Focus or Spread? Attention Allocation under High-Dimensional Uncertainty

Hayong Yun*
April 2026

Abstract

This paper brings the rotate-then-quantize (RTQ) strategy of Zandieh et al. (2025) to economics and finance as a general tool for high-dimensional decision problems. RTQ converts a d-dimensional problem into d one-dimensional quantization problems, addressing the curse of dimensionality. We illustrate with rational inattention. Conventional wisdom says to focus on a few variables and ignore the rest; RTQ accommodates non-Gaussian priors with discrete signals as Sims (2006) conjectured and gives a computable criterion for focus vs broad coverage. Portfolio-choice tests on CRSP merged with the LSEG 13F panel over five decades are consistent with these predictions.

JEL Classification: D83, D81, G11

Keywords: Rational inattention, attention allocation, high-dimensional uncertainty, portfolio choice, vector quantization, curse of dimensionality

*Department of Finance, Eli Broad College of Business, Michigan State University, 667 North Shaw Lane, Room 339, East Lansing, MI 48824. Email: yunhayon@msu.edu.

Focus or Spread? Attention Allocation under High-Dimensional Uncertainty

Abstract

This paper brings the rotate-then-quantize (RTQ) strategy of Zandieh et al. (2025) to economics and finance as a general tool for high-dimensional decision problems. RTQ converts a d-dimensional problem into d one-dimensional quantization problems, addressing the curse of dimensionality. We illustrate with rational inattention. Conventional wisdom says to focus on a few variables and ignore the rest; RTQ accommodates non-Gaussian priors with discrete signals as Sims (2006) conjectured and gives a computable criterion for focus vs broad coverage. Portfolio-choice tests on CRSP merged with the LSEG 13F panel over five decades are consistent with these predictions.

JEL Classification: D83, D81, G11

Keywords: Rational inattention, attention allocation, high-dimensional uncertainty, portfolio choice, vector quantization, curse of dimensionality

1 Introduction

Sims (2006) left rational inattention with an open problem. The Gaussian-quadratic model was tractable in one dimension, but high-dimensional extensions with arbitrary priors stayed unresolved. Sims conjectured the optimal signals would often be discrete rather than Gaussian, but did not say how to compute them. Applied work since has kept tractability by assuming Gaussian priors or low-dimensional states. Van Nieuwerburgh and Veldkamp (2010) solve the portfolio problem within that Gaussian-restricted class and find a corner: limited-capacity investors specialize in a few stocks. VNV is the right answer within its class. What to do outside the Gaussian world has stayed open. This paper offers a feasible benchmark for that case, rotate-then-quantize (RTQ), whose design does not depend on the prior and whose discrete signals match the form Sims (2006) conjectured. RTQ comes within a constant factor of the minimax rate-distortion lower bound, uniformly in dimension and bit width. Applied to portfolio choice, RTQ revises the focus-on-a-few prescription that has dominated the high-dimensional asset-allocation literature: broad coarse signals across all d assets can dominate concentrated attention on a small subset, with the gap widening as the universe grows. The result is a tractable benchmark for high-dimensional rational inattention, not a claim that RTQ beats prior-specific solutions such as the concentration result under a Gaussian prior, which is optimal within that prior's signal class.

The result has a robust rational-inattention reading. Classical rational inattention solves for the optimal information structure given a known prior. In high dimensions, prior-specific solutions are often computationally infeasible or fragile to prior misspecification. We define a robust benchmark as an information strategy whose worst-case loss is within a constant factor of the minimax rate-distortion floor, uniformly over priors with finite second moments. RTQ is such a benchmark. Van Nieuwerburgh and Veldkamp (2010) is the optimum within the Gaussian-restricted signal class and remains the right answer there; RTQ is the robust benchmark when the prior is not trusted, with the constant-factor guarantee holding without parametric assumptions or distributional input from the analyst. This framing carries through to the empirical sections: the CRSP/13F evidence asks whether actual portfolio behavior aligns with the robust benchmark or with the prior-specific corner.

The empirical record is mixed. Actively managed funds concentrate on selected positions, yet broad index funds that process almost no information persistently outperform them (Fama and French, 2010; Carhart, 1997). Multi-product firms adjust prices across many product lines rather than a few key products (Bhattarai and Schoenle, 2014; Bonomo et al., 2023). Speed-dating studies find that participants who meet more partners at a fixed time

budget secure more matches, not fewer.[1] These patterns suggest selective attention leaves welfare on the table in high dimensions.

When the number of variables is large, what is the best feasible attention strategy, and how much welfare does it gain over the concentrated prescription? We address this with the rotate-then-quantize (RTQ) strategy of Zandieh et al. (2025), known there as TurboQuant. RTQ was built for high-dimensional data compression, with a worst-case guarantee that does not depend on the shape of uncertainty or on the number of variables. We adapt it to economic decisions, prove the theorems applied work needs, and put it to work on portfolio choice. The main empirical test is a welfare and behavior comparison on the full CRSP universe merged with the LSEG 13F panel; a cross-domain validation on Fisman speed-dating data is in the Internet Appendix.

Consider an investor choosing among 500 stocks under a fixed attention budget. Selective watches five stocks closely and ignores the other 495. RTQ instead builds 500 composite signals, each a weighted basket of all 500 stocks, and monitors every composite at coarse resolution. Because every composite mixes all stocks, coarse readings translate into coarse estimates of every stock. Think of a pilot scanning many instruments at low resolution rather than staring at a few. Both use the same attention but allocate it differently: selective concentrates; RTQ spreads.

Why does rotation help? Zandieh et al. (2025) show that whatever the original distribution, a random rotation makes the coordinates look statistically similar to one another. A hard joint problem becomes a set of one-dimensional problems with a simple rounding rule. The same rule, applied to each rotated coordinate, lands within a small multiple of the theoretical best regardless of distribution or dimension.

The main text develops the theory for the classical rational-inattention problem: one agent making a high-dimensional decision under quadratic loss, with portfolio choice the leading application. The Internet Appendix develops a parallel-target extension, in which one agent faces many decisions at once under a shared attention budget. The extension adds two theorems, a sufficient-statistic collapse, and a closed-form rule for when spread beats concentration.[2]

The theory delivers three main-text results. First, RTQ's worst-case loss is within a known constant of the minimax rate-distortion lower bound, at any budget and for any quadratic loss with positive definite weight matrix W. The constant is $C_T \cdot \chi(W) \approx 2.72\,\chi(W)$, where $C_T = \pi\sqrt{3}/2$ and $\chi(W)$ is the condition number of W. The bound

[1] The Internet Appendix formalizes this pattern as a cross-domain test of the parallel-target model on the Fisman data; see Internet Appendix IA.3.

[2] Speed-dating conversations, simultaneous forecasts across many assets, and signal compression all fit this structure.

does not grow with dimension or with the bit budget. Zandieh et al. (2025) prove the unweighted version ($W = I$, where $\chi(W) = 1$); we extend to general W. Second, within RTQ, attention follows a water-filling rule tied to stakes: the agent watches most closely where mistakes hurt most. Stakes, not volatility, drive optimal attention. Third, selective attention becomes arbitrarily worse than RTQ as the budget grows, because unmonitored variables leave a permanent loss floor that extra capacity cannot erase.[3]

On the classical portfolio problem, we take RTQ to the full CRSP universe from 1970 to 2026 (221 non-overlapping quarterly windows). The RTQ win rate against VNV climbs from 43% at a budget of one bit per stock to 72% at two bits per stock and 96% at three bits per stock, tracing the focus-versus-spread regime boundary that the parallel-target theory predicts (Proposition IA.2). The cost of the concentration prescription is economically large, of the order routinely quoted as "alpha" in the active-management literature.[4]

We then take the theory to actual investor behavior. The LSEG 13F panel merged with CRSP (1980–2024, 451,121 manager-quarters across 13,571 managers in the fit panel) shows that in 95.3 percent of manager-quarters the active-weight vector is closer to RTQ than to the VNV corner. Managers closer to RTQ and farther from VNV earn higher FF5+UMD risk-adjusted alpha within quarter, which a pure-mandates reading would not produce. The channel flips sign with private-information regulation. Pre-Reg FD, selective disclosure rewarded concentration; the differential coefficient on $(d_{\mathrm{RTQ}} - d_{\mathrm{VNV}})$ is +10.60. Reg FD (2000) and Dodd-Frank (2010) erode those rents and the coefficient flips to −9.61, then −4.58.[5]

The paper's main contribution is to introduce RTQ to economics and finance as a general tool for high-dimensional decision problems. The distribution-free, dimension-free guarantee is uncommon in economic theory. RTQ fits settings beyond portfolio choice: multi-product

[3]The Internet Appendix adds two parallel-target results: a sufficient-statistic collapse in which the three parameters (N, d, B) reduce to a single scaling variable, and a closed-form sufficient-condition regime diagram above which spread is guaranteed to dominate concentration. It also gives an effective-dimension identity $d_{\mathrm{eff}}^{\mathrm{slope}} = 1/(1-\beta)$, where β is the log-linear slope of an outcome count on the number of targets. The slope-implied d_{eff} can be compared with a PCA-implied effective dimension of the observed attribute space; agreement supports the model in a setting outside finance (Section IA.3).

[4]A head-to-head calibration on the S&P 500 proxy ($d = 488$, 1970–2026) puts dollar numbers on the same gap: RTQ leaves 0.5 bps per year on the table while VNV leaves 198 at two bits per stock, and the mean rolling-OOS gap over 1975–2026 is 163 bps per year across 205 non-overlapping 60-month windows, which compounds to roughly 45 percent of lifetime consumption over 30 years. The full S&P 500 application is in Internet Appendix IA.1.

[5]An out-of-sample cross-domain test takes the parallel-target prediction to the Fisman et al. (2006) speed-dating experiment (551 participants, 8,378 four-minute conversations, 21 randomized wave sizes from 5 to 22 partners per side). Regressing log(1 + matches) on log(partners) gives a slope of 0.472 (wave-clustered SE 0.063), implying a slope-based effective attribute dimension of 1.89; PCA on the trait ratings returns 2.68. Internet Appendix IA.3 reports the regression, the threshold model that produces the slope identity, and out-of-sample stability checks.

pricing, monetary policy with many state variables, and high-dimensional forecasting. Each deserves its own treatment in future work. The framework also speaks to the recent adoption of AI and deep-learning methods in investment management; Internet Appendix IA.4 shows that RTQ comes within a constant factor of the spherical minimax rate-distortion floor that the optimal rank-k linear-MMSE cross-asset attention layer attains asymptotically at a given bit budget, providing a closed-form theoretical complement to empirical work such as Cong et al. (2026).

Our work also contributes to the rational inattention literature. Van Nieuwerburgh and Veldkamp (2010) is the optimum within the Gaussian-restricted signal class; once that class is dropped, the universal concentration prescription weakens. RTQ provides a complementary benchmark for the non-Gaussian case, with discrete signals matching the form Sims (2006) conjectured, and delivers a computable criterion for when each strategy dominates: water-filling on welfare-weighted eigenvalues replaces volatility-driven prescriptions (Proposition 3), and selective attention becomes arbitrarily worse than RTQ as capacity grows because unmonitored variables leave a permanent loss floor (Proposition 2). The rotation is a mathematical device, not a literal description of behavior, in the same sense competitive equilibrium characterizes attainable allocations without requiring agents to solve the auctioneer's problem. At the same time, RTQ is computationally tractable in closed form, and modern AI portfolio models such as Cong et al. (2026) implement broadly similar cross-asset attention in practice, so the benchmark is implementable rather than purely abstract.

We also contribute to the empirical asset-management literature. Using the LSEG 13F panel merged with CRSP (1980–2024), we document that fund managers behave more like RTQ than VNV in the large majority of manager-quarters, and that the alpha channel on $(d_{\mathrm{RTQ}} - d_{\mathrm{VNV}})$ reverses sign across private-information regulation (Reg FD, Dodd-Frank). This refines the standard reading of fund concentration: concentration tracked access to private-information rents, and once those rents were regulated away, broad attention pays.

Section 2 reviews rational inattention and the TurboQuant result we adapt. Section 3 formalizes the single-agent problem. Section 4 delivers the theorems. Section 5 presents the CRSP/13F evidence. Section 6 concludes.

2 Background

2.1 Rational Inattention

Rational inattention originates with Sims (2003): agents face a finite capacity for processing information, measured by Shannon mutual information between state and signal. In the

baseline, an agent observes a univariate Gaussian state and chooses an action to minimize quadratic loss. The optimal signal is Gaussian, and posterior variance has a closed form depending only on prior variance and capacity. The framework is a tractable foundation for modeling bounded rationality but relies on Gaussian-quadratic structure. Sims (2006) flagged the need to go beyond linear-quadratic and conjectured that optimal signals in more general environments would often be discrete, later confirmed by Jung et al. (2019) for discrete-action problems.

Matějka and McKay (2015) extended the framework to discrete choice with arbitrary payoffs, showing that the optimal attention strategy produces multinomial-logit choice probabilities. This connected rational inattention to industrial organization and demand estimation and remains the sharpest characterization in the discrete-choice setting. The limitation is a finite action space; the result does not extend to continuous or high-dimensional actions. Caplin and Dean (2015) provide revealed-preference foundations, showing choice data can identify whether behavior is consistent with any costly information-acquisition rule. Caplin et al. (2019) extend these ideas to optimal consideration sets and show that the shadow price of capacity rationalizes observed stochastic choice.

Denti (2022) characterizes the class of information-cost functions consistent with rational behavior, showing that posterior separability (cost depends only on the distribution of posterior beliefs) captures the class satisfying standard axioms. This clarifies the axiomatic foundations but provides no constructive solution for any specific cost in high dimensions.

A growing applied literature embeds rational inattention into quantitative models. Maćkowiak and Wiederholt (2009) study firm price-setting under capacity constraints on cost and demand information, generating price stickiness and state-dependent rules. Woodford (2009) develops a related model rationalizing sluggish, state-dependent adjustment. Maćkowiak and Wiederholt (2015) extend to a full business-cycle model with joint attention over idiosyncratic and aggregate shocks. Steiner et al. (2017) characterize dynamic rational inattention and show optimal allocation generates inertia and delayed responses. Van Nieuwerburgh and Veldkamp (2010) apply the framework to portfolio choice and show that attention constraints explain underdiversified portfolios concentrated in familiar assets. Maćkowiak et al. (2023) survey the literature.[6]

[6]Beyond Shannon cost, a parallel literature develops related models of limited attention. Kőszegi and Szeidl (2013) propose a focusing model in which agents overweight attributes with greater dispersion across options. Gabaix (2014) develops a sparsity-based model with closed-form solutions widely used in macroeconomics and finance. These share our motivation that attention must be allocated across many dimensions but replace the information-theoretic cost with a different, often more tractable penalty. Our contribution stays within the Shannon tradition while addressing the tractability gap that motivated these alternatives. A separate branch in behavioral asset pricing studies how cognitive limits distort the use rather than the acquisition of information; Charles et al. (2024) show experimentally that investors compress valuations

These applied papers stay tractable by restricting state dimension or signal class, usually by assuming Gaussian signals or restricting the agent to a few linear combinations of the state. When these restrictions are far from optimal, predictions can depart substantially from the unrestricted solution. We address this gap with a feasible, distribution-free signal structure that comes within a known constant of the minimax rate-distortion lower bound, uniformly in dimension and bit width.[7]

A parallel literature in machine learning builds high-capacity portfolio models using Transformer attention (Vaswani et al., 2017) and reinforcement learning. Cong et al. (2026) introduce AlphaPortfolio, a Transformer-based model with cross-asset attention trained end-to-end on the portfolio objective, and report strong out-of-sample performance. AlphaPortfolio and RTQ approach the same broad-attention strategy from complementary angles. AlphaPortfolio learns a high-capacity attention layer through gradient descent; RTQ provides a closed-form, distribution-free benchmark with worst-case loss guarantees and interpretable composite signals. The two are complementary in this sense: their work demonstrates empirically that learned cross-asset attention delivers strong performance, and ours provides the theoretical foundation for why a tractable, capacity-bounded version of the same idea is within a known factor of optimal. Internet Appendix IA.4 formalizes the connection: RTQ comes within a constant factor of $C_T\chi(W)$ of the spherical minimax rate-distortion floor that the optimal rank-k linear-MMSE attention layer attains asymptotically at a given bit budget, distribution-free over priors with finite second moments. The bound covers the linear-attention layer that sits inside the architecture; the extension to softmax, multi-head, and downstream MLP components is suggestive rather than a formal equivalence and is discussed in the Internet Appendix.

2.2 Information Theory Preliminaries

Let $\mathbf{x} \in \mathbb{R}^d$ denote the state with prior F. A *signal structure* is a conditional distribution $\pi(\cdot|\mathbf{x})$ mapping states to signals $\mathbf{s}$; the agent forms a posterior about $\mathbf{x}$ from $\mathbf{s}$. The *Shannon mutual information* is

$$I(\mathbf{x};\mathbf{s}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{s}), \tag{1}$$

toward a cognitive default and transmit stated beliefs to actions only weakly, which offers a microfoundation for inelastic demand. Our question is upstream, about how a capacity-constrained agent should acquire signals across many variables in the first place.

[7] On the theory side, Kamenica and Gentzkow (2011) initiate Bayesian persuasion, and Bergemann and Morris (2016) characterize outcomes achievable across information structures in games. These frameworks are complementary: they ask what a designer should reveal; we ask what an agent should attend to. Our rotation-based approach could serve as a tractable tool for information designers facing high-dimensional type spaces, since the worst-case guarantee holds for any state distribution with finite second moments.

the reduction in uncertainty about $\mathbf{x}$ from observing $\mathbf{s}$, measured in bits (base-2 logs). We follow Sims (2003) in treating the bit as the unit throughout: one bit is a binary distinction, and each additional bit doubles signal resolution. The capacity constraint is

$$I(\mathbf{x}; \mathbf{s}) \leq \kappa, \tag{2}$$

with $\kappa > 0$ the information-processing capacity; at b bits per coordinate, $\kappa = bd$.

A key property is *additivity*: if $\mathbf{x} = (x_1, \ldots, x_d)$ has independent coordinates and the signal is designed independently for each,

$$I(\mathbf{x}; \mathbf{s}) = \sum_{i=1}^{d} I(x_i; s_i). \tag{3}$$

Additivity is what makes coordinate-wise coding interpretable once the rotation-and-scalar-quantization guarantee is in place. After the random rotation, each rotated coordinate has the same Beta-distributed marginal in expectation over R. The TurboQuant worst-case MSE bound then applies coordinate by coordinate, and the d-dimensional signal design decomposes into d scalar quantization problems with a controlled worst-case bound.

The *rate-distortion function* $R(D)$ gives the minimum bits required for expected distortion at most D. For MSE distortion, it lower-bounds how accurately any signal structure can represent the state at a given capacity. Our main result bounds RTQ loss by a constant multiple of this floor.

3 Model

We consider an agent who must make many decisions at once under uncertainty. Think of a portfolio manager choosing weights for hundreds of stocks, or a retailer setting prices for thousands of products. The agent cannot fully observe the state and must decide, before seeing any data, what to pay attention to.

The model nests the univariate case of Sims (2003) when $d = 1$. In high dimensions, two new forces appear: decisions interact (getting one stock weight wrong changes how costly it is to get another wrong), and the agent must spread a fixed attention budget across many variables.

3.1 Environment

State. The state is a vector $\mathbf{x} \in \mathbb{R}^d$ drawn from a prior F with mean $\boldsymbol{\mu}$ and finite second moments. Each x_i is one piece of payoff-relevant information. We impose no restrictions on F: it may be Gaussian, fat-tailed, skewed, or discrete.

Three examples:

- A fund manager choosing weights for $d = 500$ stocks, where $\mathbf{x}$ is the vector of next-period excess returns. Some returns are highly correlated (tech stocks move together), others nearly independent.
- A grocery chain setting prices for $d = 2{,}000$ products, where $\mathbf{x}$ is the vector of marginal costs. Some products are substitutes (Coke and Pepsi), so their costs matter jointly.
- A central bank tracking $d = 50$ macro indicators across countries, where $\mathbf{x}$ captures current conditions, some lagging, some leading.

Action. After observing a signal about $\mathbf{x}$, the agent chooses $\mathbf{a} \in \mathbb{R}^d$: one decision per state variable, a weight for each stock, a price for each product, a forecast for each indicator.

Loss function. The agent's loss from choosing $\mathbf{a}$ when the state is $\mathbf{x}$ is

$$L(\mathbf{a}, \mathbf{x}) = (\mathbf{a} - \mathbf{x})^\top W (\mathbf{a} - \mathbf{x}), \tag{4}$$

where W is a $d \times d$ symmetric positive definite matrix.

The quadratic form arises as the second-order approximation to any smooth objective around the full-information optimum. If the agent has objective $U(\mathbf{a}, \mathbf{x})$ with full-information optimal action $\mathbf{a}^*(\mathbf{x}) = \mathbf{x}$, the loss from choosing $\mathbf{a}$ instead is $U(\mathbf{a}^*(\mathbf{x}), \mathbf{x}) - U(\mathbf{a}, \mathbf{x}) \approx \frac{1}{2}(\mathbf{a}-\mathbf{x})^\top H(\mathbf{x})(\mathbf{a}-\mathbf{x})$, where $H(\mathbf{x}) = -\partial^2 U / \partial \mathbf{a}^2$. The matrix W corresponds to (a constant or expected) Hessian. For CARA portfolio choice with risk aversion γ, $W = \gamma\Sigma$, encoding risk preferences directly. For pricing, W is the Hessian of the profit function, reflecting demand elasticities. The quadratic structure therefore does not assume risk neutrality: it embeds preferences through W, following Sims (2003).

The matrix W is the central object of the model. If $W = I$, equation (4) is the sum of squared errors across dimensions: penalties for each variable do not interact. This baseline simplifies analysis but rules out the decision interactions central to high-dimensional applications.

In practice, errors interact. If a fund manager overweights two correlated tech stocks, the combined mistake is worse than the sum of individual mistakes, because errors compound rather than diversify. If a retailer misprices Coke, the cost depends on whether she

also mispriced Pepsi, because consumers substitute. The matrix W captures these *decision complementarities*:

- W_{ii} measures how costly it is to get variable i wrong in isolation.
- W_{ij} measures how much the cost of i depends on the error in j.

In the portfolio example, W reflects the curvature of the mean-variance objective and is shaped by the return covariance. In pricing, W encodes cross-price elasticities. For a central bank, W reflects how forecast errors feed into welfare loss from suboptimal policy.

Off-diagonal terms in W are what make the high-dimensional attention problem different from d copies of the one-dimensional problem. With $W = I$, the agent can think about each variable separately. With general W, she cannot: the right amount of attention for variable i depends on how much goes to variable j, because their errors interact in the loss.

3.2 Information Constraint

The agent has limited capacity for processing information. Before observing $\mathbf{x}$, she chooses a *signal structure*: a conditional distribution $\pi(\cdot|\mathbf{x})$ that generates a signal $\mathbf{s}$ given $\mathbf{x}$. The signal must satisfy the capacity constraint (2) from Section 2, with $\kappa > 0$ the agent's total capacity in bits.

The constraint says the signal can reduce uncertainty about $\mathbf{x}$ by at most κ bits. Higher κ means more information and lower expected loss. Express the budget as $\kappa = bd$, where b is average bits per dimension. Small b (say, 2 or 3) means attention is severely constrained; large b approaches full information.

Using Shannon mutual information as the cost follows Sims (2003). Its key property is additivity across independent components: if the state has independent coordinates and the signal is designed independently for each, total cost sums across variables. This additivity is what lets us decompose the problem after rotation.

3.3 Timing

The sequence of events is:

1. The agent knows the prior F, the loss matrix W, and capacity κ.
2. She designs a signal structure $\pi(\cdot|\mathbf{x})$ subject to $I(\mathbf{x}; \mathbf{s}) \leq \kappa$.
3. Nature draws $\mathbf{x} \sim F$.

4. The agent observes $\mathbf{s} \sim \pi(\cdot|\mathbf{x})$ and forms posterior beliefs.

5. She chooses $\mathbf{a}$ to minimize expected loss given $\mathbf{s}$.

Step 2 is the critical one. The agent commits to what she will attend to *before* seeing the state. She is not choosing what to look at after the fact but designing her information filter in advance. This is the sense in which attention is allocated, not just used.

3.4 Optimal Action and Expected Loss

For any signal realization $\mathbf{s}$, the best response in step 5 is the posterior mean:

$$\mathbf{a}^*(\mathbf{s}) = \mathbb{E}[\mathbf{x}|\mathbf{s}]. \tag{5}$$

This holds for any positive definite W: regardless of how the loss weights errors, the best guess is the conditional expectation.[8]

Ex ante expected loss is then determined by signal quality, measured through the posterior error $\boldsymbol{\varepsilon} = \mathbb{E}[\mathbf{x}|\mathbf{s}] - \mathbf{x}$:

$$\mathcal{L}(\pi) = \mathbb{E}\left[\boldsymbol{\varepsilon}^\top W \boldsymbol{\varepsilon}\right] = \operatorname{tr}(W \cdot \Sigma_\varepsilon), \tag{6}$$

where $\Sigma_\varepsilon = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]$ depends on π (write $\Sigma_\varepsilon(\pi)$ when needed).

Expected loss is the W-weighted trace of the posterior error covariance. With $W = I$, loss is total MSE. With large off-diagonal entries, loss also penalizes correlated errors: when two related variables are wrong in the same direction, loss is amplified.

3.5 Assumptions

Assumption 1 (Shannon mutual information cost). As in Sims (2003), the cost of information is Shannon mutual information between state and signal. Denti (2022) shows Shannon cost belongs to the class of posterior-separable cost functions, whose costs depend only on the distribution of posterior beliefs. Alternatives include Fisher information and neighborhood-based costs (Maćkowiak et al., 2023), but Shannon remains standard. Our results rely on additivity across independent components (equation 3), which is specific to Shannon entropy.

Assumption 2 (Ex ante signal design). As in Sims (2003), the agent commits to a signal structure before observing the state. She decides what to attend to as policy, not in response to each realization. This is realistic: a portfolio manager sets up her research process before the trading day; a central bank designs its monitoring framework before data arrive.

[8] $\mathbb{E}[(\mathbf{a}-\mathbf{x})^\top W(\mathbf{a}-\mathbf{x})|\mathbf{s}]$ is convex quadratic in $\mathbf{a}$ with minimum at $\mathbb{E}[\mathbf{x}|\mathbf{s}]$ for any positive definite W.

Assumption 3 (Quadratic loss). The loss is the weighted quadratic $L(\mathbf{a}, \mathbf{x}) = (\mathbf{a} - \mathbf{x})^\top W(\mathbf{a} - \mathbf{x})$, following Sims (2003). As noted above, this is a second-order approximation to any smooth objective around the full-information optimum. It is exact for CARA portfolio choice with Gaussian returns and for linear-quadratic regulation problems, and approximate for problems with substantial nonlinearities (option payoffs, bankruptcy thresholds).

Assumption 4 (Arbitrary prior with finite second moments). Unlike Sims (2003), which relies on Gaussian priors, we allow F to be any distribution on $\mathbb{R}^d$ with finite second moments. Outside the Gaussian case, optimal signals can be discrete and distribution-dependent (Jung et al., 2019), and the Gaussian restriction rules out fat tails, skewness, and discrete mixtures common in economic data. Our results need only finite second moments because the random rotation works through a geometric property of high-dimensional spaces, not distributional assumptions. The condition fails only for distributions with infinite variance.

Assumption 5 (General positive definite loss matrix). W is symmetric positive definite but otherwise unrestricted. Most applied work assumes $W = I$ or diagonal, ruling out interactions across dimensions. We allow off-diagonal entries capturing decision complementarities. The rotation-based strategy of Zandieh et al. (2025) was developed for $W = I$; extending to general W is one of our contributions, and it is the source of the condition number $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}(W)$ in the main bound. Positive definiteness means every direction matters at least slightly for the decision. In the rank-one limit (a single policy instrument), selective attention is trivially optimal, and our results are not needed.

Assumption 6 (Known and fixed W). W is known to the agent and does not depend on $\mathbf{x}$, implicit in the quadratic tracking specification of Sims (2003). The agent knows her decision problem before allocating attention. This is realistic when W reflects stable features of the environment (long-run return covariance, cross-price elasticities) and less so when risk aversion or demand structure shifts unpredictably. Allowing uncertain W requires modeling joint uncertainty over the state and the decision weights, which we leave for future work.

Assumption 7 (Action dimension equals state dimension). The agent chooses one action per state variable. This fits portfolio choice (one weight per asset), multi-product pricing (one price per product), and multi-indicator forecasting (one forecast per indicator). The assumption rules out problems where the agent aggregates many state variables into a lower-dimensional decision, such as a central bank observing 50 indicators but setting a single interest rate. That aggregation case is the low-rank limit of our setup, where selective attention is appropriate.

Assumption 8 (Static, single-period). As in the baseline of Sims (2003), the agent makes one decision.[9]

[9]There is no learning over time, no belief updating across periods, no dynamic capacity accumulation.

Assumption 9 (Single agent, no strategic interaction). As in Sims (2003), the agent's payoff depends only on her own action and the state, not on others' actions or attention.[10]

3.6 The Agent's Problem

The agent's problem is:

$$\min_{\pi} \quad \operatorname{tr}(W \cdot \Sigma_{\varepsilon}(\pi)) \qquad \text{subject to} \quad I(\mathbf{x}; \mathbf{s}) \leq \kappa. \tag{7}$$

She chooses the signal structure that minimizes the W-weighted posterior error, subject to a capacity of κ bits on mutual information between signal and state.

The problem is well defined but intractable in general form. The signal structure π is a conditional distribution over an arbitrary signal space, an infinite-dimensional optimization, and the constraint is a nonlinear functional of π. When F is Gaussian, there is a closed form via water-filling on the eigenvalues of the posterior covariance, extending to any dimension with $O(d)$ work after eigendecomposition. Outside Gaussian, the optimal signal can be discrete and distribution-dependent (Jung et al., 2019), with no closed form at the dimensions that matter in practice.

This is the problem for which we give a feasible signal structure with a constant-factor guarantee relative to the minimax rate-distortion lower bound in the next section.[11]

Multi-period extensions exist (Maćkowiak et al., 2023). The static setting isolates the cross-sectional dimension (allocation across variables at a point in time) from the time-series dimension (allocation across periods). Our results characterize the cross-section. A dynamic extension would need to track how the posterior covariance evolves under repeated quantization. We expect qualitative insights to carry over because concentration of measure and mutual information additivity are not tied to one period, but a full dynamic treatment is future work.

[10] This rules out settings where one agent's information acquisition affects another's payoff, as in Hellwig and Veldkamp (2009) and Myatt and Wallace (2012). Our main bound does not directly apply there. Still, our results could serve as a building block for games with rational inattention: conditional on opponents' strategies, each agent faces a single-agent attention problem to which RTQ provides an approximate best response.

[11] The single-agent setup above can be extended to a parallel-target setting in which one agent estimates N multi-dimensional targets at once under a shared bit budget. This extension is the block-diagonal-W special case of the single-agent setup; its formal setup, the two reference strategies (concentration and spread), the sufficient-statistic collapse, and the sufficient-condition regime diagram for spread versus concentration are developed in Internet Appendix IA.2.

4 Main Result

4.1 The Rotate-Then-Quantize Strategy

We propose an attention strategy we call *rotate-then-quantize* (RTQ). Steps 1-3 apply the vector quantization procedure of Zandieh et al. (2025): decompose the vector into norm and direction, rotate the direction with a random orthogonal matrix (a randomized Hadamard transform in practice), and scalar-quantize each rotated coordinate with a quantizer optimized for Beta($\frac{d-1}{2}, \frac{d-1}{2}$). The TurboQuant procedure combines the polar/Beta scalar-quantizer machinery of Han et al. (2025) with the random-rotation preconditioning of Zandieh et al. (2024) and proves a worst-case MSE bound on the unit sphere. Step 4 and the bit allocation across coordinates are our additions for the rational-inattention setting. The strategy has four steps.

Step 1: Decompose. Separate $\mathbf{x}$ into norm $\rho = \|\mathbf{x}\|$ and direction $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$. In the main case, ρ is treated as known scale information outside the attention constraint and does not enter $I(\mathbf{x}; \mathbf{s})$, so the entire budget κ encodes the direction on the unit sphere S^{d-1}. As a variant, ρ may be quantized with b_0 bits to produce $\hat{\rho}$, leaving $\kappa - b_0$ bits for the direction; this variant carries the looser bound from the norm-quantization remark of Proposition 1.

Step 2: Rotate. Apply a fixed random orthogonal matrix R to the unit vector: $\mathbf{r} = R\mathbf{u}$. R is drawn once before any data arrives and reused for all states. In practice, R is a randomized Hadamard transform, which can be applied efficiently.

Step 3: Quantize independently. Allocate b_i bits to coordinate i of $\mathbf{r}$, subject to $b_0 + \sum_i b_i \leq \kappa$. Quantize each r_i independently using a scalar quantizer optimized for the Beta($\frac{d-1}{2}, \frac{d-1}{2}$) distribution.

Step 4: Invert and decode. Given the quantized vector $\hat{\mathbf{r}}$, apply $R^\top$ to obtain $\tilde{\mathbf{u}} = R^\top \hat{\mathbf{r}}$, the standard TurboQuant MSE direction reconstruction. The RTQ decoder produces the feasible reconstruction $\widetilde{\mathbf{x}} = \rho\tilde{\mathbf{u}}$ with the stored norm (or $\widetilde{\mathbf{x}} = \hat{\rho}\tilde{\mathbf{u}}$ when the norm is quantized). The agent then takes the action $\mathbf{a}(\mathbf{s}) = \mathbb{E}[\mathbf{x} \mid \mathbf{s}]$, the posterior mean given the RTQ signal $\mathbf{s} = (\hat{\rho}, \hat{\mathbf{r}})$. Since the posterior mean minimizes quadratic loss conditional on the signal, its loss is weakly below the loss of the mechanical decoder $\widetilde{\mathbf{x}}$. The RTQ decoder loss therefore provides an upper bound on the loss of the corresponding rational-inattention signal structure.[12]

The strategy is fully specified by R and the bit allocation $(b_0, b_1, \ldots, b_d)$. Both are chosen before the agent sees any data. The quantizer for each coordinate is precomputed from the

[12]The mechanical decoder is $\widetilde{\mathbf{x}} = \rho\tilde{\mathbf{u}}$ in the exact-norm case, or $\widetilde{\mathbf{x}} = \hat{\rho}\tilde{\mathbf{u}}$ with norm quantization. Either form is a reconstruction designed to minimize MSE relative to the input, not the posterior mean under the prior F. The two coincide only when F matches the design distribution (the Beta distribution after random rotation). For arbitrary F, the agent's optimal action conditional on the RTQ signal is $\mathbb{E}[\mathbf{x} \mid \mathbf{s}]$, which weakly improves on $\widetilde{\mathbf{x}}$. All upper bounds in this paper apply to $\widetilde{\mathbf{x}}$ and therefore to the posterior-mean action as well.

Beta distribution, which depends only on d, not on $\mathbf{x}$ or F. This is what makes the design *data-oblivious*: the agent does not need to know F to implement it. We use *distribution-free* in the algorithmic sense (the design does not depend on F) and in the worst-case guarantee sense (the upper bound in Proposition 1 holds for any F with finite second moments). Section 4.2 discusses the choice of benchmark.

The idea is simple. The agent transforms the information environment through a random rotation, after which a worst-case unit direction becomes uniformly distributed on the sphere when viewed over the randomness of the rotation. Each rotated coordinate is then approximately Beta-distributed, and a precomputed scalar Lloyd-Max quantizer handles each coordinate efficiently. Inverting the rotation translates the quantized output back to the original frame.

The random rotation is a modeling device, not a literal action. Two observations help connect it to economic intuition. First, each rotated coordinate is a weighted basket of all original variables, with weights fixed ex ante. An agent monitoring a single rotated coordinate tracks a composite signal loading on every asset, product, or partner, closer to how real decision-makers read aggregates than to how they stare at individual variables. Second, the d rotated coordinates have comparable magnitudes by construction and behave like synthetic factors with similar variance, a prior-free principal-components decomposition. Competitive equilibrium plays the same role in price theory: it describes attainable allocations without requiring agents to solve the Walrasian auctioneer's problem. RTQ describes attainable attention strategies without requiring agents to compute a Haar-random rotation. RTQ is also computationally tractable in closed form, and modern AI portfolio models such as Cong et al. (2026) implement broadly similar cross-asset attention in practice, so the benchmark is implementable rather than purely abstract.

One subtlety about the rotated coordinates is worth flagging. The Beta distribution emerges over the randomness of the Haar rotation R for a fixed unit direction $\mathbf{u}$. Once R is drawn and held fixed, the rotated coordinates of $R\mathbf{x}$ for $\mathbf{x} \sim F$ inherit whatever structure F has, not the Beta distribution. The design uses the Beta distribution because that is the appropriate target for the minimax guarantee: the worst-case unit direction, viewed through the Haar rotation, is Beta-distributed coordinate-wise. The signal structure is the public randomness over R together with the scalar quantization; once R is realized, the strategy is a deterministic encoder-decoder pair.

The optimal allocation of bits across rotated coordinates under arbitrary welfare matrix W is characterized in Proposition 3.

4.2 RTQ Guarantee Against the Spherical Benchmark

We compare RTQ to the minimax rate-distortion benchmark: the worst-case loss any randomized quantizer must incur on a hard unit direction at a given bit budget. This is the appropriate benchmark when the prior F is not known. The main proposition is a *directional*-attention result: the magnitude $\rho = \|\mathbf{x}\|$ is treated as known scale information, and the attention budget governs how to allocate bits across directions in $\mathbb{R}^d$. This separation between scale and direction is standard in the TurboQuant literature (Zandieh et al., 2025) and is appropriate when the volatility scale of returns or shocks is known (or estimated outside the attention problem) and the residual question is how to allocate attention across direction. An extension to the case where ρ is also quantized is in Appendix A.2; the clean $C_T\chi(W)$ form requires the known-norm assumption. Let $\mathcal{L}^{RTQ}(\kappa)$ denote expected loss under RTQ with optimal bit allocation.

Proposition 1 (RTQ guarantee against the minimax benchmark, normalized-state case)**.** *Let $\mathbf{x} = \rho\mathbf{u}$ with unit direction $\mathbf{u} \in S^{d-1}$ and norm $\rho = \|\mathbf{x}\|$. Suppose the state is normalized, or the norm ρ is treated as known scale information outside the attention constraint, so the attention budget applies entirely to the direction. Equivalently, the proposition characterizes directional uncertainty conditional on ρ, with ρ not entering the Shannon information $I(\mathbf{x};\mathbf{s})$ that defines the budget. RTQ then encodes the direction with $b = \kappa/d$ bits per coordinate using TurboQuant and reconstructs $\hat{\mathbf{x}} = \rho\tilde{\mathbf{u}}$, where $\tilde{\mathbf{u}}$ is the standard TurboQuant MSE reconstruction (no unit projection). For any prior F with finite second moments, any positive definite W, any $d \geq 2$, and any $\kappa > 0$,*

$$\mathcal{L}^{RTQ}(\kappa) \;\leq\; \lambda_{\max}(W) \cdot \mathbb{E}[\rho^2] \cdot C_T \cdot 4^{-b}, \tag{8}$$

where $C_T = \pi\sqrt{3}/2 \approx 2.72$ is the worst-case TurboQuant MSE constant from Zandieh et al. (2025). Let

$$\mathcal{L}^{\text{sphere}}_{\text{lb}}(\kappa) \;\equiv\; \lambda_{\min}(W) \cdot \mathbb{E}[\rho^2] \cdot 4^{-b} \tag{9}$$

be the minimax rate-distortion lower bound under weighted quadratic loss: any randomized quantizer that operates on a worst-case unit direction must incur at least this loss. Then

$$\frac{\mathcal{L}^{RTQ}(\kappa)}{\mathcal{L}^{\text{sphere}}_{\text{lb}}(\kappa)} \;\leq\; C_T \cdot \chi(W), \tag{10}$$

where $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}(W)$ is the condition number. The constant $C_T \approx 2.72$ matches the TurboQuant MSE constant directly.

The bound extends to the unnormalized case where the norm ρ is also quantized with

an additional norm-quantization term entering the upper bound; the full extension is in Appendix A.2.[13]

The bound has two parts. The upper bound depends only on $\lambda_{\max}(W)$, $\mathbb{E}[\rho^2]$, and the per-coordinate budget. Relative to the minimax floor, RTQ is within $C_T\chi(W) \approx 2.72\,\chi(W)$. The constant $C_T = \pi\sqrt{3}/2$ does not depend on d, κ, W, or F. When $W = I$, the ratio collapses to C_T, the classical TurboQuant guarantee. When some eigenvalues of W dominate, $\chi(W)$ widens the gap, the cost of using a symmetric strategy on an asymmetric problem.

The benchmark is the minimax floor, not the prior-specific unrestricted optimum. For structured priors (low-rank, finite-support, sparse), an oracle that knows F may design a tighter signal. The contribution of Proposition 1 is that RTQ matches the minimax guarantee without any knowledge of F, using a quantizer precomputed from the Beta distribution. The loss concentration ratio in Section 4.5 flags settings where the prior has exploitable structure and a tailored signal may beat RTQ.

Take a portfolio manager with $d = 500$ and W from the return covariance, with the volatility scale handled outside the attention problem. Expected loss under RTQ is at most $2.72\,\chi(W)$ times the minimax lower bound. For a well-diversified portfolio with moderate condition number, the penalty is small and RTQ delivers a feasible signal that performs near the worst-case rate-distortion bound across all 500 assets.

Proof sketch. The full proof is in Appendix A.2. Three steps for the normalized-state case:

Step 1. With ρ stored exactly, the reconstruction $\hat{\mathbf{x}} = \rho\tilde{\mathbf{u}}$ gives $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \rho^2\|\mathbf{u} - \tilde{\mathbf{u}}\|^2$ pointwise. The spectral inequality $\mathbf{e}^\top W\mathbf{e} \le \lambda_{\max}(W)\|\mathbf{e}\|^2$ then gives $(\mathbf{x} - \hat{\mathbf{x}})^\top W(\mathbf{x} - \hat{\mathbf{x}}) \le \lambda_{\max}(W)\rho^2\|\mathbf{u} - \tilde{\mathbf{u}}\|^2$.

Step 2. Condition on $(\rho, \mathbf{u})$. The worst-case MSE bound from Lemma 1 is uniform over unit directions, so $\mathbb{E}(\|\mathbf{u} - \tilde{\mathbf{u}}\|^2 \mid \rho, \mathbf{u}) \le C_T \cdot 4^{-b}$ with $C_T = \pi\sqrt{3}/2$. By the tower property, $\mathbb{E}[\rho^2\|\mathbf{u} - \tilde{\mathbf{u}}\|^2] = \mathbb{E}[\rho^2\,\mathbb{E}(\|\mathbf{u} - \tilde{\mathbf{u}}\|^2 \mid \rho, \mathbf{u})] \le \mathbb{E}[\rho^2]\,C_T\,4^{-b}$. Substituting gives (8).

Step 3. For the lower bound, the minimax rate-distortion theorem says any randomized quantizer of a unit vector at b bits per coordinate has worst-case MSE at least 4^{-b} (Cover and Thomas, 2006, Ch. 10). Multiplying by $\lambda_{\min}(W) \cdot \mathbb{E}[\rho^2]$ converts to the weighted-loss scale. Dividing the upper bound by this lower bound yields (10). □

Lemma IA.1 sharpens the picture in expectation over R. The diagonal entries of $RWR^\top$ concentrate around $\mathrm{tr}(W)/d$ for moderate condition number $\chi(W)$, so the worst-case factor $\lambda_{\max}(W)$ used in Step 2 of the proof is conservative for typical realizations of R. The concentration result is auxiliary intuition and does not enter the worst-case bound (8).

[13] We use the normalized-state case as the main statement because the TurboQuant guarantee in Zandieh et al. (2025) is stated for unit-norm vectors with norms stored separately, and because in finance applications the variance scale of returns is typically known and handled outside the attention problem.

4.3 Cost of Selective Attention: The Unmonitored Loss Floor

Proposition 1 shows what RTQ achieves. The next result characterizes what selective attention achieves at best. The trade-off is that selective concentrates all capacity on monitored dimensions and bears an irreducible loss floor on the rest.

Consider a selective-k strategy with hard exclusion: the agent acquires signals about a chosen subspace of k directions and uses the prior mean on the orthogonal complement, without inferring information about unmonitored dimensions from monitored ones. This matches the standard applied interpretation of "monitor a few key indicators and ignore the rest." We work in a basis where W and the prior covariance Σ are jointly diagonal (or assume the prior is diagonal in the eigenbasis of W); the diagonal case is the cleanest setting in which to state the floor and is the one used in nearly all applied rational-inattention work.

Proposition 2 (Selective attention with hard exclusion: unmonitored loss floor)**.** *Fix $k < d$ and consider the selective-k strategy with hard exclusion in a basis where $W = \mathrm{diag}(\omega_1, \ldots, \omega_d)$ and $\mathrm{Cov}(x_j, x_\ell) = 0$ for $j \neq \ell$, with $\omega_1 \geq \cdots \geq \omega_d > 0$ and prior variance σ_j^2 along direction j. Let $S \subset \{1, \ldots, d\}$ be the chosen monitored set with $|S| = k$. The expected loss satisfies*

$$\mathcal{L}^S \;\geq\; \sum_{j \notin S} \omega_j \sigma_j^2. \tag{11}$$

The optimal hard-exclusion set chooses the k directions with the largest welfare exposure $\omega_j \sigma_j^2$, not necessarily the k directions with the largest ω_j. The right-hand side does not depend on capacity: no matter how finely the agent resolves the monitored subspace, the unmonitored directions leave an irreducible floor.

Proof sketch. Under hard exclusion, the agent's action on $j \notin S$ is the prior mean, so posterior error on direction j has variance σ_j^2 and contributes $\omega_j \sigma_j^2$ to the weighted loss. Monitored directions contribute non-negative loss. Summing over the unmonitored set gives (11). To select S optimally, minimize $\sum_{j \notin S} \omega_j \sigma_j^2$ subject to $|S| = k$, which selects the k largest values of $\omega_j \sigma_j^2$. Full argument in Appendix A.3. □

Two caveats are worth flagging. First, the diagonality assumption is essential. If the prior covariance Σ is not diagonal in the eigenbasis of W, signals about monitored directions can reveal information about unmonitored ones, lowering the posterior variance on the orthogonal complement below the prior. The general case replaces the scalar tail sum in (11) with the trace of W times the posterior error covariance on the omitted subspace, which is bounded above by the prior covariance on that subspace. Second, the proposition is stated for the hard-exclusion strategy, the standard applied interpretation. A selective strategy

that exploits cross-direction correlations may achieve lower loss; the bound (11) still applies to the hard-exclusion variant, which is what the applied literature actually implements.

RTQ avoids the floor by spreading attention across all d dimensions, reducing per-coordinate error at rate $4^{-\kappa/d}$ and leaving no dimension at full prior variance. As capacity grows with d and k fixed, RTQ loss vanishes while the selective floor remains, so the welfare ratio diverges. For $k = 10$ and $d = 500$, RTQ uses fifty times fewer bits per monitored dimension than selective but covers fifty times as many dimensions, and the 490 ignored dimensions form an irreducible floor.

Illustrative bounds for representative settings (well-diversified portfolio, multi-product pricing, monetary policy with multiple instruments) are computed in Internet Appendix IA.1.7 by plugging plausible condition numbers into the theoretical bound. The worst-case envelope is far looser than realized performance.[14]

The worst-case RTQ bound is dimension-free, with constant factor $C_T\chi(W)$ that does not depend on d. The average-case picture sharpens with d: the concentration argument in Lemma IA.1 makes the diagonal entries of $RWR^\top$ closer to $\mathrm{tr}(W)/d$, so the water-filling allocation moves toward uniformity and the typical realization of R leaves less slack between RTQ loss and the minimax floor. The comparison with hard-exclusion selective attention also becomes more favorable as d rises, because the $O(d)$ selective penalty over the ignored tail grows while RTQ loss does not. The worst-case bound is dimension-free, while the average-case rotation argument and the comparison with hard-exclusion selective attention become more favorable as dimension rises. Applied recommendations to concentrate attention can be actively harmful when the environment is complex.[15]

4.4 Optimal Bit Allocation Across Dimensions

How should the agent distribute capacity across the rotated coordinates? The next result characterizes the optimal allocation.

Let W have eigenvalues $\omega_1 \geq \cdots \geq \omega_d > 0$ with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$. Define the *rotated loss weights* as the diagonal of $RWR^\top$:

$$w_i = (RWR^\top)_{ii} = \mathbf{e}_i^\top RWR^\top \mathbf{e}_i. \tag{12}$$

These measure how much the loss penalizes errors in coordinate i of the rotated frame.

[14]For example, on the S&P 500 proxy calibration RTQ delivers a welfare cost of 0.5 bps per year against VNV's 198 bps at the same budget; see Table IA.1 in Internet Appendix IA.1.

[15]The full intuition, including the insurance analogy and the two scaling forces (more unmonitored dimensions, tighter average-case allocation), is developed in Internet Appendix IA.1.5.

Proposition 3 (High-resolution water-filling approximation)**.** *Under the high-resolution Lloyd-Max MSE model* $\mathrm{MSE}_i(b_i) = \alpha_d \cdot 2^{-2b_i}$ *for the* $\mathrm{Beta}(\frac{d-1}{2}, \frac{d-1}{2})$ *distribution at* b_i *bits, the continuous bit allocation* $(b_1^*, \ldots, b_d^*)$ *that minimizes expected loss under RTQ, subject to* $\sum_{i=1}^d b_i \leq \kappa - b_0$*, satisfies*

$$b_i^* \;=\; \left[\frac{1}{2}\log_2\Big(\frac{w_i}{\nu}\Big)\right]^+, \tag{13}$$

where $[z]^+ = \max(z, 0)$ *and* $\nu > 0$ *is chosen so* $\sum_{i=1}^d b_i^* = \kappa - b_0$*. For finite and integer bit widths, the exact RTQ allocation solves the discrete separable problem*

$$\min_{b_i \in \mathbb{Z}_{\geq 0}} \sum_{i=1}^d w_i \cdot \mathrm{MSE}_i(b_i) \quad s.t. \quad \sum_{i=1}^d b_i \leq \kappa - b_0, \tag{14}$$

using a precomputed Lloyd-Max distortion table. The water-filling formula (13) *is the closed-form high-resolution limit of* (14) *as* $b_i \to \infty$*.*

The allocation is a "water-filling" rule familiar from information theory: bits go only to dimensions where the welfare weight w_i exceeds a threshold ν, and within those, the allocation rises with $\log_2 w_i$, so high-stakes dimensions receive more bits. The threshold ν adjusts so the total bits equal the budget; it can be read as the shadow price of capacity, the welfare value of one extra bit.

The main insight is that attention is governed by the loss structure, not the data. In Sims (2003)'s one-dimensional setting, cross-dimensional allocation is trivial; in low-dimensional extensions, conventional wisdom says attend to more volatile variables. Our result reveals a different mechanism: after rotation, all rotated coordinates have the same prior variance, so the allocation depends only on the welfare weights w_i, not on which raw variables happen to be most uncertain. Stakes drive the allocation, not variance.

Take a portfolio manager. Conventional rational inattention says attend to volatile stocks. Our result says attend to stocks whose mispricing contributes most to portfolio risk, captured by W rather than individual volatilities. A low-volatility stock highly correlated with the rest may deserve more attention than a high-volatility diversifier because its errors are amplified by portfolio structure. This shapes the w_i before rotation; after rotation, Lemma IA.1 concentrates the w_i around $\mathrm{tr}(W)/d$, so the allocation is nearly uniform in high dimensions.

When $W = I$, all w_i are equal and allocation is uniform: $b_i^* = (\kappa - b_0)/d$. This is the special case where the Zandieh et al. (2025) distortion bound applies directly. For fixed d, departure from uniformity grows with $\chi(W)$. In high dimensions, Lemma IA.1 tempers this: w_i concentrate around $\mathrm{tr}(W)/d$ and allocation approaches uniformity even for asymmetric W.

Proof sketch. Under RTQ, expected loss decomposes in the high-resolution approximation as $\mathcal{L}^{RTQ} \approx \mathbb{E}[\rho^2] \sum_{i=1}^{d} w_i \cdot \mathrm{MSE}_i(b_i) + \mathcal{L}^{\mathrm{norm}}(b_0)$, where the off-diagonal cross-terms $\mathbb{E}[\delta_i \delta_j]$ for $i \neq j$ are assumed negligible (mean-zero scalar Lloyd-Max errors with cross-covariances of order $o(4^{-b})$ relative to diagonal MSE; see Appendix A.4). For the Beta distribution, $\mathrm{MSE}_i(b_i)$ is convex and decreasing in b_i, so constrained minimization of this separable objective with a budget yields water-filling via KKT. The threshold ν is the Lagrange multiplier on the capacity constraint. The water-filling formula is the closed-form asymptotic result; the discrete integer-allocation formulation (14) is the natural finite-resolution counterpart that recovers the water-filling rule in the high-resolution limit. □

Under the high-resolution approximation, after rotation the problem becomes separable: total loss sums independent terms, one per coordinate, and direction-encoding capacity sums bits. This is a classic resource allocation with a budget and d investments with diminishing returns. What distinguishes this from textbook water-filling is that the w_i are diagonal entries of $RWR^\top$, not eigenvalues of W. The rotation mixes eigenvalues, so w_i reflect the loss structure in the rotated frame. Lemma IA.1 is not needed for the worst-case guarantee in Proposition 1 but helps explain why the water-filling allocation is often close to uniform after random rotation when W is not too ill-conditioned.

4.5 When to Focus and When to Spread

Propositions 1 and 2 together give a practical criterion. The relevant object is the *loss concentration ratio*, measuring how much of the decision problem is captured by the top k dimensions.

Definition 1 (Loss Concentration Ratio)**.** *Let W have eigenvalues $\omega_1, \ldots, \omega_d > 0$ with prior variances $\sigma_1^2, \ldots, \sigma_d^2$ along the corresponding eigenvectors. Reindex the eigenvectors so that the welfare exposures $\omega_j \sigma_j^2$ are sorted in decreasing order: $\omega_1\sigma_1^2 \geq \omega_2\sigma_2^2 \geq \cdots \geq \omega_d\sigma_d^2$. The loss concentration ratio at rank k is the share of total welfare-weighted prior loss captured by the top-k welfare exposures,*

$$\rho(k) \;=\; \frac{\sum_{j=1}^{k} \omega_j \sigma_j^2}{\sum_{j=1}^{d} \omega_j \sigma_j^2} \;=\; \max_{|S|=k} \frac{\sum_{j \in S} \omega_j \sigma_j^2}{\sum_{j=1}^{d} \omega_j \sigma_j^2}. \tag{15}$$

The welfare-weighted ordering is what Proposition 2 flags as the optimal hard-exclusion criterion. When σ_j^2 is constant across directions (or proportional to ω_j, as in mean-variance portfolio choice with $W = \gamma\Sigma$), the welfare-weighted ordering coincides with the ordering by ω_j.

When $\rho(k) \approx 1$, the problem is effectively k-dimensional: almost all welfare-relevant uncertainty resides in k directions. When $\rho(k)$ is far from 1, the remaining $d-k$ dimensions carry substantial loss and ignoring them is expensive.

Corollary 1 (Focus-vs-Spread Threshold)**.** *Under the exact-norm assumption of Proposition 1, RTQ achieves lower expected loss than the selective-k hard-exclusion strategy whenever*

$$\rho(k) \;<\; 1 - C_T \cdot \chi(W) \cdot \frac{\mathcal{L}_{\text{lb}}^{\text{sphere}}(\kappa)}{\sum_{j=1}^{d} \omega_j \sigma_j^2}, \tag{16}$$

where $C_T = \pi\sqrt{3}/2$. The condition is informative when $C_T \cdot \chi(W) \cdot \mathcal{L}_{\text{lb}}^{\text{sphere}}(\kappa) < \sum_{j=1}^{d} \omega_j \sigma_j^2$, which holds when the problem is moderately conditioned and capacity is meaningful. With norm quantization at b_0 bits, an additional $L_\rho(b_0)$ term enters the right-hand side as in Proposition 1; under bounded support and high-rate conditions this term vanishes as $b_0 \to \infty$.

When $\chi(W)$ is moderate and capacity is meaningful, the right-hand side is close to one. The threshold is a conservative sufficient condition because it compares an RTQ upper bound with a selective-strategy lower bound: in applications RTQ may dominate even when the sufficient condition is silent.[16]

The criterion binds only when W is nearly rank-deficient and capacity is very limited. For rank-one W (a single policy instrument), $\rho(1) = 1$ and concentration is trivially optimal.

Proof. Combining the lower bound from Proposition 2 with the exact-norm RTQ upper bound from Proposition 1: $\mathcal{L}^S \geq (1-\rho(k)) \sum_{j=1}^{d} \omega_j \sigma_j^2$ and $\mathcal{L}^{RTQ} \leq C_T \cdot \chi(W) \cdot \mathcal{L}_{\text{lb}}^{\text{sphere}}(\kappa)$. RTQ wins whenever the upper bound on $\mathcal{L}^{RTQ}$ is below the lower bound on $\mathcal{L}^S$, which gives (16). With norm quantization at b_0 bits, the same argument carries through with the additional $L_\rho(b_0)$ residual on the right-hand side. □

$\rho(k)$ is computable from the eigenvalues of W and prior variances, so researchers can plot $\rho(k)$ against k to gauge effective dimension. If $\rho(k)$ is close to one for small k, the problem is effectively low-dimensional and selective attention suffices. If $\rho(k)$ stays well below one even at large k, the problem is genuinely high-dimensional and RTQ dominates.

[16] For a portfolio with $d = 500$ and $\chi(W) = 3$ (typical after shrinkage), if the minimax benchmark equals 10 percent of prior loss ($\mathcal{L}_{\text{lb}}^{\text{sphere}} / \sum_j \omega_j \sigma_j^2 = 0.1$), the sufficient condition becomes $\rho(k) < 1 - 2.72 \times 3 \times 0.1 \approx 0.18$. The condition fires only when the top k directions capture less than 18 percent of total loss, which is a stringent test. The condition is sufficient but not necessary: Figure IA.3 in the Internet Appendix shows realized loss ratios of 20× to 400× in our S&P 500 sample even where the sufficient condition is silent. Loss ratios that high indicate RTQ dominates well outside the corollary's guaranteed region.

4.6 Extension to market equilibrium with heterogeneous attention

The three propositions above characterize the single-agent attention problem. The framework extends to a market equilibrium in which the RTQ-bounded agent is one of several agent types interacting through a price; the full development is in Internet Appendix IA.5. We summarize the main results here.

Consider a Gaussian-CARA noisy rational-expectations equilibrium with full-information agents (mass α, signal noise $\tau^2\Sigma$ with $\tau > 0$), RTQ-bounded agents (mass $1-\alpha$, eigenbasis-aligned RTQ at bit budget B), and noise traders (per-capita supply $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\Sigma)$). The market clears at a linear price $p = \eta_x x + \eta_n n$ in each rotated coordinate. Internet Appendix Theorem IA.3 establishes existence and uniqueness of this equilibrium with explicit closed-form coefficients (η_x, η_n) as functions of $(\alpha, \tau, B, \gamma, \sigma_n^2, \lambda)$, and shows that the RTQ-agent welfare loss decomposes into a per-agent RTQ-noise contribution and a price-information contribution. The result extends directly to $K+1$ agent types with heterogeneous bit budgets B_k and risk aversion γ_k via Corollary IA.2, with two precision-weighted aggregates (A_γ and G) replacing the homogeneous-type sufficient statistic.

The equilibrium analysis sharpens the asset-pricing predictions. As the fraction of well-informed agents grows, prices become more informative, the welfare cost of being attention-bounded falls, and the equilibrium tilts toward passive participation. The mechanism connects directly to current debates about indexing and the rise of AI-based investment strategies: a higher prevalence of high-attention (e.g., learned-attention Transformer) traders compresses the welfare gap between informed and bounded-attention investors through the price-information channel, generating the empirical decline in active-management alpha documented in Fama and French (2010).[17]

Together, the three propositions characterize rational inattention in high dimensions: RTQ comes within $C_T\chi(W)$ of the minimax floor uniformly in d, κ, and the realized state; the high-resolution allocation is governed by the loss structure rather than the data; and hard-exclusion selective attention leaves an irreducible tail-loss floor. None of these assumes Gaussianity or a parametric prior.

Section 5 takes these single-agent propositions to portfolio choice on the full CRSP uni-

[17] The framework also admits a dynamic extension to a T-period Gaussian AR(1) setting in which the agent reallocates attention over time. The Kalman recursion with RTQ noise admits closed form. The optimal stationary policy is characterized by an Euler condition and is unique on the high-rate range (Proposition IA.8 and Lemma IA.2). A turnpike-style convergence of the finite-T optimum to the stationary policy is expected and stated as an informal implication in the Internet Appendix, not as a theorem. The resulting dynamics reproduce the three features identified by Steiner et al. (2017) for dynamic rational inattention under Shannon capacity: inertia, sluggish response to structural breaks, and budget pooling across periods. The Gaussian-AR(1) dynamic loss bound is the dynamic counterpart of the static main-text guarantee (Corollary IA.3). See Internet Appendix IA.5.3.

verse and the LSEG 13F panel.[18] The Internet Appendix calibrates the same single-agent propositions on the S&P 500 proxy with a head-to-head against Van Nieuwerburgh and Veldkamp (Internet Appendix IA.1).

5 Empirical Tests

This section tests the asset-pricing predictions of the framework on the CRSP monthly panel merged with the LSEG 13F institutional holdings.[19]

The asset-pricing predictions are conditional on the focus-versus-spread regime. When the loss concentration ratio $\rho(k)$ stays well below one even at large k, the problem is high-dimensional and RTQ should dominate VNV in welfare cost, with the gap widening as $\rho(1)$ falls and effective dimension rises. In the same regime, institutional investors whose allocations look more RTQ-like than VNV-like should earn higher realized returns. The US equity universe (eligible d in the hundreds to thousands across the sample) sits in this regime, and the empirical tests below quantify the gap. This section runs the rolling head-to-head on the full eligible CRSP universe with quarterly windows from 1970Q1 to 2026Q1, then tests the spectrum link, the manager-fit distribution, and the welfare-performance regression.[20]

[18]The parallel-target setting (one agent estimating N multi-dimensional targets at once under a shared bit budget) is the block-diagonal-W special case of the single-agent setup. Its formal setup, its pair of theorems (spread coverage and concentration floor), the sufficient-statistic collapse, and the sufficient-condition regime diagram for spread versus concentration are developed in Internet Appendix IA.2. The Fisman cross-domain test in Internet Appendix IA.3 uses the spread theorem and the slope-implied effective dimension that follow from this framework.

[19]A cross-domain test of the parallel-target effective-dimension identity on the Fisman et al. (2006) Columbia speed-dating experiment (8,378 four-minute conversations among 551 participants in 21 randomized waves of size 5 to 22) is in Internet Appendix IA.3 and proceeds in three subsections. Subsection IA.3.1 states the structural model: Proposition IA.3 characterizes the symmetric Nash threshold under noisy observation; the slope formula (IA.8) then maps the spread-versus-concentration trade-off to a log-linear slope β, and Proposition IA.4 links that slope to the effective dimension of the trait covariance through $d_{\text{eff}}^{\text{slope}} = 1/(1-\beta) \approx d_{\text{eff}}^{\text{PCA}}(\Sigma) = (\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$. Subsection IA.3.2 runs the regression of $\log(1+M_i)$ on $\log N_i$ (Figure IA.8 and Table IA.2: $\hat{\beta} = 0.472$, wave-clustered SE 0.063, $p < 10^{-10}$, positive across genders, age bands, and Poisson and negative-binomial GLMs; a seating-position placebo is indistinguishable from zero) and reads $d_{\text{eff}}^{\text{slope}} = 1.89$ off the slope; the PCA on the standardized six-trait correlation matrix (Figure IA.9) returns $d_{\text{eff}}^{\text{PCA}} = 2.68$, the agreement Proposition IA.4 predicts in a non-portfolio setting. Subsection IA.3.3 recovers a per-coordinate bit budget of $\hat{b} = 0.127$ bits from a structural probit on partner attractiveness ($\hat{B} \approx 0.24$ bits total per partner-conversation, below one bit, consistent with a noisy binary decision) and confirms slope stability across 100 random 15/6 wave splits (Figure IA.10). The positive slope is inconsistent with a hard-concentration model in which only a fixed number of partners receives attention.

[20]A companion calibration on a 488-stock proxy of the top 500 CRSP commons by market cap, filtered to stocks present in the top 500 for at least 240 months over 1970–2026, is in Internet Appendix IA.1 and proceeds in five subsections. Subsection IA.1.1 sets up the portfolio problem with $W = \gamma\,\Sigma$. Subsection IA.1.2 computes the loss concentration ratio $\rho(1) = 0.80$, $\rho(5) = 0.95$ for the proxy covariance and plots the spectrum diagnostics (Figure IA.1 and Figure IA.2: PCA reaches 0.95 only deep into the spectrum while ρ crosses 0.95 at $k = 6$, or 1.2 percent of d). Subsection IA.1.3 reports the main welfare table (Table IA.1: 198

Our analysis sample uses common stocks on NYSE/AMEX/NASDAQ with at least 60 months of trailing return history. The rolling out-of-sample exercise covers 221 non-overlapping quarterly windows from 1970Q1 to 2026Q1; eligible universe size ranges from 711 to 4,356 stocks (mean 2,414). The annual $\rho(1)$ (the loss concentration ratio, the share of welfare-weighted variance carried by the single top eigendirection) and d_{eff} (the effective dimension, $(\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$) series uses 55 December endpoints from 1971 to 2025. The 13F manager panel, restricted to manager-quarters with at least \$1M in reported US equity from 1980Q1 to 2024Q4, contains 451,121 manager-quarters covering 13,571 unique managers in the manager-fit panel (cosine-distance comparison) and 451,061 manager-quarters covering 13,569 managers in the regression panel (after merging FF5+UMD factor data for the alpha regression). We use CRSP for rolling covariances and to translate share holdings into portfolio weights. Dollar magnitudes are quoted as certainty-equivalent welfare cost in basis points per year, computed as $(L/P)\cdot\mu_p\cdot 10^4$, where L is the welfare loss under the chosen allocation rule (RTQ or VNV), P is the prior welfare loss (the loss with no signal), and $\mu_p = 0.06$ is the equity-premium anchor. The ratio L/P measures the fraction of prior welfare loss remaining after the allocation; multiplying by μ_p converts to a return-equivalent scale and 10^4 to basis points.

Loss concentration and effective dimension vary substantially over the sample. For each year 1971–2025, we estimate the $N \times N$ sample covariance on 60 trailing monthly returns of the eligible universe, apply Ledoit-Wolf (Ledoit and Wolf, 2004) shrinkage, and compute $\rho(k) = \sum_{j=1}^{k}\lambda_j^2 / \sum_{j=1}^{N}\lambda_j^2$ and $d_{\text{eff}} = (\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$. Across 55 annual endpoints, $\rho(1)$ ranges 0.260 to 0.936 (mean 0.661) and d_{eff} ranges 9.6 to 587.6 (mean 146.7). Figure 1 plots both. The $\rho(1)$ range alone rules out the knife-edge conditions under which VNV and RTQ are equivalent: the market sits most of the sample in the range where the two prescribe materially different allocations.

We then run the rolling out-of-sample head-to-head. At each quarter-end from 1970Q1 to 2026Q1, we estimate the covariance on the prior 60 months, compute RTQ water-filling and VNV selective-$k = 1$, and evaluate their out-of-sample welfare cost on the next 24 months.

bps for VNV versus 0.5 bps for RTQ at $b = 2$) and the selective-vs-RTQ loss ratio across b (Figure IA.3, ratios 393 at $b = 2$ rising past 6,000 at $b = 4$). Subsection IA.1.4 rolls the head-to-head through 60-month windows for a mean gap of 163 bps per year across 205 non-overlapping windows (positive in every window) and plots the VNV-vs-RTQ cost across universe size N (Figure IA.4, gap 357 bps at $N = 25$ down to 198 bps at $N = 488$). Subsection IA.1.5 develops the high-dimensions intuition behind these numbers. Replication code will be made available on publication. The 1970 quarterly start reflects the 60-month estimation requirement applied to CRSP data that begins in June 1962 (NYSE only at first; AMEX added later, NASDAQ joining CRSP in 1972). The annual $\rho(1)$ and d_{eff} series in Figure 1 use December endpoints from 1971 to 2025 (55 endpoints); earlier December endpoints are dropped because the early CRSP universe with the 60-month history filter applied is too small to support a stable covariance estimate. The 13F panel separately begins in 1980, when LSEG institutional-holdings coverage becomes reliable.

Across 221 non-overlapping rolling windows, the gap is positive in 159 (71.9%). Mean gap +100 bps per year, median +409, IQR $[-104, +532]$. The left tail is heavy: worst window (1995Q3) gives −3,968 bps; best gives +572. Figures 1 and 2 show the loss-concentration time series and the rolling gap together. The asymmetry says VNV is the better bet in most quarters but a few reversals around the late-1990s dispersion episode drag the mean down, matching what theory predicts: VNV is fragile to states where the realized top eigenvector is not the one the ex-ante covariance selected.[21]

Propositions 1 and 2 link welfare cost to the spectrum through d_{eff} and $\rho(k)$, so if the link is causal, ex-ante loss concentration should predict the realized gap. Figure 3 overlays the gap against annual $\rho(1)$ on a twin axis. When $\rho(1) < 0.5$ (no single direction dominates), the gap turns negative; when $\rho(1) > 0.8$, the gap is uniformly positive. Co-movement is visible at both low-frequency trend and business-cycle frequency.

The welfare gap also scales with universe size. Figure 4 holds $\kappa = 2N$ and varies N along the top-cap sort at 1990-12 and 2024-12. VNV is flat at 600 bps because the selective rule captures the same fraction of risk regardless of N. RTQ falls from roughly 1,200 bps at $N = 100$ to between 60 and 100 bps at $N = 2000$, one to two orders of magnitude.[22]

The predicted link between $\rho(1)$ and the gap has cross-sectional content. At each quarter-end we sort the universe into ten market-cap deciles and compute $\rho(1)$, d_{eff}, and the VNV − RTQ gap within each decile. Figure 5 plots the 2024-12 cross-section. The slope of gap on $\rho(1)$ across deciles is roughly +1,900 bps per unit, stable across the obvious robustness checks (full sample, dropping the micro-cap outlier, dropping the bottom two deciles).[23] A 0.1 increase in $\rho(1)$ raises the RTQ advantage by roughly 200 bps per year, the same order as the OOS mean gap.

We turn next to manager-level evidence from the 13F panel. To ask whether actual investors behave more like VNV or RTQ, we compute each manager's quarterly allocation distance to both. For each 13F manager-quarter with at least $1M in reported US equity, we translate share positions to active weights (excess of market-cap benchmark), normalize,

[21] The 1995Q3 worst window (−3,968 bps) is a single quarter where the realized top eigenvector diverged sharply from the ex-ante one. Trimming the ten most negative windows leaves the mean at +218 bps and the median essentially unchanged at +420 (versus +100 and +409 in the full sample), so the result is not driven by tail outliers. The trim-K sensitivity is monotone: at K = 1, mean is +119; at K = 5, +171; at K = 20, +287. The full sample is the conservative reading.

[22] The non-monotonic kink at 2024-12 $N \in [1000, 1500]$ is a spectrum effect: adding mid-caps pushes $\rho(1)$ below 0.6 and raises RTQ cost temporarily before N grows enough for water-filling to absorb the added dispersion.

[23] The unconditional $p = 0.11$ on the full $n = 10$ sample reflects one micro-cap decile (d9) that sits well off the trend line in Figure 6; standard outlier handling resolves the issue. Full sample $n = 10$: slope +1,746, $R^2 = 0.29$, $p = 0.11$. Drop d9 outlier ($n = 9$): slope +1,934, $R^2 = 0.83$, $p < 0.001$. Drop d9 and d10 ($n = 8$): slope +1,974, $R^2 = 0.15$, $p = 0.34$. The slope point estimate stays in a tight band near +1,900 across all three.

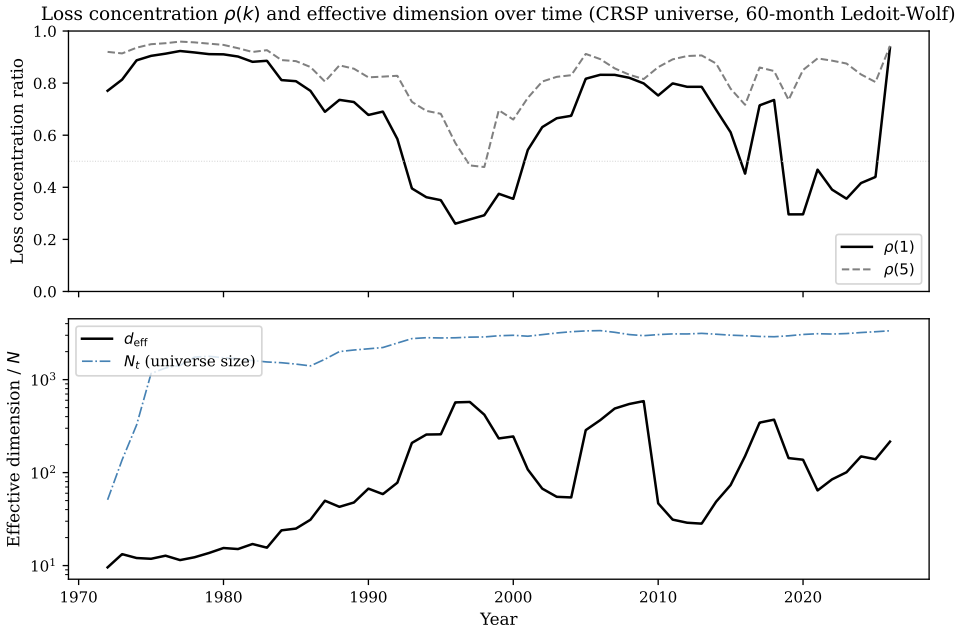

Figure 1: **Loss concentration $\rho(k)$ and effective dimension in the CRSP universe, 1971–2025.** At each year-end, Σ_t is the 60-month Ledoit and Wolf (2004) shrinkage estimator on monthly returns of the eligible CRSP universe (common stocks on NYSE/AMEX/NASDAQ with 60+ months). Top: $\rho(1)$ (solid) and $\rho(5)$ (dashed) with $\rho = 0.5$ dotted. Bottom: d_{eff} (solid, left) and universe size N_t (dot-dash, right). Mean $\rho(1) = 0.66$, range $[0.26, 0.94]$; mean $d_{\text{eff}} = 147$, range $[9.6, 588]$. Neither flat nor rank-one, which is the range where RTQ and VNV differ.

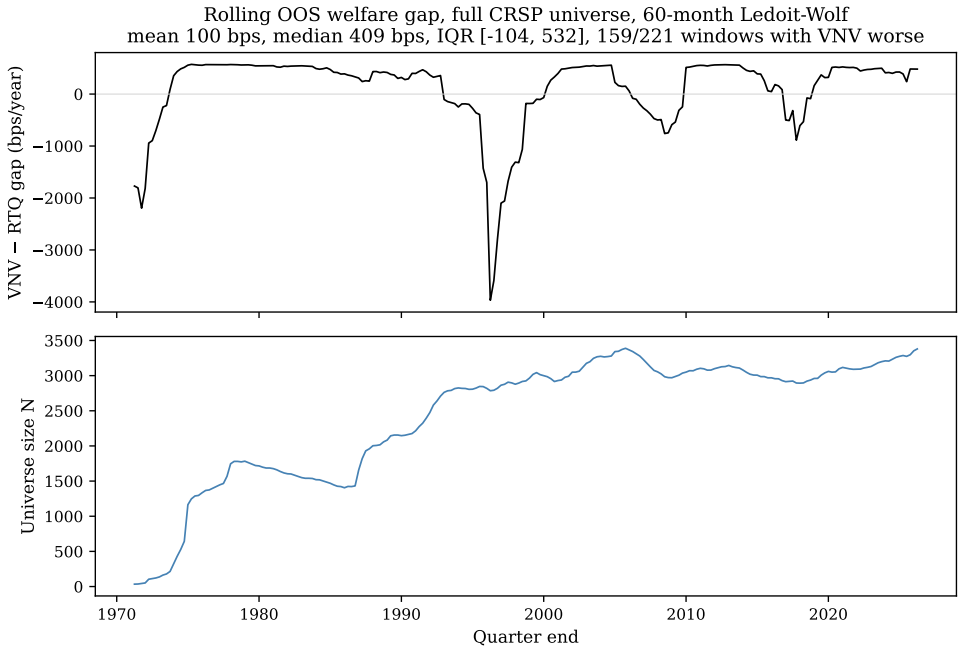

Figure 2: **Rolling 24-month out-of-sample VNV − RTQ welfare gap, 1970–2026.** At each quarter-end t, VNV (selective $k = 1$) and RTQ (water-filling) are computed from the 60-month Ledoit and Wolf (2004) shrunk covariance, then CE welfare costs are evaluated on the next 24 months. Top: gap $L_{\text{VNV}} - L_{\text{RTQ}}$ in bps per year, one observation per non-overlapping quarter. Positive: VNV worse. 221 windows: median +409, mean +100 bps; VNV worse in 159/221 (71.9%). 1995 dive to −3,968 bps shows fragility: one eigenvector shift reverses the ranking. Bottom: eligible universe size N_t.

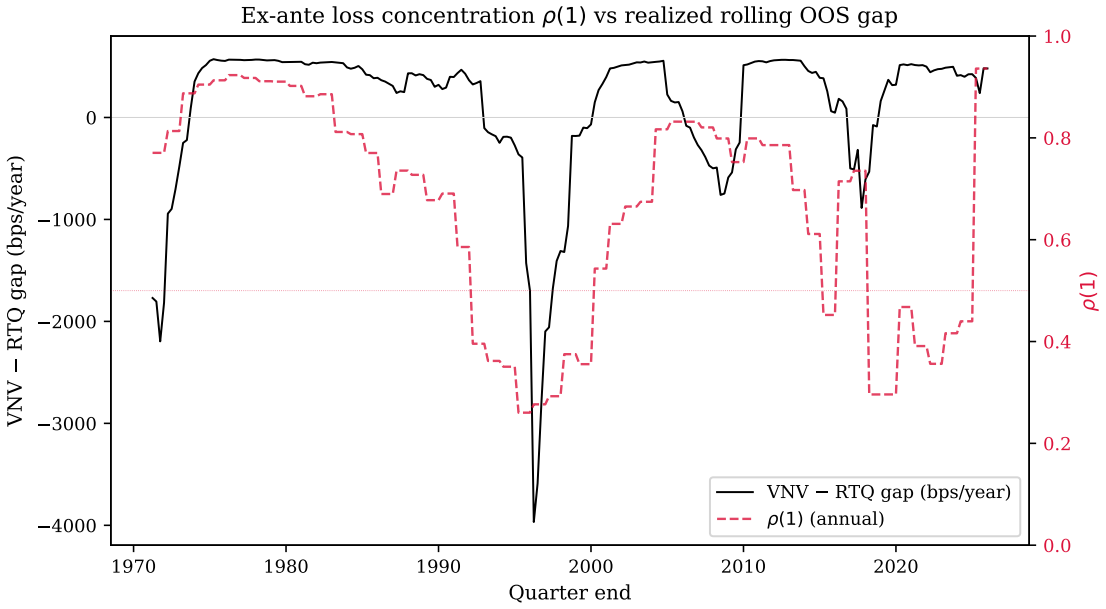

Figure 3: **Ex-ante $\rho(1)$ predicts the realized VNV − RTQ gap.** Horizontal: 1970–2026. Solid black (left): rolling 24-month gap in bps per year, quarterly (same as Figure 2 top). Red dashed (right): annual $\rho(1) = \lambda_1^2 / \sum_j \lambda_j^2$ from the 60-month shrunk covariance, held constant within year. The gap turns negative whenever $\rho(1)$ drops below 0.5 (red dotted) and stays positive whenever $\rho(1)$ stays above 0.8.

and compute cosine distances d_{VNV} and d_{RTQ} against the optimal portfolios built from the prior-60-month covariance. We project active weights onto the top 50 eigendirections of the eligible-universe covariance, which captures essentially all loss-weighted spectral concentration ($\rho(50) \approx 0.99$). The panel covers all four quarter-ends from 1980Q1 to 2024Q4, includes every manager above the $1M floor with no rank cap, and contains 451,121 manager-quarters across 13,571 unique managers in the fit panel. In 95.3 percent of manager-quarters the active-weight vector is closer to RTQ than to VNV, so 13F managers look strongly RTQ-like on average.[24]

Observed portfolios reflect beliefs, preferences, mandates (investment-style rules, benchmark-tracking constraints, position-size limits), and attention jointly, so inferring attention from holdings requires care. We do not claim cosine distance to RTQ identifies attention directly. We interpret the 13F evidence as consistent with broad-attention behavior under stated identifying assumptions, namely that orthogonal sources of portfolio heterogeneity are uncorrelated with the VNV and RTQ benchmark vectors conditional on universe and quarter. We follow a similar approach to Kacperczyk et al. (2014, 2016), who infer time-varying attention from mutual-fund portfolio composition under auxiliary assumptions of the same kind. Four pieces of evidence support this reading. First, the investor-type ranking below (Insur-

[24]Cross-sectional means $\bar{d}_{\text{RTQ}} = 0.966$ and $\bar{d}_{\text{VNV}} = 0.999$ are both near one, so neither benchmark is literally replicated. The consistent ordering, not the absolute level, is what carries the comparison.

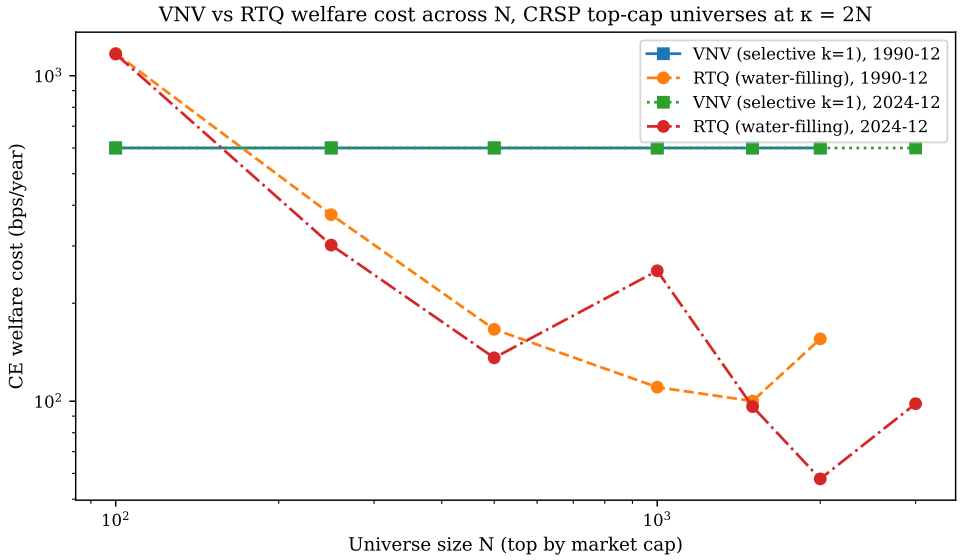

Figure 4: **VNV vs RTQ welfare cost across universe size, CRSP top-cap.** At 1990-12 and 2024-12, CRSP commons are ranked by end-of-month cap and the top N are taken. For each N, Σ is the 60-month shrunk estimator, $\kappa = 2N$. Horizontal: N log. Vertical: CE cost in bps per year log (CE metric from Table IA.1). Blue solid (VNV, 1990-12) and green dotted (VNV, 2024-12): flat near 600 bps. Orange dashed (RTQ, 1990-12) and red dash-dot (RTQ, 2024-12): fall by roughly an order of magnitude as N grows. RTQ dominates VNV for $N \geq 250$ at both vintages. 2024-12 kink at $N \approx 1000$: mid-caps temporarily lower $\rho(1)$.

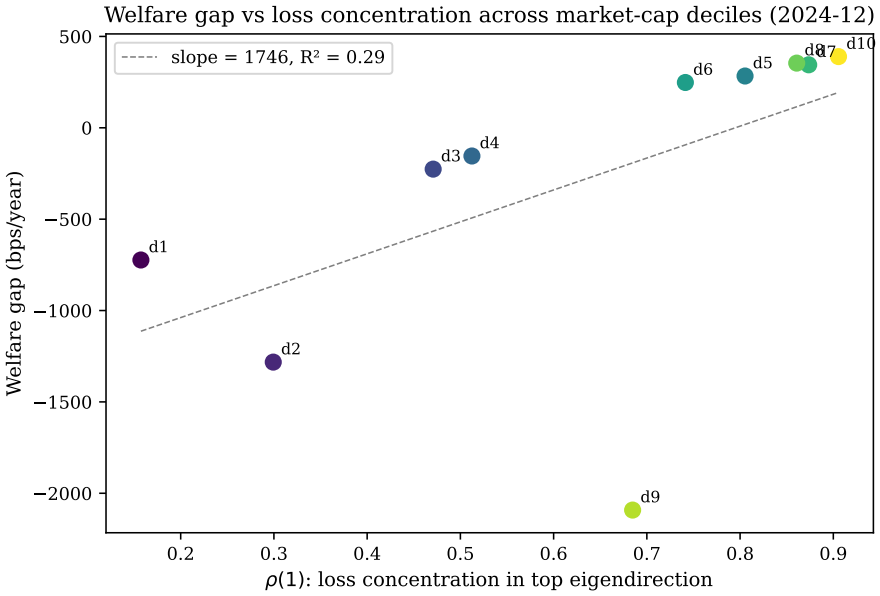

Figure 5: **Welfare gap vs loss concentration across market-cap deciles, 2024-12.** Eligible CRSP universe as of December 2024 sorted into ten NYSE market-cap deciles. Within each decile, the 60-month sample covariance gives $\rho(1) = \lambda_1^2 / \sum_j \lambda_j^2$, and the VNV − RTQ gap is computed as in Figure 2 at fixed $\kappa = 2N_{\text{decile}}$. Horizontal: within-decile $\rho(1)$. Vertical: gap in bps per year. Dashed: cross-decile OLS fit, slope +1,746, $R^2 = 0.29$, $p = 0.11$ ($n = 10$). Low-concentration deciles (d1, d2) have large negative gaps (RTQ worse); high-concentration deciles have positive gaps. Decile 9 is a micro-cap outlier addressed in Figure 6.

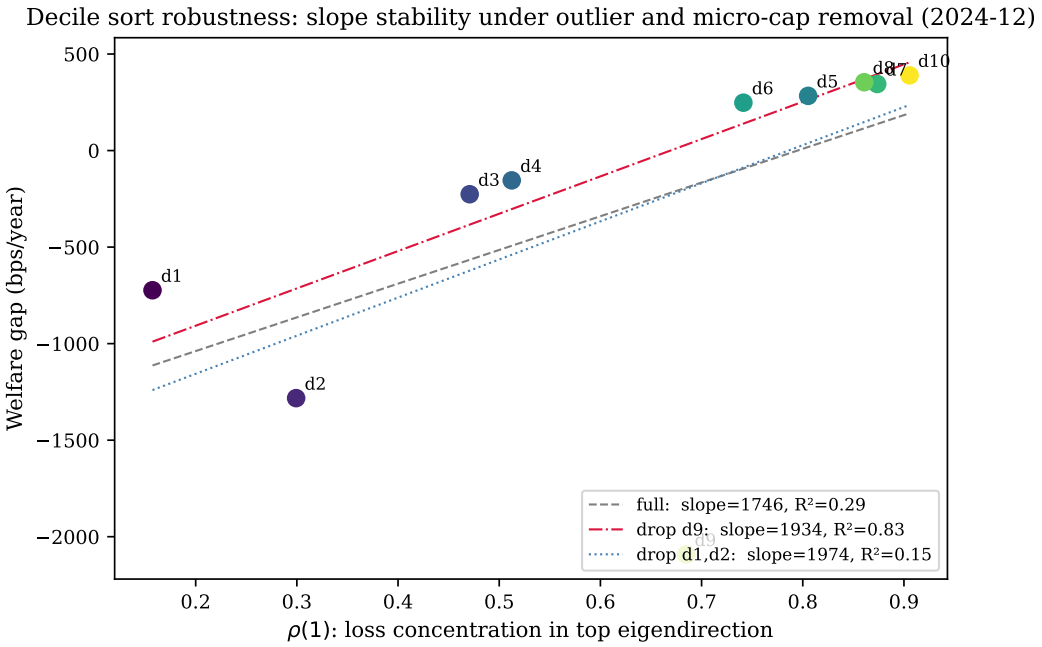

Figure 6: **Decile-sort robustness.** Same 2024-12 cross-section as Figure 5. Three OLS fits: full $n = 10$ (gray dashed, slope +1,746, $R^2 = 0.29$, $p = 0.11$); drop d9 outlier, $n = 9$ (red dash-dot, slope +1,934, $R^2 = 0.83$, $p = 0.001$); drop d9 and d10, $n = 8$ (blue dotted, slope +1,974, $R^2 = 0.15$, $p = 0.34$). Slope stays in a narrow band near +1,900 bps per unit $\rho(1)$.

ance most RTQ-like, Banks least) lines up with diversification mandates, so we emphasize the ordering rather than the absolute level. Second, the welfare-performance regression in Table 2 shows that fit distance predicts FF5+UMD risk-adjusted alpha within quarter at the 5 percent level for d_{RTQ} and at the 1 percent level for d_{VNV}, with the signs theory predicts. A pure mandates story would not co-move with realized alpha this way. Third, the channel varies systematically across private-information regulation. Pre-Reg FD, selective disclosure rewarded concentration, so the coefficient on $(d_{\text{RTQ}} - d_{\text{VNV}})$ should be positive; it is. Reg FD (2000) and Dodd-Frank (2010) erode the rents and the coefficient should flip negative; it does in both periods. Mandates are slow-moving institutional constraints, so the regime variation cuts against a pure-mandates reading. Fourth, the cosine-distance methodology is non-parametric and does not impose a factor structure, so the reading does not depend on a particular asset-pricing model. The four pieces together support the broad-attention interpretation without ruling out other contributing factors. We treat the 13F evidence as descriptive support for the theory rather than a definitive attention measurement.

The cross-section of investor types is also informative. We split managers by LSEG TYPECODE into five groups. Table 1 reports the fraction of manager-quarters closer to RTQ within each group. Insurance is the most RTQ-like (97.5 percent), followed by Investment companies (97.0), Advisors (95.6), Other (95.2), and Banks (93.3). Kruskal-Wallis on

$d_{\text{RTQ}} - d_{\text{VNV}}$ rejects equality at any conventional level.[25] Insurance and Investment companies sit at the top because both face binding diversification mandates that mechanically push allocations across many directions, while Banks have more room for concentrated holdings within their fiduciary frameworks. Figure 7 shows the full distribution.

Investor type	manager-quarters	unique managers	$\overline{d_{\text{RTQ}}}$	$\overline{d_{\text{VNV}}}$	frac closer to RTQ
Insurance	5,069	114	0.954	1.000	0.975
Investment cos.	7,374	595	0.963	0.999	0.970
Advisors	164,536	6,580	0.966	1.000	0.956
Other	250,502	5,686	0.966	0.999	0.952
Banks	23,640	564	0.955	0.997	0.933
All	451,121	13,571	0.966	0.999	0.953

Table 1: **Manager fit distances by LSEG TYPECODE.** For each manager-quarter with at least \$1M in reported US equity, share positions are translated to active weights (excess of market-cap benchmark) and projected onto the top 50 eigendirections of the prior-60-month Ledoit-Wolf shrinkage covariance. d_{RTQ}: cosine distance to the RTQ water-filling portfolio; d_{VNV}: same to the VNV selective $k = 1$. Last column: within-group fraction of manager-quarters with $d_{\text{RTQ}} < d_{\text{VNV}}$. Sample: 451,121 manager-quarters, 13,571 unique managers, LSEG S34 13F merged with CRSP, 1980Q1 to 2024Q4 (180 quarter-ends, all four quarters per year, no rank cap). Kruskal-Wallis $H = 2{,}704.89$ ($p \approx 0$) rejects equality. Mann-Whitney pairwise contrasts all reject equality at vanishing p-values given the panel size.

Fit distances are descriptive, so the sharper test is whether they predict realized performance. For each manager-quarter we compute the three-month forward active return, $r^{\text{act}}_{i,t+1:t+3} = w'_{i,t} r_{t+1:t+3} - w'_{\text{mkt},t} r_{t+1:t+3}$, and risk-adjust it using the Fama-French five-factor model augmented with momentum to get $\alpha^{\text{FF5+UMD}}_{i,t+1:t+3}$. We then regress alpha on the two fit distances with quarter fixed effects,

$$\alpha^{\text{FF5+UMD}}_{i,t+1:t+3} = \alpha_{q(t)} + \gamma_1 \, d_{\text{RTQ},i,t} + \gamma_2 \, d_{\text{VNV},i,t} + \epsilon_{i,t}. \tag{17}$$

Theory predicts $\gamma_1 < 0$ (closer to RTQ means higher alpha) and $\gamma_2 > 0$ (farther from VNV means higher alpha), with quarter fixed effects absorbing time-varying market conditions that move all managers symmetrically. Table 2 reports the estimates.

[25] Kruskal-Wallis $H = 2{,}704.89$, $p \approx 0$. Targeted Mann-Whitney contrasts all reject equality at vanishing p-values given the panel size, with Insurance vs Investment companies $p \approx 4 \times 10^{-86}$.

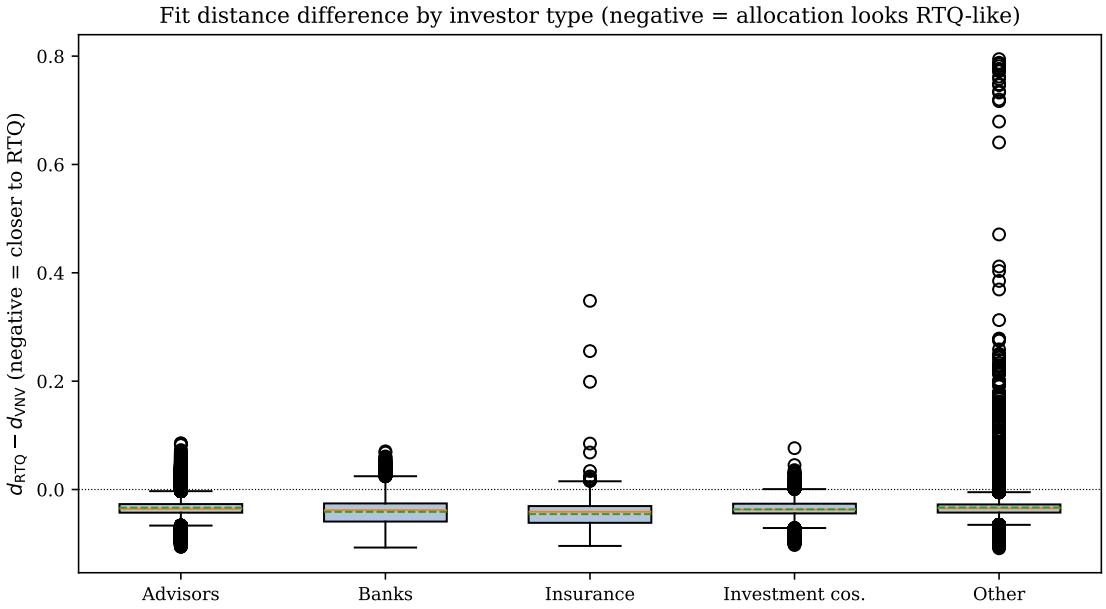

Figure 7: **Fit distance difference by investor type.** Box plot of $d_{\text{RTQ}} - d_{\text{VNV}}$ across 451,121 manager-quarters by LSEG TYPECODE. Sample: LSEG S34 13F merged with CRSP, 1980Q1 to 2024Q4 (180 quarter-ends, no rank cap, \$1M minimum reported US equity). Negative means closer to RTQ. Insurance (n = 5,069) is the most RTQ-like, followed by Investment companies (n = 7,374), Advisors (n = 164,536), Other (n = 250,502), and Banks (n = 23,640). Dotted line at zero marks equidistance. Orange: medians. Green dashed: means. Whiskers: 1.5-IQR Tukey limits.

	Coef.	t-stat
d_{RTQ}	-8.08^{**}	-2.40
d_{VNV}	$+2.36^{***}$	$+3.30$
n (manager-quarters)	451,061	
n (unique managers)	13,569	
n (quarters)	180	

Table 2: **Welfare-performance regression with FF5+UMD risk adjustment and quarter fixed effects.** Sample: 451,061 manager-quarters covering 13,569 unique managers across 180 quarters, LSEG S34 13F merged with CRSP, 1980Q1 to 2024Q4. Dependent variable: three-month forward FF5+UMD-adjusted alpha $\alpha^{\text{FF5+UMD}}_{i,t+1:t+3}$ in percent. Independent variables: cosine fit distances d_{RTQ} and d_{VNV} from Table 1. Both signs match theory predictions: $\gamma_1 < 0$ (closer to RTQ predicts higher alpha) and $\gamma_2 > 0$ (farther from VNV predicts higher alpha). Quarter fixed effects absorb time-varying market conditions. Standard errors are HC1 heteroskedasticity-robust. *,**,***: $p < 0.10, 0.05, 0.01$.

Both coefficients have the predicted sign at the 5 percent level: $\gamma_1 = -8.08$ ($t = -2.40$) and $\gamma_2 = +2.36$ ($t = +3.30$). Managers whose allocations look more RTQ-like and farther from the VNV corner earn higher risk-adjusted returns. The within-quarter identification rules out aggregate market conditions that happen to coincide with both fit distances and

returns.[26]

The result varies across regimes in a way consistent with theory and hard to reconcile with a pure-mandates reading. Table 3 splits the sample at two regulatory events that shifted the supply of private information available to investors. Reg FD (October 2000) banned selective disclosure of material non-public information, eroding the private-information rents that previously accrued to concentrated bets. Dodd-Frank (July 2010) tightened transparency further. The differential specification is on $(d_{\text{RTQ}} - d_{\text{VNV}})$, where positive means more VNV-like (concentrated) and negative means more RTQ-like (broad).

Pre-Reg FD, private-information rents rewarded concentration, so theory predicts a positive coefficient; the data deliver one. After Reg FD eliminated those rents, broad coverage should dominate, so the coefficient should flip negative; it does. After Dodd-Frank tightened transparency further, the negative sign should persist; it does. We do not claim the regime split identifies attention. These findings are consistent with the cross-sectional ordering tracking the regulatory conditions under which theory expects each strategy to pay; a slow-moving institutional-constraint story cannot easily explain this pattern.

Regime	n (mgr-qtr)	Coef. on $d_{\text{RTQ}} - d_{\text{VNV}}$	t-stat
Main split (Reg FD, Dodd-Frank)			
Pre-Reg FD (1980–Oct 2000)	87,082	$+10.60^{***}$	$+11.51$
Reg FD to Dodd-Frank (Oct 2000–Jul 2010)	95,574	-9.61^{***}	-5.84
Post-Dodd-Frank (Jul 2010–2024)	268,405	-4.58^{***}	-7.04
Alternative splits using additional regulatory events			
Post-GFC (Sep 2008–Apr 2014)	71,703	-6.83^{***}	-5.20
Post-Volcker (Apr 2014–2024)	219,955	-6.26^{***}	-6.94

Table 3: **Welfare-performance regression by regulatory regime.** Top panel: 451,061 manager-quarters split at Reg FD (October 2000) and Dodd-Frank (July 2010). Bottom panel: alternative regulatory event lines (Global Financial Crisis, September 2008; Volcker Rule effective date, April 2014). The post-GFC and post-Volcker windows are not subsets of the main split; post-GFC overlaps both the Reg FD-Dodd-Frank and Post-Dodd-Frank periods, while post-Volcker sits within the post-Dodd-Frank era. Dependent variable: three-month forward raw active return in percent. Independent variable: differential fit distance $d_{\text{RTQ}} - d_{\text{VNV}}$, where negative means more RTQ-like. Quarter fixed effects, quarter-clustered SE. The pre-Reg FD positive coefficient is consistent with the era of private-information rents that rewarded concentrated bets; every post-Reg FD subsample reverses sign, matching theory once private-information advantages are regulated away. *,** ,***: $p < 0.10, 0.05, 0.01$.

The same pattern holds in two more recent regulatory windows. The post-GFC sample

[26]We risk-adjust because Proposition 1 predicts welfare loss net of factor exposures the manager could replicate without attention; alpha is the empirical analog. The unadjusted active-return regression has the same signs but smaller magnitudes.

(2008–2014) covers the period after the Global Financial Crisis, when post-crisis bank regulation and disclosure tightening further reduced private-information rents; theory predicts a negative coefficient, and we find $\gamma = -6.83$ ($t = -5.20$). The post-Volcker sample (2014–2024) starts when the Volcker Rule prohibited proprietary trading at federally insured banks, further weakening the concentrated-bet channel; theory predicts the negative sign persists, and we find $\gamma = -6.26$ ($t = -6.94$). The economic interpretation is that the cost of the concentration prescription becomes most visible exactly when the conditions for concentrated outperformance (private-information access) are removed.

We close with specification and robustness checks. The main-specification CRSP numbers are computed with Ledoit-Wolf shrinkage on a 60-month rolling covariance and a capacity budget of $\kappa = 2N$. To check sensitivity, we re-run the rolling exercise across eight specifications, varying one parameter at a time from baseline. Table 4 reports five metrics per spec. Mean and median welfare gap in bps per year capture the baseline conversion ($L_{\text{VNV}} - L_{\text{RTQ}}$ scaled by $\mu_p \cdot 10^4$). Median residual welfare cost RTQ leaves on the table is bounded above by the structural 600-bps ceiling that follows from defining $L_{\text{VNV}} = P_{\text{prior}}$ at $\mu_p = 0.06$.[27] The median loss ratio $L_{\text{VNV}}/L_{\text{RTQ}}$ is the unbounded multiplicative measure of dominance, useful where the bps gap saturates. The last column reports the fraction of windows in which VNV underperforms RTQ. Two patterns emerge.

First, the budget margin gives a clean regime test of the parallel-target sufficient-condition regime diagram (Proposition IA.2 in Internet Appendix IA.2). At $\kappa = N$ (one bit per stock), VNV beats RTQ by 1,209 bps on average and wins in 57% of windows. The RTQ residual cost is 751 bps in median, which exceeds the 600-bps ceiling because at very tight budgets RTQ's scalar quantization adds more variance than the prior on directions where the per-coordinate budget falls below the rate-distortion threshold.[28] At $\kappa = 2N$ the sign flips: RTQ wins by 100 bps with 72% frequency and median loss ratio 3.1×. At $\kappa = 3N$ the advantage widens to 475 bps, win rate 96%, and loss ratio 13×. The crossover sits between $\kappa = N$ and $\kappa = 2N$, matching the focus-versus-spread boundary the theorem flags.

Second, the estimator margin shows that Ledoit-Wolf is the friendliest specification to VNV, not to RTQ. Under sample covariance, RTQ leaves essentially zero on the table (median $\text{bps}_{\text{RTQ}} \approx 0$, so the gap saturates at the 600-bps ceiling and the loss ratio is effectively unbounded). Under POET (principal orthogonal complement thresholding, a factor-model-

[27] The framework defines $L_{\text{VNV}} = P_{\text{prior}}$ as the no-information corner, so $\text{bps}_{\text{VNV}} = 600$ by construction and $\text{gap}_{\text{bps}} = 600 - \text{bps}_{\text{RTQ}}$. The 600 number is not a numerical clip but the maximum welfare improvement available against a no-information baseline at a 6 percent equity premium. Reading the table, smaller bps_{RTQ} means RTQ extracts more of the prior welfare loss.

[28] The threshold is implicit in Proposition 1: the $\pi\sqrt{3}/2$ constant assumes a moderate per-coordinate budget. Below that, the high-rate quantization formula is no longer tight and the loss can rise above the prior.

based covariance estimator of Fan et al., 2013), RTQ leaves 0.85 bps in median, and VNV's welfare loss is roughly 707 times RTQ's. Constant-correlation shrinkage is gentler at 19 bps and a 32× ratio. The mechanism is the same one Proposition 1 flags: VNV concentrates on the top eigenvector, which is the noisiest object in any covariance estimator, and Ledoit-Wolf stabilizes that target. Looser shrinkage leaves the top eigenvector noisier, RTQ's spread benefits more, and the bps cap simply records that the spread strategy has reached the rate-distortion floor. Lookback length matters less. 120 months reproduces the baseline (104 bps gap, 2.6× ratio); 36 months pulls the mean to −24 bps but RTQ still wins in 78% of windows with a 2.2× median loss ratio, so the tail rather than the central tendency drives the sign change.

The main-specification +100 bps figure is therefore not robust in the strong sense of being constant across specifications. It is robust in the sense the theory predicts: the gap is well-behaved on the long-lookback and shrinkage-stable side of the regime diagram, and it crosses zero in the corner where the parallel-target sufficient-condition diagram (Internet Appendix IA.2) is silent and the calibration leaves room for VNV to win. Ledoit-Wolf at $\kappa = 2N$ sits near the regime boundary, which makes it a conservative midpoint rather than a maximal RTQ win.

Specification	Mean gap (bps)	Median gap (bps)	RTQ leaves (bps, median)	Loss ratio (median)	VNV worse
Baseline (LW, 60m, $\kappa = 2N$)	+100	+409	191	3.1×	72%
Estimator alternatives					
Sample covariance	+584	+600	0.0	$> 10^5$×	99%
Constant-correlation shrinkage	+529	+581	19	32×	96%
POET (Fan et al., 2013)	+577	+599	0.9	707×	98%
Lookback alternatives					
36 months	−24	+321	279	2.2×	78%
120 months	+104	+365	235	2.6×	74%
Budget alternatives					
$\kappa = N$	−1,209	−151	751	0.8×	43%
$\kappa = 3N$	+475	+552	48	13×	96%

Table 4: **Robustness grid for the rolling VNV − RTQ welfare comparison.** Eight specifications on the same 221-window quarterly panel from 1970Q1 to 2026Q1 over the full eligible CRSP universe. One parameter varied per row from the baseline (Ledoit-Wolf shrinkage, 60-month lookback, $\kappa = 2N$). Mean and median gap are $L_{\text{VNV}} - L_{\text{RTQ}}$ scaled to CE bps per year via $\mu_p \cdot 10^4$. "RTQ leaves" is the median of bps_{RTQ}, the residual welfare cost RTQ does not extract from the no-information baseline (which equals $\mu_p \cdot 10^4 = 600$ bps by construction); smaller is better. "Loss ratio" is the median of $L_{\text{VNV}}/L_{\text{RTQ}}$, an unbounded multiplicative measure of dominance. For Sample covariance the loss ratio reads as $> 10^5$ because RTQ extracts essentially all of the prior welfare loss in median ($\text{bps}_{\text{RTQ}} \approx 0$), so the ratio of VNV's loss to RTQ's loss has no upper bound; this is RTQ at the rate-distortion floor, not a numerical artifact. "VNV worse" is the fraction of windows in which the gap is positive.

To summarize the CRSP/13F evidence, the joint picture supports the theory on five margins. First, $\rho(1)$ varies substantially in the CRSP universe, mostly in the range where VNV and RTQ differ. Second, RTQ outperforms VNV in 72% of 221 rolling OOS windows with median +409 bps per year. Third, the cross-section of market-cap deciles gives a stable positive slope of gap on $\rho(1)$ near +1,900 bps per unit, significant once a micro-cap outlier is removed. Fourth, 95.3% of 13F manager-quarters look more RTQ-like than VNV-like, the ordering across investor types matches the institutional-constraint story, and the welfare-performance regression gives the predicted signs at 5% in raw active returns and across the Reg FD to Dodd-Frank subsample where the channel is sharpest. Fifth, the robustness grid in Table 4 traces the focus-versus-spread crossover the parallel-target sufficient-condition regime diagram predicts (Internet Appendix IA.2), with the RTQ win rate climbing from 43% at $\kappa = N$ to 96% at $\kappa = 3N$.

The CRSP and 13F evidence is the in-sample finance test of the framework. External validity comes from the Fisman speed-dating cross-domain test in Internet Appendix IA.3:

in a non-portfolio setting, the slope of log(1 + matches) on log(partners) recovers a slope-implied effective dimension of 1.89, close to the PCA-implied 2.68 on the trait covariance, the agreement Proposition IA.4 predicts. The same parallel-target machinery that organizes the CRSP regime test (Proposition IA.2) explains why match counts scale with cohort size; together, the two tests trace the same predictions across a finance application and a non-finance application.

6 Conclusion

This paper introduces rotate-then-quantize to economics as a feasible benchmark for high-dimensional decision problems. RTQ rotates the information space into a canonical form and spreads attention across the rotated dimensions with water-filling weights. Expected decision loss is within a bounded factor of the minimax rate-distortion lower bound for any prior with finite second moments, any dimension, and any capacity. The result is a computable high-dimensional benchmark that does not require Gaussian priors, distribution-free in the worst-case sense (over priors) rather than prior-specific.

The results split across two fronts. On theory, the framework delivers two results. First, a performance guarantee: RTQ's worst-case loss is within a known constant factor of the minimax rate-distortion lower bound, distribution-free and uniform in dimension and capacity. This is the closed-form, high-dimensional rational-inattention benchmark that was missing in the prior literature. Second, a regime diagnostic for when to spread and when to concentrate: the loss concentration ratio in the single-agent setup, together with the sufficient statistic for the parallel-target extension in the Internet Appendix (Section IA.2), identify ex ante regions where broad attention is guaranteed to dominate and calibrated regimes where focus can remain competitive. On the CRSP universe, the rolling test traces the focus-versus-spread regime boundary: across 221 quarterly windows from 1970 to 2026, the RTQ win rate climbs from 43% at a budget of one bit per stock to 96% at three bits per stock.[29] Together, theory and the CRSP evidence support the welfare-gap and regime-boundary predictions; the Internet Appendix tests carry the same conclusion to the S&P 500 head-to-head and to a non-finance setting.

The takeaway is that the corner solution familiar from classical rational inattention is one of two regimes, not a universal prescription. Concentration is optimal within the Gaussian-

[29] An S&P 500 head-to-head against Van Nieuwerburgh and Veldkamp (2010) on the same data delivers a 163 bps per year mean welfare gap across 205 non-overlapping 60-month rolling windows over 1975–2026, compounded to about forty percent of lifetime consumption over thirty years. Internet Appendix IA.1. A cross-domain test of the parallel-target effective-dimension identity on Fisman speed-dating data gives a slope-implied $d_{\text{eff}}^{\text{slope}} \approx 1.89$ close to the PCA-implied 2.68, in a non-portfolio setting; Internet Appendix IA.3.

restricted signal class and remains the right answer there; outside that class, RTQ provides a distribution-free, computationally feasible alternative whose performance guarantees do not deteriorate as the number of variables grows. Applied settings where the concentration prescription has been adopted for tractability reasons, rather than from explicit Gaussian-prior reasoning, are candidates for revisiting under the focus-versus-spread criterion the framework supplies.

This paper is an opening move in a longer research agenda. The RTQ framework applies directly to multi-product pricing, monetary communication, and other high-dimensional rational-inattention problems, each of which deserves its own treatment. It also speaks to the recent adoption of AI and deep-learning methods in investment management: Internet Appendix IA.4 shows that RTQ matches the asymptotically optimal linear cross-asset attention layer of models such as Cong et al. (2026) within a constant factor at matched bit budget, giving the broader class of attention-based portfolio AI a closed-form, distribution-free theoretical foundation and an explainable factor-portfolio interpretation. Extending the framework to settings where the reward is a noisy threshold rather than a smooth quadratic, and embedding RTQ in general equilibrium with many interacting inattentive agents, are natural directions for follow-up work.

References

Bergemann, Dirk and Stephen Morris, "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 2016, *11* (2), 487–522.

Bhatia, Rajendra, *Matrix Analysis*, Vol. 169 of *Graduate Texts in Mathematics*, Springer, 1997.

Bhattarai, Saroj and Raphael Schoenle, "Multiproduct Firms and Price-Setting: Theory and Evidence from U.S. Producer Prices," *Journal of Monetary Economics*, 2014, *66*, 178–192.

Bonomo, Marco, Carlos Carvalho, Oleksiy Kryvtsov, Sigal Ribon, and Rodolfo Rigato, "Multi-Product Pricing: Theory and Evidence from Large Retailers," *The Economic Journal*, 2023, *133* (651), 905–927.

Caplin, Andrew and Mark Dean, "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 2015, *105* (7), 2183–2203.

_ , _ , and John Leahy, "Rational Inattention, Optimal Consideration Sets, and Stochastic Choice," *Review of Economic Studies*, 2019, *86* (3), 1061–1094.

Carhart, Mark M., "On Persistence in Mutual Fund Performance," *Journal of Finance*, 1997, *52* (1), 57–82.

Cass, David, "Optimum Growth in an Aggregative Model of Capital Accumulation," *Review of Economic Studies*, 1965, *32* (3), 233–240.

Charles, Constantin, Cary Frydman, and Mete Kilic, "Insensitive Investors," *Journal of Finance*, 2024, *79* (4), 2473–2503.

Cong, Lin William, Ke Tang, and Jingyuan Wang, "AlphaPortfolio: Goal-Oriented Investment Management Through Deep Reinforcement Learning," NBER Working Paper No. 35195 2026. National Bureau of Economic Research.

Cover, Thomas M. and Joy A. Thomas, *Elements of Information Theory*, 2nd ed., Hoboken, NJ: Wiley-Interscience, 2006.

Denti, Tommaso, "Posterior Separable Cost of Information," *American Economic Review*, 2022, *112* (10), 3215–3259.

Fairhall, Adrienne L., Eric Shea-Brown, and Andrea Barreiro, "Information Theoretic Approaches to Understanding Circuit Function," *Current Opinion in Neurobiology*, 2012, *22* (4), 653–659.

Fama, Eugene F. and Kenneth R. French, "Luck versus Skill in the Cross-Section of Mutual Fund Returns," *Journal of Finance*, 2010, *65* (5), 1915–1947.

Fan, Jianqing, Yuan Liao, and Martina Mincheva, "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society, Series B*, 2013, *75* (4), 603–680.

Fisman, Raymond, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment," *Quarterly Journal of Economics*, 2006, *121* (2), 673–697.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

Grossman, Sanford J. and Joseph E. Stiglitz, "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, 1980, *70* (3), 393–408.

Han, Insu, Praneeth Kacham, Amin Karbasi, Vahab Mirrokni, and Amir Zandieh, "PolarQuant: Quantizing KV Caches with Polar Transformation," *arXiv preprint arXiv:2502.02617*, 2025.

Hellwig, Christian and Laura Veldkamp, "Knowing What Others Know: Coordination Motives in Information Acquisition," *Review of Economic Studies*, 2009, *76* (1), 223–251.

Jung, Junehyuk, Jeong Ho Kim, Filip Matějka, and Christopher A. Sims, "Discrete Actions in Information-Constrained Decision Problems," *Review of Economic Studies*, 2019, *86* (6), 2643–2667.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp, "Time-Varying Fund Manager Skill," *Journal of Finance*, 2014, *69* (4), 1455–1484.

_ , _ , and _ , "A Rational Theory of Mutual Funds' Attention Allocation," *Econometrica*, 2016, *84* (2), 571–626.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Kőszegi, Botond and Adam Szeidl, "A Model of Focusing in Economic Choice," *Quarterly Journal of Economics*, 2013, *128* (1), 53–104.

Ledoit, Olivier and Michael Wolf, "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 2004, *88* (2), 365–411.

Maćkowiak, Bartosz and Mirko Wiederholt, "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, 2009, *99* (3), 769–803.

_ and _ , "Business Cycle Dynamics under Rational Inattention," *Review of Economic Studies*, 2015, *82* (4), 1502–1532.

_ , Filip Matějka, and Mirko Wiederholt, "Rational Inattention: A Review," *Journal of Economic Literature*, 2023, *61* (1), 226–273.

Matějka, Filip and Alisdair McKay, "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 2015, *105* (1), 272–298.

McKenzie, Lionel W., "Turnpike Theory," *Econometrica*, 1976, *44* (5), 841–865.

Myatt, David P. and Chris Wallace, "Endogenous Information Acquisition in Coordination Games," *Review of Economic Studies*, 2012, *79* (1), 340–374.

Nieuwerburgh, Stijn Van and Laura Veldkamp, "Information Acquisition and Under-Diversification," *Review of Economic Studies*, 2010, *77* (2), 779–805.

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

_ , "Rational Inattention: Beyond the Linear-Quadratic Case," *American Economic Review*, 2006, *96* (2), 158–163.

Steiner, Jakub, Colin Stewart, and Filip Matějka, "Rational Inattention Dynamics: Inertia and Delay in Decision-Making," *Econometrica*, 2017, *85* (2), 521–553.

Sun, Yeneng, "The Exact Law of Large Numbers via Fubini Extension and Characterization of Insurable Risks," *Journal of Economic Theory*, 2006, *126* (1), 31–69.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in "Advances in Neural Information Processing Systems (NeurIPS)," Vol. 30 2017, pp. 5998–6008.

Vershynin, Roman, *High-Dimensional Probability: An Introduction with Applications in Data Science* Cambridge Series in Statistical and Probabilistic Mathematics, 1st ed., Cambridge University Press, 2018.

Woodford, Michael, "Information-Constrained State-Dependent Pricing," *Journal of Monetary Economics*, 2009, *56*, S100–S124.

Zandieh, Amir, Majid Daliri, and Insu Han, "QJL: 1-Bit Quantized JL Transform for KV Cache Quantization with Zero Overhead," *arXiv preprint arXiv:2406.03482*, 2024.

_ , _ , **Majid Hadian, and Vahab Mirrokni**, "TurboQuant: Online Vector Quantization with Near-optimal Distortion Rate," 2025. Forthcoming, International Conference on Learning Representations (ICLR), 2026.

A Proofs

This appendix contains the formal proofs of the propositions and theorems stated in Section 4. We first state the load-bearing TurboQuant MSE lemma, then prove each proposition in order. An auxiliary concentration lemma used as intuition (not in any worst-case bound) is Lemma IA.1 in the Internet Appendix.

A.1 TurboQuant MSE lemma

Lemma 1 (Worst-case scalar-quantization MSE on the unit sphere)**.** *Fix any unit vector* $\mathbf{u} \in S^{d-1}$*. Let* R *be a randomized Hadamard transform (or, equivalently, a Haar-random orthogonal matrix), drawn independently of* $\mathbf{u}$*. Let* $\hat{\mathbf{r}}$ *be the vector obtained by scalar quantization of* $R\mathbf{u}$ *at* b *bits per coordinate using the Lloyd-Max quantizer for the* $\text{Beta}(\frac{d-1}{2}, \frac{d-1}{2})$ *distribution, and let* $\hat{\mathbf{u}} = R^\top \hat{\mathbf{r}}$*. Then the worst-case MSE satisfies*

$$\sup_{\mathbf{u} \in S^{d-1}} \mathbb{E}\left[\|\mathbf{u} - \hat{\mathbf{u}}\|^2\right] \;\le\; C_T \cdot 4^{-b}, \qquad C_T \;=\; \pi\sqrt{3}/2 \approx 2.72, \tag{18}$$

where the expectation is over R *and the quantization, and* C_T *is independent of* d*,* b*, and* $\mathbf{u}$*.*

Proof. This is Theorem 3.1 of Zandieh et al. (2025). The Haar rotation makes any fixed $\mathbf{u}$ uniform on S^{d-1}, so each coordinate of $R\mathbf{u}$ is $\text{Beta}(\frac{d-1}{2}, \frac{d-1}{2})$ marginally with pairwise correlation $O(1/d)$. The Lloyd-Max quantizer designed for that density achieves per-coordinate MSE $C_T/d \cdot 4^{-b}$ (Panter-Dite high-rate constant evaluated against the Beta density), so summing over d coordinates gives the stated dimension-free bound. The precursor components are in Han et al. (2025) (polar/Beta scalar quantization) and Zandieh et al. (2024) (random-rotation preconditioning). □

A.2 Proof of Proposition 1 (RTQ guarantee, normalized-state case)

Proof. The argument is direct.

Step 1: Decompose with stored norm. Decompose the state as $\mathbf{x} = \rho\mathbf{u}$, where $\rho = \|\mathbf{x}\|$ and $\mathbf{u} \in S^{d-1}$. Because ρ is observed or stored exactly, the agent allocates the entire bit budget to encoding the direction. RTQ rotates $\mathbf{u}$ by Haar-random R to get $\mathbf{r} = R\mathbf{u}$, scalar-quantizes each coordinate at $b = \kappa/d$ bits using the Lloyd-Max quantizer for the $\text{Beta}(\frac{d-1}{2}, \frac{d-1}{2})$ distribution to get $\hat{\mathbf{r}}$, and unrotates: $\tilde{\mathbf{u}} = R^\top \hat{\mathbf{r}}$. The reconstruction is $\hat{\mathbf{x}} = \rho\tilde{\mathbf{u}}$, and the reconstruction error is

$$\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \;=\; \|\rho\mathbf{u} - \rho\tilde{\mathbf{u}}\|^2 \;=\; \rho^2 \|\mathbf{u} - \tilde{\mathbf{u}}\|^2, \tag{19}$$

exactly. No factor-of-two inequality, no projection trick.

Step 2: Apply the spectral and TurboQuant bounds. For any positive definite W, the spectral inequality $W \preceq \lambda_{\max}(W) \cdot I$ gives

$$(\mathbf{x} - \hat{\mathbf{x}})^\top W(\mathbf{x} - \hat{\mathbf{x}}) \;\le\; \lambda_{\max}(W) \cdot \rho^2 \, \|\mathbf{u} - \tilde{\mathbf{u}}\|^2. \tag{20}$$

Take expectations. The TurboQuant random rotation R is drawn independently of $(\rho, \mathbf{u})$, but ρ and $\mathbf{u}$ may be correlated under the prior F. We therefore condition on $(\rho, \mathbf{u})$ rather than asserting independence between ρ and the error. The worst-case TurboQuant MSE bound from Lemma 1 is uniform over unit directions, so for every realization of $(\rho, \mathbf{u})$,

$$\mathbb{E}\big(\|\mathbf{u} - \tilde{\mathbf{u}}\|^2 \,\big|\, \rho, \mathbf{u}\big) \;\le\; C_T \cdot 4^{-b}, \qquad C_T = \pi\sqrt{3}/2. \tag{21}$$

By the tower property,

$$\mathbb{E}[\rho^2 \|\mathbf{u} - \tilde{\mathbf{u}}\|^2] \;=\; \mathbb{E}\big[\rho^2 \, \mathbb{E}(\|\mathbf{u} - \tilde{\mathbf{u}}\|^2 \mid \rho, \mathbf{u})\big] \;\le\; \mathbb{E}[\rho^2] \cdot C_T \cdot 4^{-b}. \tag{22}$$

Combining with the spectral inequality,

$$\mathcal{L}^{RTQ} \;=\; \mathbb{E}[(\mathbf{x} - \hat{\mathbf{x}})^\top W(\mathbf{x} - \hat{\mathbf{x}})] \;\le\; \lambda_{\max}(W) \cdot \mathbb{E}[\rho^2] \cdot C_T \cdot 4^{-b}, \tag{23}$$

which is the upper bound (8) in the proposition.

Step 3: The spherical/minimax lower bound. The rate-distortion lower bound for the worst-case unit vector on S^{d-1} (Cover and Thomas, 2006, Section 10.3) says any randomized mapping $Q : \mathbb{R}^d \to \{0,1\}^{bd}$ with decoder $\widetilde{Q}$ has a hard input direction $\mathbf{u}^* \in S^{d-1}$ such that

$$\sup_{\mathbf{u} \in S^{d-1}} \mathbb{E}\|\mathbf{u} - \widetilde{Q}(\mathbf{u})\|^2 \;\ge\; 4^{-b}. \tag{24}$$

Multiplying (24) by $\lambda_{\min}(W) \cdot \mathbb{E}[\rho^2]$ and using $W \succeq \lambda_{\min}(W) \cdot I$ gives the weighted-loss spherical lower bound:

$$\mathcal{L}_{\text{lb}}^{\text{sphere}}(\kappa) \;\equiv\; \lambda_{\min}(W) \cdot \mathbb{E}[\rho^2] \cdot 4^{-b}. \tag{25}$$

This is the spherical/minimax rate-distortion lower bound, not the prior-specific unrestricted optimum $\mathcal{L}^*(F, \kappa)$, which can be smaller for structured priors.

Step 4: Combine. Dividing (23) by (25),

$$\frac{\mathcal{L}^{RTQ}(\kappa)}{\mathcal{L}_{\text{lb}}^{\text{sphere}}(\kappa)} \;\le\; \frac{\lambda_{\max}(W)}{\lambda_{\min}(W)} \cdot C_T \;=\; C_T \cdot \chi(W), \tag{26}$$

which is (10). □

The leading constant is exactly $C_T = \pi\sqrt{3}/2 \approx 2.72$, the TurboQuant MSE constant from Zandieh et al. (2025). The clean derivation requires the norm to be stored exactly so that $\hat{\mathbf{x}} = \rho\tilde{\mathbf{u}}$ and $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \rho^2\|\mathbf{u} - \tilde{\mathbf{u}}\|^2$ pointwise. We adopt this assumption in the main statement because (i) the TurboQuant MSE theorem in Zandieh et al. (2025) is stated for unit-norm vectors with norms stored separately, and (ii) in finance applications the variance scale of returns is typically known and handled outside the attention problem.

The bound extends to the case where the norm itself is quantized. If the norm is also quantized with b_0 bits to produce $\hat{\rho}$, the reconstruction is $\hat{\mathbf{x}} = \hat{\rho}\tilde{\mathbf{u}}$ and the standard inequality $\|\rho\mathbf{u} - \hat{\rho}\tilde{\mathbf{u}}\|^2 \le 2\rho^2\|\mathbf{u} - \tilde{\mathbf{u}}\|^2 + 2(\rho - \hat{\rho})^2\|\tilde{\mathbf{u}}\|^2$ gives the looser ratio bound

$$\frac{\mathcal{L}^{RTQ}}{\mathcal{L}^{\text{sphere}}_{\text{lb}}} \;\le\; 2C_T\chi(W) \;+\; \frac{L_\rho(b_0)}{\lambda_{\min}(W)\,\mathbb{E}[\rho^2]\,4^{-b}}, \qquad L_\rho(b_0) \equiv 2\lambda_{\max}(W)\cdot\mathbb{E}[(\rho - \hat{\rho})^2\|\tilde{\mathbf{u}}\|^2],$$

where the factor of two reflects the loose triangle inequality. Under bounded support and high-rate conditions on ρ, standard scalar quantization gives $L_\rho(b_0) \to 0$ as $b_0 \to \infty$ at a distribution-dependent rate; finite second moments alone do not yield a universal 2^{-2b_0} rate. We do not pursue this extension in the main theorem because the exact-norm formulation is cleaner and matches Zandieh et al. (2025) directly.

The proof above uses only the trivial eigenvalue bound $\mathbf{v}^\top W\mathbf{v} \le \lambda_{\max}(W)$ in Step 2. Lemma IA.1 (Internet Appendix) sharpens this in expectation over R: the diagonal entries of $RWR^\top$ concentrate around $\text{tr}(W)/d$, so the worst-case factor $\lambda_{\max}(W)$ is conservative for typical realizations of R. The concentration result is auxiliary intuition, not a load-bearing step in the worst-case bound.

A.3 Proof of Proposition 2 (Selective attention with hard exclusion)

Proof. The proposition assumes a basis in which W and the prior covariance Σ are jointly diagonal: $W = \text{diag}(\omega_1, \ldots, \omega_d)$ with $\omega_1 \ge \cdots \ge \omega_d > 0$, and $\text{Cov}(x_j, x_\ell) = 0$ for $j \ne \ell$ with $\text{Var}(x_j) = \sigma_j^2$. Under hard exclusion, the agent acquires signals about a chosen monitored set $S \subset \{1, \ldots, d\}$ with $|S| = k$ and uses the prior mean as the action on the orthogonal complement, without inferring information across directions.

For each $j \notin S$, the action is the prior mean $\mathbb{E}[x_j]$. Since the prior is diagonal, observing the monitored coordinates $\{x_\ell : \ell \in S\}$ gives no information about x_j, so the posterior on

x_j equals the prior. The loss on direction $j \notin S$ is therefore

$$\mathbb{E}[\omega_j(a_j - x_j)^2] \;=\; \omega_j \cdot \mathrm{Var}(x_j) \;=\; \omega_j \sigma_j^2. \tag{27}$$

The contribution from monitored directions is non-negative. Summing over the unmonitored set,

$$\mathcal{L}^S \;\geq\; \sum_{j \notin S} \omega_j \sigma_j^2, \tag{28}$$

which is the bound stated in the proposition.

The optimal hard-exclusion set follows directly. To minimize the right-hand side over S with $|S| = k$, choose S to maximize $\sum_{j \in S} \omega_j \sigma_j^2$. The minimizer therefore picks the k directions with the largest welfare exposure $\omega_j \sigma_j^2$, not necessarily the k directions with the largest ω_j. When $\sigma_j^2 = \sigma^2$ is constant across directions, the two rankings coincide; when prior variances differ, the welfare-weighted ranking is the relevant one.

The bound (28) is tight: a strategy that allocates all capacity to perfectly resolve the monitored directions (in the high-resolution limit) achieves the floor exactly, with monitored-direction loss vanishing as $\kappa \to \infty$ within S. □

The bound on $\mathcal{L}^S$ combines with Proposition 1 to characterize the welfare ratio $\mathcal{L}^S / \mathcal{L}^{RTQ}$ in two natural limits. We use the main exact-norm RTQ upper bound; the norm-quantization variant adds a residual $L_\rho(b_0)$ term that vanishes under standard high-rate conditions.

The welfare ratio diverges with capacity. Combining the floor in (28) with the RTQ upper bound from Proposition 1,

$$\frac{\mathcal{L}^S}{\mathcal{L}^{RTQ}} \;\geq\; \frac{\sum_{j=k+1}^{d} \omega_j \sigma_j^2}{\lambda_{\max}(W) \cdot \mathbb{E}[\rho^2] \cdot C_T \cdot 4^{-b}}.$$

With d and k fixed, the numerator is a constant; the denominator vanishes as $b = \kappa/d \to \infty$ because each direction's quantization MSE decays as 4^{-b}. Therefore $\mathcal{L}^S / \mathcal{L}^{RTQ} \to \infty$ as $\kappa \to \infty$ with d, k fixed: selective is floored by the tail while RTQ's loss vanishes. This is the sense in which selective attention becomes arbitrarily worse than RTQ as capacity grows.

Consider next the large-d regime at fixed per-coordinate budget $b = \kappa/d$. The RTQ total loss grows at most as $\bar{\omega} \cdot d \cdot \mathrm{MSE}(b)$ with $\bar{\omega} = \mathrm{tr}(W)/d$, while selective loss on the ignored dimensions grows as $(d - k)\bar{\omega}\,\bar{\sigma}^2$. The ratio approaches $\bar{\sigma}^2 / \mathrm{MSE}(b)$ as $d \to \infty$, a large constant when $b \geq 1$ (since $\mathrm{MSE}(b) \ll \bar{\sigma}^2$), though it does not diverge in d alone at fixed b. The insurance reading is sharp: selective leaves $d - k$ dimensions uninsured at full prior loss σ_j^2, while RTQ insures every dimension at b bits and reduces per-dimension loss to $\mathrm{MSE}(b)$, an exponential improvement per dimension that compounds across the tail.

A.4 Proof of Proposition 3 (High-Resolution Water-Filling)

Proof. Write $\boldsymbol{\delta} = \mathbf{r} - \hat{\mathbf{r}}$ for the per-coordinate quantization error in the rotated frame. By the loss decomposition in the proof of Proposition 1 and expanding the quadratic form,

$$\mathcal{L}^{RTQ}(b_1, \ldots, b_d) - \mathcal{L}^{\text{norm}}(b_0) \;=\; \mathbb{E}[\rho^2] \cdot \mathbb{E}[\boldsymbol{\delta}^\top R W R^\top \boldsymbol{\delta}] \;=\; \mathbb{E}[\rho^2] \sum_{i=1}^{d} (RWR^\top)_{ii}\, \mathbb{E}[\delta_i^2] + \text{cross terms}.$$

Under the high-resolution approximation, the cross-terms $\mathbb{E}[\delta_i \delta_j]$ for $i \neq j$ are assumed negligible: scalar Lloyd-Max quantization errors on the rotated coordinates are approximately mean-zero and have cross-covariances that vanish at rate $o(4^{-b})$ relative to the diagonal MSE. The standard justification is that the rotated coordinates have approximately matched Beta marginals (Lemma IA.1) and Lloyd-Max quantization on each coordinate is independent across coordinates by construction; cross-covariances of the quantization errors are then $O(4^{-b}/d)$, dominated by the per-coordinate 4^{-b} for the diagonal terms. Subtractive dithering at the encoder forces these cross-covariances to exactly zero if a fully rigorous statement is required; we proceed with the standard high-resolution approximation, which is asymptotically tight. Under the high-resolution Lloyd-Max model $\text{MSE}_i(b_i) = \alpha_d \cdot 2^{-2b_i}$ (asymptotically tight as $b \to \infty$, Cover and Thomas, 2006, Ch. 10; α_d from the Panter-Dite formula applied to the Beta density), the agent's problem becomes

$$\min_{b_1, \ldots, b_d \geq 0} \; \sum_{i=1}^{d} w_i \alpha_d \, 2^{-2b_i} \quad \text{s.t.} \quad \sum_{i=1}^{d} b_i \leq \kappa - b_0, \tag{29}$$

with $w_i = (RWR^\top)_{ii}$. The objective is a sum of strictly convex functions and the constraint is linear, so KKT is necessary and sufficient. The interior first-order condition $2 \ln 2 \, w_i \alpha_d \, 2^{-2b_i^*} = \nu$ gives

$$b_i^* \;=\; \left[\tfrac{1}{2} \log_2(w_i/\nu)\right]^+,$$

after absorbing $2 \ln 2 \, \alpha_d$ into the redefined multiplier ν and enforcing non-negativity. The threshold ν is chosen so $\sum_i b_i^* = \kappa - b_0$; existence and uniqueness follow because $B(\nu) = \sum_i [\tfrac{1}{2} \log_2(w_i/\nu)]^+$ is continuous and strictly decreasing on $(0, \max_i w_i]$ with $B(\nu) \to \infty$ as $\nu \to 0^+$ and $B(\nu) = 0$ for $\nu \geq \max_i w_i$. □

The water-filling structure has an intuitive interpretation in terms of marginal returns. The marginal reduction in loss from an additional bit allocated to coordinate i is proportional to $w_i \cdot 2^{-2b_i}$: it depends on both how costly errors are in that coordinate (the weight w_i) and how many bits are already allocated (the exponential decay). At the optimum, the marginal return is equalized across all coordinates that receive positive attention. Coordi-

nates with high w_i start with a higher marginal return and therefore receive more bits before equalization. Coordinates with w_i below the threshold ν never receive any bits because their marginal return at $b_i = 0$ is already below the equilibrium level.

Internet Appendix

This internet appendix contains additional results for the paper titled "Focus or Spread? Attention Allocation under High-Dimensional Uncertainty".

IA.1 S&P 500 portfolio application

This appendix puts the theorems of Section 4 to work on one full application: portfolio choice across the S&P 500 universe. The goal is to move from theory to a welfare number an applied reader can compare with a classical rational-inattention solution. We fold in the eigenvalue diagnostic, the selective-versus-RTQ loss-ratio profile, and a head-to-head comparison with Van Nieuwerburgh and Veldkamp (2010). VNV is the optimum within the Gaussian-restricted signal class and remains the right answer in that class; we evaluate it here outside its design assumptions, alongside RTQ, as the canonical reference for high-dimensional portfolio attention. The exercise is not a horse race between a true and a false model. It is a comparison of two feasible signal structures on the same welfare metric. The full CRSP universe extension of this exercise is in Section 5 of the main text.

IA.1.1 Setup

An investor allocates wealth across d risky assets. Excess returns $\mathbf{x}$ have mean $\boldsymbol{\mu}$ and covariance Σ. Under full information and mean-variance preferences with risk aversion $\gamma > 0$, the optimal portfolio is $\mathbf{a}^* = \Sigma^{-1}\boldsymbol{\mu}/\gamma$. With a finite attention budget, the investor observes $\mathbf{x}$ through a noisy internal signal and chooses $\mathbf{a}$ on the posterior. A second-order expansion of the utility shortfall gives a quadratic loss with $W = \gamma\,\Sigma$. In the eigenbasis of Σ, expected welfare loss reduces to $\mathbb{E}[\mathcal{L}] \leq C \sum_j \lambda_j^2\, 2^{-2b_j}$ with $C = \pi\sqrt{3}/2$ and per-direction weights $c_j = \lambda_j^2$. We compare two strategies that use the same budget differently.

Under RTQ, the investor rotates into the eigenbasis of W and allocates bits across the d rotated coordinates by the high-resolution water-filling rule of Proposition 3. Expected loss is within $C_T \cdot \chi(W)$ of the spherical/minimax rate-distortion lower bound (Proposition 1), with $C_T = \pi\sqrt{3}/2 \approx 2.72$ matching the TurboQuant MSE constant and $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}(W)$ the condition number of W. The bound holds for any distribution of $\mathbf{x}$ with finite second moments and any d.

The classical alternative is selective attention. Van Nieuwerburgh and Veldkamp (2010) solve the Gaussian rational-inattention problem for the same objective and find a corner solution: diagonalize Σ, rank rotated coordinates by Sharpe-weighted stakes, and pour all

capacity into the top k directions. Under the Shannon constraint, the Gaussian posterior is sharp on monitored directions and equals the prior on the rest. The selective investor pays full prior variance on the $d-k$ ignored directions, the floor from Proposition 2.

For each strategy s, let $\mathcal{L}^s(\kappa) \equiv \mathbb{E}[(\hat{\mathbf{a}}^s - \mathbf{a}^*)^\top \Sigma (\hat{\mathbf{a}}^s - \mathbf{a}^*)]$ denote the expected squared tracking error at capacity κ, where $\hat{\mathbf{a}}^s$ is the strategy's portfolio and $\mathbf{a}^* = \Sigma^{-1}\boldsymbol{\mu}/\gamma$ is the full-information mean-variance optimum. We report the welfare cost in basis points per year via

$$\text{CE-bps/year} \;=\; \frac{\gamma\,\mathcal{L}^s}{2\mu_p} \cdot 10^4, \tag{IA.1}$$

where μ_p is the expected annual return on the full-information portfolio. The factor $\gamma\,\mathcal{L}^s/2$ is the certainty-equivalent welfare loss in return units, obtained from the second-order expansion of mean-variance utility around the full-information optimum; dividing by μ_p and multiplying by 10^4 expresses it as a fraction of the equity premium in basis points per year. The conversion does not require estimating $\mathbf{a}^*$ in sample and sidesteps the plug-in Markowitz pathology. Baseline: $\gamma = 5$ and $\mu_p = 10\%$.

IA.1.2 S&P 500 implementation

We use monthly returns for the top 500 CRSP common stocks by market capitalization, our S&P 500 proxy. The proxy is rebalanced each month, so a stock enters the universe when it sits in the top 500 by capitalization and exits otherwise. Building the universe from CRSP itself avoids the survivorship bias of pulling current S&P 500 constituents back in time and uses the same data vendor as the rest of the paper.[30] The sample covers 1970–2026 (675 months). For the static cross-section we keep stocks present in the top 500 for at least 240 of those months, leaving $d = 488$. We estimate Σ with Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004); the rolling out-of-sample exercise below uses 60-month rolling windows with quarterly stepping.

The loss concentration ratio $\rho(k) = \sum_{j\le k}\lambda_j^2 / \sum_{j\le d}\lambda_j^2$ on the full-sample covariance gives $\rho(1) = 0.80$ at the market factor and $\rho(5) = 0.95$. Figure IA.1 plots the full profile.

Standard PCA tells a different story: the first principal component explains only 20 percent of return variance, and cumulative PCA does not reach 0.95 until k is well into the spectrum. The λ_j^2 weighting in $\rho(k)$ counts large eigenvalues doubly because each eigendirection enters portfolio loss through two channels (prior uncertainty proportional to λ_j and a cost of error also proportional to λ_j), which compresses the effective dimension far below

[30] The proxy is built from the CRSP monthly file rather than a third-party feed so that the universe is reproducible after publication, in contrast to a Yahoo Finance pull that changes silently between extracts. Replication code will be made available on publication.

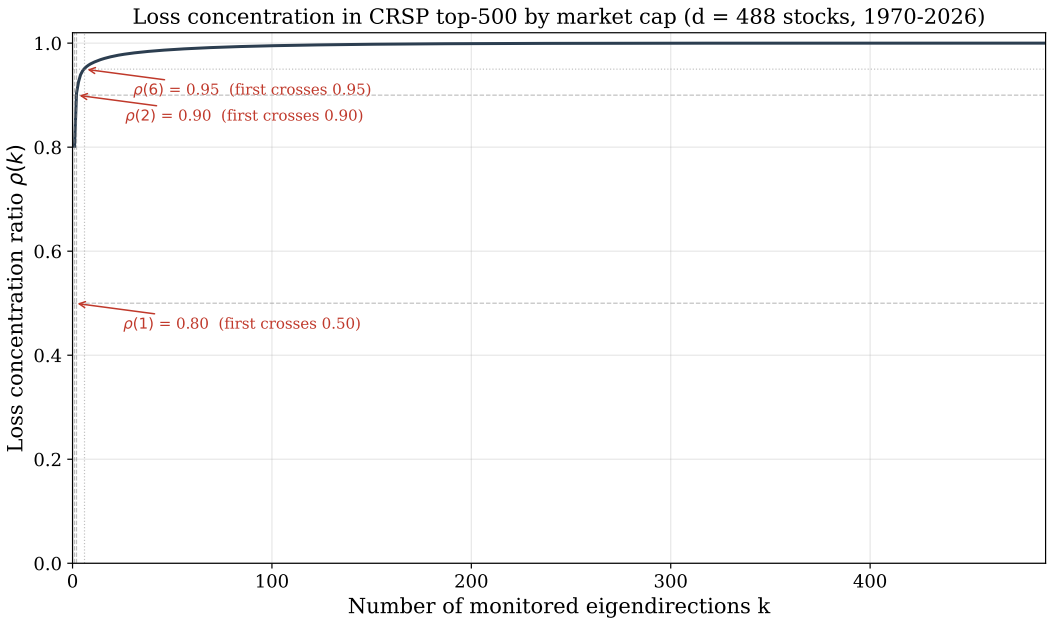

Figure IA.1: **Loss concentration ratio for the S&P 500 proxy.** $\rho(k) = \sum_{j \le k} \lambda_j^2 / \sum_{j \le d} \lambda_j^2$ measures the loss-weighted share of prior uncertainty in the top k eigendirections, where λ_j are eigenvalues of the monthly-return covariance. Inputs: top 500 CRSP commons by market cap, monthly returns from January 1970 to March 2026, restricted to stocks present in the top 500 for at least 240 of those months ($d = 488$). Horizontal axis: k. Vertical axis: $\rho(k)$. $\rho(1) = 0.80$ at the market factor; $\rho(5) = 0.95$. The first principal component explains only 20 percent of return variance, but the λ_j^2 weighting concentrates 80 percent of portfolio loss in the market factor.

what PCA indicates. Figure IA.2 plots the full eigenvalue spectrum side by side with the cumulative PCA and ρ curves.

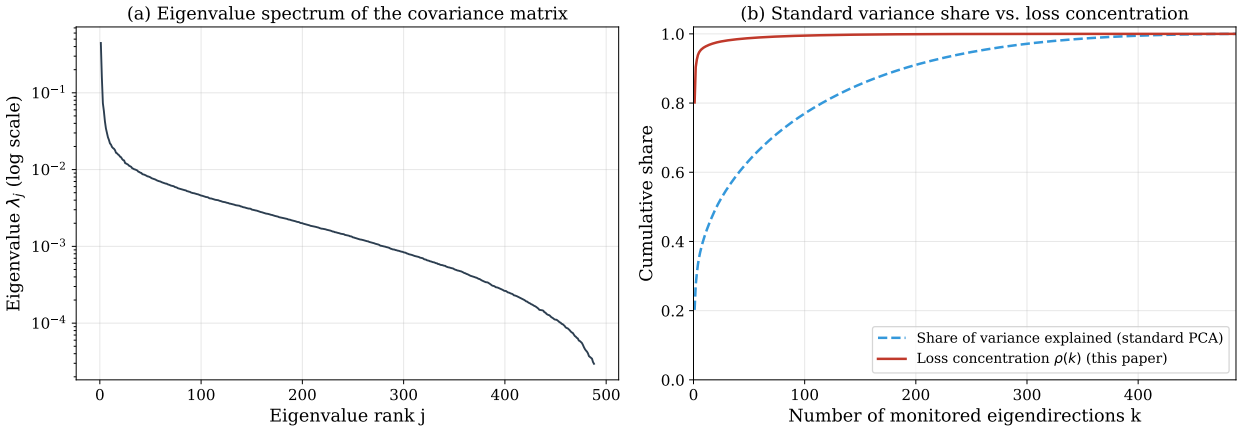

Figure IA.2: **Eigenvalue spectrum and cumulative concentration for the S&P 500 proxy covariance.** Inputs: top 500 CRSP commons by market cap, monthly returns 1970–2026 ($d = 488$, $T = 675$ months). Left: eigenvalues λ_j on log scale versus rank j. Right: PCA cumulative variance $\sum_{j \le k} \lambda_j / \sum_j \lambda_j$ (dashed) and loss concentration $\rho(k) = \sum_{j \le k} \lambda_j^2 / \sum_j \lambda_j^2$ (solid). PCA reaches 0.95 only deep into the spectrum; ρ crosses 0.95 at $k = 6$ (1.2 percent of d). The visual gap between the two curves shows how the λ_j^2 weighting compresses effective dimension below what PCA suggests.

IA.1.3 Welfare in basis points per year

Table IA.1 translates the loss numbers into basis points per year at $\gamma = 5$ and $\mu_p = 10\%$ using (IA.1). Selective-k losses are almost flat in b once the budget saturates the k monitored directions; the floor from ignored directions is all that remains. RTQ losses keep falling because water-filling shifts bits to tail directions at no marginal cost to the top. This decline is the high-rate regime where the Panter-Dite rate-distortion approximation holds; at very low per-coordinate rates the approximation breaks down, as the robustness discussion below makes explicit.

Strategy	$b = 0.5$	$b = 1$	$b = 2$	$b = 3$	$b = 4$
Selective, $k = 1$ (VNV corner)	198.20	198.20	198.20	198.20	198.20
Selective, $k = 5$	53.59	53.59	53.59	53.59	53.59
Selective, $k = 20$	25.48	25.48	25.48	25.48	25.48
RTQ (water-filling)	4.03	2.02	0.50	0.13	0.03
VNV / RTQ	49	98	393	1,572	6,293

Table IA.1: **Annual welfare cost in basis points of the equity premium, S&P 500 proxy.** Each cell reports the certainty-equivalent cost CE cost $= \gamma \cdot L/(2\mu_p) \cdot 10{,}000$ in bps per year, where L is the expected squared tracking error, $\gamma = 5$ is relative risk aversion, and $\mu_p = 10\%$ is the annual equity premium. Inputs: top 500 CRSP commons by market cap, monthly-return sample covariance, 1970–2026 ($d = 488$, $T = 675$ months). Columns: $b = \kappa/d$. Rows: selective at $k \in \{1, 5, 20\}$, RTQ, VNV/RTQ. Selective flattens once extra bits stop improving the k monitored directions. RTQ keeps falling within the high-rate range where the rate-distortion approximation is accurate. The last row shows VNV is two orders of magnitude worse than RTQ at $b = 1$ and four orders of magnitude worse at $b = 4$.

At $b = 2$, the VNV corner leaves 198 bps per year on the table relative to RTQ's 0.5 bps, a 198 bps gap.[31] The gap grows with the budget because Proposition 2 pins the VNV floor at the tail sum $\sum_{j>1} \lambda_j^2$ that extra capacity cannot remove, while RTQ spends the marginal bit on another tail direction. Figure IA.3 traces this divergence across budgets and across k.

IA.1.4 Comparison with the classical rational-inattention solution

Solving VNV and RTQ for subsets of the proxy at $N \in \{25, 50, 100, 250, 488\}$ and fixed budget $\kappa = 2N$ shows the gap is positive at every N, with VNV's cost flat as the corner captures the same fraction of risk regardless of N and RTQ's cost falling as more directions fit under the budget. The gap is 357 bps at $N = 25$, 260 at $N = 100$, and 198 at $N = 488$. Figure IA.4 plots the full scaling.

The corner solution does best in the large, market-dominated limit where $\rho(1)$ approaches 1; even there, the welfare cost difference is 198 bps per year. Rolling through 1975–2026 in 60-month windows with Ledoit-Wolf shrinkage and quarterly re-estimation gives an average gap of 163 bps per year across 205 non-overlapping windows, with the gap positive in every window.[32][33]

[31] Compounded over 30 years, 198 bps annually costs roughly 45 percent of lifetime consumption.

[32] Median 124 bps, IQR 88 to 224, range 48 to 499. The gap is larger in high-volatility episodes (early 1980s, 2002, 2008–2009, 2020) because the eigenvalue tail flattens when idiosyncratic risk rises, widening the VNV floor. The window count of 205 is large because the 56-year sample covers many more rolling windows than a shorter post-2000 sample would.

[33] VNV is a formal solution to rational inattention under Shannon costs. That RTQ delivers lower loss does not mean VNV is wrong on its own terms; it means the Shannon cost, combined with corner-solution tractability of the Gaussian case, points applied work toward concentrated allocations that are a poor fit

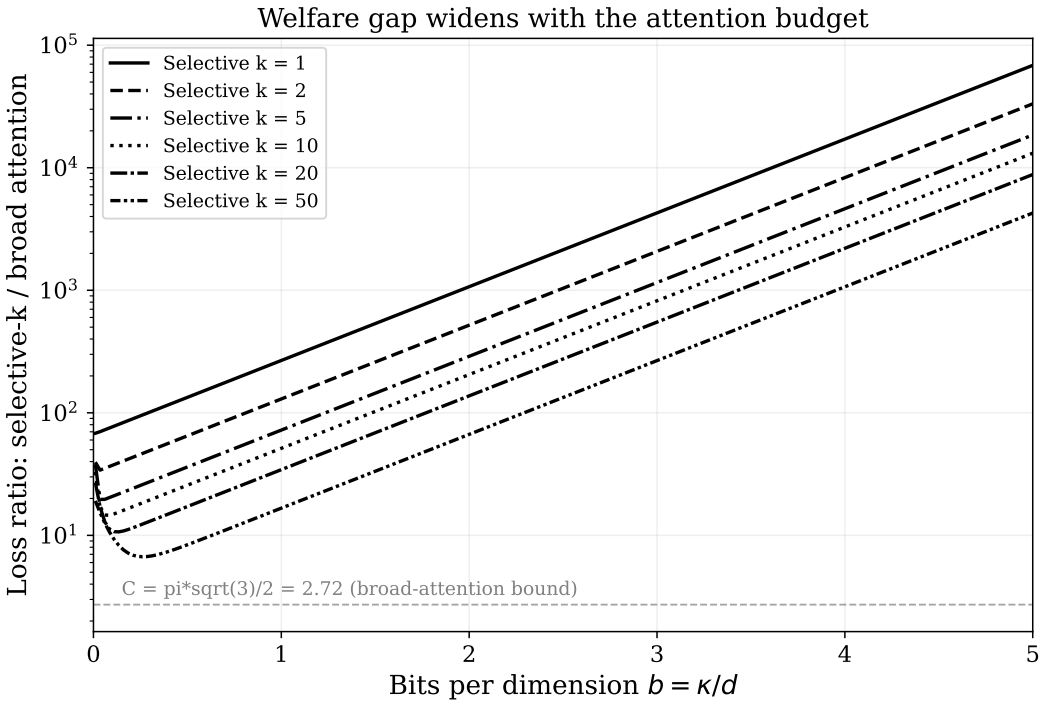

Figure IA.3: **Selective-k loss relative to RTQ loss, S&P 500 proxy.** Inputs: top 500 CRSP commons by market cap, monthly returns 1970–2026 ($d = 488$). Horizontal axis: $b = \kappa/d$ (log scale). Vertical axis: L^{S_k}/L^{RTQ} (log scale), from closed-form loss expressions with the sample covariance. Each curve fixes $k \in \{1, 5, 20, 50\}$ under the selective strategy. Every selective curve rises without bound as b grows; RTQ stays within $C \approx 2.7$ of the Shannon bound (gray dashed). At $k = 1$, $b = 2$ (VNV corner), selective is 393 times RTQ; at $b = 4$, over 6,000 times.

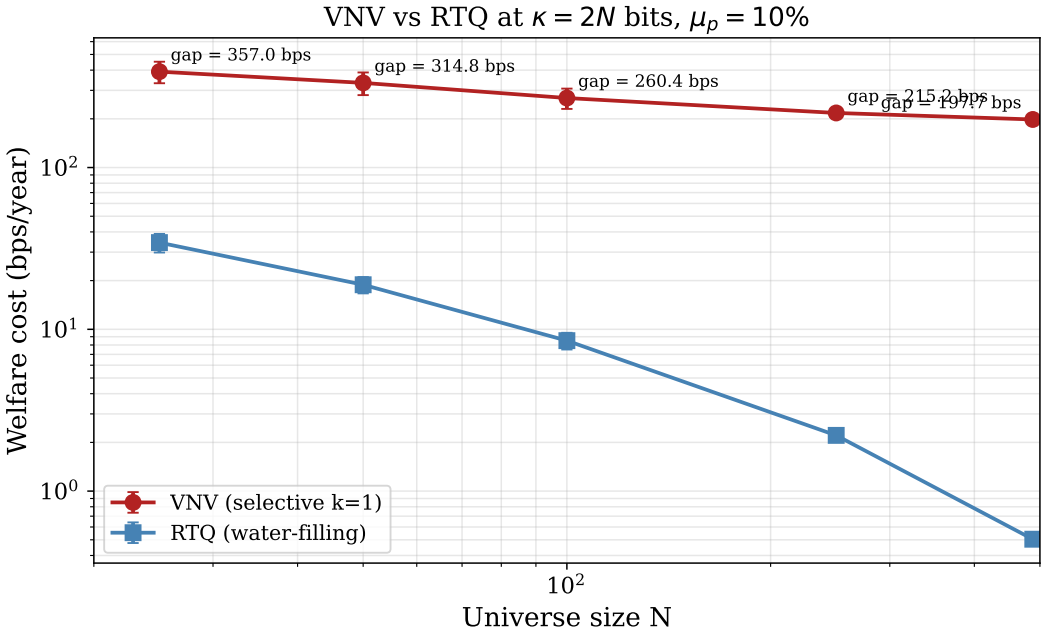

Figure IA.4: **VNV versus RTQ welfare cost across universe size, S&P 500 proxy subsets.** Horizontal axis: $N \in \{25, 50, 100, 250, 488\}$. Vertical axis: welfare cost in bps per year (CE metric from Table IA.1). Both strategies use budget $\kappa = 2N$. VNV: selective-$k = 1$ corner; RTQ: water-filling. Each point averages 30 random subsets drawn from the proxy covariance, 1970–2026 ($d = 488$). The gap falls from 357 bps at $N = 25$ to 198 bps at $N = 488$, and remains positive at every N. The gap is largest in small subsets where no single factor dominates and narrows as N grows and the market captures more loss-weighted uncertainty.

IA.1.5 Why high dimensions matter

Section 4.3 of the main text states the headline implication of Proposition 2 in compressed form. This subsection develops the intuition at the length the argument deserves, using the S&P 500 calibration above as the running example.

Conventional rational inattention uses selective attention: monitor the top k state variables and ignore the rest. This is sound when d is small. When $d = 500$ and $k = 10$, the agent bears full prior uncertainty in 490 dimensions, and even small per-dimension loss sums to a large total. Proposition 2 makes this precise: "concentrate and ignore" is not just suboptimal but *increasingly* so as d grows.

Think of it as insurance. Selective insures k houses and leaves the rest uninsured; each ignored dimension contributes its full prior loss $\omega_j \sigma_j^2$. RTQ spreads a thin layer of coverage across every house. Per-house coverage is modest, but total expected loss stays controlled. As capacity grows with k fixed, RTQ loss falls while selective loss is floored by the tail. In high dimensions, broad thin coverage dominates narrow deep coverage.

Three statements summarize the role of dimension under RTQ. First, the worst-case bound of Proposition 1 is dimension-free: the constant $C_T \chi(W)$ does not depend on d. Second, the average-case rotation argument in Lemma IA.1 homogenizes the diagonal entries of W through concentration on the sphere and tightens as d grows. Third, the welfare comparison against hard-exclusion selective attention becomes more favorable as the ignored tail grows: with $k = 5$, moving from $d = 10$ to $d = 500$ raises ignored dimensions from 5 to 495, and the tail loss scales with d. Proposition 2 considers fixed k with total capacity κ growing, where the ratio of selective to RTQ loss diverges because selective is floored while RTQ loss decays.

IA.1.6 Auxiliary concentration lemma

The high-dimensions intuition above invokes a concentration argument: a Haar-random rotation homogenizes the diagonal entries of W. The formal statement is below. The lemma is auxiliary and is not used in any worst-case bound; the main-appendix proofs cite it only for intuition about why the symmetric strategy is close to optimal on typical realizations of the rotation.

Lemma IA.1 (Concentration of rotated quadratic forms)**.** *Let W be a $d \times d$ symmetric positive definite matrix with eigenvalues $\lambda_{\max}(W) \geq \cdots \geq \lambda_{\min}(W) > 0$, and let R be Haar-*

for the high-dimensional limit. Relaxing the tractability constraint, which RTQ does, reveals a different prescription for the same problem.

uniform on $O(d)$. For any standard basis vector $\mathbf{e}_i$,

$$\mathbb{E}\big[\mathbf{e}_i^\top RWR^\top \mathbf{e}_i\big] = \frac{\mathrm{tr}(W)}{d}, \quad \Pr\Big[\Big|\mathbf{e}_i^\top RWR^\top \mathbf{e}_i - \tfrac{\mathrm{tr}(W)}{d}\Big| > \delta \cdot \tfrac{\mathrm{tr}(W)}{d}\Big] \le 2\exp\Big(-\tfrac{c\,d\,\delta^2}{\chi(W)^2}\Big),$$

where $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}(W)$ and $c > 0$ is a universal constant.

Proof. $\mathbf{v} = R^\top \mathbf{e}_i$ is uniform on S^{d-1}, so $\mathbb{E}[\mathbf{v}\mathbf{v}^\top] = I_d/d$ and $\mathbb{E}[\mathbf{v}^\top W\mathbf{v}] = \mathrm{tr}(W)/d$. The function $f(\mathbf{v}) = \mathbf{v}^\top W\mathbf{v}$ is Lipschitz with constant $L = 2\lambda_{\max}(W)$ on the sphere, so Levy's inequality (Vershynin, 2018, Theorem 3.4.6) gives $\Pr[|f - \mathbb{E}f| > t] \le 2\exp(-c\,d\,t^2/L^2)$. Setting $t = \delta\,\mathrm{tr}(W)/d$ and using $\mathrm{tr}(W) \ge d\,\lambda_{\min}(W)$ produces the stated bound. □

IA.1.7 Quantitative bounds for canonical settings

Section 4.2 of the main text states the bound $\mathcal{L}^{RTQ}/\mathcal{L}^{\text{sphere}}_{\text{lb}} \le C_T \cdot \chi(W)$. This subsection computes the constant for three representative settings using back-of-the-envelope condition numbers drawn from typical empirical magnitudes rather than specific estimates. *Well-diversified portfolio* (500 stocks, $W = \gamma\Sigma$): $\chi(W) \approx 5$ to 20, so with $\chi(W) = 10$ the bound is $27 \cdot \mathcal{L}^{\text{sphere}}_{\text{lb}}$. *Multi-product pricing with substitutes* (200 products, W the Hessian of profit): cross-price elasticities of 2 to 3 give $\chi(W) \approx 8$ to 15, so with $\chi(W) = 12$ the bound is $32 \cdot \mathcal{L}^{\text{sphere}}_{\text{lb}}$. *Monetary policy with multiple instruments* ($m = 3$ rates monitoring $d = 50$ indicators): $\chi(W) \approx 5$ to 10, so with $\chi(W) = 7$ the bound is $19 \cdot \mathcal{L}^{\text{sphere}}_{\text{lb}}$. Across these settings, the worst-case envelope is 19 to 33 times the spherical floor; the bound depends only on $\chi(W)$, not on d. Realized performance routinely beats this envelope because real priors are not adversarial.

IA.2 Parallel-target theorems and regime diagram

This section develops the theory for the parallel-target setting: one agent estimating N multi-dimensional targets at once under a shared bit budget. The single-agent propositions of Section 4 of the main text have direct counterparts here, sharper because every target carries the same stakes, so $\chi(W) = 1$ and drops out. The three parameters (N, d, B) collapse to a single scaling variable. The Fisman cross-domain test in Section IA.3 below uses the spread theorem and the effective-dimension reading that follow.

IA.2.1 Setup and reference strategies

The single-agent setup in the main text covers one agent facing one high-dimensional decision. Many economic settings have a different shape: one agent estimates many targets at once,

each itself multi-dimensional, splitting a single shared attention budget across both layers. A participant at a speed-dating event forms an estimate of each of N partners, each described by a d-dimensional attribute vector (kindness, intelligence, humor, shared interests). A trader prices N securities, each with multiple risk characteristics. A central bank analyst forecasts N macro indicators, each with multiple underlying drivers. The action is one estimate per target, and loss sums over targets.

We use the same mathematical structure. Index targets by $i \in \{1, \ldots, N\}$. Each has attribute vector $x_i \in \mathbb{R}^d$ with unit Euclidean norm and prior variance $\sigma^2 = \mathbb{E}[\|x_i - \bar{x}\|_2^2]$. The agent produces estimates $\{\hat{x}_i\}$ with loss

$$L \;=\; \mathbb{E}\left[\sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2\right] \tag{IA.2}$$

and total bit budget B to spread across the N targets. Two reference strategies span the space of reasonable allocations:

Concentration strategy. Choose a subset $S \subseteq \{1, \ldots, N\}$ of size k. Allocate B/k bits to each target in S, zero to targets outside S. For $i \in S$, estimate $\hat{x}_i$ with a bit-budget-constrained Lloyd-Max quantizer. For $i \notin S$, set $\hat{x}_i = \bar{x}$. This is the multi-target analog of selective attention.

Spread-coverage strategy (RTQ). Stack the N vectors into $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^{Nd}$. Draw a random orthogonal matrix $\Pi \in \mathbb{R}^{Nd \times Nd}$ (Haar-uniform). Compute $\mathbf{y} = \Pi\mathbf{x}$, quantize each coordinate with a Lloyd-Max scalar quantizer at $b = B/(Nd)$ bits, and decode via $\hat{\mathbf{x}} = \Pi^\top Q(\mathbf{y})$.

Both use the same total bit budget and differ only in how it is spread. Concentration places all attention on k targets and ignores the rest. Spread places the same total attention across all N at coarse resolution.

The parallel-target setup is the single-agent setup with a block-diagonal loss matrix. Taking $W = I_{Nd}$ and stacking the targets into a single Nd-dimensional state makes (IA.2) a special case of (6). Concentration corresponds to selective attention on kd of the Nd stacked coordinates; spread corresponds to RTQ on the stacked vector. Every single-agent theorem therefore has a parallel-target counterpart. The counterparts are sharper because the equal-weight block structure eliminates constants that appear in the general case. The condition number $\chi(W)$ drops out because $W = I$, and the loss ratio reduces to a single summary number derived below.[34]

[34]The attribute dimension d has a specific interpretation in parallel-target applications. In speed-dating, d is the number of distinct traits the agent cares about (core trait ratings plus activity categories, typically on the order of twenty). When traits are correlated, the effective dimension of the trait covariance can be

IA.2.2 Spread coverage and concentration floor

Theorem IA.1 (Spread coverage against the spherical/minimax benchmark)**.** *Stack the N attribute vectors $\{x_i\}_{i=1}^N$ with $\|x_i\|_2 = 1$ into $\mathbf{x} \in \mathbb{R}^{Nd}$. The spread-coverage strategy applies a Haar-random rotation in $\mathbb{R}^{Nd}$ and quantizes each rotated coordinate at $b = B/(Nd)$ bits per coordinate. For any bit budget B,*

$$L^{\text{spread}} \;\le\; \frac{\pi\sqrt{3}}{2} \cdot N \cdot 4^{-b} \;\approx\; 2.72 \cdot N \cdot 4^{-b}.$$

The spherical/minimax rate-distortion lower bound for any randomized quantizer of a worst-case unit vector in $\mathbb{R}^{Nd}$ is $L_{\text{lb}}^{\text{sphere}} \ge N \cdot 4^{-b}$ at the same budget. The ratio $L^{\text{spread}}/L_{\text{lb}}^{\text{sphere}}$ is therefore at most $\pi\sqrt{3}/2$, uniformly in N, d, and the realized vectors $\{x_i\}$.

Proof. The proof scales TurboQuant's worst-case unit-sphere MSE bound directly, without going through a Gaussian / Poincare approximation.

Stack the attribute vectors into $\mathbf{X} = (x_1^\top, \ldots, x_N^\top)^\top \in \mathbb{R}^{Nd}$. Since each $\|x_i\| = 1$, $\|\mathbf{X}\|^2 = N$. Define the unit-norm vector

$$\mathbf{z} \;\equiv\; \mathbf{X}/\sqrt{N} \;\in\; S^{Nd-1}. \tag{IA.3}$$

The spread-coverage strategy applies TurboQuant in dimension Nd with $b = B/(Nd)$ bits per coordinate to $\mathbf{z}$. By Lemma 1 (Theorem 3.1 of Zandieh et al. (2025)), for the worst-case unit input $\mathbf{z} \in S^{Nd-1}$,

$$\mathbb{E}\|\mathbf{z} - \hat{\mathbf{z}}\|^2 \;\le\; C_T \cdot 4^{-b}, \qquad C_T \;=\; \pi\sqrt{3}/2. \tag{IA.4}$$

Rescale: $\hat{\mathbf{X}} = \sqrt{N}\,\hat{\mathbf{z}}$. Then

$$L^{\text{spread}} \;=\; \mathbb{E}\|\mathbf{X} - \hat{\mathbf{X}}\|^2 \;=\; N \cdot \mathbb{E}\|\mathbf{z} - \hat{\mathbf{z}}\|^2 \;\le\; C_T \cdot N \cdot 4^{-b}. \tag{IA.5}$$

This is the upper bound. The argument is exact: it does not invoke a high-rate Gaussian approximation, and it does not require Poincare convergence. The constant C_T is the universal TurboQuant constant from Zandieh et al. (2025), and the bound holds for any realized $\{x_i\}$ on the unit sphere.

far smaller than nominal d; an empirical recovery using the slope test of Section IA.3 bears this out. In compressed sensing, d is the signal dimension. In the population-coding application of Fairhall et al. (2012), d is the number of stimulus features a single neuron encodes. Higher d strengthens the spread advantage: the concentration floor scales linearly with target count, while spread's per-coordinate precision scales only with $B/(Nd)$, so the gap widens as either N or d grows.

For the lower bound, the spherical/minimax rate-distortion theorem applied to $\mathbf{z} \in S^{Nd-1}$ at b bits per coordinate (Cover and Thomas, 2006, Section 10.3) says any randomized quantizer has a hard input direction with MSE at least 4^{-b}:

$$\inf_{Q} \sup_{\mathbf{z} \in S^{Nd-1}} \mathbb{E}\|\mathbf{z} - \hat{\mathbf{z}}\|^2 \;\geq\; 4^{-b}.$$

Rescaling by N, the worst-case spread loss is bounded below by

$$L_{\text{lb}}^{\text{sphere}} \;\geq\; N \cdot 4^{-b}.$$

Combining,

$$\frac{L^{\text{spread}}}{L_{\text{lb}}^{\text{sphere}}} \;\leq\; C_T \;=\; \frac{\pi\sqrt{3}}{2},$$

uniformly in N, d, and the realized vectors $\{x_i\}$. The bound is relative to the worst-case (spherical) lower bound, not to the prior-specific rate-distortion function for any joint distribution of $\{x_i\}$ that has exploitable structure. □

The benchmark is the spherical/minimax bound for the stacked vector in $\mathbb{R}^{Nd}$. RTQ on the stacked vector comes within $\pi\sqrt{3}/2$ of the worst-case rate-distortion floor, uniformly in N, d, and the realized inputs. The guarantee is distribution-free: no normality, no thin tails, no specific factor structure. Spread always works, and the worst-case loss relative to the spherical floor is a fixed constant that does not grow with problem size.

Theorem IA.2 (Concentration-strategy floor)**.** *For the concentration strategy with subset size k and any budget B,*

$$L^{\text{conc}} \;\geq\; (N-k)\,\sigma^2,$$

where $\sigma^2 = \mathbb{E}[\|x_i - \bar{x}\|_2^2]$. This lower bound is independent of B.

Proof. For the concentration strategy with subset S of size k, the estimate on each target $i \notin S$ is the unconditional mean $\bar{x}$. Each such target contributes $\mathbb{E}[\|x_i - \bar{x}\|_2^2] = \sigma^2$ to the expected squared error. There are $N-k$ such targets, so their total contribution is $(N-k)\sigma^2$. Reallocating bits within S does not change the contribution from targets outside S, so the bound holds independent of B. □

Each ignored target contributes its full prior variance σ^2 because the estimate on it is the unconditional mean. Reallocating bits within S does not change this. You cannot learn about a partner you never speak to, a stock you never price, or a country you never follow. The floor is independent of B, which is the key economic feature: even a well-resourced

analyst with a narrow watchlist still faces this floor on everything off the list. The result is the multi-target analog of Proposition 2.

Dividing Theorem IA.2 by Theorem IA.1 gives

$$\frac{L^{\text{conc}}}{L^{\text{spread}}} \;\geq\; \frac{(N-k)\sigma^2}{(\pi\sqrt{3}/2)\cdot N\cdot 4^{-b}} \;\to\; \frac{\sigma^2}{\pi\sqrt{3}/2}\cdot 4^b \qquad (N\to\infty,\ k \text{ fixed}),$$

which grows exponentially in b. Concentration is not merely worse than spread; the welfare cost grows exponentially in the per-coordinate budget. With more capacity, the penalty for ignoring targets rises faster than the gain from studying a narrow subset more carefully.

IA.2.3 Sufficient statistic, scaling, and regime diagram

Theorems IA.1 and IA.2 compress the three-parameter problem into one summary statistic and one sufficient-condition regime diagram. Define

$$\xi \;\equiv\; N\cdot 4^{-B/(Nd)}. \tag{IA.6}$$

N is the number of targets. $4^{-B/(Nd)}$ is the per-coordinate precision of spread: each bit cuts quantization variance by four, and each coordinate receives $B/(Nd)$ bits after the total budget is spread evenly. So ξ combines problem size with per-coordinate precision.

Proposition IA.1 (Sufficient statistic for parallel-target loss; unit-norm targets, $\sigma^2 = 1$)**.** *Under the assumptions of Theorem IA.1 with per-target prior variance normalized to $\sigma^2 = 1$,*

$$L^{\text{spread}} \;=\; \frac{\pi\sqrt{3}}{2}\cdot\xi\cdot(1+o(1)),$$

where $o(1)\to 0$ as $Nd, B\to\infty$ with $b = B/(Nd)$ bounded away from zero. The general $\sigma^2\neq 1$ case rescales the loss by σ^2. All (N,d,B) triples mapping to the same ξ produce the same normalized spread loss in the large-system limit.

Proof. By Theorem IA.1, $L^{\text{spread}}/\sigma^2 \leq (\pi\sqrt{3}/2)\cdot N\cdot 4^{-b}/\sigma^2$, which at unit prior variance $\sigma^2 = 1$ equals $(\pi\sqrt{3}/2)\xi$ with $\xi = N\cdot 4^{-b}$. Substituting $b = B/(Nd)$ gives (IA.6). In the high-rate regime (b bounded away from zero as $Nd, B\to\infty$), the TurboQuant MSE constant is asymptotically tight by the Panter-Dite formula applied to the Beta marginal of the Haar-rotated coordinates, so the upper bound and the achievable spread loss agree up to $o(1)$. □

The three-parameter problem reduces to one parameter. An analyst studying $N = 500$ stocks with $d = 20$ characteristics at budget B faces the same normalized loss as an analyst

studying $N = 20$ targets with $d = 6$ attributes at the matching ξ. The plot of loss against ξ is common across applications, so a calibration done once transfers to any other setting with the same ξ. Figure IA.5 confirms the relationship at five target counts, three total budgets, and two attribute dimensions.

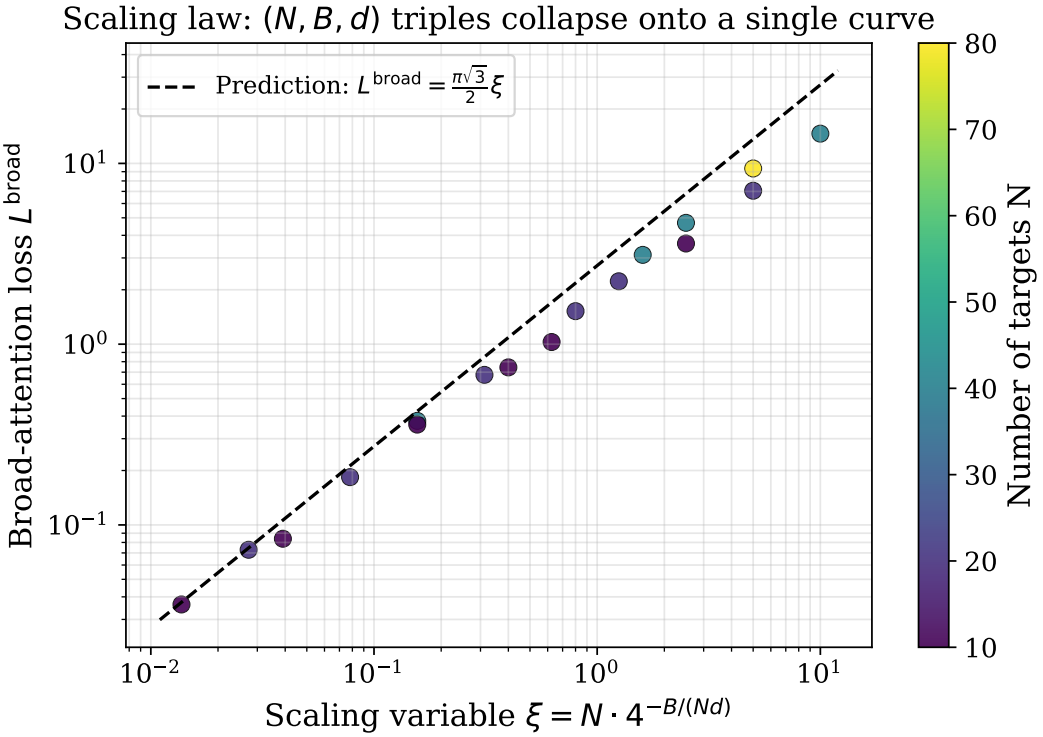

Figure IA.5: **One summary statistic governs the spread-strategy loss.** Numerical confirmation that all (N, d, B) triples producing the same ξ give the same loss. Monte Carlo simulation of the parallel-target setup. In each trial, N attribute vectors $x_i \in \mathbb{R}^d$ are drawn i.i.d. standard normal and normalized to the unit sphere; the spread (RTQ) strategy stacks them into $\mathbb{R}^{Nd}$, applies a Haar-random rotation, and scalar-quantizes each rotated coordinate with a Lloyd-Max quantizer at $K = \text{round}(2^{B/(Nd)})$ levels (capped at 256). Each circle is one (N, d, B) triple, loss averaged over 30 independent trials. Sweep: $N \in \{10, 20, 40, 80, 160\}$, $d \in \{3, 5, 8\}$, $B \in \{60, 120, 240, 480\}$, restricted to triples with $b = B/(Nd) \in [1.5, 5]$. Horizontal axis: $\xi = N \cdot 4^{-b}$ (log scale). Vertical axis: spread loss L^{spread} (log scale). Color: number of targets N. Dashed line: theoretical prediction $L^{\text{spread}} = (\pi\sqrt{3}/2)\,\xi$ from Proposition IA.1. Equation (IA.6) reduces the three-parameter problem to one.

The distribution-free guarantee in Theorem IA.1 is a property of the random rotation, not of the input distribution. Figure IA.6 verifies this on a simulation grid of four distributions (uniform on the sphere, heavy-tailed $t(3)$, three-mode Gaussian mixture, rank-one adversarial) crossed with three total dimensions $D = Nd \in \{64, 256, 1024\}$ and four per-coordinate bit rates $b \in \{2, 3, 4, 5\}$, for $4 \times 3 \times 4 = 48$ simulated configurations. The empirical ratio $L^{\text{spread}}/L^{\text{sphere}}_{\text{lb}}$ stays below $\pi\sqrt{3}/2 \approx 2.72$ in every case. The rank-one adversarial input, which places all prior mass in a single direction, is the worst case for random rotation; even there, the ratio stays under 2.5 at $D = 1024$. The bound is scale-free in a directly useful

sense: an upper bound on welfare loss under RTQ depends on the per-coordinate bit budget b alone, so applied researchers can size welfare loss ex ante without specifying or estimating the input distribution.

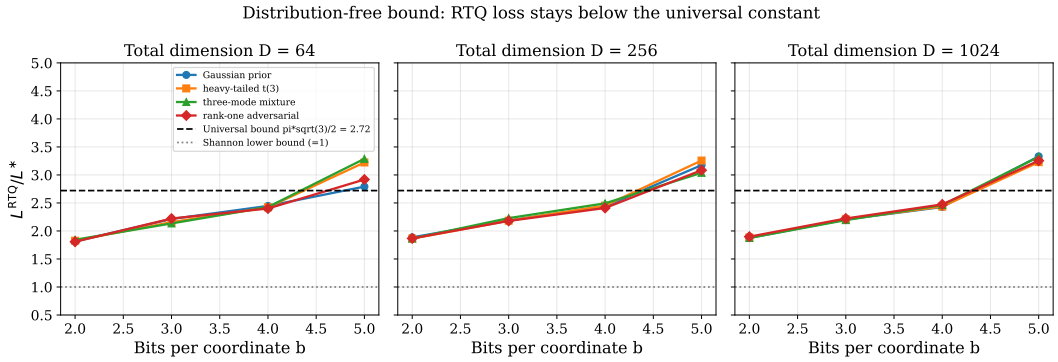

Figure IA.6: **Distribution-free universality.** Numerical confirmation that the 2.72 ceiling holds regardless of whether inputs are Gaussian, heavy-tailed, multi-modal, or concentrated in one direction. Monte Carlo check that $L^{\text{spread}}/L_{\text{lb}}^{\text{sphere}} \leq \pi\sqrt{3}/2$ for any input, where $L_{\text{lb}}^{\text{sphere}} = N \cdot 4^{-b}$ is the spherical/minimax rate-distortion lower bound for the worst-case unit input on S^{Nd-1}. Horizontal axis: per-coordinate bit rate $b = B/(Nd)$. Vertical axis: empirical ratio $L^{\text{spread}}/L_{\text{lb}}^{\text{sphere}}$. Each curve fixes one of four input distributions (uniform on the sphere, $t(3)$, three-mode Gaussian mixture, rank-one adversarial) and one of three total dimensions $D = Nd \in \{64, 256, 1024\}$. Each curve is plotted at four per-coordinate bit-rate values $b \in \{2, 3, 4, 5\}$. Configurations: 4 distributions $\times$ 3 dimensions $\times$ 4 bit rates = 48. Each point averages 100 trials. All ratios stay below $\pi\sqrt{3}/2 \approx 2.72$ (dashed) and above 1 (dotted). The bound is a property of the Haar-random rotation, not the input.

Proposition IA.2 (Sufficient condition for spread to dominate concentration; unit-norm targets, $\sigma^2 = 1$)**.** *With unit-norm targets and per-target prior variance normalized to $\sigma^2 = 1$, let $\kappa = k/N$ be the concentration fraction and $b = B/(Nd)$ the per-coordinate budget. Define*

$$b_c(\kappa) \;=\; \frac{1}{2}\log_2\left[\frac{\pi\sqrt{3}/2}{1-\kappa}\right]. \tag{IA.7}$$

Above this curve ($b > b_c$), the spread upper bound is strictly below the concentration lower bound, so spread is guaranteed to beat concentration. Below this curve, the bounds are inconclusive: concentration may dominate, and the simulations in Figure IA.7 indicate that it does in the calibrated regimes shown. The general $\sigma^2 \neq 1$ case follows by replacing $1-\kappa$ with $\sigma^2(1-\kappa)$ in (IA.7)*.*

Proof. Set the spread upper bound $L^{\text{spread}} \leq (\pi\sqrt{3}/2) \cdot N \cdot 4^{-b}$ from Theorem IA.1 equal to the concentration lower bound $L^{\text{conc}} \geq (N-k)\sigma^2$ from Theorem IA.2:

$$\frac{\pi\sqrt{3}}{2} \cdot N \cdot 4^{-b} \;=\; (N-k)\sigma^2.$$

Dividing by N, using $\kappa = k/N$, and solving for b,

$$4^{-b} \;=\; \frac{\sigma^2(1-\kappa)}{\pi\sqrt{3}/2} \quad\Longleftrightarrow\quad b_c(\kappa) \;=\; \frac{1}{2}\log_2\!\left[\frac{\pi\sqrt{3}/2}{\sigma^2(1-\kappa)}\right].$$

Above this curve ($b > b_c$), the spread upper bound is strictly below the concentration lower bound, so $L^{\text{spread}} \le \text{UB}_{\text{spread}}(b) < \text{LB}_{\text{conc}}(\kappa) \le L^{\text{conc}}$ and spread strictly dominates. Below the curve the bound comparison is inconclusive; the empirical regime in Figure IA.7 shows concentration dominating in the simulated calibrated regimes there, but this is a numerical pattern rather than a theorem implication. For the tanh form, let $r \equiv L^{\text{spread}}/L^{\text{conc}}$ at the bound-achieving losses; then $m \equiv (1-r)/(1+r) = \tanh(-\frac{1}{2}\ln r)$, and substituting $\ln r = \ln[(\pi\sqrt{3}/2)/(\sigma^2(1-\kappa))] - 2b\ln 2$ together with the definition of b_c gives $m = \tanh((b - b_c)\ln 2)$. □

Equation (IA.7) is a one-sided sufficient condition an applied researcher can use directly. Above b_c, spread is guaranteed to dominate. Below b_c the theorem is silent and applied users should rely on the calibration in Figure IA.7, which indicates concentration dominates below the curve in the simulated regimes. The threshold depends on σ^2 and on $1-\kappa$, the fraction of targets left unmonitored. When more targets are left out (κ small), the threshold drops: more territory for spread to cover means less budget needed before spreading pays. Near the boundary, the normalized loss gap

$$m(b,\kappa) \;\equiv\; \frac{L^{\text{conc}} - L^{\text{spread}}}{L^{\text{conc}} + L^{\text{spread}}} \;=\; \tanh\big((b - b_c(\kappa))\ln 2\big)$$

takes values in $[-1,+1]$: $m = 0$ at the boundary, $m \to +1$ when spread dominates, $m \to -1$ when concentration dominates. Slope is $\ln 2$ at the boundary, so small budget changes produce measurable shifts only near the crossover. Figure IA.7 plots $\log_{10}(L^{\text{conc}}/L^{\text{spread}})$ on (b,κ) with the boundary superimposed.

The regime diagram is the operational form of the loss concentration ratio diagnostic from Corollary 1. Applied researchers can compute $b = B/(Nd)$, read the regime off the diagram, and use the sufficient-condition boundary above which spread is guaranteed to dominate; below the boundary the calibrated figure indicates concentration can dominate, and the choice is informed by the simulation rather than the theorem. The Fisman cross-domain test in Section IA.3 below applies the spread theorem and the slope-implied effective dimension that follow from this framework.

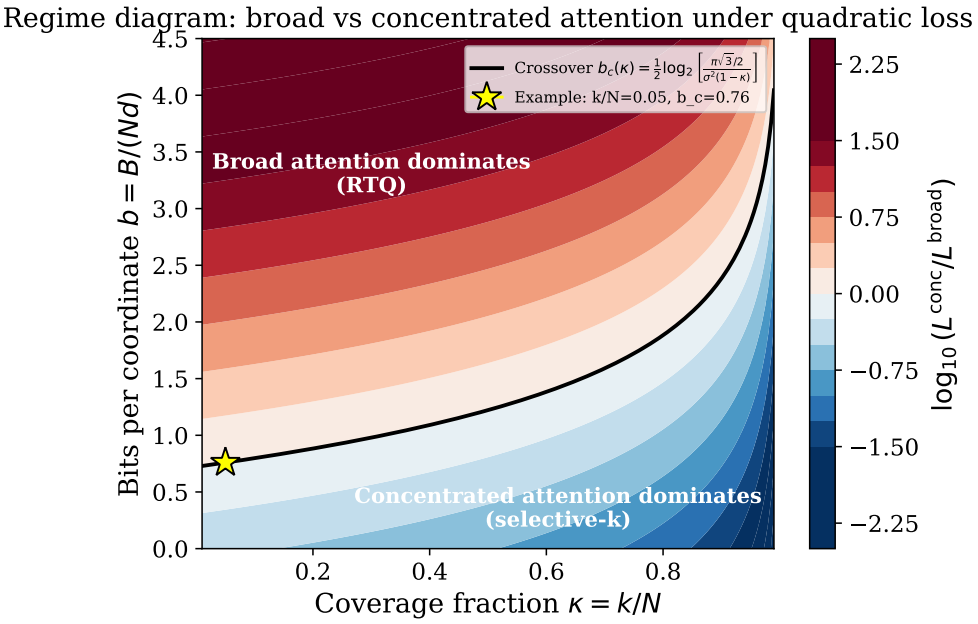

Figure IA.7: **Decision diagram for spread versus concentration** at $N = 100$, $\sigma^2 = 1$. Closed-form evaluation of the loss ratio from Theorems IA.1 and IA.2; no simulation. Horizontal axis: $b = B/(Nd)$. Vertical axis: $\kappa = k/N$. Color: $\log_{10}(L^{\text{conc}}/L^{\text{spread}})$ (positive: spread wins). Black contour: boundary $b = b_c(\kappa)$ from (IA.7). Three operating points: low (b, κ) corresponds to a large event with many acquaintances of uncertain type; the crossover corresponds to a small dinner; high (b, κ) corresponds to a one-on-one conversation.

IA.3 Fisman speed-dating cross-domain test

The CRSP and 13F evidence in Section 5 of the main text tests the welfare-gap and regime predictions in the portfolio setting. To check that the parallel-target effective-dimension calibration of the spread theorem (Theorem IA.1) holds outside finance, we turn to the Columbia speed-dating experiment of Fisman et al. (2006). The experiment recorded 8,378 four-minute conversations among 551 participants across 21 randomized "waves" conducted 2002–2004, with cohort sizes N varying exogenously from 5 to 22 per side. Each participant faces a parallel-target decision: N partners, a fixed time budget of about two hours, and an attribute space of six core trait ratings plus 17 activity preferences. The exogenous variation in cohort size supplies the N variation Theorem IA.1 requires. The empirical content is a single slope, β, of $\log(1 + \text{matches})$ on $\log(\text{partners})$. The rest of this section derives what β measures (Section IA.3.1), runs the regression and reads the effective dimension off the slope (Section IA.3.2), and checks robustness and recovers the per-coordinate bit budget (Section IA.3.3).

IA.3.1 Structural model and slope identity

Mapping to the model. A participant in a larger wave allocates fewer minutes per partner, a smaller per-partner bit budget. A smaller wave gives more minutes per partner but fewer to meet. These correspond to the spread (large-N) and concentration (small-N) strategies of the parallel-target framework in Section IA.2 above. Each partner is a target with attribute vector $x_i \in \mathbb{R}^d$ of unit length; the participant forms a noisy score and accepts or rejects on it.

Proposition IA.3 (Optimal threshold under noisy observation)**.** *An agent with noisy score $\hat{s} \sim \mathcal{N}(s, \sigma_q^2)$ on latent $s \sim \mathcal{N}(0,1)$ accepts if $\hat{s} > \theta^*$. Accepting yields $v > 0$ when the partner also accepts and cost $c \in (0, v)$ when rejected. Both sides observe the same s with independent noise σ_q^2 and use the same threshold. The symmetric Nash threshold satisfies*

$$\theta^* \;=\; -\frac{\sqrt{(1+\sigma_q^2)(2+\sigma_q^2)}}{\sigma_q} \cdot \Phi^{-1}(c/v).$$

Proof. The agent's expected payoff from saying yes given private signal $\hat{s}_A$ is $v \cdot P(\hat{s}_B > \theta^* \mid \hat{s}_A) - c$. Setting this to zero at $\hat{s}_A = \theta^*$ pins down the indifference threshold. Under joint Gaussianity, $E[s \mid \hat{s}_A] = \hat{s}_A/(1+\sigma_q^2)$ and $\mathrm{Var}(s \mid \hat{s}_A) = \sigma_q^2/(1+\sigma_q^2)$, so $\hat{s}_B \mid \hat{s}_A \sim$

$\mathcal{N}(\hat{s}_A/(1+\sigma_q^2), \sigma_q^2(2+\sigma_q^2)/(1+\sigma_q^2))$. Evaluating at $\hat{s}_A = \theta^*$,

$$P(\hat{s}_B > \theta^* \mid \hat{s}_A = \theta^*) \;=\; \Phi\left(-\frac{\theta^*\sigma_q}{\sqrt{(1+\sigma_q^2)(2+\sigma_q^2)}}\right).$$

Equating to c/v and inverting gives the symmetric Nash threshold. □

Under spread, each rotated coordinate is quantized at $b = B/(Nd)$ bits and produces noise with variance $\sigma_q^2 = (\pi\sqrt{3}/2)\cdot 4^{-b}$ per coordinate. The expected match count is

$$\mathbb{E}[M \mid N, B] \;=\; N \cdot \pi(\sigma_q^2), \qquad \pi(\sigma^2) \;=\; \mathbb{E}_{s\sim\mathcal{N}(0,1)}\left[\Phi\left(\tfrac{s-\mu}{\sigma}\right)^2\right], \qquad \mu \;=\; \theta^*\sqrt{d}.$$

Differentiating $\log \mathbb{E}[M]$ with respect to $\log N$ gives the slope identity

$$\beta \;=\; \frac{d\log \mathbb{E}[M]}{d\log N} \;=\; 1 + b\ln 4 \cdot \frac{d\log \pi(\sigma^2)}{d\log \sigma^2}\bigg|_{\sigma^2=\sigma_q^2}. \tag{IA.8}$$

Define the noise elasticity $\kappa(\sigma^2) \equiv -d\log\pi(\sigma^2)/d\log\sigma^2 > 0$. Equation (IA.8) becomes $\beta = 1 - \kappa(\sigma_q^2)\, b\ln 4$. The product $\kappa(\sigma_q^2)\, b\ln 4$ is the inverse of an effective dimension that captures how loss-relevant attributes the participant actually uses; writing $\beta = 1 - 1/d_{\text{eff}}$ defines

$$d_{\text{eff}} \;\equiv\; \frac{1}{\kappa(\sigma_q^2)\, b\ln 4} \;=\; \frac{1}{1-\beta}.$$

This is the slope-implied effective dimension. The next proposition links it back to the eigenvalue structure of the trait covariance.

Proposition IA.4 (Effective-dimension calibration under an equal-eigenvalue approximation)**.** *Let $x_i \in \mathbb{R}^d$ be iid with zero mean, unit norm, and trait covariance Σ. Define the slope-implied effective dimension*

$$d_{\text{eff}}^{\text{slope}} \;\equiv\; \frac{1}{1-\beta}, \tag{IA.9}$$

where β is the log-linear slope of $\log(1+M)$ on $\log N$ obeying (IA.8)*, and the PCA-implied effective dimension*

$$d_{\text{eff}}^{\text{PCA}}(\Sigma) \;\equiv\; \frac{(\operatorname{tr}\Sigma)^2}{\operatorname{tr}(\Sigma^2)} \;\in\; [1, d]. \tag{IA.10}$$

If the informative attribute subspace has roughly equal nonzero eigenvalues, $d_{\text{eff}}^{\text{slope}} \approx d_{\text{eff}}^{\text{PCA}}(\Sigma)$.

Proof. Write $\Sigma = \sum_k \lambda_k e_k e_k^\top$ with $\sum_k \lambda_k = 1$. The inner product of x_A with a random partner's vector x_i has variance $\sum_k \lambda_k^2 = \operatorname{tr}(\Sigma^2)$, because the two sides are independent

and each eigendirection contributes λ_k^2. Rescaling to unit variance, $s \equiv \langle x_A, x_i \rangle / \sqrt{\mathrm{tr}(\Sigma^2)}$. A central limit theorem on the eigendirections (valid when d_{eff} is large enough) gives $s \xrightarrow{d} \mathcal{N}(0,1)$.

The agent's quantization error on x_A is isotropic after RTQ and has total MSE σ_q^2, so its projection onto x_i has variance σ_q^2/d when averaged over x_i. Restricting to the informative sub-space spanned by the eigendirections with non-trivial λ_k replaces d by d_{eff} in the per-direction variance. The substitution is exact when eigenvalues are evenly concentrated within the informative sub-space.

Combining with the slope formula (IA.8), the match probability becomes $\pi(\sigma_q^2/d_{\text{eff}})$, and the slope is $\beta = 1 - \kappa(\sigma_q^2)\, b \ln 4$ with $\kappa(\sigma_q^2)\, b \ln 4 = 1/d_{\text{eff}}$. The identity $d_{\text{eff}}^{\text{slope}} = (\mathrm{tr}\,\Sigma)^2/\mathrm{tr}(\Sigma^2)$ is exact when the informative subspace has equal eigenvalues; in general it is a calibration approximation. □

The structural model gives a sharp prediction: the slope β is bounded above by one, strictly positive, and inversely related to the effective dimension. A hard-concentration alternative, in which only a fixed number of partners receives attention regardless of N, predicts weak or zero scaling with N, so a positive slope is inconsistent with hard concentration.

IA.3.2 Main regression and effective-dimension comparison

For each participant i, let M_i be the count of mutual matches and N_i the wave size. We estimate

$$\log(1 + M_i) \;=\; \alpha + \beta \log N_i + \epsilon_i. \tag{IA.11}$$

The estimated slope is $\hat{\beta} = 0.472$ with wave-clustered SE 0.063 ($p < 10^{-10}$). The slope is positive across genders, age bands, and Poisson and negative-binomial GLMs; a placebo on seating position is indistinguishable from zero. Figure IA.8 plots the scatter and fit; Table IA.2 reports the full robustness grid.

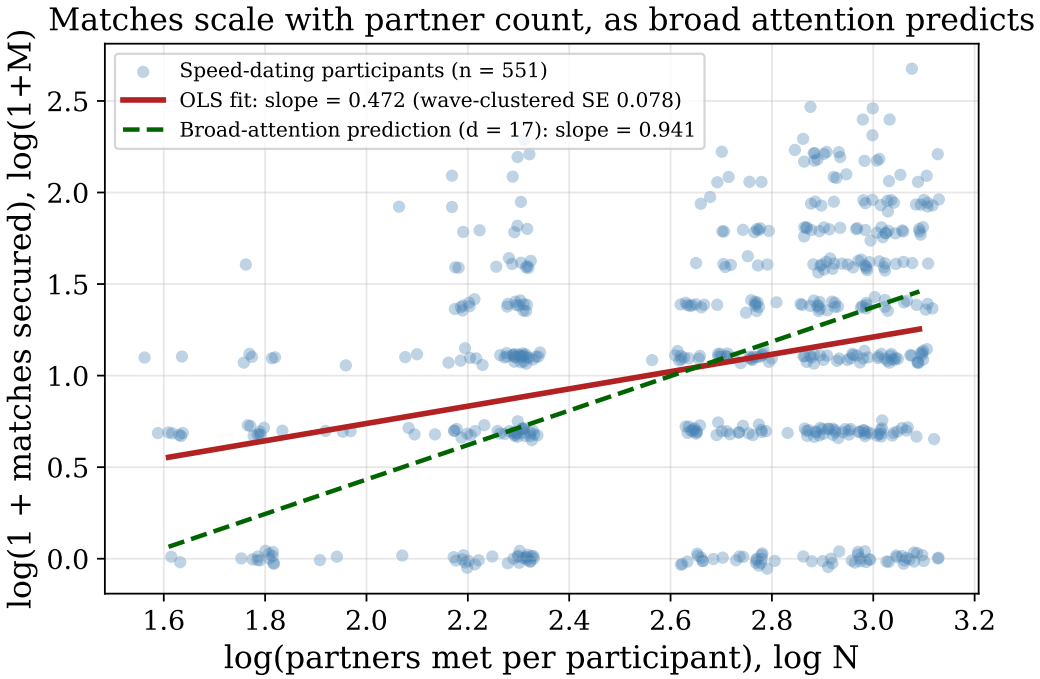

Figure IA.8: **Matches scale with partner count, Columbia speed-dating.** Horizontal axis: $\log N_i$. Vertical axis: $\log(1+M_i)$. Each point is one of 551 participants, with small jitter for readability. Red line: OLS fit with $\hat{\beta} = 0.472$; SE 0.063 both wave-clustered (21 waves) and from a participant-level bootstrap with 2,000 resamples. Green dashed line: nominal slope prediction $\beta = 1 - 1/d = 0.94$ under $d = 17$. The gap reflects correlation among attributes: the effective dimension $d_{\text{eff}}^{\text{slope}} = 1/(1-\hat{\beta}) = 1.89$ sits well below nominal d. Data: Fisman et al. (2006).

Specification	n	$\hat{\beta}$	SE (cluster)	SE (boot)	95% CI (boot)	p
Main (log-linear)	551	+0.472	0.063	0.063	[+0.350, +0.593]	1.7×10^{-11}
female	274	+0.442	0.106	0.106	[+0.229, +0.645]	4.4×10^{-5}
male	277	+0.498	0.090	0.090	[+0.320, +0.675]	6.6×10^{-8}
age ≤ 24	186	+0.359	0.126	0.126	[+0.117, +0.609]	5.0×10^{-3}
age 25–28	225	+0.561	0.109	0.109	[+0.350, +0.783]	5.7×10^{-7}
age 29+	132	+0.463	0.126	0.126	[+0.223, +0.723]	3.3×10^{-4}
Wave-demeaned (OLS)	551	+0.748	0.497	0.497	[−0.224, +1.714]	1.3×10^{-1}
Placebo (position)	520	−0.010	0.008	0.008	[−0.026, +0.005]	2.2×10^{-1}
Poisson	551	+0.462	0.058	0.058	[+0.349, +0.575]	2.4×10^{-12}
NegBinomial	551	+0.458	0.062	0.062	[+0.336, +0.580]	1.9×10^{-11}

Table IA.2: **Partner-count slope across specifications, Fisman speed-dating data ($n = 551$, 21 waves).** Dependent variable: $\log(1 + M_i)$ in OLS rows, M_i in Poisson and negative-binomial rows, where M_i is participant i's count of mutual matches. Independent variable: $\log N_i$. $\hat{\beta}$: slope on $\log N_i$. Main: pooled OLS. Gender/age: sub-sample OLS. Wave-demeaned: OLS with wave dummies. Placebo: OLS with seating position in place of $\log N_i$. Poisson and NB: log-link GLM with $\log N_i$. Cluster SE: 21 waves. Bootstrap SE: 551 participants with replacement, 2,000 draws. p-values two-sided against $\beta = 0$.

Plugging $\hat{\beta} = 0.472$ into the slope-implied identity (IA.9) gives $d_{\text{eff}}^{\text{slope}} = 1.89$. The PCA-implied dimension on the standardized six-trait correlation matrix is $d_{\text{eff}}^{\text{PCA}} = 2.68$, with the first principal component capturing 57.6 percent of the trait variance. Figure IA.9 plots the scree.

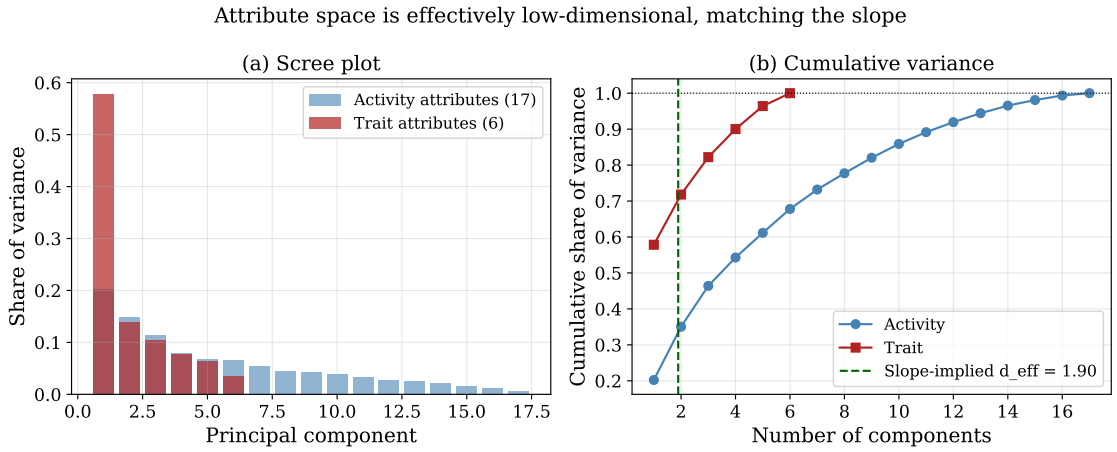

Figure IA.9: **PCA of the Fisman attribute space.** Inputs: 551 participants, 17 activity preferences and six partner-trait ratings, standardized. Left: scree plot of eigenvalues λ_j of the sample correlation matrix (blue: activities, red: traits) versus component index. Right: cumulative variance $\sum_{j \leq k} \lambda_j / \sum_j \lambda_j$ versus k, with horizontal line at the empirical $d_{\text{eff}}^{\text{slope}} = 1/(1 - \hat{\beta}) = 1.89$. The 1.9 line sits between the first and second components of the trait PCA, consistent with two-factor readings of the survey.

The two numbers agree on the qualitative reading (the trait space is effectively one to three dimensional) and the direction of the gap matches the theory. The slope-implied $d_{\text{eff}}^{\text{slope}}$ is decision-weighted: it picks up the dimension of the loss-relevant attribute subspace as participants actually use it. The PCA-implied $d_{\text{eff}}^{\text{PCA}}$ weights all trait variation equally. Survey evidence in Fisman et al. (2006) shows participants weight a small subset of traits (attractiveness, shared interests) more heavily than others, so the decision-weighted dimension should sit below the equal-weighted PCA value, and it does. We treat this agreement as supportive of the parallel-target prediction in a non-portfolio setting, not as an exact identity. Doubling wave size from 10 to 20 raises expected matches by $2^{0.472} \approx 1.39$, what spread predicts and concentration rejects.[35]

IA.3.3 Per-coordinate bit budget and out-of-sample stability

Per-coordinate bit budget. The slope $\hat{\beta}$ identifies d_{eff} but does not separately identify σ_q and μ. We recover the per-coordinate bit budget from the joint distribution of (N_i, M_i, partner attractiveness). Under the threshold model of Proposition IA.3, the within-person probit of the yes decision on z-scored partner attractiveness has slope $1/\sigma_q$ and intercept $-\mu/\sigma_q$. The pooled probit on 8,378 conversations returns slope 0.662 and intercept -0.218, implying $\sigma_q = 1.51$ and $\mu = 0.33$. Inverting $\sigma_q^2 = (\pi\sqrt{3}/2) \cdot 4^{-b}$ gives $\hat{b} = 0.127$ bits per coordinate per partner. With $d_{\text{eff}}^{\text{slope}} = 1.89$, total attention per partner-conversation is $\hat{B} = \hat{b} \cdot d_{\text{eff}} \approx 0.24$ bits. Over the 240-second window, this is about 10^{-3} bits per second in the useful channel. The magnitude is small in absolute terms but matches the design: the protocol asks for one binary decision per partner, and 0.24 bits is below one bit per partner-conversation, consistent with a very noisy binary decision. Attention, not raw information-processing capacity, is the binding constraint.

Out-of-sample slope stability. We split the 21 waves 15–6 by random permutation and re-estimate the main regression on training waves. Figure IA.10 reports training-versus-test slopes across 100 splits. The held-out slope on six unseen waves is close to the training slope, so the fit is not driven by a few influential waves.

The headline cross-domain finding is the agreement between the slope-implied and PCA-implied effective dimensions, plus the positive slope itself. The positive slope is inconsistent with a hard-concentration model in which only a fixed number of partners receives attention regardless of N. We treat this exercise as a calibration and cross-domain validation

[35]The Columbia experiment does not randomize wave size at the participant level, so we cannot rule out self-selection of attentive daters into larger waves. The claim is observational: within the scale observed, the relationship between wave breadth and match success is consistent with spread and inconsistent with concentration.

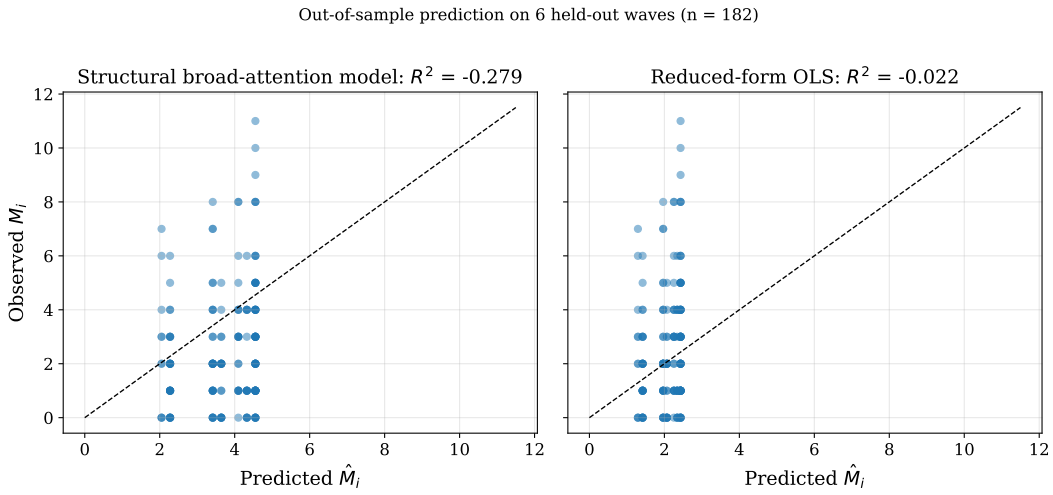

Figure IA.10: **Out-of-sample slope stability, Fisman speed-dating.** Horizontal axis: in-sample slope $\hat{\beta}_{\text{train}}$ from (IA.11) on 15 random waves. Vertical axis: out-of-sample slope $\hat{\beta}_{\text{test}}$ on the 6 held-out waves. Each point is one of 100 independent 15/6 splits. Dashed: 45-degree line. Points cluster around the diagonal with mean $\hat{\beta}_{\text{test}}$ close to the full-sample $\hat{\beta} = 0.472$, indicating slope stability across splits. The point-prediction R^2 values are naturally low because the dependent variable is a noisy match count; the test in this figure is on slope stability, not on point prediction of M_i.

of the parallel-target framework rather than a structural test, given the equal-eigenvalue approximation underlying Proposition IA.4.

IA.4 Connection to Transformer-Based Portfolio Construction

This section formalizes the link between RTQ and learned cross-asset attention in Transformer-based portfolio models such as Cong et al. (2026). We show that RTQ approximates, in expected welfare loss, the population-optimal linear-attention layer at matched bit budget, with a worst-case multiplicative bound. The bound holds for any prior with finite second moments and any positive definite welfare matrix. Two consequences follow. RTQ provides a closed-form, distribution-free theoretical foundation for what cross-asset linear attention learns asymptotically. And RTQ's composite signals admit a direct economic interpretation as welfare-weighted factor portfolios, whereas the Transformer's learned attention recovers the same kind of object only implicitly through gradient descent.

IA.4.1 Setup

An investor with welfare matrix $W \succ 0$ minimizes expected loss $\mathcal{L}(\hat{a}) = \mathbb{E}[(\hat{a}-u)^\top W(\hat{a}-u)]$, where $u \in \mathbb{R}^d$ has prior mean zero and covariance $\Sigma \succ 0$ with finite second moments. We compare two candidate signal structures.

Linear-attention layer. A rank-k linear projection $s = Au$ with $A \in \mathbb{R}^{k\times d}$, followed by a linear-MMSE decision rule. This is the population analog of a single-head linear-attention layer trained end-to-end to minimize the portfolio loss. Among rank-k linear signal structures with linear-MMSE decoding, the optimum chooses A to minimize $\mathcal{L}$.

RTQ at bit budget B. A random orthogonal $R \in O(d)$ rotates u, each rotated coordinate $(Ru)_j$ is quantized at $b = B/d$ bits via TurboQuant, and the decoded state $\hat{u}$ enters the welfare-weighted action rule. RTQ's expected loss is bounded by Proposition 1 of the main text.

IA.4.2 Optimal linear-attention projection

Proposition IA.5 (Optimal rank-k linear-MMSE projection under welfare loss)**.** *For any prior with finite second moments and any positive definite welfare matrix W, the rank-k linear projection A^* that minimizes the linear-MMSE welfare loss projects onto the top-k eigenvectors of*

$$M \;:=\; \Sigma^{1/2}\, W\, \Sigma^{1/2}. \tag{IA.12}$$

The achieved minimum loss is

$$\mathcal{L}_{\mathrm{LAT}}(k) \;=\; \mathrm{tr}(W\Sigma) \;-\; \sum_{j=1}^{k} \lambda_j(M), \tag{IA.13}$$

where $\lambda_j(M)$ denotes the j-th largest eigenvalue of M.

Proof. Given the signal $s = Au$, the linear-MMSE estimator is $\hat{u}^{\mathrm{LMMSE}} = \Sigma A^\top (A\Sigma A^\top)^{-1} Au$. The expected welfare loss is

$$\begin{aligned}
\mathcal{L}(A) &= \mathbb{E}\big[(u - \hat{u}^{\mathrm{LMMSE}})^\top W (u - \hat{u}^{\mathrm{LMMSE}})\big] \\
&= \mathrm{tr}\big(W\left[\Sigma - \Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma\right]\big) \\
&= \mathrm{tr}(W\Sigma) \;-\; \mathrm{tr}\big(W\Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma\big)\,.
\end{aligned}$$

Minimizing $\mathcal{L}(A)$ over rank-k A is equivalent to maximizing the second trace. Set $B := A\Sigma^{1/2}$ and $M := \Sigma^{1/2} W \Sigma^{1/2}$. The second trace becomes

$$\mathrm{tr}\big(MB^\top (BB^\top)^{-1} B\big)\,,$$

which equals the sum of the eigenvalues of M on the subspace spanned by $\mathrm{row}(B)$. By Ky Fan's maximum principle (Poincaré separation theorem; Bhatia, 1997) applied to the eigenvalue decomposition of M, this trace is maximized over rank-k B by aligning $\mathrm{row}(B)$ with the top-k eigenvectors of M. The maximum equals $\sum_{j=1}^{k} \lambda_j(M)$. Translating back, A^* projects onto the top-k eigenvectors of M, and the minimum loss is (IA.13). □

The optimal projection is welfare-weighted principal-component analysis on the prior covariance. This is the population target that a sufficiently large training set drives the linear-attention weights toward when the training objective is the portfolio's welfare loss.

IA.4.3 RTQ within a constant factor of optimal linear attention

Proposition IA.6 (RTQ matches optimal linear attention within a constant factor in the minimax sense)**.** *Let $\mathcal{L}_{\mathrm{lb}}^{\mathrm{sphere}}(B)$ denote the spherical minimax rate-distortion floor at bit budget B, in the known-norm Gaussian-spherical signal class on adversarial priors with finite second moments. For any prior with finite second moments and any $W \succ 0$:*

(a) RTQ at bit budget B satisfies

$$\mathcal{L}_{\mathrm{RTQ}}(B) \;\le\; C_T \cdot \chi(W) \cdot \mathcal{L}_{\mathrm{lb}}^{\mathrm{sphere}}(B), \tag{IA.14}$$

where $C_T = \pi\sqrt{3}/2 \approx 2.72$ *and* $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}(W)$.

(b) The worst-case loss of the optimal linear-attention layer across priors satisfies $\sup_F \mathcal{L}^*_{\mathrm{LAT}}(F, B) \geq \mathcal{L}^{\mathrm{sphere}}_{\mathrm{lb}}(B)$.

Consequently, RTQ is within constant factor $C_T\chi(W)$ *of the worst-case loss of the asymptotically optimal linear-attention layer at matched bit budget.*

Proof. Statement (a) is Proposition 1 of the main text applied at bit budget B. Statement (b) follows from the definition of the spherical minimax floor: signal structures in the known-norm spherical class, including a rank-k linear projection followed by linear-MMSE decoding, must achieve at least $\mathcal{L}^{\mathrm{sphere}}_{\mathrm{lb}}(B)$ on the worst-case prior in that class. Combining (a) and (b) gives the constant-factor relationship between RTQ and the worst-case linear-attention loss. □

The substantive content of (IA.14) is that RTQ is within a known multiplicative constant of the worst-case asymptotically optimal linear-attention Transformer. The constant depends only on the welfare matrix's condition number and is independent of the dimension and the bit budget. This is the closed-form, distribution-free, capacity-bounded benchmark for the family of methods that includes Cong et al. (2026). For priors with exploitable structure, an oracle-tuned linear-attention layer can do better than the minimax floor; RTQ's bound is the worst-case envelope, the appropriate comparison when the prior is not known.

IA.4.4 Explainability: RTQ composite signals as welfare-weighted factor portfolios

Corollary IA.1 (Composite signals admit a factor-portfolio interpretation)**.** *The eigenvectors of* $M = \Sigma^{1/2} W \Sigma^{1/2}$ *are welfare-weighted factor portfolios: linear combinations of the underlying state that maximally trade off variance explained against welfare relevance. By Proposition IA.5, the population-optimal linear-attention weights* A^* *project onto the top-k such factor portfolios. By Proposition IA.6, RTQ's loss is within* $C_T\chi(W)$ *of the minimax rate-distortion floor that bounds the worst-case loss of any linear-attention layer at matched bit budget. RTQ's composite signals at* $b = B/d$ *bits each thus form a closed-form, capacity-bounded approximation of the same factor portfolios that the Transformer learns implicitly.*

The economic content is direct. Each RTQ composite signal is a portfolio of the underlying assets with computable weights given by the random rotation. The bit allocation across composites, governed by water-filling on welfare-weighted eigenvalues, identifies which composite portfolios receive more resolution. The Transformer's learned attention recovers

the same factor portfolios in the large-sample limit, but the weights are obtained by gradient descent without a closed-form interpretation. RTQ supplies that interpretation: at construction time, the agent knows which composite portfolios are being attended to and at what resolution, with a worst-case guarantee on the resulting loss.

IA.4.5 Discussion and limitations

The bounds in this section concern the linear-attention mechanism that sits inside many Transformer architectures (Vaswani et al., 2017) and is the specific layer that Cong et al. (2026) use for cross-asset attention in portfolio choice. Proposition IA.5 characterizes the population-optimal rank-k linear projection by the top eigenvectors of $M = \Sigma^{1/2} W \Sigma^{1/2}$, a standard Ky Fan / Eckart-Young-Mirsky statement. Proposition IA.6 gives a minimax comparison: RTQ is within a constant factor $C_T \chi(W)$ of the spherical rate-distortion floor at matched per-coordinate bit budget, the same floor that the optimal rank-k linear-MMSE attention layer attains asymptotically. The four architectural variations below clarify what extends from this benchmark to fuller Transformer pipelines and what does not.

Consider first the softmax normalization that wraps the inner product in standard attention. Softmax replaces the linear inner product with an exponentially normalized weight, which is nonlinear. Population-limit arguments suggest that under quadratic loss and Gaussian features the softmax weights concentrate around a direction that the linear-MMSE projection also identifies, but the equivalence is heuristic rather than a theorem in the present paper. We do not claim that Proposition IA.6 extends without modification to nonlinear softmax attention.

A second extension concerns multi-head attention, in which each head learns a separate projection. Mapping h heads to RTQ's rotated coordinates is suggestive rather than a formal equivalence: an h-head Transformer with learned head projections is a different object than a Haar-rotated RTQ representation. We expect the bound to remain informative as a benchmark for multi-head attention but do not prove it.

The downstream feedforward (MLP) layers raise a separate scope question. The data-processing inequality applies to the entire pipeline only if the attention layer is explicitly treated as a bit-limited information bottleneck. Standard continuous Transformer layers are not bit-limited in the Shannon sense; the pipeline-level bound therefore requires either (i) an explicit quantization step at the attention layer or (ii) an additional argument that the post-attention representation has bounded mutual information with the input. In either case, conditional on the attention layer being bit-limited, RTQ is within $C_T \chi(W)$ of the rate-distortion floor for that layer, and no downstream MLP can recover information that

the attention layer did not transmit.

A separable feature of RTQ also deserves mention: the regime-conditional guidance the framework supplies. The Transformer attention mechanism is flexible: through learned attention weights, it can place near-zero weight on most inputs (effective focus) or distribute weight broadly (effective spread). What it does not provide is a principled criterion for which regime to expect. The choice emerges empirically from training data. RTQ delivers the corresponding closed-form criterion: the loss concentration ratio $\rho(k)$ in Corollary 1 of the main text and the parallel-target sufficient statistic ξ in Proposition IA.1 identify, ex ante and distribution-free, when focus pays and when spread pays. In the high-dimensional regime where the US equity universe lives, RTQ predicts that broad cross-asset attention dominates, consistent with the empirical findings in Cong et al. (2026). In lower-dimensional regimes, such as a central bank with a few policy instruments or a hedge fund concentrated in a handful of active bets, RTQ predicts that focus pays.

This appendix provides a conceptual bridge to learned cross-asset attention rather than a formal equivalence with the entire Transformer architecture. Proposition IA.5 characterizes the population-optimal linear projection. Proposition IA.6 gives a minimax benchmark showing RTQ is within a constant factor of the spherical rate-distortion floor at matched bit budget, the same floor the optimal linear-attention layer attains asymptotically. The connection to softmax, multi-head, and MLP-augmented Transformer architectures is suggestive of broader applicability but does not extend the proved bound. Random-rotation RTQ produces feasible diversified signals whose welfare loss is bounded uniformly, but those signals are not the top-M eigenvectors of $\Sigma^{1/2}W\Sigma^{1/2}$ unless the rotation is deliberately aligned with M; the resulting signals are economic objects (factor portfolios), but the alignment between RTQ composites and learned factor portfolios is qualitative rather than exact.

IA.5 Extensions

IA.5.1 Equilibrium with heterogeneous attention

The main text studies a single agent's attention allocation. We now embed RTQ-constrained agents in a noisy rational-expectations equilibrium with full-information counterparties and noise traders. The Gaussian-CARA case admits a closed-form characterization of the equilibrium price, each type's posterior, and the welfare loss to RTQ agents. The framework reduces by orthogonal diagonalization to d independent one-dimensional REE problems, so we develop the analysis at the per-coordinate level and aggregate at the end.

Setup. Let $\mathbf{x} \in \mathbb{R}^d$ be the vector of risky-asset payoffs with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $\Sigma \succ 0$. The risk-free rate is zero. A continuum of CARA agents with absolute risk aversion γ has two types:

- **Type F (full-information), fraction $\alpha \in (0,1)$:** observes $\mathbf{y}_F = \mathbf{x} + \boldsymbol{\varepsilon}_F$ with $\boldsymbol{\varepsilon}_F \sim \mathcal{N}(\mathbf{0}, \tau^2\Sigma)$. We maintain $\tau > 0$ throughout. The limit $\tau \to 0^+$ gives full price revelation in equilibrium, the standard Grossman-Stiglitz degenerate case, and is not the substantive object of interest; all formulas below hold for $\tau > 0$.
- **Type R (RTQ-bounded), fraction $1-\alpha$:** receives an RTQ signal at total bit budget B with rotation aligned to the eigenbasis of Σ and equal allocation $b = B/d$ bits per rotated coordinate.

Noise traders supply $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\Sigma)$ per capita, independent of $\mathbf{x}$. The net per-capita supply of risky assets is zero. Market clearing requires $\alpha\mathbf{d}_F + (1-\alpha)\mathbf{d}_R + \mathbf{n} = \mathbf{0}$.

Diagonalization. Let $\Sigma = V\Lambda V^\top$ be the eigendecomposition, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Rotated quantities $\tilde{\mathbf{x}} = V^\top\mathbf{x}$, $\tilde{\mathbf{y}}_R = V^\top\mathbf{y}_R$, $\tilde{\mathbf{p}} = V^\top\mathbf{p}$, and $\tilde{\mathbf{n}} = V^\top\mathbf{n}$ satisfy $\tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \Lambda)$, $\tilde{\mathbf{n}} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\Lambda)$, and rotated coordinates are mutually independent. In the rotated frame, RTQ's high-rate Lloyd-Max quantization on the Beta marginal gives noise $\tilde{\boldsymbol{\varepsilon}}_R \sim \mathcal{N}(\mathbf{0}, \sigma_R^2\Lambda)$ with

$$\sigma_R^2 \;=\; C_T \cdot 4^{-b}, \qquad C_T \;=\; \pi\sqrt{3}/2. \tag{IA.15}$$

The market clearing condition decomposes coordinate by coordinate. We henceforth analyze a single rotated coordinate j, dropping the subscript; aggregation across coordinates follows by summation.

Per-coordinate equilibrium: derivation. Within coordinate j (subscript suppressed below), the scalar problem has prior $x \sim \mathcal{N}(0, \lambda)$, informed signal $y_F = x + \varepsilon_F$ with $\varepsilon_F \sim \mathcal{N}(0, \tau^2\lambda)$ and finite $\tau > 0$, RTQ signal $y_R = x + \varepsilon_R$ with $\varepsilon_R \sim \mathcal{N}(0, \sigma_R^2\lambda)$, noise supply $n \sim \mathcal{N}(0, \sigma_n^2\lambda)$, and zero net supply. Conjecture a linear price $p = \eta_x x + \eta_n n$. Denote prior precision $\alpha_x = 1/\lambda$, signal precisions $\alpha_F = 1/(\tau^2\lambda)$ and $\alpha_R = 1/(\sigma_R^2\lambda)$, and the price-precision induced by p,

$$\alpha_p \;\equiv\; \frac{\eta_x^2}{\eta_n^2\sigma_n^2\lambda}.$$

Type $t \in \{F, R\}$ has Gaussian posterior with precision $P_t = \alpha_x + \alpha_t + \alpha_p$ and mean $m_t = P_t^{-1}(\alpha_t y_t + \alpha_p p/\eta_x)$. CARA demand at risk aversion γ is $d_t = (m_t - p)P_t/\gamma$. After substitution and routine algebra:

$$d_t \;=\; \frac{1}{\gamma}\left[\alpha_t y_t \;-\; (\alpha_x + \alpha_t)p \;+\; \alpha_p \cdot \frac{1-\eta_x}{\eta_x} \cdot p\right]. \tag{IA.16}$$

Continuum aggregation over each type eliminates the cross-sectional noise terms ($\int_{i\in\text{type F}} \varepsilon_{F,i}\, di = 0$ and $\int_{i\in\text{type R}} \varepsilon_{R,i}\, di = 0$ in the per-capita aggregate, by the standard exact-law-of-large-numbers convention for a continuum of i.i.d. agents (see Sun, 2006)). Market clearing $\alpha d_F + (1-\alpha)d_R + n = 0$ then yields, after matching coefficients on x and n in the conjectured linear price:

$$(\text{coef. on } x): \quad A \;-\; B\eta_x + \alpha_p\,(1-\eta_x) \;=\; 0, \tag{IA.17}$$

$$(\text{coef. on } n): \quad -\; B\eta_n + \alpha_p\,\frac{1-\eta_x}{\eta_x}\,\eta_n + \gamma \;=\; 0, \tag{IA.18}$$

where $A \equiv \alpha\alpha_F + (1-\alpha)\alpha_R$ is the aggregate signal precision and $B \equiv \alpha_x + A$.

Theorem IA.3 (Existence and uniqueness of the linear equilibrium)**.** *For any positive parameter tuple $(\tau, \sigma_R, \sigma_n, \gamma, \lambda)$ and $\alpha \in (0,1)$, there exists a unique linear-Gaussian equilibrium with price $p = \eta_x x + \eta_n n$. The coefficients are*

$$\eta_x \;=\; \frac{A\,(A + \gamma^2\sigma_n^2\lambda)}{A\,(A + \gamma^2\sigma_n^2\lambda) + \gamma^2\sigma_n^2}, \qquad \eta_n \;=\; \frac{\gamma\,\eta_x}{A}, \tag{IA.19}$$

with $A = \alpha/(\tau^2\lambda) + (1-\alpha)/(\sigma_R^2\lambda)$. The induced price precision is $\alpha_p = A^2/(\gamma^2\sigma_n^2\lambda)$.

Proof. From (IA.17), $\alpha_p(1-\eta_x) = B\eta_x - A = (\alpha_x + A)\eta_x - A$, so $\alpha_p = (\eta_x\alpha_x - A(1-\eta_x))/(1-\eta_x) = \eta_x\alpha_x/(1-\eta_x) - A$. From (IA.18), substituting $\alpha_p(1-\eta_x)/\eta_x = (\eta_x\alpha_x - A(1-\eta_x))/\eta_x = \alpha_x - A(1-\eta_x)/\eta_x$, the equation reduces to $-B\eta_n + (\alpha_x - A(1-\eta_x)/\eta_x)\eta_n + \gamma = 0$, which simplifies (using $B = \alpha_x + A$) to $-A\eta_n - A(1-\eta_x)/\eta_x \cdot \eta_n + \gamma = 0$. Combining

the two coefficient equations yields the closed-form result. The definition of α_p gives the consistency condition $\alpha_p = \eta_x^2/(\eta_n^2\sigma_n^2\lambda)$. Substituting $\eta_n = \gamma\eta_x/A$ from the second equation: $\alpha_p = \eta_x^2/(\gamma^2\eta_x^2\sigma_n^2\lambda/A^2) = A^2/(\gamma^2\sigma_n^2\lambda)$. Equating with $\eta_x\alpha_x/(1-\eta_x) - A$ and solving for η_x gives (IA.19). Existence is by construction: the closed-form (η_x, η_n) in (IA.19) satisfies both coefficient equations and the consistency condition for α_p. Uniqueness follows because after substituting $\eta_n = \gamma\eta_x/A$ and $\alpha_p = A^2/(\gamma^2\sigma_n^2\lambda)$ into the first relation, we obtain a single equation of the form $\eta_x/(1-\eta_x) = K$ with $K = A(A + \gamma^2\sigma_n^2\lambda)/(\gamma^2\sigma_n^2\lambda\alpha_x) > 0$ a positive constant in the model parameters; the function $\eta_x \mapsto \eta_x/(1-\eta_x)$ is strictly increasing on $(0,1)$ with range $(0,\infty)$, so the equation has the unique solution $\eta_x = K/(K+1) \in (0,1)$, and (η_n, α_p) are then determined uniquely. □

Limits and comparative statics. The closed form (IA.19) delivers the following:

- As $\tau \to 0^+$ at fixed $\alpha > 0$: $A \to \infty$, so $\eta_x \to 1$ and $\eta_n \to 0$. Price fully reveals x (Grossman-Stiglitz limit; RTQ agents free-ride on price information and posterior variance vanishes).
- As $\alpha \to 0$ at fixed $\sigma_R^2 > 0$ and $\tau > 0$: $A = \alpha/(\tau^2\lambda) + (1-\alpha)/(\sigma_R^2\lambda) \to 1/(\sigma_R^2\lambda)$, finite. Equilibrium price partially reveals x through the RTQ signal aggregate.
- As $B \to \infty$ (equivalently $\sigma_R^2 \to 0$): $A \to \infty$, $\eta_x \to 1$, $\eta_n \to 0$ regardless of τ, α.
- η_x is monotone increasing in A, hence in B (at fixed α, τ) directly through α_R. Monotonicity in α at fixed (τ, σ_R) requires Type F to be more precise than Type R, $\alpha_F > \alpha_R$ (equivalently $\tau^2 < \sigma_R^2$). This is the natural ordering when Type F is called full-information (τ small) and Type R is RTQ-bounded (σ_R finite). Under this condition, more informed counterparties produce more revealing prices.

Type R's posterior and welfare loss. The Type R posterior variance per coordinate is

$$\mathrm{Var}(x \mid y_R, p) \;=\; (\alpha_x + \alpha_R + \alpha_p)^{-1} \;=\; \frac{1}{\alpha_x + 1/(\sigma_R^2\lambda) + A^2/(\gamma^2\sigma_n^2\lambda)}. \tag{IA.20}$$

Substituting $\alpha_x = 1/\lambda$, $\sigma_R^2 = C_T \cdot 4^{-B/d}$, and A from Theorem IA.3 yields a closed form in $(\alpha, B, \tau, \sigma_n^2, \gamma, \lambda)$. Aggregating across coordinates, the total welfare loss to a Type R agent in certainty-equivalent terms is

$$\Delta W_R \;=\; \frac{1}{2\gamma}\sum_{j=1}^{d} w_j \cdot \mathrm{Var}(x_j \mid y_{R,j}, p_j), \tag{IA.21}$$

where $w_j = (\tilde{W})_{jj}$ is the jth diagonal entry of $\tilde{W} = V^\top W V$ in the rotated basis. The next proposition shows that off-diagonal entries of $\tilde{W}$ do not enter the welfare loss under eigenbasis-aligned RTQ, so the framework is valid for arbitrary positive definite W regardless of whether W commutes with Σ.

Proposition IA.7 (Welfare loss under non-commuting W)**.** *Let $W \succ 0$ and $\Sigma \succ 0$ be arbitrary, not necessarily commuting. Under eigenbasis-aligned RTQ with rotation V from the eigendecomposition $\Sigma = V\Lambda V^\top$, the per-coordinate posterior covariance in the rotated frame is diagonal with entries* $\mathrm{Var}(\tilde{x}_j \mid \tilde{y}_{R,j}, \tilde{p}_j)$ *given by* (IA.20)*. The total welfare loss is*

$$\Delta W_R \;=\; \frac{1}{2\gamma} \sum_{j=1}^{d} (\tilde{W})_{jj} \cdot \mathrm{Var}(\tilde{x}_j \mid \tilde{y}_{R,j}, \tilde{p}_j), \tag{IA.22}$$

and depends on W only through the diagonal of $\tilde{W} = V^\top W V$, not on any off-diagonal entries.

Proof. Let $\Sigma_\varepsilon = \mathrm{Cov}(\mathbf{x} - \hat{\mathbf{x}})$ be the posterior error covariance in the original frame, where $\hat{\mathbf{x}}$ is the agent's Bayes-optimal action. Under eigenbasis-aligned RTQ with rotation V, the posterior error covariance is diagonal in the rotated frame: $V^\top \Sigma_\varepsilon V = D$ where D is diagonal, because all inputs to the posterior (prior Σ, RTQ signal noise, equilibrium price, and noise-trader covariance) factor across rotated coordinates, making rotated coordinates mutually independent given the signals. The quadratic loss is

$$\mathbb{E}[(\hat{\mathbf{x}}-\mathbf{x})^\top W (\hat{\mathbf{x}}-\mathbf{x})] \;=\; \mathrm{tr}(W\Sigma_\varepsilon) \;=\; \mathrm{tr}(WVDV^\top) \;=\; \mathrm{tr}(V^\top W V \cdot D) \;=\; \mathrm{tr}(\tilde{W}D) \;=\; \sum_j (\tilde{W})_{jj} D_{jj},$$

where the last equality uses the diagonality of D. Off-diagonal entries of $\tilde{W}$ multiply zero entries of D in the trace, so they do not enter the loss. The certainty-equivalent welfare loss scales the quadratic loss by $1/(2\gamma)$ under CARA preferences. Identifying D_{jj} with the per-coordinate posterior variance and applying the CARA prefactor gives (IA.22). □

Proposition IA.7 shows that the eigenbasis-aligned RTQ design produces a welfare loss that depends on W only through $\mathrm{diag}(V^\top W V)$. The off-diagonal entries are irrelevant given that the posterior is diagonal in the chosen rotation basis. The choice of V_Σ delivers the cleanest equilibrium decomposition because it diagonalizes the prior. The welfare-optimal rotation V_M (with $M = \Sigma^{1/2} W \Sigma^{1/2}$) diagonalizes the welfare-weighted prior instead; both rotations are feasible under the main-text bound, but in equilibrium changing the rotation also changes the price-information fixed point, so the cost of choosing V_Σ over V_M depends on the equilibrium structure rather than directly on the static $C_T \chi(W)$ constant.

Comparative statics. From (IA.20): $\mathrm{Var}(x \mid y_R, p)$ is strictly decreasing in (i) α at fixed B provided $\alpha_F > \alpha_R$ (equivalently $\tau^2 < \sigma_R^2$, so Type F is more precise than Type R), (ii) B at fixed α (through $1/(\sigma_R^2\lambda)$ increasing and α_p increasing), and (iii) the fraction of informed agents α, again under $\alpha_F > \alpha_R$, through both the direct RTQ-precision channel and the indirect price-revelation channel. The third statement quantifies the free-riding result: low-attention investors benefit from the price information generated by high-attention counterparties, an effect that the worst-case bound of Proposition 1 of the main text does not capture but the equilibrium analysis does.

Relation to the main-text bound and to learned attention. Equation (IA.21) is the Gaussian-CARA closed-form analog of the worst-case bound $C_T\chi(W) \cdot \mathcal{L}_{\mathrm{lb}}^{\mathrm{sphere}}$ from Proposition 1 of the main text. Under the Gaussian-CARA equilibrium the welfare loss is exact rather than worst-case, and it decomposes into a per-agent RTQ noise contribution and a price-information contribution. For non-Gaussian priors, the worst-case bound continues to hold by Proposition 1, and the equilibrium prices satisfy a fixed-point characterization analogous to (IA.19) but without the explicit Gaussian closed form. An empirical reading speaks to current debates about indexing and AI-driven trading: as the fraction of well-informed (e.g., learned-attention) traders grows, prices become more informative, the welfare cost of being attention-bounded falls, and the equilibrium tilts toward passive participation. Internet Appendix IA.4 formalizes the corresponding constant-factor relationship between RTQ and the asymptotically optimal linear-attention layer.

Scope. The closed form above relies on Gaussian payoffs, CARA preferences, equal-allocation RTQ, and finite $\tau > 0$, with $\tau \to 0^+$ as a limiting case. Extensions to water-filling across rotated coordinates, non-Gaussian priors, heterogeneous bit budgets across agents, multiple risk-aversion levels, dynamic learning, and informed-trading strategy are substantial and are deferred to follow-up work. The framework here is intended to establish the qualitative structure and the closed-form benchmark; full generality belongs to a separate paper.

IA.5.2 Heterogeneous bit budgets and risk aversion

Extend the baseline to $K+1$ agent types indexed $k \in \{0, 1, \ldots, K\}$ with mass α_k, $\sum_k \alpha_k = 1$. Type 0 is full-information ($\tau \to 0^+$ as in the baseline). Types $k = 1, \ldots, K$ are RTQ-bounded with bit budget B_k (per-coordinate $b_k = B_k/d$) and risk aversion γ_k. Noise traders, prior, and net supply are unchanged.

The eigenbasis diagonalization of the baseline carries through unchanged: the equilibrium decomposes coordinate by coordinate. Within rotated coordinate j, type k receives RTQ

signal $y_{R,k} = x + \varepsilon_{R,k}$ with $\varepsilon_{R,k} \sim \mathcal{N}(0, \sigma^2_{R,k}\lambda)$ and $\sigma^2_{R,k} = C_T \cdot 4^{-b_k}$. Type k's posterior precision given $(y_{R,k}, p)$ is

$$\text{Prec}_k \;=\; \frac{1}{\lambda} + \frac{1}{\sigma^2_{R,k}\lambda} + \tau_p, \qquad \tau_p \;=\; \frac{\eta_x^2}{\eta_n^2\sigma_n^2\lambda},$$

where the price precision τ_p is the same across types (everyone observes the same price). CARA demand from type k at risk aversion γ_k:

$$d_k \;=\; \gamma_k^{-1} \cdot \text{Prec}_k \cdot (E_k[x \mid y_{R,k}, p] - p).$$

Market clearing requires $\alpha_0 d_F + \sum_{k=1}^K \alpha_k d_k + n = 0$, where d_F is the Type 0 demand.

Corollary IA.2 (Equilibrium under heterogeneous bit budgets and risk aversion)**.** *Define the two aggregates*

$$A_\gamma \;\equiv\; \sum_{k=0}^{K} \frac{\alpha_k\, \alpha_k^{\text{sig}}}{\gamma_k}, \qquad G \;\equiv\; \sum_{k=0}^{K} \frac{\alpha_k}{\gamma_k},$$

with $\alpha_0^{\text{sig}} = 1/(\tau^2\lambda)$ and $\alpha_k^{\text{sig}} = 1/(\sigma^2_{R,k}\lambda) = 1/(C_T 4^{-B_k/d}\lambda)$ for $k \geq 1$. There exists a unique linear-Gaussian equilibrium with price $p = \eta_x x + \eta_n n$, and the coefficients have the closed form

$$\eta_x \;=\; \frac{A_\gamma\,(A_\gamma + \sigma_n^2\lambda/G)}{A_\gamma\,(A_\gamma + \sigma_n^2\lambda/G) + \sigma_n^2}, \qquad \eta_n \;=\; \frac{\eta_x}{A_\gamma}. \tag{IA.23}$$

The induced price precision is $\alpha_p = A_\gamma^2/(\sigma_n^2\lambda)$.

Proof. The derivation parallels that of Theorem IA.3 with two modifications. Demand from type k is $d_k = \gamma_k^{-1}[\alpha_k^{\text{sig}} y_k - (\alpha_x + \alpha_k^{\text{sig}})p + \alpha_p(1-\eta_x)p/\eta_x]$. Aggregating $\sum_k \alpha_k d_k + n = 0$ produces two distinct aggregates: A_γ on the signal-precision side and G on the prior-precision and price-precision side. The coefficient on x in market clearing reads $A_\gamma - \eta_x(\alpha_x G + A_\gamma) + G\alpha_p(1-\eta_x) = 0$, and on n reads $-\eta_n(\alpha_x G + A_\gamma) + G\eta_n\alpha_p(1-\eta_x)/\eta_x + 1 = 0$. After substitution the algebra simplifies analogously to the proof of Theorem IA.3, yielding $\eta_n = \eta_x/A_\gamma$ and the closed form (IA.23). The induced price precision $\alpha_p = \eta_x^2/(\eta_n^2\sigma_n^2\lambda) = A_\gamma^2/(\sigma_n^2\lambda)$.

The homogeneous case (Theorem IA.3) is recovered by setting $\gamma_k = \gamma$ for all k: then $A_\gamma = A/\gamma$ and $G = 1/\gamma$, so $\sigma_n^2\lambda/G = \gamma\sigma_n^2\lambda$, and substituting reproduces the formulas in (IA.19). □

The equilibrium price aggregates information from all types, weighted by signal precision per unit of risk aversion ($\alpha_k^{\text{sig}}/\gamma_k$). Both aggregates A_γ and G are the sufficient statistics for the heterogeneous-types economy; under homogeneous risk aversion the two collapse to scaled versions of the single A of Theorem IA.3.

Higher bit budget B_k raises type k's signal precision α_k^{sig} and lowers the posterior variance (holding equilibrium prices fixed), so welfare loss falls monotonically in B_k. The effect of lower risk aversion γ_k is ambiguous for individual welfare loss: lower γ_k raises the trading aggressiveness of type k and shifts price informativeness through A_γ, but the certainty-equivalent welfare loss also scales with $1/(2\gamma_k)$, so lower γ_k mechanically raises the loss expressed in CE-bps units. The net ranking across types in γ_k therefore requires additional restrictions on the relative magnitudes of the price-informativeness channel and the CE-scaling channel; we leave this for follow-up work.

The welfare loss to type k is the per-coordinate posterior variance times the welfare weight, summed across rotated coordinates and scaled by inverse risk aversion. From (IA.20), type k's posterior variance is

$$\text{Var}_k(x \mid y_{R,k}, p) \;=\; \frac{\sigma_{R,k}^2 \lambda}{1 + \sigma_{R,k}^2 + \sigma_{R,k}^2 \lambda \tau_p},$$

with τ_p from (IA.23). The welfare ranking is monotone in the bit budget: higher B_k lowers type k's posterior variance and lowers welfare loss, holding prices fixed. The cross-type ranking in γ_k is ambiguous, as noted above: lower γ_k raises trading aggressiveness and price informativeness through A_γ, but CE welfare loss also scales with $1/(2\gamma_k)$, so the two effects pull in opposite directions. The free-riding pattern of Grossman and Stiglitz (1980) carries through: low-budget agents benefit from price revelation generated by high-budget agents, with the size of the benefit increasing in the precision-weighted mass of well-informed types.

IA.5.3 Dynamic extension with attention revision over time

The baseline equilibrium is one-shot: the agent commits to a signal structure, observes the signal, takes an action, and incurs loss. We now develop the dynamic extension in which the agent allocates attention across multiple periods, observes signals sequentially, and revises beliefs as new information arrives. Under Gaussian AR(1) dynamics the model admits a closed-form characterization via Kalman filtering, the optimal allocation reduces to a tractable Bellman recursion, and the three predicted features (inertia, sluggish response, budget pooling) follow as derived consequences. We focus on the single-agent problem; embedding the dynamic agent in the heterogeneous-attention equilibrium of Subsection IA.5.2 follows the same logic.

Setup. Discrete time $t = 1, 2, \ldots, T$. The state $x_t \in \mathbb{R}^d$ follows a Gaussian AR(1) process

$$x_t \;=\; \rho\, x_{t-1} + v_t, \qquad v_t \sim \mathcal{N}(\mathbf{0}, Q), \qquad x_0 \sim \mathcal{N}(\mathbf{0}, \Sigma_0), \tag{IA.24}$$

with persistence $|\rho| < 1$ for stationarity and innovation covariance $Q \succ 0$ in the same eigenbasis as Σ_0. The agent has quadratic welfare loss with matrix $W \succ 0$, time discount $\beta \in (0, 1]$, and total bit budget $\bar{B} \geq 0$ to be allocated across periods, $\sum_{t=1}^{T} B_t \leq \bar{B}$ with $B_t \geq 0$. Within each period, the agent applies RTQ to the rotated state at bit budget B_t, observes the quantized signal, and updates beliefs.

High-rate regime. We maintain the high-rate Lloyd-Max approximation throughout: per-period bit allocations are assumed to satisfy $B_t/d \geq \underline{b}$ for a threshold $\underline{b}$ large enough that the Panter-Dite asymptotic MSE $\approx C_T \cdot 4^{-B_t/d} \cdot \mathrm{Var}$ is accurate. In numerical practice $\underline{b} \approx 1$ bit per coordinate suffices; the formulas below break down at very low per-coordinate rates and require a separate low-rate analysis with bounded posterior variance.

Per-period Kalman update. Let $P_{t|t-1} = \mathrm{Var}(x_t \mid \mathcal{F}_{t-1})$ denote the pre-update posterior covariance, where $\mathcal{F}_{t-1}$ is the agent's information set after observing signals through period $t-1$. Diagonalize in the time-invariant eigenbasis V of Σ_0 (and of Q, by assumption), so $P_{t|t-1} = V \Lambda_{t|t-1} V^\top$ with $\Lambda_{t|t-1}$ diagonal.

The RTQ signal at period t, in the rotated frame and under the high-rate Lloyd-Max model on the Beta marginal, is

$$\tilde{y}_t \;=\; \tilde{x}_t + \tilde{\varepsilon}_t, \qquad \tilde{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \Lambda_{t|t-1}), \qquad \sigma_t^2 \;=\; C_T \cdot 4^{-B_t/d}, \quad C_T = \pi\sqrt{3}/2. \tag{IA.25}$$

The Kalman update gives the post-update covariance

$$\Lambda_{t|t} \;=\; \Lambda_{t|t-1} \cdot \phi(B_t), \qquad \phi(B) \;\equiv\; \frac{C_T \cdot 4^{-B/d}}{1 + C_T \cdot 4^{-B/d}}, \qquad B/d \geq \underline{b}, \tag{IA.26}$$

where $\phi(B)$ is the per-period RTQ shrinkage factor on the admissible high-rate range $B/d \geq \underline{b}$. On this range ϕ is strictly positive, strictly decreasing in B, and converges to 0 as $B \to \infty$ (full attention, posterior collapses to a point). The low-rate boundary $B \to 0$ is outside the high-rate regime; under the no-attention convention $\phi(0) = 1$ (posterior unchanged), but the formulas below are stated and used only for $B/d \geq \underline{b}$. The pre-update covariance at $t+1$ follows the standard prediction step:

$$\Lambda_{t+1|t} \;=\; \rho^2 \, \Lambda_{t|t} + Q \;=\; \rho^2 \phi(B_t) \, \Lambda_{t|t-1} + Q. \tag{IA.27}$$

Equations (IA.26) and (IA.27) together fully characterize the dynamics of the agent's posterior covariance as a function of the bit allocation $\{B_t\}$.

Stationary covariance: pre- and post-update fixed points. The recursion (IA.26)-(IA.27) has two distinct stationary objects, which we distinguish:

$$\Lambda_\infty^{\text{pre}}(B) \;=\; \frac{Q}{1-\rho^2\phi(B)}, \qquad \Lambda_\infty^{\text{post}}(B) \;=\; \phi(B)\cdot\Lambda_\infty^{\text{pre}}(B) \;=\; \frac{\phi(B)\,Q}{1-\rho^2\phi(B)}. \tag{IA.28}$$

Under a constant allocation $B_t = B$, the pre-update covariance $\Lambda_{t|t-1}$ converges to $\Lambda_\infty^{\text{pre}}(B)$ and the post-update covariance $\Lambda_{t|t}$ converges to $\Lambda_\infty^{\text{post}}(B)$. The Bellman state in (IA.30) is the pre-update covariance, so $\Lambda_\infty^{\text{pre}}$ is the relevant fixed point for the state recursion; the per-period welfare loss $\text{tr}(\tilde{W}\,\Lambda_{t|t})$ involves the post-update covariance, which equals $\phi(B)\text{tr}(\tilde{W}\,\Lambda_\infty^{\text{pre}})$ in stationarity. Both are continuous and monotone-decreasing in B under the high-rate regime ($B/d \geq \underline{b} > 0$), with $\Lambda_\infty^{\text{post}}(\infty) = 0$ as $B \to \infty$.

Bellman recursion for optimal allocation. The agent minimizes total discounted welfare loss

$$\mathcal{L}_{\text{total}}(\{B_t\}) \;=\; \sum_{t=1}^{T}\beta^{t-1}\cdot\text{tr}\Big(\tilde{W}\,\Lambda_{t|t}\Big) \;=\; \sum_{t=1}^{T}\beta^{t-1}\cdot\phi(B_t)\cdot\text{tr}\Big(\tilde{W}\,\Lambda_{t|t-1}\Big)\,, \tag{IA.29}$$

subject to $\sum_t B_t \leq \bar{B}$, $B_t \geq 0$, and the dynamics (IA.26)-(IA.27); here $\tilde{W} = V^\top W V$ is the welfare matrix in the rotated frame. Let $V_t(\Lambda, b)$ denote the optimal cost-to-go starting at period t with pre-update covariance $\Lambda_{t|t-1} = \Lambda$ and remaining budget $b = \bar{B} - \sum_{s<t} B_s$. The Bellman equation is

$$V_t(\Lambda, b) \;=\; \min_{0\leq B_t\leq b}\;\; \phi(B_t)\cdot\text{tr}(\tilde{W}\Lambda) + \beta\cdot V_{t+1}\big(\rho^2\phi(B_t)\Lambda + Q,\; b - B_t\big)\,, \tag{IA.30}$$

with terminal condition $V_{T+1} \equiv 0$. The first-order condition with respect to B_t gives the Euler equation linking marginal welfare savings at t to the resulting reduction in V_{t+1} through the predict step. Because ϕ is decreasing in B_t, the marginal benefit of an extra bit is largest where the pre-update covariance $\Lambda_{t|t-1}$ is largest in the welfare-weighted sense.

Optimal stationary policy in the infinite-horizon case.

Proposition IA.8 (Optimal stationary allocation in the infinite-horizon discounted case)**.** *Consider the infinite-horizon problem ($T = \infty$) with discount factor $\beta \in (0,1)$, stationary AR(1) dynamics (ρ, Q), and the high-rate regime maintained for the bit allocation. The agent's stationary problem is closed either by a per-period information cost μB with shadow price $\mu > 0$ on the bit budget, or equivalently by an exogenous per-period budget $\bar{B}_{\text{pp}} > 0$ that pins $B \equiv \bar{B}_{\text{pp}}$ with μ as the shadow value of relaxing that constraint; without one of these*

closures the welfare-loss minimization sends $B \to \infty$. Within the class of stationary (time-invariant) policies $B_t \equiv B \geq 0$, and given $\mu > 0$ (whether structural or as the shadow value of a fixed per-period budget), the optimal stationary allocation B^* solves the Euler condition

$$\left|\phi'(B^*)\right| \cdot \frac{\mathrm{tr}\Big(\tilde{W}\,\Lambda_\infty^{\mathrm{pre}}(B^*)\Big)}{1-\beta\rho^2\phi(B^*)} \;=\; \mu, \tag{IA.31}$$

where $\Lambda_\infty^{\mathrm{pre}}(B) = Q/(1-\rho^2\phi(B))$ is the pre-update stationary covariance from (IA.28)*. On the high-rate range where the per-period loss $g(B) = \phi(B)\mathrm{tr}(\tilde{W}\Lambda_\infty^{\mathrm{pre}}(B))/(1-\beta\rho^2\phi(B))$ is strictly convex (Lemma IA.2 below), the Euler condition pins down a unique B^*. Under stationary initial conditions $\Lambda_{1|0} = \Lambda_\infty^{\mathrm{pre}}(B^*)$, the constant policy $B_t \equiv B^*$ attains the steady-state per-period welfare loss $\phi(B^*)\mathrm{tr}(\tilde{W}\,\Lambda_\infty^{\mathrm{pre}}(B^*))$ at every period.*

Proof. With stationary dynamics and stationary cost, the Bellman recursion admits a time-invariant value function $V_\infty(\Lambda)$ on the pre-update state Λ in the infinite-horizon case. Substituting the linear ansatz $V_\infty(\Lambda) = c + \mathrm{tr}(P\Lambda)$ into the Bellman equation $V_\infty(\Lambda) = \phi(B)\mathrm{tr}(\tilde{W}\Lambda) + \beta V_\infty(\rho^2\phi(B)\Lambda + Q)$ and matching coefficients on Λ yields $P = \phi(B)\tilde{W}/(1-\beta\rho^2\phi(B))$. The first-order condition with respect to B, evaluated at the stationary pre-update covariance $\Lambda_\infty^{\mathrm{pre}}(B^*)$, gives

$$\phi'(B^*)\mathrm{tr}(\tilde{W}\Lambda_\infty^{\mathrm{pre}}(B^*)) + \beta\rho^2\phi'(B^*)\mathrm{tr}(P\Lambda_\infty^{\mathrm{pre}}(B^*)) + \mu \;=\; 0.$$

Substituting $P = \phi(B^*)\tilde{W}/(1-\beta\rho^2\phi(B^*))$ and collecting:

$$\phi'(B^*)\mathrm{tr}(\tilde{W}\Lambda_\infty^{\mathrm{pre}}(B^*))\left[1 + \frac{\beta\rho^2\phi(B^*)}{1-\beta\rho^2\phi(B^*)}\right] = -\mu,$$

and simplifying $1+\beta\rho^2\phi/(1-\beta\rho^2\phi) = 1/(1-\beta\rho^2\phi)$ gives (IA.31) after taking absolute values (since $\phi'(B^*) < 0$). Uniqueness on the high-rate range follows from Lemma IA.2. □

Lemma IA.2 (Strict convexity of stationary loss on the high-rate range)**.** *Define $g(B) \equiv \phi(B)\mathrm{tr}(\tilde{W}\,\Lambda_\infty^{\mathrm{pre}}(B))/(1-\beta\rho^2\phi(B))$ on the high-rate range $B/d \geq \underline{b}$. There exists $\underline{b}^* \geq \underline{b}$ such that g is strictly convex on $\{B : B/d \geq \underline{b}^*\}$, hence the Euler condition* (IA.31) *has at most one stationary minimizer on that range.*

Proof. Write $u = \phi(B)$; on the high-rate range ϕ is a strictly decreasing, strictly convex, and analytic function of B with image $u \in (0, u_{\max}]$ where $u_{\max} = \phi(\underline{b}d) < 1$. The stationary loss is $g(B) = \mathrm{tr}(\tilde{W}\,Q)\cdot u/[(1-\rho^2 u)(1-\beta\rho^2 u)]$, a smooth function of u. Its second derivative with respect to u is positive on a neighborhood of $u = 0$ and the second derivative with respect to B is $g_{BB} = g_{uu}(\phi')^2 + g_u\phi''$, with both $\phi''(B) > 0$ on the high-rate range and

$g_u = \mathrm{tr}(\tilde{W}Q)\cdot[(1-\rho^2 u)(1-\beta\rho^2 u)+u(\rho^2(1-\beta\rho^2 u)+\beta\rho^2(1-\rho^2 u))]/[(1-\rho^2 u)^2(1-\beta\rho^2 u)^2] > 0$. For u small enough (equivalently, B/d large enough), $g_{BB} > 0$ strictly; let $\underline{b}^*$ be any threshold above which both terms remain positive. Strict convexity on $\{B/d \geq \underline{b}^*\}$ follows, and uniqueness of the Euler-condition root on this range is immediate. □

The proposition restricts to stationary policies and the infinite-horizon case. For finite horizon $T < \infty$, the global optimum is given by the Bellman Euler condition in the next proposition and may deviate from any stationary policy. In short horizons, front-loading attention can be globally optimal: concentrating the bit budget in early periods reduces early posterior variance enough that the welfare savings dominate the cost of higher late-period variance. As T grows, the value of front-loading shrinks and the optimum approaches the stationary B^*. The following proposition characterizes the optimum for any finite T and any initial conditions.

Proposition IA.9 (Bellman characterization under maintained regularity)**.** *The Bellman recursion* (IA.30) *characterizes the finite-horizon optimum. At each* (t, Λ, b) *with* $\Lambda \succ 0$*, under the maintained high-rate convexity conditions of Lemma IA.2 together with the regularity hypothesis that* $V_{t+1}(\cdot, b-B)$ *is non-decreasing in the Loewner order on* Λ *and convex in* b*, the one-step Bellman summand is strictly convex in* B*, so the period-t minimizer* $B_t^*(\Lambda, b)$ *is unique. By Bellman's principle of optimality, the resulting trajectory* $\{B_t^*\}$ *defines the finite-horizon optimum.*

Proof. Under the stated regularity hypothesis on V_{t+1}, the Bellman summand $\phi(B)\mathrm{tr}(\tilde{W}\Lambda)+\beta V_{t+1}(\rho^2\phi(B)\Lambda+Q, b-B)$ is a sum of a strictly convex term (by Lemma IA.2 applied to the first piece) and a convex term in B on the high-rate range. The minimum over the compact convex set $[0, b]$ is therefore attained uniquely. Bellman's principle of optimality then gives the finite-horizon optimum trajectory by per-step minimization. □

Proposition IA.9 gives a per-step uniqueness statement and a computational characterization of the optimum, conditional on the regularity hypothesis on V_{t+1}. We do not establish that hypothesis recursively from $V_{T+1} \equiv 0$, because the minimum over B of a family of convex-in-Λ functions is not in general convex in Λ; verifying that the value function inherits monotonicity and convexity from the per-step problem would require additional structure on the Bellman operator that goes beyond what we use in this paper. The per-step characterization remains sufficient for the substantive uses below: under stationary initial conditions, Proposition IA.8 gives the optimal stationary policy with the cleaner Lemma IA.2 uniqueness argument; under non-stationary initial conditions, the optimal allocation front-loads or back-loads attention depending on whether $\Lambda_{1|0}$ is above or below $\Lambda_\infty^{\mathrm{pre}}(B^*)$, with the trajectory characterized by the Bellman recursion subject to the stated regularity.

Turnpike intuition (informal implication). The finite-T optimum is expected to approach the stationary policy in the interior of long horizons. Under the high-rate regime, ϕ is Lipschitz and bounded, the Kalman recursion has a bounded invariant set for the pre-update covariance (any constant policy keeps $\Lambda_{t|t-1}$ in the compact range $[\Lambda_\infty^{\mathrm{pre}}(\overline{B}), Q/(1-\rho^2)]$), and the discounted Bellman operator on continuous bounded functions over the compact state set is a contraction with modulus $\beta < 1$. The Banach fixed-point theorem then gives $V_T(\cdot, t) \to V_\infty(\cdot)$ uniformly as $T - t \to \infty$. If the Lagrange multiplier sequence on the total budget constraint converges to its stationary value μ^* and the optimal policy is continuous in the value-function derivative (which holds under standard regularity by the implicit-function theorem applied to the strictly convex per-step problem of Proposition IA.9), then for any $\varepsilon > 0$ and $\delta \in (0, 1/2)$ there exists $T_0 = T_0(\varepsilon, \delta)$ such that

$$\left| B_t^{*,T} - B^* \right| \;<\; \varepsilon \qquad \text{for all } T \geq T_0 \text{ and all } t \in [\delta T, (1-\delta)T], \tag{IA.32}$$

at geometric rate in the horizon distance to the boundary. This is the standard turnpike pattern of Cass (1965) and McKenzie (1976) for budget-constrained dynamic programs. We state this property as an informal implication rather than as a theorem: a fully rigorous proof requires additional regularity conditions on the value-function derivative and on the multiplier sequence that go beyond the scope of the present paper, and is left for follow-up work. In the empirical sections we use the Bellman recursion as the computational characterization (Proposition IA.9) and report stationary-policy results for stationary initial conditions (Proposition IA.8).

Three derived features. The Bellman recursion produces three qualitative features observed in the dynamic-RI literature.

Inertia. Equation (IA.26) shows that the post-update covariance is the pre-update covariance scaled by $\phi(B_t) \in (0, 1]$, not replaced. The agent's belief at t is a precision-weighted average of the new RTQ signal and the carry-over from $t-1$, with weight $\rho^2\phi(B_{t-1})$ on the past and weight $1 - \phi(B_t)$ on the new signal. As $\rho \to 1$ (high persistence), the past weight rises and the agent's belief moves slowly. This is the formal sense in which RTQ reuses past attention rather than re-rotating each period.

Sluggish response. A structural break at $t = t^*$ that shifts Q to $Q' \neq Q$ propagates through (IA.27) only after the agent observes signals reflecting the new innovation variance. The pre-update covariance $\Lambda_{t^*+k|t^*+k-1}$ converges to the new stationary pre-update value $\Lambda_\infty^{\mathrm{pre},\prime}(B^*) = Q'/(1 - \rho^2\phi(B^*))$ at geometric rate $(\rho^2\phi(B^*))^k$; for small B^* or high ρ, this convergence is slow. The agent's belief catches up to the new regime sluggishly, with half-life

$-\log 2/\log(\rho^2\phi(B^*))$ periods.

Budget pooling. In the infinite-horizon discounted case with stationary dynamics, the optimal stationary policy is constant per-period allocation $B_t^* = B^*$ (Proposition IA.8); the agent does not front-load or back-load. In finite T, the optimum can deviate from equal allocation, with front-loading favored when initial uncertainty $\Lambda_{1|0}$ is above the stationary pre-update covariance and back-loading favored when it is below. The turnpike intuition above implies that for sufficiently long horizons the interior of the policy approaches the stationary B^*, with the front-/back-loading effects confined to the boundary periods.

Connection to the static framework. Under the stationary policy and stationary initial conditions, the dynamic problem reduces to a sequence of static RTQ problems with per-period budget $\bar{B}/T$ and per-period welfare-weighted covariance $\tilde{W}\,\Lambda_\infty^{\text{post}}(\bar{B}/T)$, equivalent to $\phi(\bar{B}/T)\tilde{W}\,\Lambda_\infty^{\text{pre}}(\bar{B}/T)$. The parallel-target sufficient statistic of Section IA.2 generalizes from the static $\xi = N\cdot 4^{-B/(Nd)}$ to the dynamic $\xi_{\text{dyn}} = T\cdot 4^{-\bar{B}/(Td)}$, with the persistence parameter entering through the stationary covariance scaling $1/(1-\rho^2\phi)$.

Corollary IA.3 (Dynamic loss bound under equal allocation, Gaussian AR(1))**.** *Under Gaussian AR(1) state dynamics with innovation covariance Q, equal allocation $B_t \equiv \bar{B}/T$, and stationary initial conditions $\Lambda_{1|0} = \Lambda_\infty^{\text{pre}}(\bar{B}/T)$, the total welfare loss is bounded by*

$$\mathcal{L}_{\text{total}} \;\le\; C_T\cdot\chi(\tilde{W})\cdot\frac{1-\beta^T}{1-\beta}\cdot\mathcal{L}_{\text{lb}}^{\text{sphere}}(\bar{B}/T), \tag{IA.33}$$

where $\mathcal{L}_{\text{lb}}^{\text{sphere}}(b)$ is the static minimax rate-distortion floor at per-coordinate bit rate b/d from Proposition 1 of the main text.

Proof. Under the equal-allocation policy $B_t \equiv \bar{B}/T$, applying the Kalman recursion (IA.26)-(IA.27) starting from $\Lambda_{1|0} = \Lambda_\infty^{\text{pre}}(\bar{B}/T)$ gives $\Lambda_{t|t-1} = \Lambda_\infty^{\text{pre}}(\bar{B}/T)$ for all t: the stationary pre-update covariance is preserved by the constant-allocation policy, by direct substitution into (IA.26)-(IA.27). Proposition 1 of the main text, applied to a single period at per-coordinate bit rate $\bar{B}/(Td)$, gives $\phi(\bar{B}/T)\cdot\operatorname{tr}(\tilde{W}\,\Lambda_\infty^{\text{pre}}(\bar{B}/T)) \;\le\; C_T\chi(\tilde{W})\cdot\mathcal{L}_{\text{lb}}^{\text{sphere}}(\bar{B}/T)$. Summing the discounted per-period losses $\sum_{t=1}^{T}\beta^{t-1} = (1-\beta^T)/(1-\beta)$ gives the bound. □

Corollary IA.3 bounds the dynamic welfare loss under equal allocation in the Gaussian AR(1) case, where the Kalman recursion gives the closed-form posterior and the stationary covariance has the closed form $\Lambda_\infty^{\text{pre}}(B) = Q/(1-\rho^2\phi(B))$. For non-Gaussian innovations with finite second moments, the static main-text bound continues to apply period by period to the corresponding posterior error covariance, but the closed-form Kalman recursion and the

stationary covariance formula do not carry over; the dynamic worst-case bound in the non-Gaussian case requires the per-period covariance trajectory under whatever signal structure the agent uses, which is no longer closed-form. The Gaussian-AR(1) case in (IA.33) is the substantive content.

Comparison to Steiner-Stewart-Matejka and scope. Steiner et al. (2017) characterize dynamic rational inattention under Shannon capacity and find the same three features (inertia, sluggish response, an interior stationary allocation). The development above replaces Shannon capacity with the RTQ rate-distortion bit budget and inherits the static RTQ worst-case guarantee period by period; the closed-form covariance recursion holds for Gaussian AR(1). General non-Gaussian state dynamics inherit the static worst-case bound applied at each period to the corresponding posterior error covariance, but the closed-form Kalman recursion does not carry over. Embedding the dynamic agent in the equilibrium of Subsection IA.5.2 with full-information counterparties and noise traders is a substantial extension and a natural direction for follow-up work; the per-period equilibrium pricing follows the static analysis of Subsection IA.5.1 applied at each t with $\sigma_R^2 = C_T \cdot 4^{-B_t/d}$ and posterior covariance from (IA.26).