goal products, and two nuisance coordinates, followed by additive Gaussian sensing noise. The latent-dynamics model is

$$14 \to 48 \text{ (tanh)} \to 32 \text{ (tanh)} \to 4,$$

with two heads: $4 \to 32$ (ReLU) $\to 4$ for next-state prediction, and $4 \to 32$ (ReLU) $\to 4$ for action logits. It is trained with Adam (learning rate 2×10^{-3}, weight decay 10^{-5}), batch size 256, for 28 fixed epochs on 14,400 samples in the ten-seed rerun. The training loss is

$$\mathcal{L}_{\text{dyn}} = \text{MSE}(\widehat{s}_{t+1}, s_{t+1}) + 0.5 \ \text{CE}(\widehat{a}_t, a_t) + 0.002 \, \| \, \text{offdiag} \, \text{Cov}(z_t) \|_F^2.$$

No early stopping or validation-based model selection is used.

After training, the encoder and heads are frozen. Each latent coordinate is standardized using the training mean and standard deviation, clipped to $[-3, 3]$, and quantized independently with 2^k uniform levels for $k \in \{1, \ldots, 4\}$; $k = 0$ maps the standardized coordinate to zero, i.e. to its training mean after de-standardization. Thus $K_i = 4$ for each of the four latent coordinates, while the maximum useful total capacity is $\sum_i K_i = 16$.

The latent consequence predictor receives

$$[\cos\theta_g, \sin\theta_g, \sigma_{\text{obs}}, \text{onehot}(i), k/4]$$

and uses two tanh hidden layers of width 48 with a scalar output. It is trained with Adam (3×10^{-3}, 10^{-5} weight decay) for 1000 fixed steps in the ten-seed rerun. Consequence curves are estimated from 12 training contexts (four cardinal goals $\times$ three sensing-noise levels) with 260 samples per context in the ten-seed rerun. The primary evaluation averages over 36 context–budget cells (12 contexts $\times$ unseen budgets $\{6, 10, 14\}$). The harder OOD test uses four diagonal goal directions, three unseen sensing-noise levels, and the same three unseen budgets, again totaling 36 cells.

External-control regret. From a quantized latent, the action head selects $\widehat{a}_t = \arg\max_a \ell_a(z_t)$. External evaluation uses the deterministic transition

$$v_{t+1} = 0.88 v_t + 0.18 e_{\widehat{a}_t}, \qquad x_{t+1} = x_t + v_{t+1},$$

and one-step cost

$$C(\mathbf{r}) = \mathbb{E}\left[\|x_{t+1} - g\|_2^2 + 0.05 \|v_{t+1}\|_2^2 \right].$$

The reported external-control regret is

$$C(\mathbf{r}_{\text{method}}) - \min_{\mathbf{r}:\, 0 \le r_i \le 4,\ \sum_i r_i \le B} C(\mathbf{r}).$$

C.4 Randomness, aggregation, and model selection

The motivating falsification studies use five seeds unless stated otherwise. RQ3 and RQ4 use fixed seeded Monte Carlo datasets and report aggregation over hundreds of held-out allocation conditions. RQ5 uses ten independently trained latent models. The first five use predetermined seeds $\{11, 22, 33, 44, 55\}$. To audit seed selection, five additional seeds are sampled without replacement from $[1000, 100000)$ with NumPy Generator master seed 20260925, yielding $\{37296, 75951, 20341, 16764, 62029\}$. Initial states and observation-noise streams are paired across methods. All training schedules are fixed in advance; no early stopping or test-set model selection is used.

C.5 Controlled coupling stress test

To quantify the boundary of the separable approximation, evaluation adds

$$D_\lambda(\mathbf{r}) = \sum_i d_i D_i(r_i; o_i) + \lambda C(\mathbf{r}),$$

where C is a bounded pairwise interaction penalty unknown to the allocator. The separable allocator is unchanged; only the evaluation oracle sees the coupling. We report exact allocation agreement and normalized objective regret as λ increases.

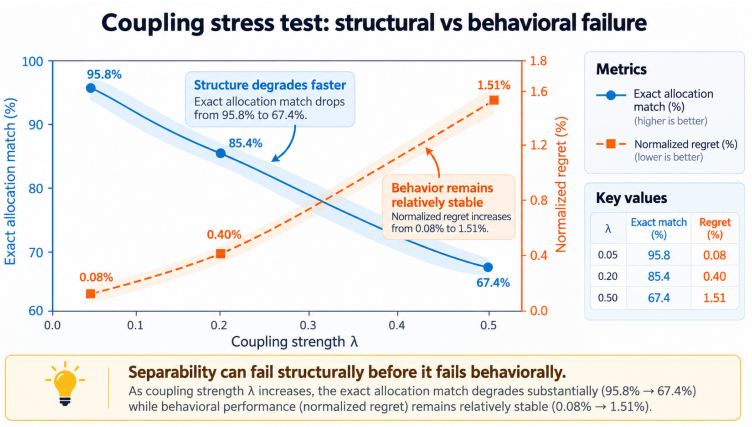


Figure 4: Controlled coupling stress test. Structural agreement degrades faster than behavioral regret.

D Quantizer Robustness

We test whether the controlled allocation conclusion depends on the fixed clipping range responsible for the low-bit pathology. In addition to the original fixed-range uniform quantizer on $[-3, 3]$, we evaluate (i) an MSE-calibrated uniform scalar quantizer whose clipping scale is selected from training posterior-mean samples and (ii) a Lloyd–Max scalar quantizer. The quantizer audit is a separate controlled rerun using five independently trained neural seeds for both Learned-DP and Direct-MLP+Slack. It is not numerically identical to the RQ4 fairness rerun because it uses regenerated Monte Carlo curves and a different retraining realization; only within-audit method comparisons are interpreted.

The original fixed-range curves violate adjacent-rate monotonicity in roughly 15% of tested pairs, with a worst increase above 7. Under calibrated uniform and Lloyd–Max quantization, the violation rate falls to 1.5–5% and the worst increase is below 1.5×10^{-4}, consistent with finite Monte Carlo error. Thus the dramatic low-bit nonmonotonicity is not universal. Importantly, Learned-DP retains lower mean regret than Direct-MLP+Slack for both calibrated quantizers, although the gap is substantially smaller than under the pathological fixed range. The architectural conclusion therefore survives reasonable scalar quantization, while the magnitude of the original gap is quantizer-dependent.

Table 5: Separate five-seed joint-OOD quantizer audit. Values are mean ± SD. These numbers are not directly interchangeable with the independently regenerated RQ4 fairness rerun.

Quantizer	Method	Exact match	Regret
Fixed-range	Learned-DP	$64.00 \pm 7.89\%$	0.0981 ± 0.0458
Fixed-range	Direct-MLP+Slack	$36.00 \pm 1.98\%$	6.898 ± 1.086
Calibrated uniform	Learned-DP	$60.44 \pm 3.30\%$	$\mathbf{0.0270 \pm 0.0029}$
Calibrated uniform	Direct-MLP+Slack	$51.33 \pm 8.49\%$	0.0442 ± 0.0196
Lloyd–Max	Learned-DP	$61.00 \pm 3.13\%$	$\mathbf{0.0276 \pm 0.0026}$
Lloyd–Max	Direct-MLP+Slack	$56.33 \pm 7.13\%$	0.0379 ± 0.0110

E Representation-Specific Mechanism Ablation

This audit asks which consequence structure finite representation allocation must preserve: whether each direction can be summarized by a static scalar, whether a generic rate-dependent shape suffices, or whether the general discrete curve $D_i(k; o_i, h)$ is needed. We use the existing RQ1–RQ2 controlled pipeline and change only the information supplied to the allocator.

Compared mechanisms. **FullCurve-DP** receives the complete measured table $D_i(k; o_i)$ and serves as the full-mechanism reference. **StaticImportance** collapses direction i to the single score

$$s_i = d_i\big[D_i(0) - D_i(8)\big],$$

which is independent of the currently consumed rate. **ScalarLinear-DP** retains only the endpoint amplitude $D_i(0) - D_i(8)$ and reconstructs

$$\widetilde{D}_i(k) = D_i(0) - \frac{k}{8}\big[D_i(0) - D_i(8)\big],$$

so the marginal value is constant across rate. **GenericHighRate-DP** retains the same scalar amplitude but imposes the generic shape

$$\widetilde{D}_i(k) = D_i(8) + \big[D_i(0) - D_i(8)\big]2^{-2k}.$$

All DP variants use the same discrete optimizer; only the representation of consequence differs.

Table 6: Mechanism ablation on the original RQ1 fixed-range pipeline. Entries are exact allocation match. Static scalar/linear surrogates cannot reproduce the changing marginal structure across budgets.

Method	$B = 2$	$B = 4$	$B = 6$	$B = 8$
FullCurve-DP	1.000	1.000	1.000	1.000
GenericHighRate-DP	0.500	0.000	0.167	0.000
ScalarLinear-DP	0.667	1.000	0.167	0.000
StaticImportance	0.667	1.000	0.167	0.000

To directly measure scarcity-dependent changes in preference, we compare the sign of pairwise allocation differences between $B = 2$ and $B = 8$. FullCurve-DP changes pairwise preference in 55.6% of tested direction pairs, compared with 11.1% for GenericHighRate-DP and 0% for both ScalarLinear-DP and StaticImportance. The original RQ1 scarcity signature $\rho(r, d) - \rho(r, o)$ also changes sign and magnitude for FullCurve-DP, while remaining fixed at 0.144 for the static/linear scalar mechanisms. This specific observability-to-relevance reversal is an illustrative property of the fixed-range pipeline rather than a universal law.

Under calibrated uniform quantization, FullCurve-DP remains the reference oracle, GenericHighRate-DP attains 66.7% overall exact match, and both ScalarLinear-DP and StaticImportance attain only 12.5%. The generic high-rate model therefore captures some

Table 7: Calibrated-uniform robustness audit: mean regret to the full-curve oracle. The low-bit fixed-range pathology is removed, yet scalar/static mechanisms still fail increasingly as budget grows.

Method	$B = 2$	$B = 4$	$B = 6$	$B = 8$
FullCurve-DP	0	0	0	0
GenericHighRate-DP	0.0652	0.0290	0.0188	0.0003
ScalarLinear-DP	0.1604	0.6285	0.9270	1.0700
StaticImportance	0.1604	0.6285	0.9270	1.0700

diminishing-return structure, but it is still a fixed-shape prior. By contrast, a static importance scalar cannot change marginal ordering with consumed capacity at all. The robust mechanism conclusion is

scarcity-sensitive allocation requires a consequence model richer than static importance.

The ablation therefore establishes a minimum structural requirement: static importance is insufficient because it cannot express changing marginal consequence. It does not establish that a full flexible curve is necessary; the parsimony audit below tests that stronger hypothesis directly.

F Unified Factorization Benchmark

The main unified comparison uses the calibrated-uniform consequence table and the same train/test split for every structured method. Training observability values are

$$\{0.12, 0.20, 0.35, 0.55, 0.90, 1.50, 2.50, 4.50\},$$

held-out observability values are

$$\{0.16, 0.28, 0.70, 1.20, 3.20\},$$

training relevance uses permutations of $[0.5, 1.5, 4.0]$, held-out relevance uses permutations of $[0.8, 2.4, 5.0]$, seen budgets are $\{2, 4, 6, 8\}$, and held-out budgets are $\{3, 5, 7\}$. The test set is the same 540 joint held-out conditions for all methods.

StaticImportance+DP receives the true held-out endpoint amplitude $D_i(0; o_i) - D_i(8; o_i)$ but reconstructs a linear consequence law, making it a strong structural diagnostic of constant marginal value. ConditionalParametric predicts (c_i, a_i, α_i) and uses

$$\widehat{D}_i(k; o_i) = c_i(o_i) + a_i(o_i) 2^{-\alpha_i(o_i)k}.$$

FullConditional uses the original two-layer 48-unit tanh predictor over $(\log o_i, k/8, \mathbb{I}[k = 1], \mathbb{I}[k = 0])$. Both learned consequence models use the same optimization schedule and five seeds $\{11, 22, 33, 44, 55\}$. Direct-MLP+Slack uses three 96-unit ReLU layers and an explicit fourth slack output; under this unified run its training exact fit is approximately 96%.

Held-out curve prediction is slightly more accurate for FullConditional (MAE/RMSE 0.01033/0.01204) than for ConditionalParametric (0.01042/0.01238), yet allocation quality is worse. This establishes that small differences in pointwise consequence-prediction error do not necessarily order decision quality.

The previously reported 97.22% exact match belongs to the independent fixed-range RQ4 fairness rerun, whereas 60.44% belongs to the independently regenerated calibrated-quantizer robustness audit. These results remain useful within their own protocols but are not used for the primary cross-method ranking.

G Learned-Latent Decision-Relevance Boundary Diagnostic

The original RQ5 checkpoints and per-coordinate consequence curves were not persisted in the available experiment artifacts. We therefore independently reconstruct the stated

RQ5 architecture and protocol using the same ten seeds, rather than claiming byte-identical reuse of the original trained models. The reconstructed action-head training accuracy is $92.37\% \pm 0.28\%$, compared with approximately 94–95% in the original RQ5 runs. This discrepancy indicates that the reconstruction is not training-identical to the original RQ5 realization; we therefore use it only to test the qualitative decision-relevance boundary and do not combine its numerical results with the RQ5 control evaluation.

For each of four latent coordinates, four cardinal goal contexts, three sensing-noise levels, and rates $k = 0, \ldots, 4$, we measure the one-step external-control consequence while holding non-target coordinates at full precision. Endpoint-normalized curves are considered only after auditing the effect size $|D_i(0) - D_i(4)|$. Across minimum endpoint-gap thresholds $0, 0.0025, 0.005, 0.01, 0.02$, the mean pairwise normalized-shape distance is

$$3.108,\ 0.724,\ 0.471,\ 0.265,\ 0.214,$$

while the retained fraction of curves remains $100\%, 87.7\%, 80.4\%, 70.8\%, 64.0\%$. Thus measurable shape variation persists after removing weak-effect curves.

Decision relevance is assessed separately. We compare pairwise marginal ordering using

$$\Delta_i(k; c) = D_i(k-1; c) - D_i(k; c)$$

and count a comparison only if the competing marginal values differ by at least ϵ in both contexts. The all-unit reversal rate is

$$16.82\%,\ 12.34\%,\ 9.63\%,\ 8.40\%$$

for $\epsilon = 0, 0.001, 0.0025, 0.005$. Before inspecting this diagnostic, we set 10% as a practical go/no-go stopping rule for effect-size-filtered marginal-order reversal; it is not a theoretically privileged constant. Independently of that threshold, the reversal rate decreases monotonically from 16.82% to 8.40% as near-tie comparisons are removed. We therefore classify the result as Case 2: function-space variation exists, but decision-relevant marginal-priority reversal is not robust. No further ConditionalParametric-versus-FullConditional learned-latent experiment is run.

H Ten-Seed RQ5 Control Evaluation

The RQ5 action target is generated by the explicit greedy expert in eq. (31). Each latent model is trained for 28 fixed epochs on 14,400 transitions, and the latent consequence predictor is trained for 1000 fixed steps from 260 samples per training context. The original five predetermined neural seeds are $\{11, 22, 33, 44, 55\}$. To check that the result is not an artifact of manually chosen seeds, we independently repeat the full experiment with five seeds sampled without replacement using NumPy Generator master seed 20260925: $\{37296, 75951, 20341, 16764, 62029\}$. We report their combined ten-seed result. No early stopping or validation-based model selection is used.

The encoder and action head are frozen after training. The primary endpoint uses $H = 1$, unseen budgets $B \in \{6, 10, 14\}$, and 12 familiar goal/noise contexts. Representation regret is measured relative to the same trained model with full-precision latent inference.

Table 8: Ten-seed RQ5 horizon sweep. Values are mean ± SD of per-seed aggregate representation regret. The allocation is chosen once and then frozen for the entire rollout.

Horizon	Static Learned-DP	Uniform	Lower-regret cells
$H = 1$	0.0228 ± 0.0090	0.0456 ± 0.0208	56.1%
$H = 10$	6.16 ± 3.63	9.17 ± 6.16	51.4%
$H = 25$	102.57 ± 148.60	124.38 ± 191.91	47.5%
$H = 50$	977.48 ± 1893.21	823.92 ± 1771.88	45.8%

For the normalized visualization in fig. 3, Uniform is set to 1.0 at each horizon and the plotted Learned-DP value is the ratio of aggregate mean regrets,

$$\rho_H = \frac{\overline{R}_{\text{static-LDP}}(H)}{\overline{R}_{\text{Uniform}}(H)}. \tag{32}$$

The resulting values are

$$\rho_1 = 0.50, \qquad \rho_{10} = 0.67, \qquad \rho_{25} = 0.82, \qquad \rho_{50} = 1.19.$$

Thus a one-shot allocation is most valuable locally; when it is frozen and reused as the trajectory evolves, its relative advantage shrinks, with the aggregate mean crossing above Uniform at $H = 50$. Because the RQ5 consequence predictor is conditioned on episode-level goal direction and sensing noise rather than the full evolving state, this experiment should be interpreted as a *frozen-allocation horizon diagnostic*. The observed decay is consistent with allocation staleness and motivates a state-conditioned receding re-allocation extension, but does not establish staleness as the sole cause or claim that dynamic re-allocation has already been evaluated.

Independent random-seed repeat. The second five-seed run alone yields $H = 1$ regret 0.0199 ± 0.0068 for Learned-DP versus 0.0394 ± 0.0213 for Uniform, reproducing the one-step trend of the original five-seed run. The corresponding action-head training accuracies range from 94.31% to 94.97%.

Joint context/sensing shift. Combining both five-seed runs, we additionally test four diagonal goal directions and three unseen sensing-noise levels with unseen budgets. Learned-DP obtains 0.0478 ± 0.0235 one-step representation regret, while Uniform obtains 0.0447 ± 0.0218. The familiar-context advantage therefore disappears under simultaneous context and sensing shift.

I Additional Results

I.1 Direct-MLP capacity sanity check

The large RQ4 gap raises a natural baseline-capacity question: could Direct-MLP+Slack degrade under joint OOD simply because the width-96 network is underparameterized? We therefore perform an auxiliary architectural-capacity audit that keeps the direct-policy formulation unchanged while doubling the hidden width from 96 to 192. Both variants use three ReLU hidden layers, the explicit slack output, the same Paper-B Gaussian distortion construction, the same relevance/observability/budget train–test grids, and neural seeds $\{11, 33, 55\}$. The ordered observation tuples and retraining realization are independently regenerated, so this audit is not numerically interchangeable with the primary five-seed RQ4 rerun and is interpreted only as a within-audit capacity comparison.

Table 9: Auxiliary three-seed Direct-MLP+Slack capacity audit. Doubling hidden width substantially improves in-distribution fit but only modestly changes joint-OOD allocation quality. Values are mean ± SD where applicable.

Width	Train exact	Joint-OOD exact	Active	Mean ℓ_1	Regret
96	94.83%	$51.91 \pm 4.11\%$	67.35%	1.853	1.290 ± 0.211
192	99.48%	$54.51 \pm 2.42\%$	68.83%	1.818	1.216 ± 0.247

Increasing width from 96 to 192 raises training exact fit by 4.65 percentage points, reaching 99.48%, but joint-OOD exact match improves by only 2.60 points and mean regret decreases by only 0.074. Hence the larger network fits the training allocation map nearly perfectly without qualitatively closing the joint-OOD gap. This audit does not prove that no alternative direct-policy architecture could improve transfer, but it makes simple under-capacity of the width-96 baseline an unlikely primary explanation for the observed degradation.

I.2 Additional RQ4 diagnostics

The main text reports aggregate RQ4 performance. Budget-specific results preserve the same qualitative conclusion: Learned-DP maintains low regret across the jointly held-out conditions, while the direct baseline degrades under combined relevance, observability, and

budget shift. Because the direct baseline already degrades under held-out d/o with seen budgets, this gap cannot be attributed to budget extrapolation alone.