COVERAGE-GATED COALITIONAL CODING FOR MULTI-MODAL DISCRETE TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete multimodal codes, such as semantic IDs, neural codecs, unified tokenizers, are distributed: each token observes only a subset of modalities, and synergistic targets vanish unless every constituent is present. Existing analyses typically treat the tokenizer as a centralized quantizer of a fused vector and report a single rate-distortion curve. That formulation conceals an operational constraint. A missing factor is information-theoretically invisible, independent of decoder skill; even after every factor is covered, recovery remains a finite-rate problem. We develop a coverage-gated theory of this constraint. For independent Gaussian factors and a product interaction, local channels with information budgets r_i attain an exact one-letter MMSE given by a product of scalar Gaussian gains. Average-rate codes replace that product by its concave envelope, which for order $k \geq 2$ has a genuine setup phase (0.906 bits per modality at $k = 2$). Allocating a global budget across an access hypergraph of such edges is coalitional water-filling, which strictly dominates order-scalar V/rank rules; finite tokens inherit a $2L\epsilon$ plug-in guarantee. Seeded operational checks recover these predictions. On Flickr8k captions with frozen OpenCLIP residual tokens, exact-budget allocation is compared with MMQ-style shared/specific schedules and a peak-rate analogue of bursting.

1 INTRODUCTION

Recently, multimodal systems have adopted short discrete codes for retrieval, generation, and unified modeling. These methods, ranging from semantic IDs for recommendation (Rajput et al., 2023; Xu et al., 2025; Zheng et al., 2025) to neural codecs (Esser et al., 2020; Zeghidour et al., 2021; D'efossez et al., 2022) and tokenizers for vision–language models (Tang et al., 2026; Zhuang et al., 2026), map continuous modality representations to a handful of K-ary tokens. Shared versus specific codebooks and token-selection heuristics are useful in practice (Dong et al., 2026; Du et al., 2026), but they do not say how a finite token budget should be split when some targets are synergistic products that vanish unless every constituent is observed. We treat that question as a distributed rate–distortion problem. Residual VQ stages correspond to singleton hyperedges; a shared codebook corresponds to an order-k edge whose target is a designed interaction of encoder embeddings—not a claim that CLIP coordinates are independent Gaussians. Independent Gaussian products are the analyzable core that produces a setup slope; on real pairs we allocate from measured curves. A tokenizer that must emit a code on every sample cannot time-share silence: the peak-rate counterpart of bursting is to leave the shared codebook closed rather than drip one joint token.

In the literature, although there is a mature theory of functional computation (Orlitsky & Roche, 2001; Misra et al., 2011), remote and CEO Gaussian coding (Wolf & Ziv, 1970; Oohama, 1998; Cover & Thomas, 2006), and vector quantization (Fang et al., 2026; Chen, 2025), the prevailing analysis of a multimodal tokenizer treats the code as a centralized quantizer of a fused vector. That formulation hides two facts. First, a token is an *access-constrained* observation: it is produced from some subset of modalities, and a product interaction that depends on a missing factor is information-theoretically invisible. Coverage is therefore not a Boolean interval between "impossible" and "decoder-dependent success," but a Hilbert-space floor on any subsequent code (Wolf & Ziv, 1970). Second, once every factor is present, recovery is still a *finite-rate* problem. Interaction diagnostics such as partial information decomposition (Liang et al., 2023; Zhao et al., 2025) and the Gaussian information bottleneck (Chechik et al., 2005) measure synergy; they do not prescribe an operational budget among hyperedges of different order.

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

To address these issues we rebuild the problem around an access hypergraph of product targets. For independent standard Gaussians and $Y = \sqrt{v}\prod_{i=1}^{k} X_i$, any product-kernel family of local auxiliaries with $I(X_i; U_i) \le r_i$ satisfies

$$\mathrm{mmse}(Y \mid U_1, \dots, U_k) = v\Big[1 - \prod_{i=1}^{k}\big(1 - 2^{-2r_i}\big)\Big], \tag{1}$$

with equality for Gaussian test channels. One missing or zero-rate factor forces full distortion at every finite rate of the others; complete coverage yields an explicit recovery curve rather than an unspecified interval $[0, v]$. Average-rate block codes can time-share complete coalitions, so the operational distortion is the concave envelope of (1). For transferable rate split equally inside a k-set, the envelope is linear up to a setup threshold solving $e^{x_k} = 1 + kx_k$, and follows the raw product thereafter. For $k = 2$ this threshold is 0.906 bits per modality, matching the equal-rate Gaussian matrix-product transition of Ordentlich & Polyanskiy (2025); for $k \ge 3$ the setup is strictly larger. Allocating a global budget across several such hyperedges is coalitional water-filling. It is not the same as water-filling the traces of concatenated ANOVA vectors, and we prove a strict separation from order-scalar V/rank rules.

A discrete tokenizer, however, assigns an exact integer number of tokens rather than a real rate, and modalities need not be independent. The continuous program therefore converts into exact-budget dynamic programming with a $2L\epsilon$ plug-in guarantee on estimated curves, while incomplete access is controlled by a conditional maximal-correlation leakage floor (Hirschfeld, 1935; Polyanskiy & Wu, 2016; Yu, 2019) that is vacuous exactly when nonlinear leakage is perfect. Seeded Monte Carlo recovers the product-kernel law, the burst-versus-drip envelope, the hypergraph gap, the plug-in bound, and the leakage floor. A Flickr8k OpenCLIP residual-token probe then tests the allocator against MMQ-style shared/specific schedules, with setup-gated versus drip joint codes, on caption retrieval.

Overall, our contributions are as follows: (i) an exact one-letter product-kernel MMSE and its operational separate-block envelope (Theorems 1–2), with the closed-form envelope in Appendix F (Theorem 18); (ii) coalition setup, equal splitting, and edge-separable hypergraph water-filling, with a strict gap to order-scalar allocation (Corollary 3, Theorem 5, Corollary 6); (iii) a plug-in $2L\epsilon$ bound for classical exact-budget DP on estimated curves, and a leakage floor under dependence (Theorems 7 and 8); (iv) hypothesis audits of each claimed-new theorem, plus a Flickr8k frozen-OpenCLIP residual-token allocation study with MMQ-style baselines and a peak-rate bursting analogue (Section 7). Standing definitions and classical primitives are collected in Appendix A; proofs and extra experimental grids follow.

2 Related Work

Distributed functional and Gaussian remote coding. Coding for computing (Orlitsky & Roche, 2001) characterizes lossless function computation on graphs; distributed functional scalar quantization (Misra et al., 2011) designs high-resolution bins for a smooth g. Remote and CEO results (Wolf & Ziv, 1970; Oohama, 1998; Cover & Thomas, 2006) reconstruct a source, or a linear Gaussian mixture, from noisy observations, reducing the problem to coding a conditional expectation. Information-theoretic views of vector, product, and scalar quantization (Fang et al., 2026) and multimodal retrieval rate–distortion (Chen, 2025) likewise concern centralized or linearly mixed targets. The closest product result is the RDF for estimating $A^\top B$ from separately quantized Gaussian matrices (Ordentlich & Polyanskiy, 2025), whose equal-rate $k = 2$ setup we recover. None of these supplies an exact finite-rate MMSE for a k-linear Gaussian product under local mutual-information budgets, an explicit multivariate envelope, or an allocator over a hypergraph of interaction edges. We use the remote-source identity as a coverage decomposition, not as a product law.

Interaction diagnostics. The Gaussian information bottleneck (Chechik et al., 2005) trades compression against a relevance variable. Partial information decomposition (Liang et al., 2023; Zhao et al., 2025) and Sobol-style sensitivity (Sobol', 1990; Saltelli, 2002) measure unique, redundant, and synergistic information. Hirschfeld–Gebelein–Rényi correlation and strong data-processing inequalities (Hirschfeld, 1935; Polyanskiy & Wu, 2016; Yu, 2019) bound nonlinear dependence. Those tools diagnose interaction. They do not prescribe an operational token budget for a product

target, and a raw-covariance robustness claim fails under parity-type leakage. We use the conditional Hilbertian form only as a leakage coefficient, and only after the product-kernel law is in place.

Discrete tokenizers and coverage heuristics. TIGER (Rajput et al., 2023) and MMQ (Xu et al., 2025) instantiate shared versus specific codebooks for semantic IDs; related generative-retrieval designs include UTGR (Zheng et al., 2025). Neural codecs (Esser et al., 2020; Zeghidour et al., 2021; D'efossez et al., 2022) and unified multimodal tokenizers (Tang et al., 2026; Zhuang et al., 2026) discretize continuous signals under a total rate. MMTok and CoIn select vision tokens by set-coverage heuristics (Dong et al., 2026; Du et al., 2026). Those notions of coverage are architectural or subset-selection criteria. Ours is an access constraint on a synergistic rate–distortion law. The Gaussian model says to open a shared (order-k) codebook when the shadow price falls to the setup slope; Section 7 tests that recommendation as an allocator on measured CLIP curves, not as a theorem about MMTok.

3 The Product-Kernel Law and Its Operational Envelope

Fix independent $X_1, \dots, X_k \sim \mathcal{N}(0,1)$ and $Y = \sqrt{v}\prod_{i=1}^k X_i$, with $k \geq 1$ and $v \geq 0$. Certified access is Definition 9; the Hilbert decomposition and scalar Gaussian information–MMSE relation are Lemmas 13–14. The one-letter object is the product gain

$$F_k(\boldsymbol{r}) := \prod_{i=1}^k \left(1 - 2^{-2r_i}\right), \qquad \boldsymbol{r} \in \mathbb{R}_+^k. \tag{2}$$

Theorem 1 (Exact one-letter product-kernel law)**.** *Suppose the local auxiliaries satisfy the product kernel*

$$P_{X_1,\dots,X_k,U_1,\dots,U_k} = \prod_{i=1}^k \big\{\phi(dx_i) P_{U_i|X_i}(du_i \mid x_i)\big\}, \tag{3}$$

where ϕ is standard Gaussian measure. Put $D_i = \mathrm{mmse}(X_i \mid U_i)$. Then

$$\mathrm{mmse}(Y \mid U_1, \dots, U_k) = v\left[1 - \prod_{i=1}^k (1 - D_i)\right], \tag{4}$$

the MMSE decoder is $\widehat{Y}_{\mathrm{B}} = \sqrt{v}\prod_{i=1}^k \mathbb{E}[X_i \mid U_i]$, and

$$\inf_{\substack{(3) \\ I(X_i;U_i) \leq r_i,\ i\in[k]}} \mathrm{mmse}(Y \mid U_1, \dots, U_k) = v\big[1 - F_k(\boldsymbol{r})\big]. \tag{5}$$

Independent Gaussian test channels $U_i = \sqrt{1 - 2^{-2r_i}}\, X_i + 2^{-r_i} N_i$ with $N_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ attain the infimum. If any $r_j = 0$, the optimum is v. At finite positive rates, equality holds if and only if every scalar channel saturates Lemma 14.

The product kernel is necessary. For $k = 2$, an independent Rademacher Q hidden from the decoder and $U_i = Q\,\mathrm{sgn}(X_i)$ gives $I(X_i; U_i) = 0$ yet $\mathbb{E}[X_1 X_2 \mid U_1, U_2] = (2/\pi) U_1 U_2$, so the product MMSE is $v(1 - 4/\pi^2) < v$. The pair shares Q, violating (3); Definition 9 does not replace a product kernel. Finite-valued one-shot auxiliaries cannot attain the nondegenerate Gaussian equality; a common time-sharing variable is excluded from (3) and is the mechanism of the next theorem.

Average-rate block codes can time-share complete coalitions, so F_k is not the operational gain for $k \geq 2$. The object allocated later is the closed envelope $\overline{\mathrm{cav}} F_k$ of Definition 10.

Theorem 2 (Exact asymptotic separate block-coding law)**.** *For each i, encoder i observes X_i^n and sends $M_i \in \mathcal{M}_{i,n}$ with separate kernels (deterministic or independent private randomization) and $\limsup_{n\to\infty} n^{-1} \log_2 |\mathcal{M}_{i,n}| \leq r_i$. The decoder estimates $Y_t = \sqrt{v}\prod_{i=1}^k X_{i,t}$ under average squared error. The exact operational distortion is*

$$D_{\mathrm{sep}}(\boldsymbol{r}) = v\left[1 - \overline{\mathrm{cav}} F_k(\boldsymbol{r})\right]. \tag{6}$$

The converse holds for arbitrary separate vector encoders, not only memoryless test channels. Achievability uses a common deterministic partition into aligned subblocks; it requires no shared encoder

randomness. A per-coordinate peak-rate constraint would change the envelope: time-sharing silence is then forbidden, so the operational object is the raw product F_k rather than $\overline{\mathrm{cav}}F_k$. For a tokenizer that must emit tokens on every sample, the counterpart of bursting is to keep an order-k codebook closed below setup instead of dripping a single shared token.

The envelope has an exact closed form, recorded as Theorem 18 in Appendix F: it vanishes if any $r_i = 0$, coincides with F_k past a contact surface, and is a ray mixture with $\mathbf{0}$ in the setup region. When rate is transferable inside a k-set, that form collapses to a scalar setup threshold—the reason to burst rather than drip, and the quantity later compared across hyperedges.

Corollary 3 (Equal split and coalition setup threshold)**.** *If a total instantaneous rate B is freely transferable among the k local encoders, then*

$$u_k(B) := \max_{\substack{r_i \geq 0 \\ \sum_i r_i \leq B}} F_k(\boldsymbol{r}) = \left(1 - 2^{-2B/k}\right)^k, \tag{7}$$

and for $B > 0$ the unique maximizer is $r_i = B/k$.

For $k = 1$, u_1 is already concave. For $k \geq 2$, let $x_k > 0$ be the unique positive solution of

$$e^{x_k} = 1 + kx_k, \tag{8}$$

and define

$$B_k^0 := \frac{kx_k}{2\ln 2}, \qquad r_k^0 := \frac{x_k}{2\ln 2}.$$

Then

$$\operatorname{cav} u_k(B) = \begin{cases} \dfrac{B}{B_k^0} u_k(B_k^0), & 0 \leq B \leq B_k^0, \\ u_k(B), & B \geq B_k^0. \end{cases} \tag{9}$$

Equivalently,

$$\overline{\mathrm{cav}}F_k(B/k, \ldots, B/k) = \operatorname{cav} u_k(B).$$

Below setup, an optimal average-rate code activates a fraction B/B_k^0 of coordinates at r_k^0 bits per modality and leaves the others inactive.

Remark 4 (Relation to Gaussian matrix-product quantization)**.** *For $k = 2$, the per-modality setup rate is $r_2^0 = x_2/(2\ln 2) \approx 0.906$ bits. This is the same numerical threshold appearing in the equal-rate Gaussian matrix-product rate–distortion function of Ordentlich & Polyanskiy (2025). The present development differs in three operational respects: it treats general k-linear products, it allows unequal local rates and an exact multivariate envelope, and it allocates across a hypergraph of interaction edges rather than a single matrix product.*

4 Coalitional Water-Filling on an Access Hypergraph

A tokenizer typically maintains several interactions at once: unimodal residuals, pairwise shared codes, and higher-order joints. Each interaction is an edge of an access hypergraph, and Theorem 2 supplies that edge's operational gain. We work in the edge-separable architecture of Definition 11: a finite hypergraph $\mathcal{G} = (V, \mathcal{E})$ of product targets $Y_{e,t} = \sqrt{v_e} \prod_{i \in e} X_{i,t}$, a dedicated incidence stream $M_{e,i}$ of rate $r_{e,i}$ for every $i \in e$, and reconstruction of Y_e^n only from $(M_{e,i})_{i \in e}$. For $e \in \mathcal{E}$, write

$$H_e(\boldsymbol{r}_e) := \overline{\mathrm{cav}}F_{|e|}((r_{e,i})_{i \in e}),$$

the envelope of Definition 10 on that edge. Thus H_e is exactly the operational gain of Theorem 2.

Theorem 5 (Edge-stream coalitional water-filling)**.** *In the architecture of Definition 11, impose node budgets $B_i \geq 0$ and a global incidence-rate budget $B_0 \geq 0$:*

$$\sum_{e \ni i} r_{e,i} \leq B_i \quad (i \in V), \qquad \sum_{e \in \mathcal{E}} \sum_{i \in e} r_{e,i} \leq B_0.$$

The exact asymptotic edge-separable explained variance is the concave program

$$\mathcal{G}^*_{\mathrm{edge}} = \max_{\boldsymbol{r} \geq 0} \left\{ \sum_{e \in \mathcal{E}} v_e H_e(\boldsymbol{r}_e) : \textit{the displayed budgets hold} \right\}, \tag{10}$$

and the minimum total distortion is $\sum_e v_e - \mathcal{G}^*_{\mathrm{edge}}$.

Assume the feasible set has nonempty relative interior. A feasible $\boldsymbol{r}^*$ *is optimal if and only if there are node prices* $\alpha_i \geq 0$ *and a global price* $\beta \geq 0$ *such that, with* $p_i = \alpha_i + \beta$,

$$\boldsymbol{r}^*_e \in \arg\max_{\boldsymbol{s}_e \geq 0} \left\{ v_e H_e(\boldsymbol{s}_e) - \sum_{i \in e} p_i s_{e,i} \right\} \quad (e \in \mathcal{E}), \tag{11}$$

and

$$\alpha_i \left(\sum_{e \ni i} r^*_{e,i} - B_i \right) = 0,$$

$$\beta \left(\sum_e \sum_{i \in e} r^*_{e,i} - B_0 \right) = 0.$$

For an edge with all $r^*_{e,i} > 0$, *let* $a = 2 \ln 2$ *and let* A_k, τ *be as in Theorem 18. Define the contact vector*

$$\boldsymbol{c}_e = \begin{cases} \boldsymbol{r}^*_e, & A_{|e|}(\boldsymbol{r}^*_e) \leq 1, \\ \tau(\boldsymbol{r}^*_e)\boldsymbol{r}^*_e, & A_{|e|}(\boldsymbol{r}^*_e) > 1. \end{cases}$$

At every positive incidence, marginal-price equalization reads

$$p_i = \frac{v_e a F_{|e|}(\boldsymbol{c}_e)}{e^{a c_{e,i}} - 1}. \tag{12}$$

At boundary incidences the corresponding supergradient inequality replaces equality.

This theorem is exact only for the explicitly edge-separable architecture. For an unrestricted code that couples several edge sources into one node message, (10) *is an achievable edge-separable benchmark, not an asserted converse.*

The KKT system uses the envelope H_e, not the raw product and not the spectral water-filling of Lemma 17. In particular, an order-k edge remains silent until the global shadow price falls to its setup slope from Corollary 3. Order-scalar rules, which force every incidence of a given cardinality to share one rate, are a strict restriction of (10).

Corollary 6 (Strict separation from order-scalar allocation)**.** *Let an* order-scalar *policy require* $r_{e,i} = q_{|e|}$ *for all incidences belonging to edges of the same cardinality. Whenever such a policy is feasible, the optimum in* (10) *weakly dominates its best value. The dominance can be strict even for singleton edges.*

Specifically, take two singleton edges with only a global budget $B_0 = B > 0$, *weights* $v_1 = 1$ *and* $v_2 = \varepsilon$, *and* $0 < \varepsilon < 2^{-2B}$. *The full edge-specific optimum assigns all rate to edge* 1 *and has gain* $1 - 2^{-2B}$, *whereas the order-scalar policy assigns* $B/2$ *to each and has gain* $(1 + \varepsilon)(1 - 2^{-B})$, *which is strictly smaller.*

5 From Average Rates to Finite Tokens

The envelope program of Theorem 5 is a continuous average-rate law. A discrete tokenizer instead assigns an exact integer number of K-ary tokens. Lemmas 15 and 16 (Appendix A) already give the exact combinatorial optimizer, and they apply to arbitrary measured curves, not only to Gaussian products. Two operational facts remain. First, the setup region of Corollary 3 can make raw product-token gains increase before they diminish, so greedy is not a substitute for the dynamic program unless diminishing returns are verified. Second, the curves used in practice are estimated, so the optimizer needs a plug-in guarantee.

Theorem 7 (Plug-in regret and exact-recovery margins)**.** *Suppose estimated marginal gains satisfy*

$$|\widehat{\Delta}_j(q) - \Delta_j(q)| \leq \epsilon \quad \textit{for all } j,\ q \leq L.$$

Let n^* *maximize the true total gain and let* $\widehat{n}$ *maximize the estimated total gain, each over allocations using at most* L *tokens. Then*

$$\sum_j G_j(n^*_j) - \sum_j G_j(\widehat{n}_j) \leq 2L\epsilon. \tag{13}$$

If the true optimizer is unique and its objective gap to every other feasible allocation is $\gamma > 2L\epsilon$, then every plug-in optimizer equals n^*.

Under the diminishing-gains assumptions of Lemma 16, a sharper structural condition is available. If exactly L tokens are used and the L-th selected true marginal exceeds the largest unselected true marginal by more than 2ϵ, then the plug-in allocation exactly recovers n^*.

The $2L\epsilon$ bound is the only novel guarantee in this section; the dynamic program and the greedy lemma are classical. Together they convert the continuous coalitional program into an exact-budget tokenizer: measure validation curves, run Lemma 15 (or greedy after a diminishing-gain check), and invoke Theorem 7 for estimation error.

6 Dependent Modalities

Sections 3–4 assume independent Gaussian factors, so a missing factor of a product is orthogonal to every proper subset. If modalities are dependent, incomplete access can leak the missing innovation. The product law then ceases to be a converse, but the Hilbert decomposition of Lemma 13 still applies. The contraction of that innovation is the coefficient $\chi(B, C \mid A)$ of Definition 12, the conditional Hilbertian form of maximal correlation (Polyanskiy & Wu, 2016; Yu, 2019), not a substitute for the product RDF. For a target support S and an access set T with $S \nsubseteq T$, write $A = S \cap T$, $B = S \setminus T$, and $C = T \setminus S$.

Theorem 8 (Dependent-modality leakage floor)**.** *Let L_0^2 denote centered square-integrable variables, and let $Y_S \in L_0^2(\sigma(X_S))$ have variance $v_S > 0$. Define*

$$P_A Y_S := \mathbb{E}[Y_S \mid X_A], \qquad Z_S := Y_S - P_A Y_S, \qquad \delta_{S,T} := \frac{\|P_A Y_S\|_2^2}{v_S}.$$

For any representation $U - X_T - Y_S$, with $\chi(B, C \mid A)$ as in Definition 12,

$$\|\mathbb{E}[Y_S \mid U]\|_2^2 \le \big[\delta_{S,T} + \chi(B, C \mid A)^2(1 - \delta_{S,T})\big] v_S, \tag{14}$$

$$\mathrm{mmse}(Y_S \mid U) \ge (1 - \delta_{S,T})\big[1 - \chi(B, C \mid A)^2\big] v_S. \tag{15}$$

If Y_S is hierarchically centered against the overlap ($P_A Y_S = 0$), then $\delta_{S,T} = 0$ and the floor is $(1 - \chi^2)v_S$. Conditional independence $X_B \perp X_C \mid X_A$ gives $\chi = 0$; perfect nonlinear leakage gives $\chi = 1$, where a nontrivial universal floor is impossible.

Under independence of the missing factor, $\chi = 0$ and Theorem 8 recovers the coverage floor of Lemma 13: incomplete access cannot reduce product MMSE. Under parity-type leakage, $\chi = 1$ and no universal floor survives, which is why a raw-covariance robustness theorem is false. Theorem 8 is therefore the boundary of the product-kernel theory, not an extension of F_k to dependent sources.

7 Experiments

7.1 Setup

We test the operational claims of Sections 3–6. Section 7.2 checks the product-kernel MMSE and the burst-versus-drip envelope, including a high-repeat audit of two previously noisy cells. Section 7.3 compares coalitional water-filling with order-scalar and V/rank rules. Section 7.4 checks the plug-in $2L\epsilon$ guarantee. Section 7.5 checks the incomplete-access leakage floor. Section 7.6 is the tokenizer-facing study: frozen OpenCLIP residual tokens on Flickr8k captions, trained interaction heads, MMQ-style baselines, and setup-gated versus drip joint codes. Appendix C records a CIFAR-10 class-name engineering check. Protocols for the main-text experiments are in Appendix B; additional grids follow in Appendix D.

7.2 Product kernel and the operational envelope

Theorem 1 has three operational consequences that a formula identity cannot replace. First, a product kernel is necessary: shared randomness can leak the interaction while every local mutual information is essentially zero. Second, the MMSE decoder is the product of posterior means, not a linear map

Table 1: Theorem 1: kernel, decoder, and channel. Unclipped MSE. Shared-Q uses $U_i = Q\,\mathrm{sgn}(X_i)$ with hidden Rademacher Q. Gaussian rows use independent test channels. Silencing zeros one local rate at companion rates $0.25, 0.5, 1, 2$ bits.

Protocol	Prediction	Empirical	Check
Shared Q, $k = 2$, binned local MI	$1 - 4/\pi^2 = 0.595$	0.584	formula in 95% CI
binned $I(X_i; U_i)$ (bits)	≈ 0	$3{\times}10^{-4}$	no local information
Product kernel at the same zero rates	1	0.985	no local recovery
OLS on (U_1, U_2) under shared Q	bilinear, not linear	0.985	linear decoder fails
Gaussian $k = 2$, $r = (0.5, 0.5)$, Bayes product	0.750	0.747	matches
naive product of observations	$>$ Bayes	0.999	decoder excess
OLS on observations	≈ 1	0.997	not linear
Gaussian $k = 2$, $r = (1, 1)$, Bayes product	0.438	0.439	matches
naive product of observations	$>$ Bayes	0.500	decoder excess
Gaussian $k = 3$, $r = (0.75, 0.75, 0.75)$, Bayes	0.730	0.720	matches
Sign channel, $k = 2$, $I = 1$ bit	0.595	0.595	matches
Gaussian, $k = 2$, $I = 1$ bit	0.438	0.439	beats sign
4-ary quantizer, $k = 2$, nominal 2 bits	$>$ Gaussian 0.121	0.260	no Gaussian equality
Sign channel, $k = 3$, $I = 1$ bit	0.742	0.747	matches
Gaussian, $k = 3$, $I = 1$ bit	0.578	0.581	beats sign
4-ary quantizer, $k = 3$, nominal 2 bits	$>$ Gaussian 0.176	0.366	no Gaussian equality
Zero rate on one factor (12 cells)	1	0.98-1.01	$11/12$ CIs contain 1

Table 2: Burst versus drip. Envelope/raw columns are Theorems 2–18; burst/drip are unclipped Monte Carlo MSE. Setup: 0.906 and 1.373 bits/modality for $k = 2, 3$.

k	r/r_k^0	envelope D	raw D	burst	drip
2	0.25	0.872	0.927	0.878	0.933
2	0.50	0.744	0.782	0.744	0.784
2	1.00	0.488	0.488	0.491	0.490
3	0.25	0.846	0.946	0.849	0.947
3	0.50	0.692	0.769	0.697	0.772
3	1.00	0.384	0.384	0.384	0.383

of the observations. Third, Gaussian test channels strictly beat matched-rate non-Gaussian locals, because only the Gaussian channel saturates the equality condition.

Table 1 shows all three. Hidden Rademacher Q with $U_i = Q\,\mathrm{sgn}(X_i)$ has binned local MI $3{\times}10^{-4}$ bits, yet Bayes MSE 0.584 sits on the formula $1 - 4/\pi^2 = 0.595$. The same zero local rates under a product kernel remain at MSE 0.985, and OLS on (U_1, U_2) cannot see the bilinear leakage. At equal rate 0.5 bits the Bayes product decoder matches 0.750 theory (0.747); the naive product of channel outputs does not (0.999), nor does OLS on the observations (0.997). A sign channel at 1 bit per modality is exactly the non-Gaussian product 0.595; a Gaussian channel at the same mutual information reaches 0.439. The same decoder gap and Gaussian equality gap hold at $r = (1, 1)$, at $k = 3$, and for matched-rate 4-ary quantizers. Silencing is sharp: zeroing any one local rate keeps MSE at 1 in $11/12$ cells.

Theorems 2 and 18 then replace the raw one-letter product by its concave envelope. A burst encoder time-shares the contact profile (a complete coalition on a fraction of samples, silence on the rest). A drip encoder spends the same average rate on every sample. Table 2 is the operational distinction. Below setup, burst tracks the envelope and strictly beats drip, which tracks the raw product; the gap is largest at the smallest fraction of setup ($k = 3$, $r/r_k^0 = 0.25$: burst 0.849 versus drip 0.947). At and above setup the envelope coincides with the raw product, and the two codes agree. For $k = 1$ there is no setup region, so burst and drip coincide (Appendix D).

7.3 Hypergraph water-filling is not order-scalar

Theorem 5 allocates the envelopes of Section 3 on an edge-separable hypergraph with an exact global budget. Corollary 6 is the minimal gap: two singletons with $v = (1, 0.01)$ and $B = 1$. Coalitional

Table 3: Mixed hypergraph, exact global budget B. Distortion $\sum_e v_e - \mathcal{G}$. Coalitional burst MSE in parentheses.

B	coalitional D	order-scalar D	$V/\mathrm{rank}\ D$	uniform D
1.0	1.72 (1.72)	1.96	1.86	2.15
2.5	1.35 (1.35)	1.62	1.45	1.79
7.0	0.70 (0.71)	1.06	0.78	1.12

Table 4: Corollary 6: two singleton edges, $B = 1$, $v = (1, 0.01)$.

Policy	Allocation	Gain
Coalitional (Theorem 5)	$(1, 0)$	0.750
Order-scalar ($r_{e,i} = q_{\lvert e\rvert}$)	$(1/2, 1/2)$	0.505

Table 5: Plug-in regret and combinatorial exactness.

Protocol	Prediction	Outcome
Plug-in MC curves, $L = 4$, five repeats	regret $\leq 2L\epsilon$	regret 0, $5/5$ recover DP
Greedy with diminishing-gain check, raw pair	refuse greedy	raises
Four arbitrary curves, $L = 8$, 165 allocations	DP = brute force	$D = 16.53$, difference 0
Diminishing arbitrary curves, greedy vs. DP	greedy exact	coincide, $D = 17.50$
Perturbation ϵ, margin $> 2L\epsilon$	recover n^*	$30/30$
Perturbation ϵ, margin $< 2L\epsilon$	bound inactive	$30/30$ recovered anyway

water-filling spends the whole budget on the strong edge (gain 0.750); any order-scalar policy must split $B/2$ each (gain 0.505). Table 4 is that instance.

Table 3 places the same allocator on a mixed hypergraph (strong/weak singleton, strong/weak pair, triple; total variance 2.47). Coalitional distortion is strictly below order-scalar, V/rank reverse water-filling of the v_e, and uniform at every budget. The gap is not a tuning artifact: order-scalar cannot silence the weak singleton without silencing the strong one, and V/rank opens higher-order edges too early. Empirical burst MSE matches the envelope program. Which edges open, and where drip is worse because an open edge is still in its setup region, are in Appendix D.

7.4 Plug-in allocation recovers the exact-budget optimizer

Theorem 7 is the operational token guarantee: estimated curves with uniform error ϵ incur regret at most $2L\epsilon$, and recover n^* when the true margin exceeds that bound. Table 5 is that test. Monte Carlo plug-in curves at $L = 4$ recover the true DP allocation in $5/5$ repeats with regret 0. DP matches brute force on four arbitrary curves ($D = 16.53$, 165 exact-budget allocations). When the true margin exceeds $2L\epsilon$, recovery is $30/30$. The classical DP and greedy lemmas used by the theorem, including greedy failure on raw $k \geq 2$ product tokens, are in Appendix D.

7.5 Leakage under dependence

Theorem 8 bounds MMSE by $(1 - \delta)(1 - \chi^2)v$ under incomplete access. Table 6 uses $Y = X_1X_2$ with access $T = \{1, 3\}$ and $X_3 = \rho X_2 + \sqrt{1 - \rho^2}N$, so $\delta = 0$ and $\chi = |\rho|$. Pairwise Pearson of Y with an observed coordinate stays near zero in every row (0.003–0.006), including at $\rho = 0.8$ where uncompressed MSE 0.362 matches the floor 0.360. A raw-covariance robustness claim is therefore false; the leakage floor is not. $K = 8$ tokens of the access set respect the floor and sit above it, as a finite codebook must. Identity-suite maps (quadratic, parity) are in Appendix D: Pearson remains ≈ 0 while Bayes leakage is perfect.

7.6 Frozen OpenCLIP residual tokens on Flickr8k

This subsection tests the *allocator*, not the Gaussian product converse. Experiment details are in App. B.6. Allocation uses validation reconstruction only. Retrieval is mean image↔text $R@1$ on quantized unimodal embeddings (joint tokens are not used at unimodal query time). Table 7 reports mean±std for the product head. Continuous CLIP $R@1$ is 0.718. Linear CCA between modalities is $\chi \approx 0.86$, so Theorem 8 applies as a warning: image and text are dependent and the product-kernel converse is not claimed.

The first joint token's validation gain is 0.13, below the best unimodal token (0.22). DP therefore keeps the shared codebook closed at $L \leq 2$ and opens it at $L = 4$ for every head (Hadamard, product,

Table 6: Theorem 8. $Y = X_1X_2$, access $T = \{1, 3\}$, $\chi = |\rho|$. Pearson is $\max_i |\mathrm{corr}(Y, X_{T,i})|$. Floor is $(1 - \chi^2)v$ with $v = 1$.

Design	floor	uncomp. MSE	$K = 8$ MSE	Pearson
Independent incomplete ($T = \{1\}$)	1	0.999	0.999	0.003
Noisy copy $\rho = 0$	1	0.995	0.995	0.006
Noisy copy $\rho = 0.8$	0.360	0.362	0.431	0.005
Perfect copy $\rho = 1$	0	0	0.108	0.006

Table 7: Flickr8k, frozen OpenCLIP, trained residual product, $K = 128$, three seeds. Retrieval is mean image↔text $R@1$. Joint column is the fraction of seeds in which DP spends at least one joint token. Continuous CLIP $R@1 = 0.718$; chance on 800 test images is 0.00125.

L	DP	setup-gated	drip	MMQ-fixed	specific-only	DP joint
2	$\mathbf{0.0234}_{\pm 0.0009}$	$0.0234_{\pm 0.0009}$	0.0013	0.0013	$0.0234_{\pm 0.0009}$	0/3
4	$0.0300_{\pm 0.0022}$	$0.0234_{\pm 0.0009}$	$0.0300_{\pm 0.0022}$	$0.0300_{\pm 0.0022}$	$\mathbf{0.0335}_{\pm 0.0025}$	3/3
6	$0.0335_{\pm 0.0025}$	$0.0335_{\pm 0.0025}$	$0.0356_{\pm 0.0022}$	$0.0505_{\pm 0.0100}$	$\mathbf{0.0548}_{\pm 0.0108}$	3/3
8	$0.0335_{\pm 0.0025}$	$0.0335_{\pm 0.0025}$	$0.0564_{\pm 0.0106}$	$0.0564_{\pm 0.0106}$	$\mathbf{0.0592}_{\pm 0.0095}$	3/3

MLP). That answers the CIFAR joint-silence question by measurement: at low L the interaction is a weaker reconstruction target than another text or image stage, not a test of the Gaussian setup equation. Forcing a joint token anyway—MMQ-fixed or drip—collapses $L = 2$ retrieval from 0.023 to chance 0.0013. That is the peak-rate analogue of Corollary 3: do not drip a shared codebook below the point where its measured gain wins. At large L, reconstruction DP opens the joint branch and attains the best total MSE (1.10 versus 1.32 specific-only at $L = 8$), while retrieval, which does not consume joint tokens, prefers specific-only. A val-A/val-B certificate gives $\hat{\epsilon} = 0.013$ at $L = 4$, $2L\hat{\epsilon} = 0.105$, and observed plug-in regret 0. Dedicated joint RVQ reconstructs Y more tightly than forming Y from reused unimodal codes (product-head gap 0.25 MSE). Average-rate bursting on the joint branch, with fallback equal to the training mean of Y, beats drip at a full on-fraction (0.36 versus 0.42) and loses at 0.25 because the fallback is crude.

These outcomes are allocator findings on measured CLIP residual-token curves, not a verification of the Gaussian product law. At a per-sample budget the integer analogue of keeping an order-k codebook closed below setup is visible: when the first joint token's validation gain loses to a unimodal stage, reconstruction DP leaves the joint branch silent ($L = 2$) and dripping a shared token collapses retrieval to chance.

8 Conclusion

We developed an operational finite-rate theory of coverage-gated coding for Gaussian product targets, then tested the resulting allocator on frozen OpenCLIP residual tokens. A missing factor of a product is invisible independently of decoder skill; complete coverage still leaves a quantitative rate–distortion law. Local channels with information budgets r_i attain an exact product-kernel MMSE, with equality for Gaussian test channels. Average-rate codes follow the concave envelope of that product—with a setup phase of 0.906 bits per modality at order $k = 2$—and a global budget is split by coalitional water-filling, which strictly dominates order-scalar V/rank rules. Finite tokens inherit a $2L\epsilon$ plug-in guarantee when ϵ is observed. Seeded audits recover the formulas, including a high-repeat check of two previously noisy one-letter cells. On Flickr8k, dripping a shared token at $L = 2$ collapses caption retrieval to chance, while reconstruction DP keeps the joint codebook closed until $L = 4$. That is an allocator result on measured CLIP curves, not a claim that natural image–text pairs obey the product-kernel MMSE. Trained residual/semantic-ID tokenizers at MMQ scale remain open.

AI use statement

In this work, we used generative AI tools for help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, assist in the writing

of proofs and interpret results. We have not used generative AI tools for formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses and implement methods, and Generate synthetic data sets, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for summarize or analyse existing literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, 6(6):165–188, 2005. URL http://jmlr.org/papers/v6/chechik05a.html.

Thomas Y. Chen. Rate-distortion limits for multimodal retrieval: Theory, optimal codes, and finite-sample guarantees. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4206–4215, 2025. URL https://api.semanticscholar.org/CorpusID:281315876.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006. ISBN 978-0-471-24195-9. doi: 10.1002/047174882X.

Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022. URL https://api.semanticscholar.org/CorpusID:253097788.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. MMTok: Multimodal coverage maximization for efficient inference of VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GvPdSWZT31.

Chenxi Du, Yongheng Deng, Jiani Liu, Yujia Zhang, Xi Chen, and Ju Ren. Coin: Coverage and informativeness-guided token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10492–10501, June 2026.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, 2020. URL https://api.semanticscholar.org/CorpusID:229297973.

Xianghong Fang, Wenlong Mou, Yuan Yuan, Dehan Kong, and Tim G. J. Rudner. A unifying view of vector, product and scalar quantization: An information-theoretic perspective, 2026. URL https://openreview.net/forum?id=4HvbTbO4hI.

H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935. doi: 10.1017/S0305004100013517.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Russ R Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393. Curran Associates, Inc., 2023. doi: 10.52202/075280-1192. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/575286a73f238b6516ce0467d67eadb2-Paper-Conference.pdf.

Vinith Misra, Vivek K Goyal, and Lav R. Varshney. Distributed scalar quantization for computing: High-resolution analysis and extensions. *IEEE Transactions on Information Theory*, 57(8):5298–5325, 2011. doi: 10.1109/TIT.2011.2158882.

Y. Oohama. The rate-distortion function for the quadratic gaussian ceo problem. *IEEE Transactions on Information Theory*, 44(3):1057–1070, 1998. doi: 10.1109/18.669162.

Or Ordentlich and Yury Polyanskiy. Optimal quantization for matrix multiplication. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6, 2025. doi: 10.1109/ISIT63088.2025.11195544.

A. Orlitsky and J.R. Roche. Coding for computing. *IEEE Transactions on Information Theory*, 47(3): 903–917, 2001. doi: 10.1109/18.915643.

Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks, 2016. URL https://arxiv.org/abs/1508.06025.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Hieu Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. *ArXiv*, abs/2305.05065, 2023. URL https://api.semanticscholar.org/CorpusID:258564854.

Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297, 2002. ISSN 0010-4655. doi: https://doi.org/10.1016/S0010-4655(02)00280-1. URL https://www.sciencedirect.com/science/article/pii/S0010465502002801.

Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

Lv Tang, Tianyi Zheng, Bo Li, and Xingyu Li. InfoTok: Information-theoretic regularization for capacity-constrained shared visual tokenization in unified MLLMs, 2026. URL https://arxiv.org/abs/2602.01554.

J. Wolf and J. Ziv. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Transactions on Information Theory*, 16(4):406–411, 1970. doi: 10.1109/TIT.1970.1054469.

Yi Xu, Moyu Zhang, Chenxuan Li, Zhihao Liao, Haibo Xing, Hao Deng, Jinxin Hu, Yu Zhang, Xiaoyi Zeng, and Jing Zhang. Mmq: Multimodal mixture-of-quantization tokenization for semantic id generation and user behavioral adaptation. *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, 2025. URL https://api.semanticscholar.org/CorpusID:280700429.

Lei Yu. On conditional correlations, 2019. URL https://arxiv.org/abs/1811.03918.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. URL https://api.semanticscholar.org/CorpusID:236149944.

Wenyuan Zhao, Adithya Balachandran, Chao Tian, and Paul Liang. Partial information decomposition via normalizing flows in latent gaussian distributions. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 4495–4530. Curran Associates, Inc., 2025. doi: 10.52202/085713-0159. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/06f484b5e9ce4e4b0e88ac0f29239201-Paper-Conference.pdf.

Bowen Zheng, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. Universal item tokenization for transferable generative recommendation. *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. URL https://api.semanticscholar.org/CorpusID:277622182.

Shaobin Zhuang, Yuang Ai, Jiaming Han, Weijia Mao, Xiaohui Li, Fangyikang Wang, Xiao Wang, Yan Li, Shanchuan Lin, Kun Xu, Zhenheng Yang, Huaibo Huang, Xiangyu Yue, Hao Chen, and Yali Wang. Uniwetok: An unified binary tokenizer with codebook size 2128 for unified multimodal large language model. *ArXiv*, abs/2602.14178, 2026. URL https://api.semanticscholar.org/CorpusID:285615326.

A PRELIMINARIES

This appendix records the standing definitions used in the main text and the classical primitives those theorems invoke. All logarithms and mutual informations are in bits. Random elements take values in standard Borel spaces, so regular conditional distributions exist. The results of Sections 3–6 are operational coding statements, not empirical claims. The lemmas below are classical; they isolate what is *not* new.

The distributed-computation context includes coding for computing and distributed functional quantization (Orlitsky & Roche, 2001; Misra et al., 2011). The closest matrix-product quantization result is Ordentlich & Polyanskiy (2025), recovered in Corollary 3 as the equal-rate $k = 2$ special case.

A.1 STANDING DEFINITIONS

The four objects below are the only definitions used by the main theorems. Definition 9 is the access gate of Section 3. Definition 10 is the operational envelope of Theorem 2 (closed form: Theorem 18). Definition 11 is the architecture of Section 4. Definition 12 is the leakage coefficient of Section 6.

Definition 9 (Certified coalition access)**.** *Let $X = (X_1, \dots, X_M)$, let Y be a centered random element of a separable real Hilbert space H, and assume $\mathbb{E}\|Y\|_{\mathsf{H}}^2 < \infty$. A representation U has* certified access *to coalition $C \subseteq [M]$ if*

$$U - X_C - (Y, X_{[M]\setminus C})$$

is a Markov chain. This collection-level condition permits private encoder randomness. When several representations are asserted to be separate and local, absence of hidden shared randomness requires the stronger product factorization in (3)*; the Markov condition alone does not imply that factorization. For deterministic tokens it reduces to $\sigma(U) \subseteq \sigma(X_C)$.*

Average-rate block codes can time-share complete-coalition profiles, so the one-letter product F_k of (2) is not the operational gain for $k \geq 2$.

Definition 10 (Closed upper concave envelope)**.** *For $\boldsymbol{r} \in \mathbb{R}_+^k$, define*

$$\overline{\mathrm{cav}} F_k(\boldsymbol{r}) := \lim_{\epsilon \downarrow 0} \sup_{\substack{J<\infty,\ \lambda_j \geq 0,\ \sum_j \lambda_j = 1 \\ \boldsymbol{\rho}^{(j)} \in \mathbb{R}_+^k \\ \sum_j \lambda_j \rho_i^{(j)} \leq r_i + \epsilon,\ i \in [k]}} \sum_{j=1}^{J} \lambda_j F_k(\boldsymbol{\rho}^{(j)}). \tag{16}$$

Equivalently, this is the least upper-semicontinuous, coordinatewise nondecreasing concave majorant of F_k.

A tokenizer typically maintains several product interactions at once. The following architecture charges every incidence separately, so the program of Section 4 is exact rather than a relaxation of an unrestricted network code.

Definition 11 (Edge-separable access-hypergraph architecture)**.** *Let $\mathcal{G} = (V, \mathcal{E})$ be a finite hypergraph. Let $X_{i,t} \sim \mathcal{N}(0,1)$, independently over nodes $i \in V$ and samples t, and define*

$$Y_{e,t} := \sqrt{v_e} \prod_{i \in e} X_{i,t}, \qquad v_e > 0.$$

Node i observes X_i^n. An edge-separable code *assigns and charges a dedicated submessage $M_{e,i}$ to every incidence $i \in e$, uses independent private encoder randomness across nodes, transmits the concatenation of these separately charged streams, and reconstructs Y_e^n only from $(M_{e,i})_{i \in e}$. The incidence rate is $r_{e,i} = \limsup_n n^{-1} \log_2 |\mathcal{M}_{e,i,n}|$, and rates of concatenated streams are added. This definition deliberately forbids reusing one coded stream for several edges and excludes unrestricted cross-edge or cross-target coupled messages.*

When modalities are dependent, incomplete access can leak a missing innovation. The following coefficient is the contraction of that innovation.

Definition 12 (Conditional HGR contraction for an S-innovation)**.** *Fix a target support S and an access set T with $S \not\subseteq T$. Put*

$$A = S \cap T, \qquad B = S \setminus T, \qquad C = T \setminus S.$$

The conditional Hilbertian maximal-correlation coefficient is

$$\chi(B, C \mid A) := \sup_{\substack{0 \neq h \in L^2(\sigma(X_A, X_B)) \\ \mathbb{E}[h \mid X_A] = 0}} \frac{\|\mathbb{E}[h \mid X_A, X_C]\|_2}{\|h\|_2}. \tag{17}$$

It lies in $[0, 1]$. This operator form is the conditional HGR coefficient relevant to innovation prediction (Polyanskiy & Wu, 2016; Yu, 2019).

A.2 Hilbert decomposition and indirect RDF

Any subsequent code of an observation W splits squared error into three orthogonal terms. Under certified access (Definition 9), the first term is the price of the access set; it cannot be reduced by decoder skill.

Lemma 13 (Hilbert projection decomposition and indirect RDF)**.** *Let W be any observation, let $U - W - Y$, and let $\widehat{Y} = d(U) \in L^2(\mathsf{H})$. Define*

$$m(W) := \mathbb{E}[Y \mid W], \qquad q(U) := \mathbb{E}[m(W) \mid U] = \mathbb{E}[Y \mid U].$$

Then the following exact orthogonal decomposition holds:

$$\begin{aligned} \mathbb{E}\|Y - \widehat{Y}\|_{\mathsf{H}}^2 = {} & \underbrace{\mathbb{E}\|Y - m(W)\|_{\mathsf{H}}^2}_{\textit{coverage/observation loss}} \\ & + \underbrace{\mathbb{E}\|m(W) - q(U)\|_{\mathsf{H}}^2}_{\textit{compression loss}} + \underbrace{\mathbb{E}\|q(U) - \widehat{Y}\|_{\mathsf{H}}^2}_{\textit{decoder excess}}. \end{aligned} \tag{18}$$

In particular, the MMSE decoder is $q(U)$; MMSE itself is not decoder-dependent.

Define the centralized indirect rate–distortion function

$$R^{\mathrm{ind}}_{Y \leftarrow W}(D) := \inf_{\substack{P_{U|W},\, d \\ \mathbb{E}\|Y - d(U)\|^2 \le D}} I(W; U)$$

and the ordinary squared-error RDF of $m = m(W)$,

$$R_m(d) := \inf_{\substack{P_{\widehat{m}|m} \\ \mathbb{E}\|m - \widehat{m}\|^2 \le d}} I(m; \widehat{m}).$$

With

$$\Delta_W := \mathbb{E}\|Y - m(W)\|^2,$$

the exact indirect-source identity is

$$R^{\mathrm{ind}}_{Y \leftarrow W}(D) = R_m(D - \Delta_W), \tag{19}$$

where the right side is $+\infty$ for $D < \Delta_W$. This is the squared-error remote-source reduction underlying classical indirect source coding (Wolf & Ziv, 1970; Cover & Thomas, 2006).

For coalition access $W = X_C$ in the sense of Definition 9, the first term of (18) is the irreducible price of the access set. The second term measures what the finite representation discards after access is fixed, and the third term isolates decoder misspecification. The product theorems of Section 3 evaluate the first two terms in closed form when Y is a Gaussian product and the auxiliaries are local.

A.3 Scalar Gaussian information–MMSE

The one-letter product law is a tensorization of the scalar Gaussian rate–distortion equality.

Lemma 14 (Gaussian information–MMSE inequality)**.** *If $X \sim \mathcal{N}(0,1)$ and U is any standard-Borel random element with $I(X;U) < \infty$, then*

$$\operatorname{mmse}(X \mid U) := \mathbb{E}\big[(X - \mathbb{E}[X \mid U])^2\big] \geq 2^{-2I(X;U)}. \tag{20}$$

Equality holds if and only if, with $\mu(U) = \mathbb{E}[X \mid U]$ and $D = \operatorname{mmse}(X \mid U)$,

$$X = \mu(U) + Z, \qquad Z \sim \mathcal{N}(0, D), \qquad Z \perp U.$$

In the equality case $\mu(U) \sim \mathcal{N}(0, 1 - D)$. This is the scalar Gaussian RDF equality condition (Cover & Thomas, 2006).

A.4 Exact finite-token allocation

A tokenizer assigns an integer number of K-ary tokens, not a real rate. The following two lemmas are the classical combinatorial facts used in Section 5. Let J denote the number of branches (edges, spectra, or empirically measured token streams). A token has alphabet size K and nominal rate $\log_2 K$. Let $G_j(\ell)$ be the explained-variance gain from assigning $\ell \in \{0, \dots, L\}$ tokens to branch j, normalized by $G_j(0) = 0$. Arbitrary distortion curves are included by setting $G_j(\ell) = D_j(0) - D_j(\ell)$.

Lemma 15 (Exact finite-token dynamic program)**.** *For arbitrary gain curves, the exact allocation under $\sum_{j=1}^{J} n_j = L$ is obtained from*

$$\begin{aligned} V_0(0) = 0, \quad V_0(\ell) = -\infty \ (\ell > 0), \\ V_j(\ell) = \max_{0 \leq q \leq \ell} \{V_{j-1}(\ell - q) + G_j(q)\}. \end{aligned} \tag{21}$$

Then $V_J(L)$ is the optimum, and backtracking gives an optimal allocation. The direct implementation costs $O(JL^2)$ time and $O(JL)$ memory. For a "at most L" budget, maximize $V_J(\ell)$ over $\ell \leq L$.

Lemma 16 (Greedy exactness under diminishing gains)**.** *Define*

$$\Delta_j(q) := G_j(q) - G_j(q-1), \qquad q = 1, \dots, L.$$

If every sequence $\Delta_j(1), \Delta_j(2), \dots$ is nonnegative and nonincreasing, repeatedly assigning the next token to a branch with largest currently available marginal gain is exactly optimal. Coalitional setup curves need not have diminishing gains; when the condition fails, the dynamic program, not this greedy lemma, applies.

A.5 Classical Gaussian reverse water-filling

When the observation is centralized and the predictable signal is Gaussian, Lemma 13 reduces compression to reverse water-filling of a covariance spectrum. This is the correct allocator for a single predictable vector. It is *not* the allocator for distributed product targets: those are governed by the envelope program of Section 4. We record the spectral law so that order-scalar V/rank rules can be compared with a genuine Gaussian baseline rather than with a slogan.

Lemma 17 (Indirect Gaussian reverse water-filling)**.** *In Lemma 13, let $Y \in \mathbb{R}^d$ and suppose the predictable signal $m(W) = \mathbb{E}[Y \mid W]$ is Gaussian. Let its positive covariance eigenvalues be $\lambda_1 \geq \cdots \geq \lambda_q > 0$, and put $\Delta_W = \mathbb{E}\|Y - m(W)\|^2$. For i.i.d. block coding at rate R bits per source vector, the exact indirect distortion is*

$$D_{Y \leftarrow W}(R) = \Delta_W + \sum_{j=1}^{q} \min\{\lambda_j, \theta\}, \tag{22}$$

where, for $R > 0$, the unique water level $\theta \in (0, \lambda_1)$ satisfies

$$R = \frac{1}{2} \sum_{j=1}^{q} \left[\log_2 \frac{\lambda_j}{\theta}\right]^+. \tag{23}$$

At $R = 0$, one may take $\theta = \lambda_1$. The $(j+1)$-st eigenmode activates at

$$R_{\mathrm{crit}}^{(j+1)} = \frac{1}{2} \sum_{i=1}^{j} \log_2 \frac{\lambda_i}{\lambda_{j+1}}. \tag{24}$$

Tied eigenvalues activate simultaneously. This is classical Gaussian reverse water-filling (Cover & Thomas, 2006; Wolf & Ziv, 1970), applied to the predictable spectrum rather than to covariance rank or trace alone. Its geometry is related to, but is not the same optimization as, the Gaussian information bottleneck (Chechik et al., 2005).

B EXPERIMENTAL DETAILS

This appendix records the protocols behind the main-text experiments of Section 7. The main text keeps only the cells that distinguish the new theorems; the grids in Appendix D are the remainder of the same seeded runs. CIFAR-10 class-name settings are in Appendix C.

B.1 COMMON IMPLEMENTATION AND REPORTING PROTOCOL

Reference implementations. The Gaussian scripts use NumPy double precision and SciPy; they do not train a neural encoder. The Flickr study uses frozen OpenCLIP plus a small NumPy interaction head; residual codebooks are ordinary Lloyd k-means, not a VQ-VAE. No GPU was used for the reported tables.

Samples, seeds, and intervals. Unless a subsection says otherwise, Gaussian Monte Carlo uses master seed 20270829, five repeats, and 6×10^4 i.i.d. standard-normal samples per repeat. The high-repeat audit uses 80 repeats and 2×10^5 samples. Displayed 95% intervals are two-sided Student-t on the repeat means. Flickr Table 7 reports mean$\pm$sample standard deviation over three seeds, not a t-interval.

Metrics and audits. All reported reconstruction numbers are unclipped MSE: for scalars, the mean squared error; for vectors, the mean squared Euclidean residual $\mathbb{E}\|Y-\widehat{Y}\|^2$. Coverage grids that report R^2 also leave that statistic unclipped. Average-rate codes are audited by realized mean bits per sample. Token codes are audited by $\sum_j n_j = L$ exactly; rows that cannot spend the integer budget are dropped. Mutual information of discrete auxiliaries is a histogram estimate (quantile bins for mixed continuous–discrete pairs). Pearson numbers are $\max_i |\mathrm{corr}(Y, X_{T,i})|$.

B.2 PRODUCT KERNEL AND OPERATIONAL ENVELOPE

Source and channels. Independent $X_i \sim \mathcal{N}(0,1)$ and $Y = \prod_i X_i$. A Gaussian test channel of rate r bits is the scalar observation $U = \sqrt{1-2^{-2r}}X + \sqrt{2^{-2r}}N$ with $N \sim \mathcal{N}(0,1)$ independent; the Bayes decoder uses the posterior mean $\sqrt{1-2^{-2r}}U$. The product-kernel decoder is the product of those posterior means. The naive decoder multiplies the raw observations; OLS is least squares of Y on an intercept and the local observations (or local auxiliaries).

Shared-randomness counterexample. Draw a hidden Rademacher Q independent of (X_1, X_2) and set $U_i = Q\,\mathrm{sgn}(X_i)$, replacing a zero sign by Q. The Bayes product predictor is $(2/\pi)U_1U_2$. Local mutual information is estimated by binning each X_i into 24 quantile bins and computing the discrete MI with U_i.

Sign and finite-alphabet channels. A sign channel uses $\mathrm{sgn}(X_i)$ (one bit of mutual information). A 4-ary token is an equiprobable scalar quantizer of $\mathcal{N}(0,1)$ with alphabet 4 (nominal two bits); reconstruction is the bin mean. Silencing zeros one local rate while companions take $\{0.25, 0.5, 1, 2\}$ bits, for $k \in \{2,3,4\}$ (12 cells).

Burst versus drip. Drip applies the same local rates on every sample (peak-rate / memoryless product). Burst time-shares the contact profile of the concave envelope: a fraction $1/\tau$ of samples receive rate τr and the rest receive zero, with τ from the envelope program, then a random permutation of the on-set. Table 2 uses equal local rates at fractions $0.25, 0.50, 1.00$ of the symmetric setup rate for $k = 2, 3$. Realized mean bits are recorded per repeat.

B.3 HYPERGRAPH WATER-FILLING

Graph. Six independent standard Gaussians and five dedicated product edges: a strong singleton $v = 1$, a weak singleton $v = 0.03$, a strong pair $v = 0.85$, a weak pair $v = 0.04$, and a triple $v = 0.55$ (total variance 2.47). Each edge has its own Gaussian test-channel incidences; there is no codebook reuse across edges. The exact global budget is $B \in \{0.5, 1.0, 1.5, 2.5, 4.0, 7.0\}$; Table 3 reports $1.0, 2.5, 7.0$.

Policies. Coalitional rates are the continuous edge-separable water-filling program. Order-scalar searches nonnegative compositions of a 48-point simplex over cardinality classes and assigns one common incidence rate per order. V/rank treats each v_e as a rank-1 Gaussian and reverse-water-fills those scalars, placing any residual bit on the strongest edge. Uniform splits B equally across edges. Distortion is $\sum_e v_e - \mathcal{G}$ from the envelope program; empirical burst MSE is the Monte Carlo product decoder on the same edge rates. The two-singleton corollary uses $v = (1, 0.01)$ and $B = 1$.

B.4 Plug-in allocation

Monte Carlo plug-in. A singleton of weight 0.38 and a pair of weight 1 receive one-bit tokens. For each of five repeats, plug-in distortion curves at $L = 4$ are estimated with 1.5×10^4 Gaussian samples per cell by equal-splitting the token budget inside each edge (raw product, no time-sharing). DP on the estimates is scored on the closed-form raw curves. ϵ is the max absolute error of marginal gains; the theorem bound is $2L\epsilon$.

Combinatorial checks. Four arbitrary length-8 gain sequences with bases $(7, 6, 5.5, 4.5)$ are allocated at $L = 8$ by DP and by exhaustive compositions (165 allocations). A second family with geometrically decaying gains checks that greedy coincides with DP. Raw $k = 2$ product tokens are fed to greedy with a diminishing-gain verifier, which must raise. Plug-in perturbations add Uniform$[-\epsilon, \epsilon]$ noise to the arbitrary curves for 30 repeats at a margin-active ϵ and a margin-inactive ϵ.

B.5 Incomplete-access leakage

Independent $X_1, X_2 \sim \mathcal{N}(0, 1)$ and $Y = X_1X_2$. Access is $T = \{1\}$ (independent incomplete) or $T = \{1, 3\}$ with $X_3 = \rho X_2 + \sqrt{1-\rho^2}N$ at $\rho \in \{0, 0.8, 1\}$. Uncompressed predictors are the Bayes maps 0, ρX_1X_3, or X_1X_3. Token rows replace each accessed coordinate by an equiprobable $K = 8$ scalar quantizer and use the corresponding bin-mean product. Pearson is computed on the uncompressed access coordinates. Sample size 6×10^4, five repeats. Identity-suite maps (quadratic, parity) are in Appendix D.

B.6 Flickr8k OpenCLIP residual tokens

This is the tokenizer-facing study of Section 7.6. It tests the allocator on measured residual-token curves. It is not a claim that CLIP embeddings obey the Gaussian product MMSE.

Data. Images and captions are the jbrownlee Flickr8k GitHub release. There are $13{,}500$ pairs on $2{,}700$ unique images. Official split files, when present, assign images to train/val/test; the run then keeps the first $1500/400/800$ unique images in caption-file order within each official split. All five captions per retained image are kept, so the reported split is $7500/2000/4000$ pairs. Splits are by image id: no image appears in two splits.

Frozen encoder. OpenCLIP `ViT-B-32` with checkpoint `laion2b_s34b_b79k`, batch size 32, CPU. Images use the model's default preprocess; texts use the model tokenizer. Image and text embeddings are ℓ_2-normalized to dimension 512 and cached. Unique images are encoded once and reused across captions. Train-split means are subtracted from image and text before any interaction head or quantizer is fit; test reconstruction for retrieval adds those means back.

Interaction heads. Three maps $Y = f_\theta(z_I, z_T)$ are trained on centered training pairs only, then applied to val/test: (i) Hadamard $\sqrt{d}\, z_I \odot z_T$ (no training); (ii) product: two 512×64 projections, row-normalized, InfoNCE logits with temperature scale 20, 80 steps, batch $\min(256, n_{\mathrm{train}})$, learning rate 0.25 decayed by 0.4 at step 40, projection init $\mathcal{N}(0, 0.02^2)$; (iii) MLP: concatenate (z_I, z_T), ReLU hidden of width $2\times64 = 128$, output dimension 64, matched-versus-shuffled InfoNCE, 80 steps, batch $\min(128, n_{\mathrm{train}})$, learning rate 0.02 with the same decay, init $\mathcal{N}(0, 0.03^2)$. After the head, Y is globally rescaled so that its training energy $\mathbb{E}\|Y\|^2$ matches the training image energy, then re-centered. Table 7 is the product head. Linear CCA χ is estimated on training image/text after a fixed Gaussian projection to 24 dimensions (seed 0) with covariance ridge $10^{-3}I$.

Residual quantization. Each branch (image, text, joint) gets its own residual k-means bank: $K = 128$, at most four stages, 12 Lloyd iterations, one k-means++ restart, empty clusters reset to the farthest residual. Stage ℓ is fit to the training residual after stages $1{:}\ell - 1$; held-out assignments use frozen centers. Branch seeds are $\text{seed} + (b + 1)\times 100003$ with stage offsets $\times 10007$. Distortion curves are mean squared Euclidean residuals on the validation split, including the zero-stage energy.

Allocation policies. The exact integer budget is $L \in \{0, 1, 2, 4, 6, 8\}$. DP is exact-budget dynamic programming on the three validation curves. Setup-gated DP may use joint stages only in $\{0\} \cup \{2, \ldots, 4\}$. Drip forces one joint stage whenever $L \geq 1$ and DP-allocates the remainder on image/text. MMQ-fixed spends one joint stage (if $L \geq 1$) and splits the rest evenly across image and text. Specific-only DP-allocates on image and text only. Uniform, variance, and rank are capacity-aware proportional integer rules using unit weights, training $\mathbb{E}\|Y_e\|^2$, and numerical rank (singular values above $10^{-3}\sigma_{\max}$). Policies that cannot spend exactly L stages are omitted from that row.

Evaluation. Reconstruction is total test MSE summed over the three branches. Retrieval is the mean of image-to-text and text-to-image $R@1$ on *unimodal* reconstructions in CLIP space (joint tokens unused at query time). Image queries use one embedding per unique image (first caption occurrence); text queries use every caption; cosine scores; stable argsort. Continuous CLIP $R@1 = 0.718$ is the same metric on unquantized embeddings. Chance on 800 test images is $1/800 = 0.00125$.

Certificates and peak-rate probes. Val-A/val-B split the validation pairs in order. $\hat{\epsilon}$ is the max absolute curve discrepancy; plug-in regret is DP(A) scored on B minus DP(B) on B; the bound is $2L\hat{\epsilon}$. Bursting uses two joint stages on a random fraction $\{0.25, 0.5, 1\}$ of test pairs and the training mean of Y otherwise, against one joint stage on every pair (drip). Reuse fits two-stage unimodal RVQ, reconstructs Y from quantized (z_I, z_T) through the frozen head (energy-matched), and compares to a dedicated two-stage joint RVQ.

C CIFAR-10 CLASS-NAME CHECK

The CIFAR-10 harness uses class-name strings rather than captions and lightweight descriptors rather than OpenCLIP. Table 8 remains a reconstruction engineering check. At $L = 1$, DP spends the unique stage on text (1.85) while uniform/variance/rank spend it on image (2.08). DP never opens the joint branch for $L \leq 3$; as on Flickr8k, that is because the measured interaction gain loses to extra unimodal stages, not because the Gaussian setup root was tested. This table is not evidence about natural-caption tokenizers.

Table 8: CIFAR-10 class-name pilot (not captions), exact stage budget L, $K = 8$. Distortion is total test MSE. DP uses validation curves only.

	Stages (img, txt, joint)				Test distortion			
L	DP	Unif.	Var.	Rank	DP	Unif.	Var.	Rank
0	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	2.45	2.45	2.45	2.45
1	(0,1,0)	(1,0,0)	(1,0,0)	(1,0,0)	**1.85**	2.08	2.08	2.08
2	(1,1,0)	(1,1,0)	(1,1,0)	(1,0,1)	**1.47**	1.47	1.47	1.96
3	(1,2,0)	(1,1,1)	(1,1,1)	(2,0,1)	**1.32**	1.35	1.35	1.87

Data. Images are fetched from the Hugging Face dataset API and cached. Each image is paired with its class-name string, not a caption. Seed 2027 draws a stratified subset: 300 train and 100 val from the official train pool, 200 test from the official test pool.

Features. The encoder is *not* OpenCLIP. Images are resized to 16×16, concatenated with RGB mean/std and mean color gradients, projected by a seeded Gaussian matrix to dimension 64, passed through $\tanh$, and ℓ_2-normalized. Text uses signed word and character n-gram hashing (widths 2–5) into the same 64 dimensions, then ℓ_2-normalization. A bidirectional ridge map (regularization 10^{-3}) is fit on training pairs only for retrieval alignment; it is not used in the reconstruction table.

Coalitions and quantizers. Image and text branches are train-centered embeddings. The joint branch is the centered Hadamard product $\sqrt{d}\, z_I \odot z_T$. Residual k-means uses $K = 8$, at most two stages per branch, 12 Lloyd iterations, one restart. Validation curves drive exact-budget DP. Uniform/variance/rank are the same proportional integer rules as in Appendix B, with rank cutoff $10^{-3}\sigma_{\max}$. Table 8 is summed test MSE over the three branches. Image-only class accuracy is near chance; paired accuracy is label-leaky and is not a result. Caption evidence is Section 7.6, not this probe.

D Additional Experimental Results

D.1 One-letter grid and coverage

Table 9 is the identity grid behind the Gaussian rows of Table 1. A follow-up audit with 80 repeats and 2×10^5 samples places both previously out-of-interval heterogeneous cells inside the empirical 95% interval: $(0.08, 0.25, 0.65)$ has formula 0.982 and mean 0.982; $(0, 3)$ is the known boundary $\mathrm{MMSE} = 1$ with mean 1.001. Neither cell is a threat to Theorem 1. Table 10 is the discrete-token counterpart of the same access gate: a missing modality or a sign-blind codebook yields $R^2 = 0$; a union of locally uninformative tokens recovers the covered-informative curve.

Table 9: One-letter Gaussian test channels. Unclipped MSE. Symmetric cells vary k and equal local rate.

Protocol	Prediction	Empirical	Check
24 symmetric (k, r) cells	formula in every CI	max $\lvert\mathrm{err}\rvert = 0.025$	24/24
$(r_1, r_2) = (0.08, 0.25)$	0.969	0.970	in 95% CI
$(0.08, 0.25, 0.65)$	0.982	0.982	in 95% CI (80 repeats)
$(0.08, 0.25, 0.65, 1.3)$	0.985	1.004	in 95% CI
$(0.2, 1.3)$	0.798	0.794	in 95% CI
$(0.5, 1.0)$	0.625	0.629	in 95% CI
$(0.05, 2.0)$	0.937	0.933	in 95% CI
$(0.0, 3.0)$	1	1.001	in 95% CI (80 repeats)

Table 10: Finite K-ary coverage. Covered rows are population Bayes R^2; the union row is exact held-out R^2.

	$k = 2$			$k = 3$		
Access pattern	$K = 2$	$K = 4$	$K = 8$	$K = 2$	$K = 4$	$K = 8$
Missing modality	0	0	0	0	0	0
Covered, sign-blind magnitude bins	0	0	0	0	0	0
Covered, sign-preserving bins	0.405	0.741	0.893	0.258	0.637	0.844
Union of locally uninformative tokens	0.406	0.739	0.894	0.259	0.638	0.845

D.2 Burst, drip, and equal split

Table 11 extends Table 2 to $k = 1$ (no setup region: burst and drip coincide) and $k = 4$. Table 12 splits a fixed total $B = 1.2$ bits at $k = 2$. Equal local split uniquely minimizes envelope distortion; a zero coordinate forces full distortion, as predicted by the product kernel.

D.3 Hypergraph silence pattern

Table 14 is the full budget sweep; Table 13 records which coalitional edges are open. Only the strong singleton is open at $B \le 1$; the strong pair opens at $B = 1.5$; the triple opens only at $B = 7$; weak edges remain silent throughout. Drip is strictly worse wherever an open edge is still in its setup region ($B = 1.5$ and $B = 7$).

Table 11: Burst versus drip, including $k = 1$ and $k = 4$. Setup rates: $0.906, 1.373, 1.686$ bits/modality for $k = 2, 3, 4$.

k	r/r_k^0	envelope D	raw D	burst	drip
1	— ($r = 0.40$)	0.574	0.574	0.574	0.576
2	0.25	0.872	0.927	0.878	0.933
2	0.50	0.744	0.782	0.744	0.784
2	1.00	0.488	0.488	0.491	0.490
3	0.25	0.846	0.946	0.849	0.947
3	0.50	0.692	0.769	0.697	0.772
3	1.00	0.384	0.384	0.384	0.383
4	0.25	0.834	0.962	0.832	0.958
4	0.50	0.667	0.774	0.695	0.797
4	1.00	0.334	0.334	0.334	0.332

Table 12: Equal versus unequal split of a total $B = 1.2$ bits at $k = 2$.

Rate vector	envelope D	burst MSE	drip MSE
$(0.6, 0.6)$	0.661	0.659	0.683
$(0.4, 0.8)$	0.695	0.687	0.713
$(0.2, 1.0)$	0.800	0.795	0.816
$(0.05, 1.15)$	0.938	0.936	0.948
$(0.0, 1.2)$	1	0.999	0.999

D.4 Exact-budget DP versus greedy

Lemma 15 is exact for arbitrary curves. Lemma 16 is exact only under diminishing gains. Raw $k \geq 2$ products violate that hypothesis in the setup region. Table 15 uses one-bit tokens on a singleton ($v = 0.38$) and a raw pair ($v = 1$). Pair-marginals increase then decrease, so greedy is not a substitute for DP. At exact budget $L = 2$, DP returns $(0, 2)$ with distortion 0.818; greedy returns $(1, 1)$ with 0.845. After the pair is replaced by its concave envelope, greedy matches DP. Table 16 extends the same protocol through $L = 5$: after the pair is replaced by $\operatorname{cav} u_2$, greedy matches DP at every budget.

D.5 Identity-suite leakage

Table 17 records maps whose pairwise Pearson with the observed coordinates vanishes while Bayes leakage need not. Quadratic and parity have Pearson ≈ 0 and uncompressed MSE 0; a shared Gaussian factor is the linear case, where Pearson tracks χ.

E Proofs of the Preliminaries

Proof of Lemma 13. All conditional expectations below are Bochner conditional expectations in the separable Hilbert space H. The Markov relation $U - W - Y$ gives

$$\mathbb{E}[Y \mid W, U] = \mathbb{E}[Y \mid W] = m(W).$$

Conditioning once more on U yields

$$\mathbb{E}[Y \mid U] = \mathbb{E}[\mathbb{E}[Y \mid W, U] \mid U] = \mathbb{E}[m(W) \mid U] = q(U).$$

Write

$$e_0 = Y - m(W), \qquad e_1 = m(W) - q(U), \qquad e_2 = q(U) - \widehat{Y}.$$

Since $\mathbb{E}[e_0 \mid W, U] = 0$, e_0 is orthogonal in $L^2(\mathsf{H})$ to every $\sigma(W, U)$-measurable square-integrable random element, in particular to e_1 and e_2. Moreover, $\mathbb{E}[e_1 \mid U] = 0$, whereas e_2 is $\sigma(U)$-measurable, so $e_1 \perp e_2$. The identity $Y - \widehat{Y} = e_0 + e_1 + e_2$ and Pythagoras give (18). The last term is nonnegative and vanishes exactly for the posterior-mean decoder.

Table 13: Coalitional active edges on the mixed hypergraph.

B	strong singleton	weak singleton	strong pair	weak pair	triple
0.5	✓	–	–	–	–
1.0	✓	–	–	–	–
1.5	✓	–	✓	–	–
2.5	✓	–	✓	–	–
4.0	✓	–	✓	–	–
7.0	✓	–	✓	–	✓

Table 14: Mixed hypergraph, exact global budget B. Distortion $\sum_e v_e - \mathcal{G}$. Coalitional burst (drip) MSE in parentheses.

B	coalitional D	order-scalar D	$V/\mathrm{rank}\ D$	uniform D
0.5	1.97 (1.96, 1.97)	2.17	2.08	2.30
1.0	1.72 (1.72, 1.72)	1.96	1.86	2.15
1.5	1.59 (1.58, 1.62)	1.80	1.70	2.02
2.5	1.35 (1.35, 1.36)	1.62	1.45	1.79
4.0	1.02 (1.02, 1.02)	1.43	1.17	1.51
7.0	0.70 (0.71, 0.76)	1.06	0.78	1.12

It remains to prove the indirect RDF identity. For any feasible $(P_{U|W}, d)$, orthogonality of $Y - m(W)$ to every function of (W, U) gives

$$\mathbb{E}\|Y - d(U)\|^2 = \Delta_W + \mathbb{E}\|m(W) - d(U)\|^2. \tag{25}$$

The reproduction $\widetilde{m} = d(U)$ is feasible for the ordinary RDF of m at distortion $D - \Delta_W$. Since m is a function of W, the chain rule and data processing give

$$I(m; \widetilde{m}) \le I(m; U) \le I(W; U).$$

Taking the infimum over indirect codes proves

$$R^{\mathrm{ind}}_{Y \leftarrow W}(D) \ge R_m(D - \Delta_W).$$

Conversely, take any ordinary test channel $P_{\widehat{m}|m}$ with distortion at most $D - \Delta_W$. An encoder observing W can compute $m(W)$, sample $\widehat{m}$ from this test channel, set $U = \widehat{m}$, and use the identity decoder. Because $W - m - \widehat{m}$ and m is a function of W,

$$I(W; U) = I(W; \widehat{m}) = I(m; \widehat{m}).$$

Equation (25) gives total distortion at most D. Taking infima (and an approximating sequence if the infimum is not attained) proves the reverse inequality and hence (19). If $D < \Delta_W$, (25) shows that no code is feasible. □

Proof of Lemma 14. Let

$$D = \mathbb{E}[\mathrm{Var}(X \mid U)] = \mathrm{mmse}(X \mid U).$$

Using base-two differential entropy, conditional Gaussian maximum-entropy, and then Jensen's inequality,

$$\begin{aligned} h_2(X \mid U) &= \mathbb{E}[h_2(X \mid U = u)] \\ &\le \frac{1}{2}\mathbb{E}[\log_2(2\pi e\ \mathrm{Var}(X \mid U))] \\ &\le \frac{1}{2}\log_2(2\pi e D). \end{aligned}$$

Since $h_2(X) = \frac{1}{2}\log_2(2\pi e)$,

$$I(X; U) = h_2(X) - h_2(X \mid U) \ge \frac{1}{2}\log_2 \frac{1}{D}.$$

Table 15: Exact token budget L on raw product curves. Allocation is (singleton, pair). Every method spends exactly L tokens.

	raw product		pair envelope	
L	DP (alloc. / D)	greedy (alloc. / D)	greedy = DP	DP alloc.
1	$(1,0)$ / 1.095	$(1,0)$ / 1.095	yes	$(1,0)$
2	$(0,2)$ / 0.818	$(1,1)$ / 0.845	yes	$(1,1)$
3	$(1,2)$ / 0.533	$(1,2)$ / 0.533	yes	$(1,2)$

Table 16: Exact token budget L on raw product curves, full range. Allocation is (singleton, pair).

	raw product		pair envelope	
L	DP (alloc. / D)	greedy (alloc. / D)	greedy = DP	DP alloc.
1	$(1,0)$ / 1.095	$(1,0)$ / 1.095	yes	$(1,0)$
2	$(0,2)$ / 0.818	$(1,1)$ / 0.845	yes	$(1,1)$
3	$(1,2)$ / 0.533	$(1,2)$ / 0.533	yes	$(1,2)$
4	$(1,3)$ / 0.329	$(1,3)$ / 0.329	yes	$(1,3)$
5	$(1,4)$ / 0.216	$(1,4)$ / 0.216	yes	$(1,4)$

Rearranging proves (20).

Equality requires equality in both inequalities above. Therefore $X \mid U = u$ is Gaussian for P_U-almost every u, and its conditional variance is a constant D. With $\mu(U) = \mathbb{E}[X \mid U]$ and $Z = X - \mu(U)$, the conditional law of Z given $U = u$ is the same $\mathcal{N}(0, D)$ for almost every u; hence $Z \perp U$. Conversely, this independent Gaussian residual condition makes both entropy inequalities equalities. Finally, $X = \mu(U) + Z$ is a sum of independent variables and

$$\varphi_{\mu(U)}(t) = \frac{\varphi_X(t)}{\varphi_Z(t)} = \exp\left[-\frac{1-D}{2}t^2\right],$$

so $\mu(U) \sim \mathcal{N}(0, 1-D)$. □

Proof of Lemma 15. We prove by induction on j that $V_j(\ell)$ is the maximum total gain obtainable by assigning exactly ℓ tokens among the first j branches. The initialization is correct for zero branches. For $j \geq 1$, every allocation gives some $q \in \{0, \ldots, \ell\}$ tokens to branch j, and its remaining $\ell - q$ tokens have value at most $V_{j-1}(\ell - q)$ by the induction hypothesis. Maximizing over q gives the upper bound in (21); concatenating an optimizer for $V_{j-1}(\ell - q)$ with q tokens on branch j attains every candidate, giving equality. Thus $V_J(L)$ is exact. Storing a maximizing q at each state permits backtracking. There are JL states and at most $L + 1$ transitions per state, giving the claimed complexity. □

Proof of Lemma 16. Represent assigning n_j tokens to branch j by selecting the prefix

$$\Delta_j(1), \ldots, \Delta_j(n_j)$$

from its marginal-gain sequence. Because each sequence is nonincreasing, if a marginal $\Delta_j(q)$ is selected, every earlier marginal on that branch is at least as large and can also be selected. Hence a set of the L largest marginals (with any tie-breaking that respects prefixes) is a feasible allocation.

At each step, the greedy rule selects the largest marginal whose predecessors have already been selected. If a globally larger unselected marginal existed, all of its predecessors would be at least as large; the first unselected member of that prefix would be currently available and would contradict the greedy choice. Thus after L steps greedy has selected L globally largest marginals. No other allocation, being a selection of L marginals subject to the same prefix constraints, can have larger sum. Nonnegativity also shows that an at-most-L optimum may be taken to use all available positive-gain tokens. □

Proof of Lemma 17. Lemma 13 reduces the indirect problem exactly to coding the Gaussian predictable signal $m(W)$, plus the fixed distortion Δ_W. An orthogonal transform diagonalizes its

Table 17: Identity-suite leakage maps (uncompressed Bayes). Floor is $(1-\chi^2)v$ with $v=1$.

Design	χ	floor / uncomp. MSE	Pearson
Shared Gaussian factor	0.65	0.578 / 0.575	0.652
Quadratic map	1	0 / 0	0.008
Noisy parity	0.60	0.640 / 0.641	0.004
Parity	1	0 / 0	0.006

covariance without changing squared error or mutual information. It therefore remains to derive the RDF of

$$M = (M_1, \dots, M_q), \qquad M_j \text{ independent}, \quad M_j \sim \mathcal{N}(0, \lambda_j).$$

For any reproduction $\widehat{M}$, replacing it by the posterior mean $\mathbb{E}[M \mid \widehat{M}]$ cannot increase distortion or mutual information. Let

$$K = \mathbb{E}[(M - \mathbb{E}[M \mid \widehat{M}])(M - \mathbb{E}[M \mid \widehat{M}])^\top].$$

The law of total covariance gives $0 \preceq K \preceq \Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_q)$, and $\operatorname{tr} K$ is the distortion. Conditional maximum-entropy and concavity of $\log\det$ give

$$\begin{aligned} I(M; \widehat{M}) &= h(M) - h(M \mid \widehat{M}) \\ &\geq \frac{1}{2} \log_2 \frac{\det \Lambda}{\det K}. \end{aligned}$$

In the eigenbasis of Λ, Hadamard's inequality gives $\det K \leq \prod_j K_{jj}$, while $0 \leq K_{jj} \leq \lambda_j$ and $\sum_j K_{jj} = \operatorname{tr} K$. Thus the converse reduces to the finite-dimensional optimization

$$\min_{\substack{0 < d_j \leq \lambda_j \\ \sum_j d_j \leq D'}} \frac{1}{2} \sum_{j=1}^{q} \log_2 \frac{\lambda_j}{d_j},$$

where D' is distortion on the predictable signal.

The KKT conditions set all unsaturated d_j's to a common value θ, while the constraint $d_j \leq \lambda_j$ gives

$$d_j^* = \min\{\lambda_j, \theta\}.$$

The associated rate and distortion are

$$R = \frac{1}{2} \sum_j \left[\log_2 \frac{\lambda_j}{\theta} \right]^+, \qquad D' = \sum_j \min\{\lambda_j, \theta\}.$$

This converse is attainable. For every active coordinate $\lambda_j > \theta$, take the backward Gaussian test channel

$$M_j = \widehat{M}_j + Z_j, \qquad \widehat{M}_j \sim \mathcal{N}(0, \lambda_j - \theta), \quad Z_j \sim \mathcal{N}(0, \theta),$$

with independent terms; for an inactive coordinate $\lambda_j \leq \theta$, set $\widehat{M}_j = 0$, so its error is λ_j. The resulting mutual informations and distortions add and meet the converse. Shannon random coding implements these test channels asymptotically.

Adding Δ_W proves (22)–(23). For $R > 0$, the right side of (23) is continuous and strictly decreasing in $\theta \in (0, \lambda_1)$, ranging from $+\infty$ to 0, so the water level is unique. The $(j+1)$-st mode enters when $\theta = \lambda_{j+1}$. Substitution into the rate formula leaves only the first j positive terms and yields (24). Equal eigenvalues produce equal thresholds and hence simultaneous activation. □

F Proofs of the Product-Kernel Law

Proof of Theorem 1. Under the product law (3), the pairs (X_i, U_i) are mutually independent. Conditional factorization gives

$$\mathbb{E}[Y \mid U_1, \dots, U_k] = \sqrt{v} \prod_{i=1}^{k} \mathbb{E}[X_i \mid U_i]. \tag{26}$$

Let $\mu_i(U_i) = \mathbb{E}[X_i \mid U_i]$. Since $\mathbb{E}[X_i^2] = 1$, the scalar projection identity gives

$$\mathbb{E}[\mu_i(U_i)^2] = 1 - D_i.$$

The $\mu_i(U_i)$'s are independent, and $\mathbb{E}[Y^2] = v$. Therefore

$$\mathbb{E}\big[\mathbb{E}[Y \mid U_1, \dots, U_k]^2\big] = v \prod_{i=1}^{k} (1 - D_i).$$

Subtracting this explained variance from $\mathbb{E}[Y^2]$ proves (4). It also proves that the displayed posterior mean is the unique minimum-MSE decoder up to almost-sure equality.

Lemma 14 and the rate constraints imply

$$D_i \geq 2^{-2I(X_i;U_i)} \geq 2^{-2r_i}.$$

Thus $1 - D_i \leq 1 - 2^{-2r_i}$, and (4) gives the lower bound in (5).

For achievability, set $d_i = 2^{-2r_i}$, let the N_i's be independent standard Gaussians independent of all X_i, and take

$$U_i = \sqrt{1 - d_i}\, X_i + \sqrt{d_i}\, N_i.$$

Then $U_i \sim \mathcal{N}(0, 1)$, $\mathrm{Corr}(X_i, U_i)^2 = 1 - d_i$, and hence

$$I(X_i; U_i) = -\frac{1}{2} \log_2(1 - \mathrm{Corr}(X_i, U_i)^2) = -\frac{1}{2} \log_2 d_i = r_i,$$

while $\mathrm{Var}(X_i \mid U_i) = d_i$. Substitution into (4) proves achievability.

Suppose now that $v > 0$ and every r_i is finite and strictly positive. All target factors $1 - 2^{-2r_i}$ are then positive. Equality of the product in the lower-bound argument forces $D_i = 2^{-2r_i}$ for every i. Both inequalities

$$D_i \geq 2^{-2I(X_i;U_i)} \geq 2^{-2r_i}$$

must therefore be equalities. Thus $I(X_i; U_i) = r_i$, and Lemma 14 gives exactly the stated scalar equality condition. The converse follows from the same lemma.

If $r_j = 0$, then $I(X_j; U_j) = 0$, so $X_j \perp U_j$ and $\mathbb{E}[X_j \mid U_j] = 0$. Equation (26) is then zero regardless of the other channels, and the MMSE equals v. The case $v = 0$ is immediate. □

Proof of Theorem 2. Converse. Fix a blocklength n. Separate encoder kernels and independent source blocks imply that $M_1, \dots, M_k$ are mutually independent. For each modality i and coordinate t, define

$$\mu_{i,t}(M_i) := \mathbb{E}[X_{i,t} \mid M_i], \qquad q_{i,t} := \mathbb{E}[\mu_{i,t}(M_i)^2], \qquad s_{i,t} := I(X_{i,t}; M_i).$$

The same conditional-factorization argument as in (26) gives

$$\mathbb{E}[Y_t \mid M_1, \dots, M_k] = \sqrt{v} \prod_{i=1}^{k} \mu_{i,t}(M_i).$$

Consequently, the Bayes explained variance at coordinate t is

$$\mathbb{E}\big[\mathbb{E}[Y_t \mid M_1, \dots, M_k]^2\big] = v \prod_{i=1}^{k} q_{i,t}.$$

Every other decoder has at least the Bayes distortion.

Applying Lemma 14 to the scalar pair $(X_{i,t}, M_i)$ yields

$$q_{i,t} = 1 - \mathrm{mmse}(X_{i,t} \mid M_i) \leq 1 - 2^{-2s_{i,t}}. \tag{27}$$

The source coordinates are independent, so

$$I(X_{i,t}; M_i) \leq I(X_{i,t}; M_i, X_{i,1}^{t-1}) = I(X_{i,t}; M_i \mid X_{i,1}^{t-1}),$$

where the equality uses $I(X_{i,t}; X_{i,1}^{t-1}) = 0$. Summing and applying the chain rule,

$$\sum_{t=1}^{n} s_{i,t} \le I(X_i^n; M_i) \le H(M_i) \le \log_2 |\mathcal{M}_{i,n}|. \tag{28}$$

Let T be uniform on $[n]$, independently of everything, and put

$$\boldsymbol{S}_T = (s_{1,T}, \dots, s_{k,T}).$$

From (27), (28), and the definition of the closed upper concave envelope,

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\big[\mathbb{E}[Y_t \mid M_1, \dots, M_k]^2\big] &\le \frac{v}{n}\sum_{t=1}^{n} F_k(s_{1,t}, \dots, s_{k,t}) \\ &= v\,\mathbb{E}[F_k(\boldsymbol{S}_T)] \\ &\le v\,\overline{\mathrm{cav}} F_k\left(\frac{1}{n}\log|\mathcal{M}_{1,n}|, \dots, \frac{1}{n}\log|\mathcal{M}_{k,n}|\right).\end{aligned}$$

Taking the rate-constrained limit and using the closedness and coordinatewise monotonicity of the envelope proves

$$\liminf_{n\to\infty} \frac{1}{n}\mathbb{E}\|Y^n - \widehat{Y}^n\|^2 \ge v\left[1 - \overline{\mathrm{cav}} F_k(\boldsymbol{r})\right].$$

Achievability for one fixed rate profile. Fix $\boldsymbol{\rho} \in \mathbb{R}_+^k$ and $\eta > 0$. The operational Gaussian rate–distortion theorem supplies, for all sufficiently large blocklengths N, a separate code for source X_i^N with

$$\frac{1}{N}\log_2 |\mathcal{M}_i| \le \rho_i + \eta$$

and a reconstruction $\widetilde{X}_i^N$ satisfying

$$\frac{1}{N}\sum_{t=1}^{N} \mathbb{E}[(X_{i,t} - \widetilde{X}_{i,t})^2] \le 2^{-2\rho_i} + \eta.$$

Replacing each reconstruction coordinate by $\mathbb{E}[X_{i,t} \mid M_i]$ cannot increase MSE. For the resulting profile $q_{i,t} = \mathbb{E}[\mathbb{E}[X_{i,t} \mid M_i]^2]$,

$$\overline{q}_i := \frac{1}{N}\sum_{t=1}^{N} q_{i,t} \ge 1 - 2^{-2\rho_i} - \eta. \tag{29}$$

The coordinate profiles of different component codes need not be aligned. For each i, relabel the source coordinates presented to its code by a permutation π_i, and undo this relabeling at the decoder. If the π_i's are sampled independently and uniformly, then

$$\mathbb{E}_{\pi_1,\dots,\pi_k}\left[\frac{1}{N}\sum_{t=1}^{N}\prod_{i=1}^{k} q_{i,\pi_i(t)}\right] = \prod_{i=1}^{k}\left(\frac{1}{N}\sum_{t=1}^{N} q_{i,t}\right) = \prod_{i=1}^{k}\overline{q}_i. \tag{30}$$

Hence a deterministic tuple of permutations attains at least the right-hand side. With this fixed tuple and the coordinatewise product posterior-mean decoder, the explained variance per coordinate is at least

$$v\prod_{i=1}^{k}(1 - 2^{-2\rho_i} - \eta),$$

which tends to $vF_k(\boldsymbol{\rho})$ as $\eta \downarrow 0$. The permutations are fixed parts of the code and are not shared randomness.

Achievability of every mixture. Consider a finite mixture $\{(\lambda_j, \boldsymbol{\rho}^{(j)}) : j = 1, \dots, J\}$ feasible in (16). First approximate its weights by rational weights. Partition every modality block using the same deterministic subblocks of relative lengths λ_j. This alignment is essential: on subblock j, all

modalities use the code for the same profile $\boldsymbol{\rho}^{(j)}$. Within each subblock apply the deterministic permutation construction above. Concatenating the local messages gives modality-i rate

$$\sum_j \lambda_j \rho_i^{(j)} + o(1) \le r_i + o(1),$$

and explained variance

$$v \sum_j \lambda_j F_k(\boldsymbol{\rho}^{(j)}) - o(1).$$

Irrational weights follow by approximation, and taking a sequence of finite mixtures approaching the closed supremum proves achievability of (6). □

F.1 Closed-form envelope

Write $a := 2 \ln 2$ and

$$A_k(\boldsymbol{r}) := \sum_{i=1}^k \frac{a r_i}{e^{a r_i} - 1},$$

where the summand has continuous value 1 at $r_i = 0$.

Theorem 18 (Exact multivariate product-gain envelope)**.** *For every $\boldsymbol{r} \in \mathbb{R}_+^k$:*

1. *If some $r_i = 0$, then*
$$\overline{\mathrm{cav}} F_k(\boldsymbol{r}) = 0.$$
2. *If every $r_i > 0$ and $A_k(\boldsymbol{r}) \le 1$, then*
$$\overline{\mathrm{cav}} F_k(\boldsymbol{r}) = F_k(\boldsymbol{r}).$$
3. *If every $r_i > 0$ and $A_k(\boldsymbol{r}) > 1$, there is a unique $\tau(\boldsymbol{r}) > 1$ such that*
$$A_k(\tau(\boldsymbol{r})\boldsymbol{r}) = 1,$$
and
$$\overline{\mathrm{cav}} F_k(\boldsymbol{r}) = \frac{F_k(\tau(\boldsymbol{r})\boldsymbol{r})}{\tau(\boldsymbol{r})}. \tag{31}$$

The optimizer in the third case time-shares between $\mathbf{0}$ and the contact point $\tau(\boldsymbol{r})\boldsymbol{r}$. Non-collinear mixtures cannot improve the value.

Proof of Theorem 18. It is convenient to rescale coordinates. Put

$$x_i = a r_i, \qquad f(\boldsymbol{x}) = \prod_{i=1}^k (1 - e^{-x_i}), \qquad \mathcal{A}(\boldsymbol{x}) = \sum_{i=1}^k \frac{x_i}{e^{x_i} - 1}.$$

Positive diagonal rescaling preserves mixtures and concavity, so it suffices to identify the upper concave envelope of f on $\mathbb{R}_+^k$.

Boundary. If $x_j = 0$, then $f(\boldsymbol{x}) = 0$. For every $\boldsymbol{z} \ge 0$,

$$f(\boldsymbol{z}) \le 1 - e^{-z_j} \le z_j.$$

Thus the affine function $\boldsymbol{z} \mapsto z_j$ is a global majorant that vanishes at $\boldsymbol{x}$. The least concave majorant is bounded below by $f(\boldsymbol{x}) = 0$ and above by this affine majorant, so its value is zero.

A global supporting-plane lemma. Fix an interior point $\boldsymbol{x} > 0$ with $\mathcal{A}(\boldsymbol{x}) \le 1$, and set

$$b_i = \frac{1}{e^{x_i} - 1}, \qquad c = 1 - \sum_{i=1}^k b_i x_i = 1 - \mathcal{A}(\boldsymbol{x}) \ge 0.$$

Since $\partial_i f(\boldsymbol{x}) = f(\boldsymbol{x}) b_i$, the tangent plane is

$$T_{\boldsymbol{x}}(\boldsymbol{z}) = f(\boldsymbol{x}) \left(c + \sum_{i=1}^k b_i z_i \right). \tag{32}$$

We prove that this tangent plane majorizes f globally.

On the interior of the orthant, consider

$$\mathcal{R}(\boldsymbol{z}) = \frac{f(\boldsymbol{z})}{c + \sum_i b_i z_i}.$$

Boundary limits are zero, and $\mathcal{R}(\boldsymbol{z}) \to 0$ whenever $\|\boldsymbol{z}\| \to \infty$. (When $c = 0$, the all-zero boundary limit is also zero because then $k \geq 2$ and the numerator has product order while the denominator has first order.) A positive maximum therefore occurs at an interior stationary point. If $d = c + \sum_i b_i z_i$, stationarity of $\log \mathcal{R}$ gives

$$\frac{1}{e^{z_i} - 1} = \frac{b_i}{d}, \qquad z_i = \log\left(1 + \frac{d}{b_i}\right).$$

Substitution into the definition of d shows that $d > 0$ is a root of

$$\Phi(d) := c + \sum_{i=1}^{k} b_i \log\left(1 + \frac{d}{b_i}\right) - d.$$

Now

$$\Phi(1) = c + \sum_i b_i x_i - 1 = 0, \qquad \Phi''(d) = -\sum_i \frac{b_i}{(b_i + d)^2} < 0.$$

Moreover,

$$\Phi'(1) = \sum_i \frac{b_i}{b_i + 1} - 1 = \sum_i e^{-x_i} - 1 \leq \mathcal{A}(\boldsymbol{x}) - 1 \leq 0,$$

where

$$e^{-x} \leq \frac{x}{e^x - 1} \quad (x > 0)$$

follows from $e^x(x - 1) + 1 \geq 0$. Strict concavity, together with $\Phi(0) = c \geq 0$, implies $\Phi(d) > 0$ for $0 < d < 1$; since Φ' is strictly decreasing and $\Phi'(1) \leq 0$, it implies $\Phi(d) < 0$ for $d > 1$. Hence $d = 1$ is the unique positive root. The stationary point is therefore

$$z_i = \log(1 + 1/b_i) = x_i.$$

It is the global maximizer, and

$$\mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}) = f(\boldsymbol{x}),$$

which is exactly $f(\boldsymbol{z}) \leq T_{\boldsymbol{x}}(\boldsymbol{z})$.

Points at or beyond contact. If $\mathcal{A}(\boldsymbol{x}) \leq 1$, the global affine majorant (32) touches f at $\boldsymbol{x}$. The least concave majorant is at least f and no larger than this tangent plane, hence equals $f(\boldsymbol{x})$.

Points before contact. Suppose $\mathcal{A}(\boldsymbol{x}) > 1$. Each summand $tx_i/(e^{tx_i} - 1)$ is continuous and strictly decreasing in $t > 0$, with limit zero as $t \to \infty$. There is therefore a unique $\tau > 1$ for which $\mathcal{A}(\tau\boldsymbol{x}) = 1$. Set $\boldsymbol{y} = \tau\boldsymbol{x}$. At $\boldsymbol{y}$, the constant c in (32) is zero, so its global tangent majorant is linear:

$$T_{\boldsymbol{y}}(\boldsymbol{z}) = f(\boldsymbol{y}) \sum_i \frac{z_i}{e^{y_i} - 1}.$$

Because $\sum_i y_i/(e^{y_i} - 1) = 1$,

$$T_{\boldsymbol{y}}(\boldsymbol{x}) = \frac{1}{\tau} T_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{f(\boldsymbol{y})}{\tau}.$$

This proves that the envelope is at most $f(\boldsymbol{y})/\tau$.

Conversely, every concave majorant $h \geq f$ obeys

$$h(\boldsymbol{x}) = h\left(\frac{1}{\tau}\boldsymbol{y} + \left(1 - \frac{1}{\tau}\right)\mathbf{0}\right) \geq \frac{1}{\tau} h(\boldsymbol{y}) + \left(1 - \frac{1}{\tau}\right) h(\mathbf{0}) \geq \frac{f(\boldsymbol{y})}{\tau}.$$

The upper and lower bounds agree. Returning to $\boldsymbol{r} = \boldsymbol{x}/a$ proves (31).

Finally, for any finite mixture with mean $\boldsymbol{x}$, the same global tangent plane at $\boldsymbol{y}$ gives

$$\sum_j \lambda_j f(\boldsymbol{z}^{(j)}) \leq \sum_j \lambda_j T_{\boldsymbol{y}}(\boldsymbol{z}^{(j)}) = T_{\boldsymbol{y}}(\boldsymbol{x}) = \frac{f(\boldsymbol{y})}{\tau}.$$

Thus no non-collinear mixture improves on the radial two-point mixture. □

Proof of Corollary 3. Let $a = 2\ln 2$ and $\ell(r) = \ln(1 - e^{-ar})$ for $r > 0$. Direct differentiation gives

$$\ell''(r) = -\frac{a^2 e^{ar}}{(e^{ar} - 1)^2} < 0.$$

For $B > 0$, an allocation with a zero coordinate has zero product gain, whereas the equal positive allocation has positive gain. Thus a maximizer is interior. Since F_k is increasing, it uses all of B, and strict Jensen gives

$$\log F_k(\boldsymbol{r}) = \sum_{i=1}^{k} \ell(r_i) \le k\ell(B/k),$$

with equality exactly when $r_i = B/k$. This proves (7).

We next relate vector time sharing to scalar time sharing. For every random rate vector $\boldsymbol{R}$, put $T = \sum_i R_i$. The equal-split result applied pointwise gives

$$F_k(\boldsymbol{R}) \le u_k(T).$$

If $\mathbb{E}R_i \le B/k$, then $\mathbb{E}T \le B$; hence

$$\begin{aligned}\mathbb{E}F_k(\boldsymbol{R}) &\le \mathbb{E}u_k(T)\\ &\le \mathbb{E}[\operatorname{cav} u_k(T)]\\ &\le \operatorname{cav} u_k(\mathbb{E}T)\\ &\le \operatorname{cav} u_k(B),\end{aligned}$$

where the third step is Jensen's inequality and the last uses monotonicity of the envelope. Conversely, any finite time-sharing law for T can be implemented by setting $R_i = T/k$. Therefore

$$\overline{\operatorname{cav}} F_k(B/k, \dots, B/k) = \operatorname{cav} u_k(B).$$

The same argument applies when only a transferable total average budget is imposed.

It remains to identify the scalar envelope. Put $x = aB/k$. For

$$u_k(B) = (1 - e^{-x})^k,$$

calculation gives

$$\frac{u_k'(B)}{u_k(B)} = \frac{a}{e^x - 1}, \tag{33}$$

$$u_k''(B) = \frac{a^2}{k} e^{-x}(1 - e^{-x})^{k-2}(ke^{-x} - 1). \tag{34}$$

For $k = 1$, (34) is negative for $B > 0$, so no concavification is required.

Let $k \ge 2$. Tangency of the line from the origin is

$$Bu_k'(B) = u_k(B).$$

Using $B = kx/a$ and (33), this is

$$\frac{kx}{e^x - 1} = 1,$$

equivalent to (8). The function $\psi(x) = e^x - 1 - kx$ has $\psi(0) = 0$, $\psi'(0) = 1 - k < 0$, $\psi''(x) > 0$, and tends to $+\infty$. Hence it has exactly one positive root x_k. Since $\psi(\ln k) = k - 1 - k\ln k < 0$, this root satisfies $x_k > \ln k$.

Furthermore,

$$\frac{d}{dB}\frac{u_k(B)}{B} = \frac{Bu_k'(B) - u_k(B)}{B^2}.$$

The numerator is positive for $0 < B < B_k^0$, zero at B_k^0, and negative afterward. Thus the line

$$L(B) = \frac{B}{B_k^0} u_k(B_k^0)$$

majorizes $u_k(B)$ on $[0, B_k^0]$. By (34) and $x_k > \ln k$, u_k is concave on $[B_k^0, \infty)$; the derivatives of L and u_k agree at B_k^0. The piecewise function in (9) is therefore a concave majorant.

Any concave majorant $h \ge u_k$ must, for $B = \theta B_k^0 \in [0, B_k^0]$, satisfy

$$h(B) \ge \theta h(B_k^0) + (1 - \theta)h(0) \ge \theta u_k(B_k^0) = L(B).$$

For $B \ge B_k^0$, it must be at least $u_k(B)$. The displayed piecewise majorant is therefore the least one, proving (9) and the operational setup statement. □

G PROOFS OF COALITIONAL WATER-FILLING

Proof of Theorem 5. Fix incidence rates $\boldsymbol{r}$. By Theorem 2, edge e has exact asymptotic explained variance $v_e H_e(\boldsymbol{r}_e)$ when encoded by its dedicated incidence messages. Distortion is summed over edge targets, so the gains add. Concatenating the finitely many incidence messages at each node adds their logarithmic cardinalities. Consequently every feasible rate vector in (10) is asymptotically achievable.

Conversely, Definition 11 provides a dedicated rate $r_{e,i}$ for every incidence and permits edge e's decoder to use only $(M_{e,i})_{i\in e}$. The node sources within each edge are independent, so applying Theorem 2 separately to every edge upper-bounds its gain by $v_e H_e(\boldsymbol{r}_e)$. The message-rate constraints sum to the displayed node and global constraints. Maximizing these edgewise upper bounds proves (10) and the distortion identity.

Each H_e is closed, concave, and nondecreasing by construction. The feasible set is a compact polytope after removing incidences forced to zero, so an optimizer exists. Under the stated relative-interior condition, standard concave KKT duality has no gap. Associate $\alpha_i \geq 0$ with the node constraint and $\beta \geq 0$ with the global constraint. Up to budget-dependent constants, the Lagrangian is

$$\sum_{e\in\mathcal{E}} \left[v_e H_e(\boldsymbol{r}_e) - \sum_{i\in e} (\alpha_i + \beta) r_{e,i} \right].$$

It separates over edges. Maximization of each bracket is exactly (11); primal feasibility, dual feasibility, and the displayed complementary-slackness equations complete the necessary and sufficient KKT system.

We finally derive the marginal formula. In the post-contact region $A_{|e|}(\boldsymbol{r}_e) \leq 1$, Theorem 18 gives $H_e = F_{|e|}$, so

$$\frac{\partial H_e}{\partial r_{e,i}} = \frac{a F_{|e|}(\boldsymbol{r}_e)}{e^{a r_{e,i}} - 1}.$$

In the pre-contact region, write $\boldsymbol{c} = \tau(\boldsymbol{r}_e)\boldsymbol{r}_e$. Then

$$H_e(\boldsymbol{r}_e) = \frac{F_{|e|}(\boldsymbol{c})}{\tau}.$$

The contact equation is

$$\nabla F_{|e|}(\boldsymbol{c})^\top \boldsymbol{c} = F_{|e|}(\boldsymbol{c}),$$

because its left side divided by $F_{|e|}(\boldsymbol{c})$ is $A_{|e|}(\boldsymbol{c}) = 1$. Differentiating $F(\tau\boldsymbol{r})/\tau$, all terms containing $\partial\tau/\partial r_{e,i}$ therefore cancel, leaving

$$\frac{\partial H_e}{\partial r_{e,i}} = \frac{\partial F_{|e|}}{\partial c_i}(\boldsymbol{c}) = \frac{a F_{|e|}(\boldsymbol{c})}{e^{a c_i} - 1}.$$

Stationarity at a positive incidence sets $v_e \partial_i H_e = p_i$, proving (12). At a boundary point, concave supergradient KKT conditions give the stated inequality.

The converse in the first paragraph invoked edge separation explicitly. Nothing in the argument tensorizes a message coupled across edges, which is why no unrestricted-code converse is claimed. □

Proof of Corollary 6. Order-scalar allocations form a subset of the feasible allocations in (10); weak dominance is immediate.

For the strict example, a singleton edge has

$$H_{\{i\}}(r) = 1 - 2^{-2r} = 1 - e^{-ar}.$$

The full problem is

$$\max_{r_1,r_2\geq 0,\ r_1+r_2\leq B} \{1 - e^{-a r_1} + \varepsilon(1 - e^{-a r_2})\}.$$

At $(r_1, r_2) = (B, 0)$, the marginal on the first edge is $a e^{-aB} = a 2^{-2B}$, while the marginal on the second edge at zero is $a\varepsilon$. Since $\varepsilon < 2^{-2B}$, concavity and the KKT conditions show that assigning all rate to edge 1 is the unique optimum. Its gain is $1 - 2^{-2B}$.

The order-scalar constraint forces $r_1 = r_2 = B/2$ and gives $(1+\varepsilon)(1-2^{-B})$. With $q = 2^{-B}$, the difference is

$$(1-q^2) - (1+\varepsilon)(1-q) = (1-q)(q-\varepsilon) > 0,$$

because $\varepsilon < q^2 < q$. This proves strict separation. □

H Proofs of Plug-in Allocation

Proof of Theorem 7. Telescoping the marginal errors gives, for every branch and token count,

$$|\widehat{G}_j(q) - G_j(q)| = \left|\sum_{s=1}^{q} (\widehat{\Delta}_j(s) - \Delta_j(s))\right| \le q\epsilon.$$

Therefore every feasible allocation n using at most L tokens satisfies

$$|\widehat{G}(n) - G(n)| \le \sum_j n_j \epsilon \le L\epsilon, \tag{35}$$

where $G(n) = \sum_j G_j(n_j)$ and similarly for $\widehat{G}$. Optimality of n^* for G and of $\widehat{n}$ for $\widehat{G}$, together with (35), gives

$$\begin{aligned} G(n^*) - G(\widehat{n}) &\le \widehat{G}(n^*) + L\epsilon - \widehat{G}(\widehat{n}) + L\epsilon \\ &\le 2L\epsilon. \end{aligned}$$

This proves (13).

If n^* is unique with true objective gap $\gamma > 2L\epsilon$, then for every $n \neq n^*$,

$$\widehat{G}(n^*) - \widehat{G}(n) \ge G(n^*) - G(n) - 2L\epsilon \ge \gamma - 2L\epsilon > 0.$$

Thus n^* is also the unique plug-in optimizer.

Under diminishing gains, the true optimum consists of the L largest prefix-compatible marginals by Lemma 16. If its smallest selected marginal exceeds the largest unselected marginal by more than 2ϵ, every selected estimated marginal still exceeds every unselected estimated marginal. The selected set is prefix-closed and hence feasible, so it is also the unique set of L largest estimated marginals. The plug-in allocation is therefore exactly n^*. If an at-most budget is used, a zero-valued dummy marginal can be included among the unselected marginals to certify that all selected estimated gains remain positive. □

I Proofs of the Leakage Floor

Proof of Theorem 8. Write

$$p = P_A Y_S = \mathbb{E}[Y_S \mid X_A], \qquad z = Z_S = Y_S - p.$$

Then $\mathbb{E}[z \mid X_A] = 0$, $p \perp z$, and

$$\|p\|_2^2 = \delta_{S,T} v_S, \qquad \|z\|_2^2 = (1-\delta_{S,T}) v_S. \tag{36}$$

Since $X_T = (X_A, X_C)$,

$$\mathbb{E}[Y_S \mid X_T] = p + \mathbb{E}[z \mid X_A, X_C].$$

The two terms on the right are orthogonal: by the tower property,

$$\mathbb{E}[\mathbb{E}[z \mid X_A, X_C] \mid X_A] = \mathbb{E}[z \mid X_A] = 0.$$

Definition 12, applied to $h = z$, now gives

$$\begin{aligned} \|\mathbb{E}[Y_S \mid X_T]\|_2^2 &= \|p\|_2^2 + \|\mathbb{E}[z \mid X_A, X_C]\|_2^2 && (37) \\ &\le \delta_{S,T} v_S + \chi(B, C \mid A)^2 (1-\delta_{S,T}) v_S. && (38) \end{aligned}$$

For a stochastic representation satisfying $U - X_T - Y_S$,

$$\mathbb{E}[Y_S \mid U] = \mathbb{E}[\mathbb{E}[Y_S \mid X_T] \mid U].$$

Conditional expectation is an L^2 contraction, so

$$\|\mathbb{E}[Y_S \mid U]\|_2 \le \|\mathbb{E}[Y_S \mid X_T]\|_2.$$

Combining this with (38) proves (14). Finally, the scalar projection identity

$$\mathrm{mmse}(Y_S \mid U) = v_S - \|\mathbb{E}[Y_S \mid U]\|_2^2$$

gives (15).

If $X_B \perp X_C \mid X_A$, every admissible h in (17) satisfies

$$\mathbb{E}[h \mid X_A, X_C] = \mathbb{E}[h \mid X_A] = 0,$$

so $\chi = 0$. If $\chi = 1$, the definition permits innovations that are perfectly predictable from (X_A, X_C), and no positive uniform innovation floor can be deduced. These observations establish the stated special cases. □