RATE–DISTORTION DICTIONARIES FOR RECONSTRUCTION AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying a family of β-VAE codecs at multiple rate–distortion operating points is straightforward; deciding which of them to keep is not. We study the problem of selecting a small dictionary of k frozen codecs from a larger candidate frontier so that every future request is served by its single nearest deployed atom. We show that this constrained k-medoids problem is not solvable by the classical one-dimensional clustering dynamic program, because nearest-atom assignment is provably non-monotone once matching is judged against points external to the frontier. We restore tractability by freezing, independently of k, which frontier atom each request would use on the full unreduced frontier, which makes the resulting surrogate objective exactly solvable by an $O(M^2(N+k))$ dynamic program. We then certify this surrogate: an exact identity for its gap to the true deployed loss, a geometric bound computable before optimization, and a finite-sample guarantee that extends the certificate to the underlying population objective. Experiments across three datasets show the certificate is non-vacuous and that the resulting dictionary matches or exceeds standard selection heuristics.

1 INTRODUCTION

Train enough β-VAEs on the same data, one per value of β, and the result is more useful than M independent models: a rate–distortion *frontier*, a curve of achievable operating points from which a downstream system can select. A single neural codec cannot serve every request. If a deployment must reconstruct inputs at one rate today and generate novel samples at another rate tomorrow, the natural but non-scalable answer is to keep a separate frozen codec for every operating point it might ever need. A deployed system cannot afford to keep all M codecs on hand, and it cannot retrain one on demand either; it must instead commit, in advance, to a small dictionary of k atoms and route every incoming request to whichever one fits best. The question this raises is simple to state and, as we show, considerably less simple to answer: given a large family of independently trained β-VAE codecs occupying a rate–distortion frontier, how many of them does a deployment actually need to keep, and which ones?

What makes this frontier worth curating carefully, rather than simply covered evenly, is that every codec on it is already doing two jobs at once. Feed it a real input and decode from the encoder's latent, and the result is a reconstruction. Skip the input entirely and decode from the codec's own learned prior, and the result is a generated sample. Both readouts share the same frozen decoder; only where the latent originates differs. The rate parameter used to trace the frontier governs both readouts simultaneously, and it governs them in opposite directions. A codec forced to a low rate discards most of what made a given input unique, which is precisely what allows a prior sample to decode into a coherent, generic member of the data distribution, at the cost of a poor reconstruction of that specific input. A codec allowed a high rate keeps enough of the input to reconstruct it almost exactly, but its aggregate posterior no longer matches the prior against which it was trained, so sampling from that prior decodes into something distorted rather than novel. The same rate parameter that trades off reconstruction against distortion therefore also trades off reconstruction quality against generation quality, and this second axis is one a fixed dictionary must cover deliberately. This turns dictionary selection into a genuinely two-sided problem, not a one-dimensional rate–distortion sweep, and it is why keeping the *right* k points, rather than k evenly spaced ones, is what matters.

Formally, dictionary selection is a constrained k-medoids problem: centers are restricted to points already realized on the candidate frontier, cost is squared distance under a fixed metric, and each request is served by its single nearest deployed atom. Because the frontier is sorted and one-dimensional in index, it is tempting to apply the classical ordered-partition dynamic program for one-dimensional clustering directly. This does not work, because that dynamic program assumes the points being clustered are the same ordered items being partitioned. Here they are not: requests are points *external* to the frontier, and once matching is judged in the full two-dimensional $(D, \bar{R})$ plane rather than along a single coordinate, nearest-atom assignment stops being monotone in frontier order. A request can sit closest

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 atom 3 while a neighboring request, despite lying "between" atoms 3 and 4 in distortion, is in fact closest to atom 7, because the frontier bends. No reordering of the classical recurrence recovers this structure.

Our fix: freeze first, choose atoms second. We restore tractability with a single idea: decide which frontier point each request would use *before* deciding which points survive into the dictionary. This anchor is computed once, on the full unreduced frontier, and depends neither on k nor on which subset is eventually chosen. Freezing it turns the objective into a sum over frontier segments, exactly the structure the ordered dynamic program requires, and admits an exact solution in $O(M^2(N+k))$ time.

Freezing the anchor first is also, unavoidably, an approximation of the true deployment rule. A request's frozen anchor need not survive into the final dictionary, in which case the anchor-routed surrogate reroutes it to the nearer endpoint of its segment rather than to its true nearest surviving atom. The relevant question is therefore not whether this shortcut is exact — it is not — but how far from exact it can be, and whether that distance is knowable in advance, before the dictionary is ever computed.

Contributions.

1. **An exact solver for an inexact objective.** We show that the anchor-routed surrogate is solved to global optimality by an $O(M^2(N+k))$ dynamic program, despite true nearest-atom assignment being provably non-ordered on the same frontier.
2. **A certificate, not just a heuristic.** We derive an exact identity for the gap between the surrogate and the true deployed loss, a geometric bound expressed in quantities observable before optimization, and a closed-form *a priori* regret certificate against the best achievable dictionary in the same feasible class — computable in $O(MN)$ time, before the dynamic program is ever run.
3. **From one sample to the population.** We extend the empirical certificate to a finite-sample guarantee on the true request distribution, and show that the deployed system degrades gracefully, rather than catastrophically, under request-distribution shift and under measurement error in the frontier itself.
4. **Validation across the whole chain.** Across three datasets and several request-distribution regimes, we check every link of this argument empirically: that the surrogate dominates the true loss as proven, that the certificate is not vacuous, and that the resulting dictionary matches or beats standard selection heuristics that search the same space without a certificate to show for it.

Related Work Our codecs are β-VAEs with learned priors (17), read through the rate–distortion view of (15) and rooted in the information bottleneck (1; 2; 3). We need no nested representations, unlike successive-refinement IB (5), and no convex relevance–compression curve (4): we work directly on the measured, Pareto-filtered points. One alternative conditions a single network on the rate parameter instead (23; 18; 24; 25; 26). Closer to us is work that routes requests across several learned codecs (27; 28). We ask for more: atoms are frozen before selection, one atom serves each request with no mixing, an exact optimizer chooses membership, and the result comes with an a priori regret certificate — not just a heuristic that happened to work well.

Underneath, this is a candidate-constrained facility-location / p-median problem (9; 10; 6; 12); swap local search (11) and integer programming are the natural baselines. The classical 1-D clustering DP (7; 8) needs the points being clustered to be the same items being ordered, which breaks once requests sit outside the frontier — the gap our anchor-routed surrogate closes. Bounding a relaxation's exactness gap to certify the real objective is a familiar move in approximation algorithms; what we add is a *computable* version of that gap for this bent, anisotropic geometry, plus the finite-sample step that carries it to the population loss. Wasserstein distance (14) does one simple job for us: turning Lipschitz stability of the nearest-atom distance into a bound on how much population risk moves under request-distribution shift.

2 Variable-Rate VIB Codecs: Reconstruction and Generation Modes

Each codec is a β-VAE with a learned prior, fig. 1. Read through the rate–distortion lens of (15), this is a variational information bottleneck in which the bottleneck variable is T and the relevance target is the input X itself (16; 15): the unsupervised instance of the deep VIB, with no auxiliary label and no classification head. We consider M_0 such stochastic codecs $i \in \{1, \ldots, M_0\}$, each defined by an encoder q_i, a decoder g_i, and a learned prior p_i, motivated by device–edge co-inference and split inference, where an information-bearing latent crosses a resource-constrained boundary (21; 22).

Encoder, decoder, prior. The encoder is a diagonal Gaussian

$$q_i(t \mid x) = \mathcal{N}\big(t; \mu_i(x), \mathrm{diag}(\sigma_i^2(x))\big), \tag{1}$$

with $\mu_i(x)$, $\log \sigma_i^2(x)$, $X_i \in [0,1]^{C\times H\times W}$, produced by a strided convolutional network. The decoder g_i maps a latent $T_i \in \mathbb{R}^d$ to $\hat{X}_i = g_i(T_i) \in [0,1]^{C\times H\times W}$ through a deconvolutional network with a final sigmoid. Each codec also learns a zero-mean diagonal-Gaussian prior

$$p_i(t) = \mathcal{N}\big(t; 0, \mathrm{diag}(\tau_i^2)\big), \qquad \tau_i^2 \in \mathbb{R}_+^d, \tag{2}$$

trained jointly with the encoder and decoder, in contrast to the fixed prior of the standard VAE. Crucially, g_i can be driven by a latent obtained *either* from q_i (conditioned on an input) *or* from p_i (with no input at all) — the two inference-time modes formalized in Section 3.2.

Distortion and rate. Define $D_i = \mathbb{E}\big[d_X(X, \hat{X}_i)\big]$, $d_X(x, \hat{x}) = \frac{\|x-\hat{x}\|_2^2}{CHW}$, $\bar{R}_i = \mathbb{E}_X\big[D_{\mathrm{KL}}\big(q_i(T \mid X) \,\|\, p_i(T)\big)\big]$. With the aggregate posterior $q_i(t)$, the exact decomposition (20)

$$\bar{R}_i = I_{q_i}(X; T_i) + D_{\mathrm{KL}}\big(q_i(T_i) \,\|\, p_i(T_i)\big) \geq I_{q_i}(X; T_i) \tag{3}$$

gives $I_{q_i}(X;T_i) \leq \bar{R}_i$, with equality iff $p_i = q_i$ almost everywhere — generally unattainable in the diagonal-Gaussian family; the slack $D_{\mathrm{KL}}(q_i \| p_i) > 0$ is the price of the closed form. Since both distributions are diagonal Gaussians, $\bar{R}_i = \mathbb{E}_X\left[\frac{1}{2}\sum_{\ell=1}^{d}\left(\frac{\sigma_{i,\ell}^2(X)}{\tau_{i,\ell}^2} + \frac{\mu_{i,\ell}^2(X)}{\tau_{i,\ell}^2} - 1 + \log\frac{\tau_{i,\ell}^2}{\sigma_{i,\ell}^2(X)}\right)\right]$, evaluated identically for every codec.

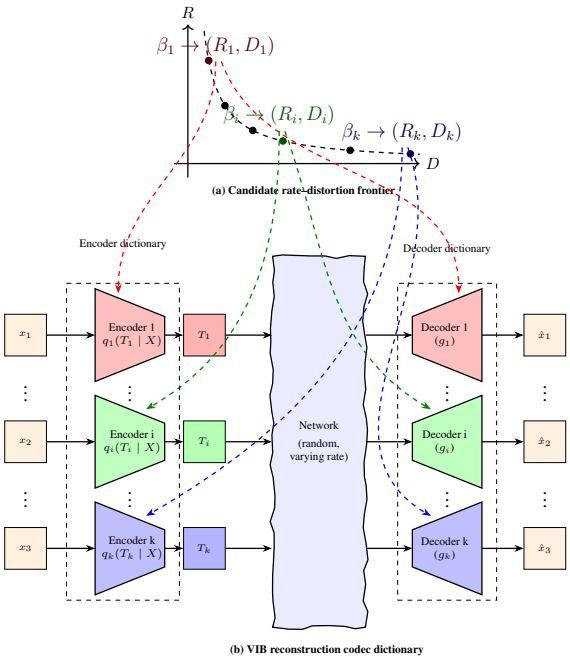


Figure 1: Frozen codec dictionary for request-aware reconstruction/generative. (a) Each independently trained encoder–decoder pair defines a point on the distortion–rate frontier $\mathcal{F}$. (b) At deployment, an external request is served by exactly one frozen atom, selected as its nearest neighbor under the matching metric.

Why a learned prior. For a *fixed* encoder, the prior minimizing $\mathbb{E}_X D_{\mathrm{KL}}(q_i(T \mid X) \| p(T))$ over all priors is the aggregate posterior $q_i(t) = \int p_{\mathrm{data}}(x) q_i(t \mid x)\, dx$. A fixed $\mathcal{N}(0, I)$ prior is generally far from this minimizer and therefore *overstates* the rate the encoder needs; letting the prior adapt tightens the rate bound. We restrict it to a zero-mean diagonal Gaussian, keeping the rate in closed form and identical in structure across codecs; richer priors (19; 18) would tighten the rate further at the cost of that uniformity, which is what makes the atoms mutually comparable.

Proposition 1 (Quantization-invariant information bound)**.** *Let $T_i \sim q_i(\cdot \mid X)$ and let $\hat{T}_i = Q(T_i)$ be any quantization of T_i obtained without further access to X. Then*

$$I(X; \hat{T}_i) \;\leq\; I_{q_i}(X; T_i) \;\leq\; \bar{R}_i. \tag{4}$$

If $\hat{T}_i$ is further losslessly entropy-coded into C_i, then for any prefix code $I(X;\hat{T}_i) \leq H(\hat{T}_i) \leq \mathbb{E}[\ell(C_i)]$, so $\bar{R}_i$ upper-bounds the information any coded representation can carry about X, though $\mathbb{E}[\ell(C_i)]$ itself depends additionally on the quantizer, the entropy model, and coding redundancy.

Proof. The chain $X \to T_i \to \hat{T}_i$ is Markov, so data processing gives $I(X;\hat{T}_i) \leq I_{q_i}(X;T_i)$; the second inequality is equation 3. For the coded length, $I(X;\hat{T}_i) \leq H(\hat{T}_i) \leq \mathbb{E}[\ell(C_i)]$ by the source coding theorem. □

Remark 1 ($\bar{R}_i$ is a variational information-rate upper bound)**.** *$\bar{R}_i$ is quantization-invariant but not a physical bit count: coded length depends additionally on the quantizer and entropy model. Our experiments perform no quantization or entropy coding, so $\bar{R}_i$ is used throughout as a variational information-rate coordinate for comparing frozen codecs, never as a measured bitrate.*

3 Rate–Distortion Frontier: Two Operating Regimes of a VIB Codec

3.1 Codec Family and the Operating-Point Frontier

Each codec i is trained from a fresh initialization on the rate–distortion Lagrangian

$$\mathcal{L}_i(\beta_i) = D_i + \beta_i \bar{R}_i, \qquad \beta_i > 0, \tag{5}$$

a β-VAE objective (17) with a pixel-space squared-error likelihood and a learned prior. Larger β_i places greater weight on the variational rate and favors lower-rate, higher-distortion operating points; smaller β_i favors lower distortion at the cost of a larger rate. Each value of β_i defines a *separate* codec — encoder, decoder, and prior trained jointly and frozen after training — so β_i enters the paper only as the training knob; the deployed operating point is the measured pair

$$P_i = (D_i, \bar{R}_i), \tag{6}$$

obtained by Monte Carlo on held-out data. Two codecs trained with different β are treated as distinct candidate atoms only through their achieved $(D_i, \bar{R}_i)$ and their frozen encoder–decoder–prior triple. Sweeping β over a grid traces the variable-rate family from which the finite candidate frontier is constructed in Section 4.

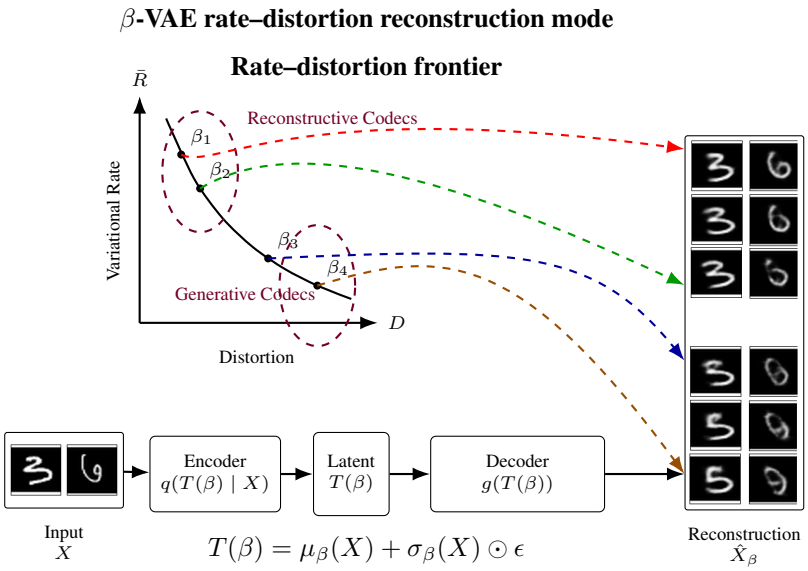


Figure 2: **Rate–distortion effect on the reconstruction and generation regimes.** Each codec minimizes the rate–distortion Lagrangian eq. (5); changing β produces different encoder–decoder pairs and therefore different operating points P_i on the empirical distortion–rate frontier. In **reconstruction mode** equation 7, the latent is drawn from $q_i(T \mid X)$: a larger rate penalty reduces the variational rate and generally yields a less faithful reconstruction, as input-specific information is removed from the latent. The image examples illustrate this joint reconstruction trade-off for MNIST.

3.2 Two Operating Regimes of the VIB Codec Family

Although every codec is trained with the single objective equation 5, its behavior *at inference* depends qualitatively on two things: the operating point P_i, and whether the decoder is driven by an input-dependent latent draw or an input-independent one. This gives two distinct inference-time modes built from the same frozen codec.

Reconstruction mode. The latent is drawn from the encoder conditioned on an input X, $T_i = \mu_i(X) + \sigma_i(X) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I_d)$

$$X \;\to\; T_i \sim q_i(T \mid X) \;\to\; \hat{X}_i = g_i(T_i). \tag{7}$$

Input-free generation mode. The latent is instead drawn from the codec's own learned prior, with no input supplied, $T_i = \tau_i \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I_d)$

$$T_i \sim p_i(T) \;\to\; \hat{X}_i = g_i(T_i). \tag{8}$$

Both modes reuse the same frozen decoder g_i; equation 7 is the deployment rule used throughout the rest of this paper, and equation 8 simply removes the conditioning path $X \to T_i$ while leaving g_i untouched — possible precisely because $\hat{X}_i$ depends on X only through T_i. Figures 2 and 3 illustrate the two modes across the same β-sweep.

In (17), they show that in the β-VAE objective $\mathcal{L} = D + \beta\bar{R}$, increasing β tightens the constraint on the information the latent may carry and generally trades reconstruction accuracy for a representation that more closely matches the prior. We observe an analogous but distinct trade-off along the same axis, now assessed *across the two modes* rather than within reconstruction alone.

Why the two modes diverge across the frontier. The mechanism is compression. At a low-rate operating point (large β_i), the encoder is forced to discard most input-specific information, so $q_i(T \mid X)$ is close to $p_i(T)$ for typical X, and the aggregate posterior $q_i(t) = \mathbb{E}_X q_i(t \mid X)$ is well matched by the learned prior. Consequently:

- In **reconstruction mode** (Figure 2), so little information about X survives in T_i that $g_i(T_i)$ cannot recover it: the reconstruction is essentially uninformative given X removed.

- In **generation mode** (Figure 3), the same compression is exactly what makes sampling $T_i \sim p_i(T)$ useful: because p_i has been forced to absorb the common structure of $q_i(T \mid X)$ across the data distribution, decoding a prior sample yields a coherent, recognizable image — new, but drawn from the same learned structure, not tied to any particular X.

At a high-rate operating point (small β_i), the situation reverses. Reconstruction mode is faithful, since T_i retains enough of X for g_i to recover it almost exactly. But generation mode degrades: the aggregate posterior $q_i(t)$ has not been compressed toward $p_i(t)$, so a prior sample $T_i \sim p_i(T)$ lands in a region of latent space the decoder was never trained to interpret without input-specific information, and $\hat{X}_i$ is no longer a clean sample close to the data — decoding is still constrained by g_i, so the result is a distorted, input-adjacent image rather than the discarded-information failure of low-rate reconstruction.

In short, the same rate parameter $\bar{R}_i$ that governs the reconstruction–distortion trade-off of Section 3.1 also governs a second, orthogonal trade-off between the two modes: low rate sacrifices reconstruction fidelity but is exactly what enables coherent input-free generation, while high rate preserves reconstruction fidelity at the cost of generation quality. This reframes each atom P_i on the frontier not only as a reconstruction operating point, but as a joint reconstruction/generation operating point, and is the observation we build on next.

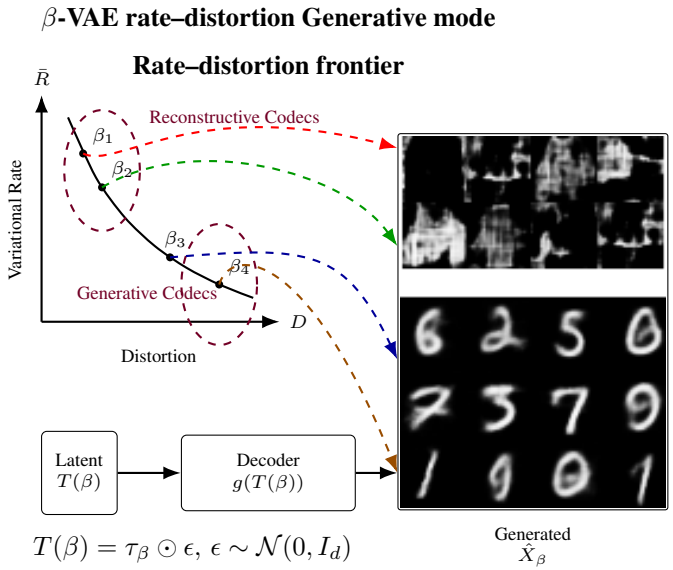


Figure 3: **Rate–distortion effect on the generative and generation regimes.** Each codec minimizes the rate–distortion Lagrangian eq. (5); changing β produces different encoder–decoder pairs and therefore different operating points P_i on the empirical distortion–rate frontier. In **generation mode** equation 8, the latent is instead drawn from the learned prior $p_i(T)$ with no input supplied: the same rate reduction that degrades reconstruction is what makes prior samples decode into coherent, recognizable images, since the aggregate posterior has been compressed toward $p_i(T)$. The image examples illustrate this joint generation trade-off for FashionMNIST in up, reconstructive codecs, and MNIST in down, generative codecs.

4 Frontier and the Selection Problem

Training M_0 codecs independently, one per β_i, and measuring each on held-out data gives a point $P_i = (D_i, \bar{R}_i)$ per codec. The encoder and decoder are frozen before selection ever begins, so representation learning and selection stay fully separate. After discarding collapsed and Pareto-dominated candidates, what remains is the *candidate frontier* $\mathcal{F} = (P_0, \ldots, P_{M-1})$, sorted and strictly monotone in D and $\bar{R}$. As established in Section 3.2, every P_i is simultaneously a reconstruction and a generation operating point; selection treats it purely as a coordinate.

A *request* $Q = (D_Q, R_Q)$ is a desired operating point, and we measure distance to the frontier with a normalized Euclidean metric d_W, so the usual triangle inequality is available whenever we need it. A dictionary $\mathcal{D}_k \subseteq \mathcal{F}^\circ$ keeps k *interior* atoms; the two frontier endpoints are never deployed, but they bookend the ordered selection path $0 = i_0 < i_1 < \cdots < i_k < i_{k+1} = M-1$ and serve as fixed guards for it. Every request is routed to its single nearest deployed atom, and this one rule — no exceptions, no baselines exempted — is what every loss, certificate, and comparison in this paper is computed against: $L_{\text{true}}(\mathcal{D}_k; \{Q_n\}) = \sum_{n=1}^{N} w_n \min_{P \in \mathcal{D}_k} d_W^2(Q_n, P)$. We also require a minimum and maximum spacing $\delta_{\min}, \delta_{\max}$ between consecutive path atoms, ruling out near-duplicates and bounding how far a request's true nearest atom can drift once endpoints are excluded. Feasible dictionaries form the class $\mathcal{H}_k$, and the problem we solve is simply $\mathcal{D}_k^\star = \arg\min_{\mathcal{D}_k \in \mathcal{H}_k} L_{\text{true}}(\mathcal{D}_k; \mathcal{Q}_{\text{tr}})$. See Appendix A for the full definitions of the matching metric, the feasible class $\mathcal{H}_k$, and the augmented DP path.

5 SURROGATE DYNAMIC PROGRAMMING

The classical 1-D k-means DP (7; 8) needs a cluster's cost to depend only on the contiguous block it covers. Our frontier does not cooperate: because it bends, a request can end up closer to an interior atom than to either atom flanking it in D-order, so no partition of the request axis reproduces true nearest-atom assignment.

We fix this by deciding, once and for all, which frontier point each request *would* use if the whole frontier were deployable. Call this its *anchor* $a(n)$, with residual $r_n = d_W(Q_n, P_{a(n)})$ (Equation (14)). The anchor never looks at k or at which subset survives, and that independence is exactly what makes the resulting objective additive over frontier segments. For a path $0 = i_0 < \cdots < i_{k+1} = M-1$, each anchor falls in exactly one segment $\mathcal{I}(i_t, i_{t+1})$ (Equation (15)), and is rerouted to whichever deployed endpoint of that segment is closer — the two frontier endpoints themselves are never deployed, only used to bound the path. This gives the *anchor-routed surrogate*

$$L_{\text{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\text{tr}}) = \sum_{n=1}^{N} w_n \, d_W^2\big(Q_n, P_{r(n)}\big), \tag{9}$$

with $r(n)$ the routed atom and $h_n := d_W(P_{a(n)}, P_{r(n)})$ the *anchor displacement* it introduces. The surrogate is not an approximation of the frontier — it is an approximation of who serves whom.

Theorem 1 (Surrogate dominates the true loss)**.** *For every dictionary $\mathcal{D}_k$, $L_{\text{true}}(\mathcal{D}_k; \mathcal{Q}_{\text{tr}}) \le L_{\text{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\text{tr}})$, since the routed atom in equation 9 is always a candidate for the true nearest-atom search but need not win it.*

The exact solver. Because equation 9 is additive over segments, it is solved by an ordinary shortest-path recurrence over feasible transitions (i, j) with $\delta_{\min} \le \Delta_{ij} \le \delta_{\max}$:

$$S(m, \ell) = \min_{0 \le m' < m} \big[S(m', \ell - 1) + C[m', m]\big], \qquad S(0, 0) = 0, \tag{10}$$

with $C[i, j]$ the cost of routing every request anchored between i and j to whichever endpoint is closer. The dictionary $\mathcal{D}_k^{\text{DP}}$ read out by backtracking $S(M-1, k+1)$ in equation 10 is exactly optimal for L_{anchor} over the feasible class $\mathcal{H}_k$:

Theorem 2 (Exact optimization of the surrogate)**.** $S(M-1, k+1) = \min_{\mathcal{D}_k \in \mathcal{H}_k} L_{\text{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\text{tr}})$*, computed in $O\big(M^2(N+k)\big)$ time — against $O\big(\binom{M-2}{k} Nk\big)$ for exhaustive search over the same class.*

What the anchor buys us is not just a fast solver: it gives an *a priori* bound on how far $\mathcal{D}_k^{\text{DP}}$ can be from the true optimal dictionary, computable before the recurrence in equation 10 is ever run. Full proofs, the segment-cost formula, and pseudocode are in Appendix B.

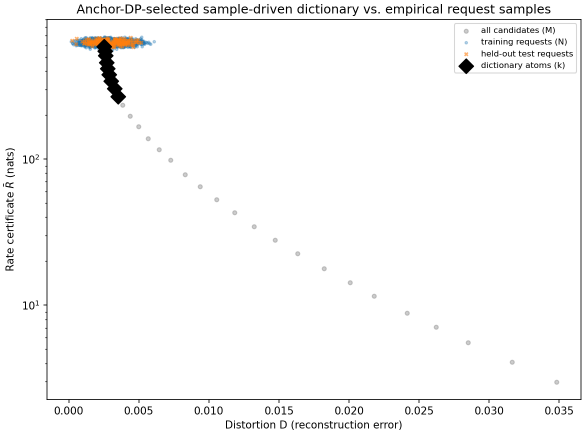
6 EXPERIMENTS

6.1 SETUP

Codecs and frontier. M_0 β-VAE reconstruction codecs with learned diagonal-Gaussian priors per dataset, one per β on a log-spaced grid, trained from fresh initialization with a convolutional encoder, deconvolutional decoder, and learned prior. The upper end of the β grid is set below the onset of posterior collapse. **Datasets:** CIFAR-10 and FashionMNIST (zero-padded to 32×32). **Multiple seeds:** the pilot sweep is repeated over independent codec-family seeds; we report variability of D_i and $\bar{R}_i$ across seeds and the Monte-Carlo standard error within a seed, converted into the ε of Proposition 7. A frontier-quality report is generated per run (ranges, decades of $\bar{R}$, monotonicity violations, per-segment deviation from collinearity, Λ_{front} and $\Delta_{0,M-1}$).

Neural Network of Each Encoder and Decoder. Each candidate codec is a convolutional VIB with a diagonal-Gaussian encoder, a deconvolutional decoder with a final sigmoid, and a learned zero-mean diagonal-Gaussian prior. The encoder applies three stride-2 convolutions (padding chosen so spatial resolution halves cleanly at each layer) followed by two linear heads producing $\mu(x)$ and $\log \sigma^2(x)$; the decoder mirrors this with a linear projection followed by three stride-2 transposed convolutions. Architecture and latent dimensionality are held fixed across the β-sweep, so every candidate on the frontier differs only in its training rate penalty β_i, not in network capacity.

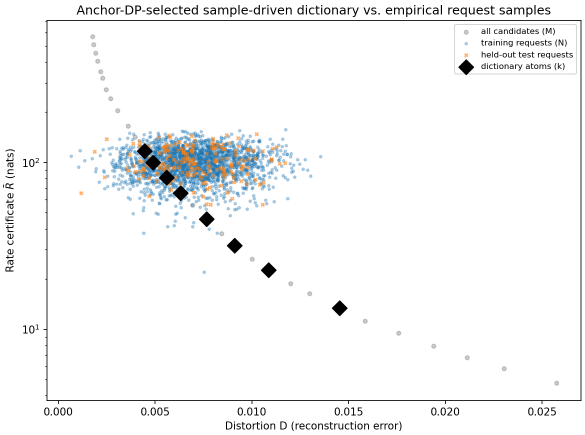
We compare our method against the following baseline selection rules; the resulting numbers are reported and discussed in Section 6.



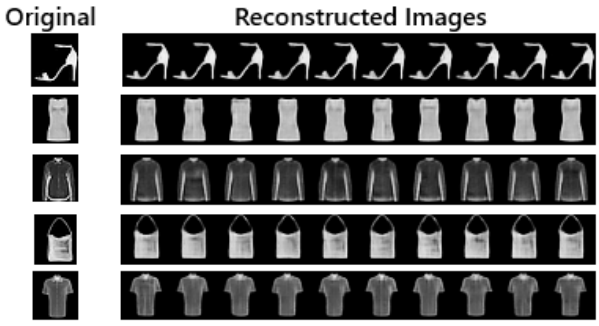
(a) Dictionaries selected by the proposed DP method for CIFAR-10



(b) Original vs. Reconstructed CIFAR-10 Images for fig. 4a

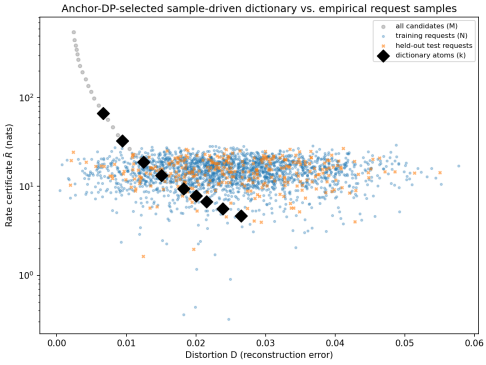


(c) Dictionaries selected by the proposed DP method for fashionmnist



(d) Original vs. Reconstructed FashionMNIST Images, fig. 4c

Figure 4: Reconstruction quality across the Cifar-10 and Fashion-MNIST benchmark datasets.

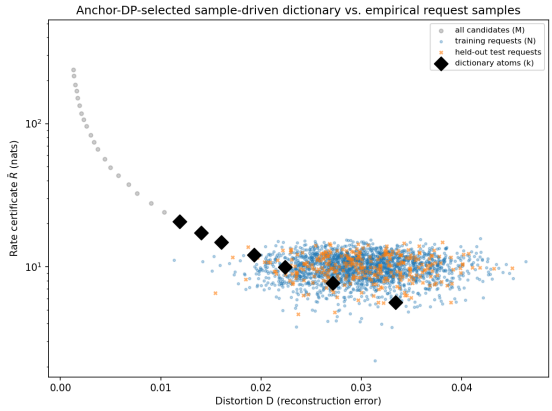


(a) Dictionaries selected by the proposed DP method for FashionMNIST



(b) 42 Generated FashionMNIST Images from Dictionary in fig. 5a

Figure 5: Generative Images for FashionMNIST Dataset.



(a) Dictionaries selected by the proposed DP method for MNIST



(b) 42 Generated MNIST Images from Dictionary in fig. 6a

Figure 6: Generative Images for MNIST Dataset.

Table 1: Mean point-distance error on held-out test requests versus dictionary size k. Random-k is reported as the mean over 30 random draws (standard deviations are omitted for space).

Dataset	Method	k=2	k=4	k=6	k=8	k=10	k=15
MNIST (single Gaussian)	Random-k	**3.13409**	1.30781	0.94961	0.46919	0.33770	0.12107
	Uniform-rank	**3.13409**	1.30799	0.53800	0.20017	0.10051	0.05101
	Uniform-arc-length	**3.13409**	0.38359	0.13363	0.06789	0.06117	0.02548
	Farthest-point	**3.13409**	1.59779	0.10276	0.10276	0.05676	0.02548
	Greedy-nearest-atom	**3.13409**	0.10697	**0.03683**	0.02580	0.02129	**0.02050**
	Proposed DP	**3.13409**	**0.09723**	**0.03683**	**0.02482**	**0.02093**	**0.02050**
MNIST (3-mixture)	Random-k	**5.42879**	0.70819	0.50832	0.32590	0.24450	0.15428
	Uniform-rank	**5.42879**	0.57995	0.40314	0.25096	0.15744	0.10685
	Uniform-arc-length	**5.42879**	0.50701	0.14227	0.12673	0.13997	0.09433
	Farthest-point	**5.42879**	0.29627	0.27033	0.17511	0.14496	0.09534
	Greedy-nearest-atom	**5.42879**	**0.24552**	0.12285	0.10380	0.09463	**0.08933**
	Proposed DP	**5.42879**	0.24805	**0.11861**	**0.10098**	**0.09409**	**0.08933**
CIFAR10	Random-k	**2.73769**	1.14734	1.00541	0.94658	0.91005	0.88993
	Uniform-rank	**2.73769**	1.02163	0.91726	0.90259	0.89341	0.88222
	Uniform-arc-length	**2.73769**	1.10792	0.94672	0.91140	0.90147	0.88790
	Farthest-point	**2.73769**	1.36524	0.94972	0.91354	0.90542	0.88794
	Greedy-nearest-atom	**2.73769**	0.98409	0.90013	0.88880	0.88320	**0.87752**
	Proposed DP	**2.73769**	**0.94672**	**0.89980**	**0.88554**	**0.88160**	**0.87752**
Fashion-MNIST	Random-k	**7.13439**	0.45527	0.17628	0.18214	0.13376	0.09802
	Uniform-rank	**7.13439**	0.19779	0.10815	0.10670	0.09529	0.08987
	Uniform-arc-length	**7.13439**	0.66924	0.22230	0.15705	0.11180	0.09817
	Farthest-point	**7.13439**	0.15534	0.12309	0.12229	0.11613	0.09600
	Greedy-nearest-atom	**7.13439**	0.12041	0.09194	0.08769	0.08580	**0.08482**
	Proposed DP	**7.13439**	**0.10612**	**0.09058**	**0.08711**	**0.08640**	**0.08482**

Interior random-k draws k interior atoms uniformly at random, providing an uninformed reference against which any structured rule should improve (Definition 6 for the full sampling and averaging protocol).

Uniform-rank spaces atoms evenly by frontier *index* rather than by geometric position (Definition 7 for the rounding convention and edge cases).

Uniform-arc-length spaces atoms evenly along the frontier's cumulative arc length, correcting for non-uniform candidate spacing (Definition 8 for the exact construction).

Farthest-point sampling greedily maximizes geometric coverage of the $(D, \bar{R})$ plane rather than targeting the deployed loss directly (Definition 9 for the greedy update and complexity).

Greedy-nearest-atom greedily minimizes the true deployed loss L_{true} itself at each step, making it the sharpest baseline for isolating surrogate slack from ordinary search suboptimality (Definition 10 for the update rule and cost analysis).

7 Conclusion

On a bent frontier, nearest-atom assignment is not ordered, so the classical 1-D clustering DP cannot select a dictionary directly. Freezing ownership on the *full* frontier before choosing a subset restores that order and makes selection exactly solvable; the frontier and the deployment rule stay untouched, and only the assignment is approximated, and that approximation is certified rather than assumed. Its per-request gap shrinks with the full-frontier residual, vanishes on an affine frontier (Theorem 6), and composes into $\mathrm{Cert}(\delta_{\max}) = 2\bar{r}\delta_{\max} + \delta_{\max}^2$: an a priori bound, a two-sided bracket on the true optimum, and a margin guarantee against any competitor (Corollary 3), all available before the optimizer runs, exactly what a direct integer-program solve for L_{true} cannot give until after it does. The same guarantee extends to the population objective and degrades gracefully under request-distribution shift (Theorem 9). Against a single codec conditioned on the rate parameter, k frozen specialists are a different way to spend the same budget, not a strictly better one; we let the comparison decide, not the claim.

AI use statement

In this work, we used generative AI tools to assist with editing and polishing the writing of some sections. We have not used generative AI tools to generate experimental results, run experiments, or produce the core theoretical derivations in this work. We take responsibility for the final content of this work, including all text, claims, and materials produced with the aid of generative AI.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint arXiv:physics/0004057*, 2000.

[2] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proc. Conf. Learn. Theory (COLT)*, 2003.

[3] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2014, pp. 501–505.

[4] A. E. Parker, T. Gedeon, and A. G. Dimitrov, "On the lack of convexity of the relevance-compression function," *arXiv preprint arXiv:2204.10957*, 2022.

[5] H. Charvin, N. Catenacci Volpi, and D. Polani, "Exact and soft successive refinement of the information bottleneck," *Entropy*, vol. 25, no. 9, p. 1355, 2023.

[6] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[7] R. Bellman, "A note on cluster analysis and dynamic programming," *Math. Biosci.*, vol. 18, no. 3–4, pp. 311–312, 1973.

[8] H. Wang and M. Song, "Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming," *The R J.*, vol. 3, no. 2, pp. 29–33, 2011.

[9] S. L. Hakimi, "Optimum locations of switching centers and the absolute centers and medians of a graph," *Oper. Res.*, vol. 12, no. 3, pp. 450–459, 1964.

[10] O. Kariv and S. L. Hakimi, "An algorithmic approach to network location problems. II: The p-medians," *SIAM J. Appl. Math.*, vol. 37, no. 3, pp. 539–560, 1979.

[11] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.

[12] M. S. Daskin, *Network and Discrete Location: Models, Algorithms, and Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2013.

[13] A. Maurer and M. Pontil, "Empirical Bernstein bounds and sample variance penalization," in *Proc. Conf. Learn. Theory (COLT)*, 2009.

[14] C. Villani, *Optimal Transport: Old and New*. Berlin, Germany: Springer, 2009.

[15] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken ELBO," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018.

[16] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[17] I. Higgins *et al.*, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[18] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[19] J. M. Tomczak and M. Welling, "VAE with a VampPrior," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2018.

[20] M. D. Hoffman and M. J. Johnson, "ELBO surgery: Yet another way to carve up the variational evidence lower bound," in *NIPS Workshop Adv. Approx. Bayesian Inference*, 2016.

[21] J. Shao and J. Zhang, "BottleNet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2020.

[22] Y. Matsubara, R. Yang, M. Levorato, and S. Mandt, "Supervised compression for resource-constrained edge computing systems," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022.

[23] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[24] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.

[25] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10532–10541.

[26] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021.

[27] Y. Wang, D. Liu, S. Ma, F. Wu, and W. Gao, "Ensemble learning-based rate-distortion optimization for end-to-end image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1193–1207, 2021.

[28] X. Hang, Z. Ge, H. Fan, C. Jia, S. Ma, and W. Gao, "Rate-distortion-complexity optimized framework for multi-model image compression," *IEEE Trans. Image Process.*, vol. 34, pp. 5385–5399, 2025.

A Operating-Point Frontier and the Selection Problem

Each point on the frontier constructed below is, as established in Section 3.2, simultaneously a reconstruction and a generation operating point; the selection problem that follows treats it purely as a coordinate $P_i = (D_i, \bar{R}_i)$, agnostic to which mode ultimately serves the request.

We train M_0 codecs independently, one per β_i, and measure by Monte Carlo on a held-out validation split $P_i = (D_i, \bar{R}_i)$. The parameter β_i is a training knob only. **The Encoder and Decoder networks are frozen before selection begins**, separating representation learning from selection. Candidates with $\bar{R}_i$ below a floor R_{collapse} (posterior collapse) are discarded, and then Pareto-dominated candidates are removed.

The *finite candidate Pareto frontier* is the sequence $\mathcal{F} = (P_0, \ldots, P_{M-1})$, $M \leq M_0$, of non-collapsed, non-dominated candidates sorted by increasing D. Since no retained point dominates another, the frontier is strictly monotone, $D_0 < D_1 < \cdots < D_{M-1}$, $\bar{R}_0 > \bar{R}_1 > \cdots > \bar{R}_{M-1}$.

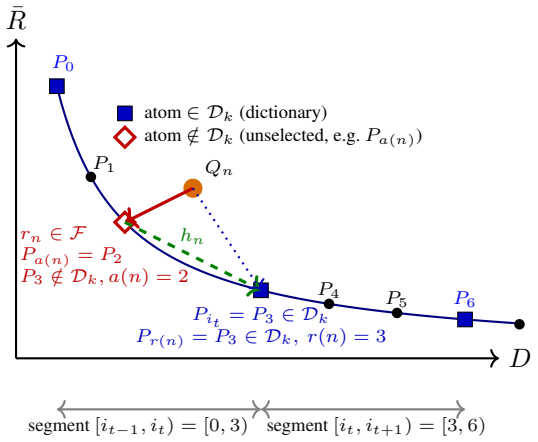


Figure 7: Illustration of the anchor-routed construction, distinguishing atoms *inside* the dictionary from those *outside* it. The request Q_n lies close to $P_{a(n)} = P_2$, the nearest point over the *complete* frontier $\mathcal{F}$, yet P_2 is **not** a member of the selected dictionary $\mathcal{D}_k = \{P_0, P_3, \dots\}$ (hollow red diamond), and is far from both dictionary atoms. Because the dictionary cannot serve P_2, the request is instead routed to the nearer *selected* endpoint of its segment, $P_{r(n)} = P_3 \in \mathcal{D}_k$ (filled blue square) — a materially worse match. The red solid line is the full-frontier residual $r_n = d_W(Q_n, P_{a(n)})$; the green dashed line is the anchor displacement $h_n = d_W(P_{a(n)}, P_{r(n)})$; the blue dotted line is the anchor-routed distance $d_W(Q_n, P_{r(n)})$ actually paid at deployment.

Requests and the Matching Metric. A *request* is a target $Q = (D_Q, R_Q)$ drawn i.i.d. from a distribution $\mathcal{Q}$ on a compact set. The squared distance to an atom is $d_W^2(Q, P) = \lambda_D \left(\frac{D_P - D_Q}{s_D}\right)^2 + \lambda_R \left(\frac{R_P - R_Q}{s_R}\right)^2$, with s_D, s_R the standard deviations over $\mathcal{F}$ and $\lambda_D, \lambda_R > 0$ fixed metric weights (distinct from the request weights w_n). With $z(P) := \left(\sqrt{\lambda_D}\frac{D_P}{s_D}, \sqrt{\lambda_R}\frac{R_P}{s_R}\right)$, $d_W(Q, P) = \|z(Q) - z(P)\|_2$, d_W is a genuine Euclidean metric, so the triangle inequality holds; the geometric theory below depends on this. The normalized gap between atoms is $\Delta_{ij} := d_W(P_i, P_j)$. A request expresses a desired operating point; we serve it by routing to the nearest atom under d_W, treating Q as a target rather than a pair of upper bounds.

Dictionary Selection Problem. Let $\mathcal{F}^\circ = \{P_1, \dots, P_{M-2}\}$ denote the interior candidates that may actually be deployed. A deployed dictionary contains exactly k interior frontier atoms, $\mathcal{D}_k = \{P_{i_1}, \dots, P_{i_k}\} \subseteq \mathcal{F}^\circ$, while the fixed frontier endpoints P_0 and P_{M-1} serve only as boundary guards for the ordered selection path. The augmented DP path is $0 = i_0 < i_1 < \cdots < i_k < i_{k+1} = M - 1$, containing $k + 2$ path nodes, k deployed atoms, and $k + 1$ consecutive segments.

Every request is served by exactly one deployed atom, $m^\star(Q) = \arg\min_{P \in \mathcal{D}_k} d_W^2(Q, P)$, $e(Q, \mathcal{D}_k) = \min_{P \in \mathcal{D}_k} d_W^2(Q, P)$, by exhaustive search over the k deployed atoms. For a weighted sample $\{(Q_n, w_n)\}$ with $w_n \geq 0$, $\sum_n w_n = 1$,

$$L_{\text{true}}(\mathcal{D}_k; \{Q_n\}) = \sum_{n=1}^{N} w_n \, e(Q_n, \mathcal{D}_k). \tag{11}$$

Every request is served by exactly one frozen codec; this single rule is used for selection targets, held-out evaluation, every baseline, and physical deployment; no other rule produces any reported number in this paper. We impose the same structural constraints on the proposed method and every baseline:

$$\mathcal{H}_k = \left\{ \mathcal{D}_k = \{P_{i_1}, \dots, P_{i_k}\} \subseteq \mathcal{F}^\circ : \begin{array}{c} 0 = i_0 < i_1 < \cdots < i_k < i_{k+1} = M - 1, \\ \delta_{\min} \leq \Delta_{i_t i_{t+1}} \leq \delta_{\max}, \quad t = 0, \dots, k \end{array} \right\}. \tag{12}$$

The two endpoints are present in every augmented DP path but are not members of $\mathcal{D}_k$ and are not available to the deployment rule. The spacing constraint prevents near-duplicate deployed atoms and, crucially, bounds the anchor displacement.

Problem 1 (Request-aware constrained dictionary selection)**.** *Given the candidate frontier $\mathcal{F}$, training requests $\mathcal{Q}_{\text{tr}}$, target dictionary size $k \in \{1, \dots, M - 2\}$, and spacing bounds $\delta_{\min}, \delta_{\max}$, find*

$$\mathcal{D}_k^\star = \arg\min_{\mathcal{D}_k \in \mathcal{H}_k} L_{\text{true}}(\mathcal{D}_k; \mathcal{Q}_{\text{tr}}), \tag{13}$$

where $\mathcal{H}_k$ is the feasible class of Equation (12)*.*

B Anchor-Routed Surrogate Dynamic Programming

The classical DP for one-dimensional k-means (7; 8) requires the cost of a cluster to depend only on the contiguous block it covers. Here it does not: because the frontier bends (is non-collinear), a request near an endpoint can be closer to an interior atom, while a request near the midpoint in D can end up farther from that atom than from an endpoint. This bend breaks the correspondence between order in D and order in nearest-atom ownership, so no partition of the request axis into contiguous blocks reproduces true nearest-atom assignment.

Definition 1 (Anchor and residual)**.** *The* natural anchor *of request Q_n on the full sorted frontier is the smallest-index minimizer*

$$a(n) = \min\Big\{ \operatorname*{arg\,min}_{m\in\{0,\dots,M-1\}} d_W^2(Q_n, P_m)\Big\}, \tag{14}$$

and the anchor residual *is*

$$r_n = d_W(Q_n, P_{a(n)}) = \min_m d_W(Q_n, P_m).$$

Both depend only on the data, never on any dictionary.

The anchor is computed once, before any subset is chosen. This independence is what makes the surrogate additively separable, and it is also the source of its slack. A path

$$0 = i_0 < \cdots < i_{k+1} = M-1$$

splits the frontier into segments $[i_t, i_{t+1}]$.

Definition 2 (Anchor interval and segment cost)**.** *For $0 \le i < j \le M-1$, define*

$$\mathcal{I}(i,j) = \begin{cases} \{i,\dots,j-1\}, & j < M-1, \\ \{i,\dots,M-1\}, & j = M-1. \end{cases} \tag{15}$$

$$C[i,j] = \begin{cases} \displaystyle\sum_{n:\, a(n)\in\mathcal{I}(0,j)} w_n\, d_W^2(Q_n, P_j), & i = 0,\ j < M-1, \\ \displaystyle\sum_{n:\, a(n)\in\mathcal{I}(i,j)} w_n \min\{d_W^2(Q_n, P_i), d_W^2(Q_n, P_j)\}, & 0 < i < j < M-1, \\ \displaystyle\sum_{n:\, a(n)\in\mathcal{I}(i,M-1)} w_n\, d_W^2(Q_n, P_i), & 0 < i,\ j = M-1. \end{cases} \tag{16}$$

Thus the first and last frontier points are boundary guards rather than deployable atoms: requests anchored in the first segment are routed to its right endpoint, requests anchored in an interior segment are routed to the nearer of its two deployed endpoints, and requests anchored in the final segment are routed to its left endpoint. For any augmented path $0 = i_0 < i_1 < \cdots < i_k < i_{k+1} = M-1$, every $m \in \{0,\dots,M-1\}$ belongs to exactly one of the $k+1$ anchor intervals $\mathcal{I}(i_t, i_{t+1})$, $t = 0,\dots,k$.

Definition 3 (Anchor-routed surrogate)**.** *For a deployed dictionary $\mathcal{D}_k = \{P_{i_1},\dots,P_{i_k}\}$ with augmented path $0 = i_0 < i_1 < \cdots < i_k < i_{k+1} = M-1$, define*

$$L_{\mathrm{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\mathrm{tr}}) = \sum_{t=0}^{k} C[i_t, i_{t+1}] = \sum_{n=1}^{N} w_n\, d_W^2\big(Q_n, P_{r(n)}\big), \tag{17}$$

where $r(n)$ is the deployable atom assigned by the segment containing $a(n)$:

$$r(n) = \begin{cases} i_1, & a(n) \in \mathcal{I}(i_0, i_1), \\ \displaystyle\operatorname*{arg\,min}_{j\in\{i_t, i_{t+1}\}} d_W^2(Q_n, P_j), & a(n) \in \mathcal{I}(i_t, i_{t+1}),\ 1 \le t \le k-1, \\ i_k, & a(n) \in \mathcal{I}(i_k, i_{k+1}), \end{cases} \tag{18}$$

with lower-index tie breaking in the middle case. The anchor displacement *is $h_n := d_W(P_{a(n)}, P_{r(n)})$. The two frontier endpoints P_0 and P_{M-1} are never elements of $\mathcal{D}_k$ and are used only as boundary guards. In fact, the surrogate does not approximate the frontier; it approximates the assignment rule.*

Theorem 3 (Surrogate dominance)**.** *For every $\mathcal{D}_k$ realized as a path and every request sample, $L_{\mathrm{true}}(\mathcal{D}_k; \mathcal{Q}_{\mathrm{tr}}) \le L_{\mathrm{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\mathrm{tr}})$. The statement holds verbatim with d_W^2 replaced by any per-pair cost $c(Q, P) \ge 0$.*

Proof. Fix n. By Definition 3, the anchor-routed atom $P_{r(n)}$ belongs to the deployed dictionary $\mathcal{D}_k$. Therefore

$$e(Q_n, \mathcal{D}_k) = \min_{P \in \mathcal{D}_k} d_W^2(Q_n, P) \le d_W^2(Q_n, P_{r(n)}).$$

Multiplying by $w_n \ge 0$ and summing over n gives $L_{\mathrm{true}}(\mathcal{D}_k; \mathcal{Q}_{\mathrm{tr}}) \le L_{\mathrm{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\mathrm{tr}})$. The same argument applies to any nonnegative per-pair cost $c(Q, P)$. □

This pointwise inequality holds for every realization of Q_n, hence, by taking expectations under $\mathcal{Q}$, also between the population quantities $L_{\mathrm{anchor}}^{\mathrm{pop}}$ and $L_{\mathrm{true}}^{\mathrm{pop}}$ (Section C.7).

Lemma 1 (Anchor displacement is controlled by spacing)**.** *Let d_W and the spacing gap Δ_{ij} be computed in the* same *coordinates. For any path in $\mathcal{H}_k$ and any request n with $a(n) \in \mathcal{I}(i_t, i_{t+1})$,*

$$h_n \le \Delta_{i_t i_{t+1}} \le \delta_{\max}. \tag{19}$$

Proof. There are three cases.

If $a(n) \in \mathcal{I}(i_0, i_1)$, then $r(n) = i_1$. Both $a(n)$ and i_1 have indices in $[0, i_1]$. By $D_0 < D_1 < \cdots < D_{M-1}$ and $\bar{R}_0 > \bar{R}_1 > \cdots > \bar{R}_{M-1}$, D increases and $\bar{R}$ decreases along that range, so for any indices $p, q \in [0, i_1]$, $|D_p - D_q| \le D_{i_1} - D_0$ and $|\bar{R}_p - \bar{R}_q| \le \bar{R}_0 - \bar{R}_{i_1}$. Since $d_W(P_p, P_q)$ is nondecreasing in $|\Delta D|$ and $|\Delta \bar{R}|$,

$$h_n = d_W(P_{a(n)}, P_{i_1}) \le d_W(P_0, P_{i_1}) = \Delta_{i_0 i_1}.$$

If $a(n) \in \mathcal{I}(i_t, i_{t+1})$ for $1 \le t \le k-1$, then $r(n) \in \{i_t, i_{t+1}\}$ and both $a(n)$ and $r(n)$ have indices in $[i_t, i_{t+1}]$. The same monotonicity argument gives $h_n \le \Delta_{i_t i_{t+1}}$.

If $a(n) \in \mathcal{I}(i_k, i_{k+1})$, then $r(n) = i_k$, and both $a(n)$ and i_k have indices in $[i_k, M-1]$, so

$$h_n = d_W(P_{a(n)}, P_{i_k}) \le d_W(P_{i_k}, P_{M-1}) = \Delta_{i_k i_{k+1}}.$$

In every case the spacing constraint gives $\Delta_{i_t i_{t+1}} \le \delta_{\max}$, hence $h_n \le \delta_{\max}$. □

The Exact Dynamic Program. Let $\mathcal{T} = \{(i, j) : i < j,\ \delta_{\min} \le \Delta_{ij} \le \delta_{\max}\}$ be the feasible transitions and set $C[i, j] = \infty$ off $\mathcal{T}$. Define $S(m, \ell)$ as the minimum anchor-routed cost of covering $0, \ldots, m$ with ℓ segments ending at m, with $S(0, 0) = 0$, $S(0, \ell) = \infty$ $(\ell > 0)$, $S(m, 0) = \infty$ $(m > 0)$:

$$S(m, \ell) = \min_{0 \le m' < m} \big[S(m', \ell - 1) + C[m', m]\big]. \tag{20}$$

The answer is $S(M-1, k+1)$, recovered by backtracking.

Theorem 4 (Exact optimization of the anchor surrogate)**.** *For any $k \in \{1, \ldots, M-2\}$, the recurrence equation 20 satisfies $S(M-1, k+1) = \min_{\mathcal{D}_k \in \mathcal{H}_k} L_{\mathrm{anchor}}(\mathcal{D}_k; \mathcal{Q}_{\mathrm{tr}})$, so $\mathcal{D}_k^{\mathrm{DP}}$ minimizes L_{anchor} over $\mathcal{H}_k$.*

Proof. By induction on ℓ, $S(m, \ell)$ equals the minimum anchor-routed cost of any feasible $(\ell+1)$-atom path from 0 to m. *Base.* A one-atom path has no segments and cost 0; the initialization gives $S(0, 0) = 0$ and $S(m, 0) = \infty$ for $m > 0$. *Step.* Take any feasible $(\ell+1)$-atom path $0 \to m$ whose second-to-last atom is m'. Its final segment costs $C[m', m]$. By Definition 2 the anchor sets of its segments are pairwise disjoint, so by Definition 3 its cost is exactly (cost of the sub-path $0 \to m'$) plus $C[m', m]$. Feasibility of $(m', m) \in \mathcal{T}$ and the value $C[m', m]$ depend only on m', m, not on the sub-path's interior, so replacing the sub-path by a cheaper feasible one lowers the total (cut-and-paste). Hence the optimum ending at m via m' is $S(m', \ell-1) + C[m', m]$, and minimizing over feasible m' gives $S(m, \ell)$. Reading out at $(M-1, k+1)$ gives the claim. As in Theorem 3, only additivity over segments and endpoint-locality of C were used, not the metric structure. □

Since $S(0, 0) = 0$ and the readout is at $M-1$, indices 0 and $M-1$ are always used as boundary guards for the augmented path, but are never included in the deployed dictionary $\mathcal{D}_k$ itself; $\mathcal{H}_k$ (Equation (12)) is defined purely over the k interior atoms.

Proposition 2 (Computational complexity)**.** *Anchors cost $O(MN)$. The segment-cost matrix costs $O(M^2 N)$ in the straightforward implementation, and the recurrence costs $O(M^2 k)$, so the full solver runs in $O\big(M^2(N+k)\big)$ time. Exhaustive search over $\mathcal{H}_k$ costs $O\big(\binom{M-2}{k} Nk\big)$ yielding a total complexity of $O\left(Nk\left(\frac{e(M-2)}{k}\right)^k\right)$.*

Proof. Each request needs one scan over M candidates for its anchor. For each of the $O(M^2)$ pairs (i, j), computing $C[i, j]$ scans the requests anchored in $\mathcal{I}(i, j)$, at most N. The recurrence has Mk states, each minimizing over at most M predecessors. Exhaustive search enumerates $\binom{M-2}{k}$ interior subsets, since the k deployed atoms are chosen from the $M - 2$ interior candidates, and evaluates each in $O(Nk)$. □

What the anchor route provides. The anchor construction does more than make the selection problem polynomial: it yields an *a priori* bound on the regret relative to the constrained optimum, providing a computable certificate of how far the deployed solution can be from the best feasible dictionary, while preserving the original frozen-codec and single-atom deployment rule.

Algorithm 1 Anchor-routed surrogate construction

Require: Frontier $\mathcal{F} = (P_0, \ldots, P_{M-1})$; requests $\{(Q_n, w_n)\}_{n=1}^N$; spacing bounds $\delta_{\min}, \delta_{\max}$
Ensure: Segment-cost matrix C (encodes $\delta_{\min}, \delta_{\max}$); residual mean $\bar{r}$

for $n = 1, \ldots, N$ **do**
  $a(n) \leftarrow \min \arg\min_{m \in \{0,\ldots,M-1\}} d_W^2(Q_n, P_m)$
  $r_n \leftarrow d_W(Q_n, P_{a(n)})$
end for
$\bar{r} \leftarrow \sum_{n=1}^N w_n r_n$
for $0 \le i < j \le M - 1$ **do**
  if $\delta_{\min} \le \Delta_{ij} \le \delta_{\max}$ **then**
    $C[i, j] \leftarrow \sum_{n:\, a(n) \in \mathcal{I}(i,j)} w_n \min\{d_W^2(Q_n, P_i),\, d_W^2(Q_n, P_j)\}$
  else
    $C[i, j] \leftarrow \infty$
  end if
end for
return $C,\ \bar{r}$

Algorithm 1 freezes request ownership on the full frontier and builds the segment-cost matrix C and residual mean $\bar{r}$. Algorithm 2 solves the resulting surrogate exactly by dynamic programming and returns the dictionary $\mathcal{D}_k^{\mathrm{DP}}$ with its a priori certificate $\mathrm{Cert}(\delta_{\max})$.

C Surrogate Fidelity and Certified Near-Optimality

This section quantifies and certifies the price of that substitution. Unless explicitly marked with a superscript "pop" or a hat, $L_{\mathrm{anchor}}(\mathcal{D}_k)$ and $L_{\mathrm{true}}(\mathcal{D}_k)$ denote the *empirical* losses on a fixed request sample $\{(Q_n, w_n)\}_{n=1}^N$; population and finite-sample-corrected quantities appear only in Section C.7, and we never mix the two without an explicit generalization step.

Proposition 3 (Full-frontier and interior-reference losses)**.** *Define the full-frontier reference loss*

$$L_{\mathrm{full}} = \sum_{n=1}^{N} w_n \min_{0 \le m \le M-1} d_W^2(Q_n, P_m), \tag{21}$$

and the endpoint-excluded interior reference loss

$$L_{\mathrm{int}} = \sum_{n=1}^{N} w_n \min_{1 \le m \le M-2} d_W^2(Q_n, P_m). \tag{22}$$

Then for every $\mathcal{D}_k \in \mathcal{H}_k$,

$$L_{\mathrm{true}}(\mathcal{D}_k) \;\ge\; L_{\mathrm{int}} \;\ge\; L_{\mathrm{full}} \;\ge\; \bar{r}^2, \qquad \bar{r} = \sum_n w_n r_n, \tag{23}$$

and we define the excess selection loss $E_{\mathrm{sel}}(\mathcal{D}_k) = L_{\mathrm{true}}(\mathcal{D}_k) - L_{\mathrm{int}} \ge 0$.

Proof. r_n is the distance from Q_n to its nearest atom on the whole frontier $\mathcal{F}$ (endpoints included), so $r_n^2 \le \min_{1 \le m \le M-2} d_W^2(Q_n, P_m) \le e(Q_n, \mathcal{D}_k)$ for every $\mathcal{D}_k \in \mathcal{H}_k$; weight and sum to get $L_{\mathrm{full}} \le L_{\mathrm{int}} \le L_{\mathrm{true}}(\mathcal{D}_k)$. The bound $L_{\mathrm{full}} \ge \bar{r}^2$ is Jensen applied to $x \mapsto x^2$. □

Algorithm 2 Certified anchor-routed dictionary design

Require: $C, \bar{r}$ from Algorithm 1; dictionary size k; $\delta_{\max}$ (used only for the certificate)
Ensure: Dictionary $\mathcal{D}_k^{\mathrm{DP}}$ and certificate $\mathrm{Cert}(\delta_{\max})$
1: $S(0,0) \leftarrow 0$; $S(m,\ell) \leftarrow \infty$ otherwise
2: **for** $\ell = 1, \dots, k+1$ **do**
3: **for** $m = 1, \dots, M-1$ **do**
4: $S(m,\ell) \leftarrow \min_{m'<m}\{S(m',\ell-1) + C[m',m]\}$
5: **end for**
6: **end for**
7: Backtrack from $(M-1, k+1)$ to obtain $0 = i_0 < i_1 < \cdots < i_k < i_{k+1} = M-1$
8: $\mathcal{D}_k^{\mathrm{DP}} \leftarrow \{P_{i_t}\}_{t=1}^{k}$
9: $\mathrm{Cert}(\delta_{\max}) \leftarrow 2\bar{r}\,\delta_{\max} + \delta_{\max}^2$
At deployment: serve each request Q by $\arg\min_{P \in \mathcal{D}_k^{\mathrm{DP}}} d_W^2(Q,P)$
10: **return** $\mathcal{D}_k^{\mathrm{DP}}$, $\mathrm{Cert}(\delta_{\max})$

L_{full} is the hypothetical lower reference obtained if every candidate frontier atom, including the two endpoints, were available for deployment. Because the deployment contract excludes the endpoints from the k-atom dictionary, the irreducible reference within the deployment class is instead L_{int}, obtained by deploying all $M-2$ interior candidates. Thus $E_{\mathrm{sel}}(\mathcal{D}_k) = L_{\mathrm{true}}(\mathcal{D}_k) - L_{\mathrm{int}}$ isolates the loss caused by selecting only k of the $M-2$ interior atoms; we report L_{full} separately to quantify the additional gap caused by excluding the two boundary atoms. We report L_{full}, L_{int}, L_{true}, E_{sel} and Cert side by side, and quote tightness as $\mathrm{Cert}/L_{\mathrm{int}}$ rather than $\mathrm{Cert}/L_{\mathrm{true}}$, which would flatter the method by including an unavoidable term. In particular, the all-interior reference $\mathcal{D}_k = \mathcal{F}^\circ$ corresponds to $k = M-2$ and has $E_{\mathrm{sel}} = 0$ and regret *exactly* zero, never negative; any apparent negative value is floating-point noise, is clipped to zero, and is caught by the runtime assertion in Algorithm 2. We also report $\mathrm{Cert}/E_{\mathrm{sel}}(\mathcal{D}_k^\star)$, the ratio against the quantity selection actually controls.

C.1 The Exact Surrogate Gap and A Geometric Bound and Stability

Definition 4 (Per-request anchor mismatch)**.** *For fixed $\mathcal{D}_k$ with true owner $m^\star(n)$ and anchor-routed atom $r(n)$,*

$$\Delta_a(Q_n, \mathcal{D}_k) = d_W^2\big(Q_n, P_{r(n)}\big) - d_W^2\big(Q_n, P_{m^\star(n)}\big) \;\geq\; 0, \tag{24}$$

non-negative because $P_{r(n)} \in \mathcal{D}_k$ and $m^\star(n)$ minimizes over $\mathcal{D}_k$; and $\Delta_a = 0$ whenever $r(n) = m^\star(n)$.

Theorem 5 (Exact surrogate-gap decomposition)**.** *For every fixed $\mathcal{D}_k$ (on the empirical sample),*

$$L_{\mathrm{anchor}}(\mathcal{D}_k) - L_{\mathrm{true}}(\mathcal{D}_k) = \sum_{n=1}^{N} w_n \Delta_a(Q_n, \mathcal{D}_k) \;\geq\; 0, \tag{25}$$

and with $\pi_{\mathrm{mismatch}} = \sum_{n: r(n) \neq m^\star(n)} w_n$ and $\Delta_a^{\max} = \max_n \Delta_a(Q_n, \mathcal{D}_k)$,

$$L_{\mathrm{anchor}}(\mathcal{D}_k) - L_{\mathrm{true}}(\mathcal{D}_k) \leq \pi_{\mathrm{mismatch}} \cdot \Delta_a^{\max}. \tag{26}$$

Proof. By Definition 3 the anchor-routed cost of request n is $d_W^2(Q_n, P_{r(n)})$, well defined because Definition 2 puts $a(n)$ in exactly one segment; by equation 11 the true cost is $d_W^2(Q_n, P_{m^\star(n)})$. Subtracting and summing with weights gives equation 25; each term is non-negative by Definition 4. For equation 26, requests with $r(n) = m^\star(n)$ contribute 0, and the others contribute at most $\Delta_a^{\max}$ each with total weight π_{mismatch}. □

Theorem 6 (Geometric surrogate-gap bound)**.** *For every request Q_n and every $\mathcal{D}_k$ realized as a path,*

$$0 \leq \Delta_a(Q_n, \mathcal{D}_k) \leq 2 r_n h_n + h_n^2, \tag{27}$$

and consequently

$$L_{\mathrm{anchor}}(\mathcal{D}_k) - L_{\mathrm{true}}(\mathcal{D}_k) \leq \sum_n w_n \big(2 r_n h_n + h_n^2\big) \leq 2 r_{\max} h_{\max} + h_{\max}^2. \tag{28}$$

Proof. Step 1. $P_{a(n)}$ minimizes $d_W(Q_n, \cdot)$ over the full frontier $\mathcal{F}$, which contains $P_{m^\star(n)}$, so $r_n \leq d_W(Q_n, P_{m^\star(n)})$. *Step 2.* d_W is Euclidean, so the triangle inequality gives $d_W(Q_n, P_{r(n)}) \leq d_W(Q_n, P_{a(n)}) + d_W(P_{a(n)}, P_{r(n)}) = r_n + h_n$. *Step 3.* Squaring Step 2 and subtracting the square of Step 1, $\Delta_a(Q_n, \mathcal{D}_k) \leq (r_n + h_n)^2 - r_n^2 = 2 r_n h_n + h_n^2$. Multiply by w_n, sum, and apply Theorem 5; the last inequality uses $\sum_n w_n = 1$ and the maxima. □

C.2 The Spacing–Fidelity Tradeoff, and Its Feasibility Limits

Define the feasible class at spacing level δ,

$$\mathcal{H}_k(\delta) = \{\mathcal{D}_k : P_0, P_{M-1} \in \mathcal{D}_k,\ \delta_{\min} \le \Delta_{i_t i_{t+1}} \le \delta\ \forall t\}. \tag{29}$$

Proposition 4 (Spacing–fidelity tradeoff). *Let* $\mathrm{OPT}_{\mathrm{true}}(\delta) := \min_{\mathcal{D}_k \in \mathcal{H}_k(\delta)} L_{\mathrm{true}}(\mathcal{D}_k)$ *and* $\mathrm{Cert}(\delta) := 2\bar{r}\delta + \delta^2$ *with* $\bar{r} \ge 0$. *For any* $\delta_1 \le \delta_2$*:*

(i) $\mathcal{H}_k(\delta_1) \subseteq \mathcal{H}_k(\delta_2)$*, and consequently* $\mathrm{OPT}_{\mathrm{true}}(\delta_1) \ge \mathrm{OPT}_{\mathrm{true}}(\delta_2)$*;*

(ii) $\mathrm{Cert}(\delta_1) \le \mathrm{Cert}(\delta_2)$.

Proof. (i) Any $\mathcal{D}_k \in \mathcal{H}_k(\delta_1)$ satisfies $\Delta_{i_t i_{t+1}} \le \delta_1 \le \delta_2$ for all t, hence also the weaker constraint defining $\mathcal{H}_k(\delta_2)$; thus $\mathcal{D}_k \in \mathcal{H}_k(\delta_2)$, giving the inclusion. Minimizing L_{true} over the larger set $\mathcal{H}_k(\delta_2)$ can only lower the minimum, so $\mathrm{OPT}_{\mathrm{true}}(\delta_1) \ge \mathrm{OPT}_{\mathrm{true}}(\delta_2)$.

(ii) Cert is increasing on $\delta \ge 0$ since $\mathrm{Cert}'(\delta) = 2\bar{r} + 2\delta \ge 0$ there, so $\delta_1 \le \delta_2$ implies $\mathrm{Cert}(\delta_1) \le \mathrm{Cert}(\delta_2)$. Making the atom spacing smaller gives you a tighter (more trustworthy) error guarantee, but it also shrinks the set of dictionaries you're allowed to pick from — so there's a tradeoff between how tight your guarantee is and how much freedom you have to find a good dictionary. □

Proposition 5 (Necessary feasibility conditions on the spacing bounds). *Let* $\Lambda_{\mathrm{front}} = \sum_{i=0}^{M-2} \Delta_{i,i+1}$ *be the arc length of the polyline through the sorted frontier. If* $\mathcal{H}_k \ne \varnothing$ *then*

$$\delta_{\max} \ \ge\ \frac{\Delta_{0,M-1}}{k+1} \qquad \textit{and} \qquad \delta_{\min} \ \le\ \frac{\Lambda_{\mathrm{front}}}{k+1}, \tag{30}$$

and consequently the certificate obeys the unavoidable lower bound

$$\mathrm{Cert}(\delta_{\max}) \ \ge\ \frac{2\bar{r}\,\Delta_{0,M-1}}{k+1} + \frac{\Delta_{0,M-1}^2}{(k+1)^2}. \tag{31}$$

Neither condition in equation 30 is sufficient: feasibility remains combinatorial, since the steps must land on actual frontier indices.

Proof. A feasible dictionary is a path $P_0 = P_{i_0}, \dots, P_{i_{k+1}} = P_{M-1}$ with $k+1$ steps. For the upper bound, the triangle inequality gives $\Delta_{0,M-1} \le \sum_t \Delta_{i_t i_{t+1}} \le (k+1)\delta_{\max}$. For the lower bound, each step spans a set of consecutive frontier edges, and by the triangle inequality $\Delta_{i_t i_{t+1}} \le \sum_{i=i_t}^{i_{t+1}-1} \Delta_{i,i+1}$; summing over disjoint index ranges gives $(k+1)\delta_{\min} \le \sum_t \Delta_{i_t i_{t+1}} \le \Lambda_{\mathrm{front}}$. Substituting the first into the increasing function Cert gives equation 31. Note Λ_{front} cannot be replaced by the chord $\Delta_{0,M-1}$, since a non-affine frontier has $\Lambda_{\mathrm{front}} > \Delta_{0,M-1}$. □

By Proposition 3, $L_{\mathrm{true}}(\mathcal{D}_k) \ge L_{\mathrm{int}}$, so the certificate is useful only when small relative to L_{int}; a sufficient regime is $\delta_{\max} \ll \bar{r}$, which together with Proposition 5 requires $k+1 \gg \Delta_{0,M-1}/\bar{r}$. We report $\mathrm{Cert}/L_{\mathrm{int}}$ for every k, since L_{int} is the irreducible reference within the endpoint-excluded deployment class; the weaker L_{full} reference is reported separately. We flag values above one as uninformative. For $k = 2$, feasibility requires $\delta_{\max} \ge \Delta_{0,M-1}/3$ (there are three gaps when $k = 2$), making the certificate potentially uninformative; the measured $\varepsilon_{\mathrm{sur}}^{(k)}$ therefore provides the relevant small-k diagnostic.

C.3 Exactness on Affine Frontiers

Theorem 7 (Zero gap on affine normalized frontiers). *If the normalized frontier is collinear, i.e.* $z(P_i) = c_0 + t_i\, v$ *with* $v \ne 0$ *and* $t_0 < \cdots < t_{M-1}$*, then for every request and every* $\mathcal{D}_k \in \mathcal{H}_k$*,* $d_W^2(Q, P_{r(Q)}) = e(Q, \mathcal{D}_k)$*, hence* $L_{\mathrm{anchor}}(\mathcal{D}_k) = L_{\mathrm{true}}(\mathcal{D}_k)$ *and* $\varepsilon_{\mathrm{sur}} = 0$.

Proof. Absorb the scale so $\|v\|_2 = 1$; then $d_W(P_i, P_j) = |t_i - t_j|$. Let $\pi = \langle z(Q) - c_0, v\rangle$ and σ^2 the squared orthogonal residual, so

$$d_W^2(Q, P_i) = (\pi - t_i)^2 + \sigma^2, \tag{32}$$

and $a(\cdot)$, $m^\star(\cdot)$ reduce to nearest-point problems in π against $\{t_i\}$; the constant σ^2 affects neither argmin. Let $a = a(Q)$ and let $\alpha = i_{t^\dagger}$, $\omega = i_{t^\dagger+1}$ be the endpoints of the unique segment of Definition 2 containing a; $r(Q)$ is the nearer of t_α, t_ω to π.

Case 1: $t_\alpha \le \pi \le t_\omega$. Any other selected t_{i_s} is $\le t_\alpha$ or $\ge t_\omega$ by monotonicity, hence at least as far from π as t_α or t_ω respectively, so the nearer of t_α, t_ω is the global nearest selected.

Case 2: $\pi < t_\alpha$. Since $\alpha \le a$, $t_a \ge t_\alpha > \pi$, so $|\pi - t_a| \ge |\pi - t_\alpha|$; but a minimizes over *all* candidates and α is one, so $|\pi - t_a| \le |\pi - t_\alpha|$. Strict monotonicity of t_i gives $a = \alpha$ (with ties in equation 14 broken toward the smaller index, this holds even if π is equidistant from two adjacent t_i). If the current segment is not the first segment, every deployed index to the right has $t \ge t_\omega > t_\alpha \ge t_{i_1} > \pi$, hence is farther, and t_α is itself deployed, so t_α is the nearest deployed atom and $r(Q) = \alpha$. If the current segment is the first segment ($\alpha = i_0 = 0$), $r(Q) = i_1 = \omega$ is forced by Definition 3; since $\pi < t_\alpha = t_0 < t_{i_1}$ places Q before every deployed atom, t_{i_1} is trivially the nearest deployed atom, so $r(Q) = \omega$ still matches $m^\star(Q)$.

Case 3: $\pi > t_\omega$. If the current segment is not the last segment, this case cannot occur under the half-open convention: $a \le \omega - 1$ by definition of $\mathcal{I}$, contradicting the argument's forced $a = \omega$. On the last segment $\mathcal{I}(i_k, M-1)$ (closed), $\omega = M-1$ and $a = M-1$ is permitted; there $r(Q) = i_k$ is forced by Definition 3. Since t_{M-1} is the largest value on the frontier and $\pi > t_{M-1}$ places Q beyond every candidate, P_{i_k} is the nearest deployed atom on this segment, since no deployed atom lies between t_{i_k} and t_{M-1} closer to π than t_{i_k} itself. Hence $r(Q) = i_k$ matches $m^\star(Q)$ here too.

In every case $r(Q)$ is the nearest selected atom. Summing with weights gives the claim. The affine case is a useful special case: it gives zero surrogate gap, while the general geometric bound quantifies the gap for non-affine frontiers. □

C.4 Uniform Transfer to the Deployed Objective

Definition 5 (Uniform empirical surrogate gap)**.** *On the fixed empirical sample* $\{(Q_n, w_n)\}_{n=1}^N$,

$$\varepsilon_{\mathrm{sur}} = \sup_{\mathcal{D}_k \in \mathcal{H}_k} \big[L_{\mathrm{anchor}}(\mathcal{D}_k) - L_{\mathrm{true}}(\mathcal{D}_k)\big] \;\ge\; 0, \tag{33}$$

non-negative by Theorem 3; we write $\varepsilon_{\mathrm{sur}}^{(k)}$ *for its dependence on* k*. This is an* empirical *quantity computed on the same finite sample used to fit* $\mathcal{D}_k^{\mathrm{DP}}$*; Section C.7 relates it to the population objective.*

Theorem 8 (Uniform surrogate transfer)**.** *Let* $\mathcal{D}_k^{\mathrm{DP}} = \arg\min_{\mathcal{H}_k} L_{\mathrm{anchor}}$ *(Theorem 4) and* $\mathcal{D}_k^\star = \arg\min_{\mathcal{H}_k} L_{\mathrm{true}}$*, both with respect to the same empirical sample. Then*

$$L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \le L_{\mathrm{true}}(\mathcal{D}_k^\star) + \varepsilon_{\mathrm{sur}}, \tag{34}$$

equivalently $E_{\mathrm{sel}}(\mathcal{D}_k^{\mathrm{DP}}) \le E_{\mathrm{sel}}(\mathcal{D}_k^\star) + \varepsilon_{\mathrm{sur}}$*, and, if* $L_{\mathrm{true}}(\mathcal{D}_k^\star) > 0$*,* $L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}})/L_{\mathrm{true}}(\mathcal{D}_k^\star) \le 1 + \varepsilon_{\mathrm{sur}}/L_{\mathrm{true}}(\mathcal{D}_k^\star)$*. If moreover the gap is relative,* $L_{\mathrm{anchor}}(\mathcal{D}_k) \le (1+\gamma) L_{\mathrm{true}}(\mathcal{D}_k)$ *for every* $\mathcal{D}_k \in \mathcal{H}_k$*, then* $L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \le (1+\gamma) L_{\mathrm{true}}(\mathcal{D}_k^\star)$*.*

Proof. Both dictionaries lie in $\mathcal{H}_k$, and

$$L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \overset{\text{Thm. 3}}{\le} L_{\mathrm{anchor}}(\mathcal{D}_k^{\mathrm{DP}}) \overset{\text{Thm. 4}}{\le} L_{\mathrm{anchor}}(\mathcal{D}_k^\star) \overset{\text{Def. 5}}{\le} L_{\mathrm{true}}(\mathcal{D}_k^\star) + \varepsilon_{\mathrm{sur}},$$

the middle step using that $\mathcal{D}_k^{\mathrm{DP}}$ minimizes L_{anchor} over $\mathcal{H}_k$ and $\mathcal{D}_k^\star \in \mathcal{H}_k$. The excess form follows by subtracting the constant L_{full} from both sides; the ratio form by dividing by $L_{\mathrm{true}}(\mathcal{D}_k^\star) > 0$; and under the relative hypothesis the last step is replaced by $L_{\mathrm{anchor}}(\mathcal{D}_k^\star) \le (1+\gamma) L_{\mathrm{true}}(\mathcal{D}_k^\star)$. □

C.5 The Computable Certificate

Corollary 1 (Computable a priori regret certificate)**.** *With spacing bound* $\delta_{\max}$ *active,*

$$\varepsilon_{\mathrm{sur}} \le 2\bar{r}\,\delta_{\max} + \delta_{\max}^2 \;\le\; 2 r_{\max}\delta_{\max} + \delta_{\max}^2, \qquad \bar{r} = \sum_n w_n r_n,\; r_{\max} = \max_n r_n, \tag{35}$$

and therefore $L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \le L_{\mathrm{true}}(\mathcal{D}_k^\star) + 2\bar{r}\,\delta_{\max} + \delta_{\max}^2$*. Both* $\bar{r}$ *and* $r_{\max}$ *are computable in* $O(MN)$ *time* before any optimizer is run*, and all quantities are empirical.*

Proof. For every $\mathcal{D}_k \in \mathcal{H}_k$ and every n, Lemma 1 gives $h_n \le \delta_{\max}$. Theorem 6 then gives $L_{\mathrm{anchor}}(\mathcal{D}_k) - L_{\mathrm{true}}(\mathcal{D}_k) \le \sum_n w_n(2r_n\delta_{\max} + \delta_{\max}^2) = 2\bar{r}\delta_{\max} + \delta_{\max}^2$. Take the supremum over $\mathcal{H}_k$; the second inequality follows from $\bar{r} \le r_{\max}$. Combine with Theorem 8. □

Corollary 2 (Certified lower bound on the unknown constrained optimum)**.** *With $\widehat{\varepsilon_{\mathrm{sur}}}$ any valid upper bound on the empirical $\varepsilon_{\mathrm{sur}}$,*

$$\max\left\{L_{\mathrm{int}},\ L_{\mathrm{anchor}}(\mathcal{D}_k^{\mathrm{DP}}) - \widehat{\varepsilon_{\mathrm{sur}}}\right\} \ \le\ \min_{\mathcal{D}_k \in \mathcal{H}_k} L_{\mathrm{true}}(\mathcal{D}_k) = L_{\mathrm{true}}(\mathcal{D}_k^\star) \ \le\ L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}). \tag{36}$$

Proof. The right inequality holds because $\mathcal{D}_k^{\mathrm{DP}} \in \mathcal{H}_k$. For the left: for every $\mathcal{D}_k \in \mathcal{H}_k$, $L_{\mathrm{true}}(\mathcal{D}_k) \ge L_{\mathrm{anchor}}(\mathcal{D}_k) - \varepsilon_{\mathrm{sur}} \ge L_{\mathrm{anchor}}(\mathcal{D}_k^{\mathrm{DP}}) - \widehat{\varepsilon_{\mathrm{sur}}}$, the second step by optimality of $\mathcal{D}_k^{\mathrm{DP}}$ for L_{anchor}; and $L_{\mathrm{true}}(\mathcal{D}_k) \ge L_{\mathrm{int}}$ by Proposition 3. Minimize over $\mathcal{H}_k$. □

C.6 Certified Margin Over Any Feasible Competitor

The transfer theorem compares $\mathcal{D}_k^{\mathrm{DP}}$ only with the constrained optimum. The following turns it into a statement about every baseline without assuming anything about how the competitor was produced.

Corollary 3 (Competitor margin)**.** *Let $\mathcal{D}_b \in \mathcal{H}_k$ be the dictionary returned by* any *method, subject only to $\mathcal{D}_b \in \mathcal{H}_k$. Then*

$$L_{\mathrm{true}}(\mathcal{D}_b) - L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \ \ge\ \left[L_{\mathrm{true}}(\mathcal{D}_b) - L_{\mathrm{true}}(\mathcal{D}_k^\star)\right] - \varepsilon_{\mathrm{sur}}. \tag{37}$$

Consequently, for any valid $\widehat{\varepsilon_{\mathrm{sur}}} \ge \varepsilon_{\mathrm{sur}}$,

$$L_{\mathrm{true}}(\mathcal{D}_b) - L_{\mathrm{true}}(\mathcal{D}_k^\star) > \widehat{\varepsilon_{\mathrm{sur}}} \quad\Longrightarrow\quad L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) < L_{\mathrm{true}}(\mathcal{D}_b). \tag{38}$$

Proof. By Theorem 8, $L_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \le L_{\mathrm{true}}(\mathcal{D}_k^\star) + \varepsilon_{\mathrm{sur}}$. Subtract this from $L_{\mathrm{true}}(\mathcal{D}_b)$ to obtain equation 37. If the bracket exceeds $\widehat{\varepsilon_{\mathrm{sur}}} \ge \varepsilon_{\mathrm{sur}}$, the right side of equation 37 is strictly positive. Thus the DP is certified to outperform any feasible competitor whose suboptimality exceeds the surrogate gap by more than the available margin. □

C.7 End-to-End Population Guarantee

We now connect the empirical statements to the population objective. This requires care: $\varepsilon_{\mathrm{sur}}$ is empirical, and it does *not* directly bound the gap between population losses without an explicit finite-sample argument.

Assumption 1 (Uniformly bounded pairwise distances)**.** *There is a finite B with $d_W^2(Q,P) \le B^2$ for every $Q \in \mathrm{supp}(\mathcal{Q})$ and every $P \in \mathcal{F}$. Because both L_{true} and L_{anchor} are averages of terms $d_W^2(Q,P)$ with $P \in \mathcal{F}$ (the true loss uses $P = P_{m^\star(Q)} \in \mathcal{D}_k \subseteq \mathcal{F}$, the surrogate uses $P = P_{r(Q)} \in \mathcal{F}$), this implies both per-request costs lie in $[0, B^2]$ for every $\mathcal{D}_k \in \mathcal{H}_k$ — stronger than bounding $e(Q, \mathcal{D}_k)$ alone, and exactly what is needed to concentrate both losses simultaneously.*

The experiments enforce this exactly: requests are drawn from the Gaussian of Section E.1 truncated to a bounded region containing the frontier, so B is computed directly.

Lemma 2 (Uniform finite-sample deviation, Hoeffding form)**.** *Under Assumption 1 , for $\eta \in (0,1)$, with probability at least $1-\eta$ over $Q_1, \dots, Q_N \sim \mathcal{Q}$ i.i.d. with uniform weights $w_n = 1/N$, simultaneously for every $\mathcal{D}_k \in \mathcal{H}_k$ and for both losses,*

$$\max\left\{\left|\widehat{L}_{\mathrm{anchor}}(\mathcal{D}_k) - L^{\mathrm{pop}}_{\mathrm{anchor}}(\mathcal{D}_k)\right|,\ \left|\widehat{L}_{\mathrm{true}}(\mathcal{D}_k) - L^{\mathrm{pop}}_{\mathrm{true}}(\mathcal{D}_k)\right|\right\} \le \varepsilon_{\mathrm{gen}}(N,\eta) := B^2\sqrt{\frac{\log(4|\mathcal{H}_k|/\eta)}{2N}}, \quad |\mathcal{H}_k| \le \tbinom{M-2}{k}. \tag{39}$$

Proof. $\mathcal{F}$, the scales s_D, s_R, and $\delta_{\min}, \delta_{\max}$ are determined by the pilot stage and independent of the request sample, so $\mathcal{H}_k$ is a finite class fixed before $\{Q_n\}$ is drawn. For fixed $\mathcal{D}_k$ and either loss, each summand lies in $[0, B^2]$ and the summands are i.i.d., so Hoeffding gives $\Pr[|\widehat{L} - L^{\mathrm{pop}}| > \epsilon] \le 2\exp(-2N\epsilon^2/B^4)$. A union bound over $|\mathcal{H}_k|$ dictionaries and two losses gives failure probability at most $4|\mathcal{H}_k|\exp(-2N\epsilon^2/B^4)$; solving for ϵ yields equation 39. □

Corollary 4 (Sample-complexity form)**.** *Using $\binom{n}{r} \le (en/r)^r$ with $n = M-2$, $r = k$,*

$$\varepsilon_{\mathrm{gen}}(N,\eta) = O\left(B^2\sqrt{\frac{k\log((M-2)/k) + \log(1/\eta)}{N}}\right), \qquad N = \Omega\left(\frac{B^4\left[\,k\log((M-2)/k) + \log(1/\eta)\,\right]}{\epsilon^2}\right) \tag{40}$$

suffices for $\varepsilon_{\mathrm{gen}} \le \epsilon$.

Proof. $\log|\mathcal{H}_k| \le \log\binom{M-2}{k} \le k\log\big(e(M-2)/k\big) = O(k\log((M-2)/k))$; substitute into equation 39 and solve $\varepsilon_{\mathrm{gen}} \le \epsilon$ for N. □

Enlarging the candidate frontier costs only *logarithmically* in requests; enlarging the deployed dictionary costs *linearly*, reflecting that k, not M, is the combinatorial complexity of the class the DP searches. This ties the optimization complexity of Proposition 2 to the statistical complexity of the same problem.

Proposition 6 (Empirical Bernstein deviation)**.** *Under Assumption 1 , let $\widehat{V}(\mathcal{D}_k)$ denote the unbiased* empirical *variance of the per-request cost $c(Q,\mathcal{D}_k) \in [0,B^2]$ on the sample, for the relevant loss. With $\kappa = \log(4|\mathcal{H}_k|/\eta)$, an empirical Bernstein inequality (13) gives, with probability at least $1-\eta$, simultaneously over $\mathcal{H}_k$ and both losses,*

$$\big|\widehat{L}(\mathcal{D}_k) - L^{\mathrm{pop}}(\mathcal{D}_k)\big| \;\le\; \sqrt{\frac{2\widehat{V}(\mathcal{D}_k)\,\kappa}{N}} + \frac{7B^2\kappa}{3(N-1)} \;=:\; \widehat{\varepsilon_{\mathrm{gen}}}(\mathcal{D}_k;N,\eta). \tag{41}$$

Proof. Apply the empirical Bernstein bound of (13) to variables in $[0,B^2]$, in both directions, then union bound over the $|\mathcal{H}_k|$ dictionaries and the two losses, which is absorbed into κ. □

We report equation 41 directly rather than substituting an empirical variance into a population-variance Bernstein inequality, which would not be a valid bound. When the losses are small, $\sqrt{\widehat{V}\kappa/N} \ll B^2\sqrt{\kappa/N}$ and equation 41 is much sharper than equation 39; Corollary 5 holds with $\varepsilon_{\mathrm{gen}}$ replaced by $\sup_{\mathcal{D}_k\in\mathcal{H}_k}\widehat{\varepsilon_{\mathrm{gen}}}(\mathcal{D}_k;N,\eta)$, and we report both numbers.

Corollary 5 (End-to-end population guarantee)**.** *Let $\mathcal{Q}_{\mathrm{tr}} = \{(Q_n,w_n)\}_{n=1}^N$ be the i.i.d. training sample with uniform weights used to fit $\mathcal{D}_k^{\mathrm{DP}}$, and let $\varepsilon_{\mathrm{sur}}$ be the* empirical *uniform surrogate gap on this same sample. Under Assumption 1 , with probability at least $1-\eta$,*

$$L^{\mathrm{pop}}_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) \le \min_{\mathcal{D}_k\in\mathcal{H}_k} L^{\mathrm{pop}}_{\mathrm{true}}(\mathcal{D}_k) + \varepsilon_{\mathrm{sur}} + 2\varepsilon_{\mathrm{gen}}(N,\eta). \tag{42}$$

Proof. Let $\mathcal{D}^{\mathrm{pop}} \in \arg\min_{\mathcal{D}_k\in\mathcal{H}_k} L^{\mathrm{pop}}_{\mathrm{true}}(\mathcal{D}_k)$, a population quantity not depending on $\mathcal{Q}_{\mathrm{tr}}$. Then

$$\begin{aligned}
L^{\mathrm{pop}}_{\mathrm{true}}(\mathcal{D}_k^{\mathrm{DP}}) &\le L^{\mathrm{pop}}_{\mathrm{anchor}}(\mathcal{D}_k^{\mathrm{DP}}) && \text{(Thm. 3, pointwise, integrated over } \mathcal{Q})\\
&\le \widehat{L}_{\mathrm{anchor}}(\mathcal{D}_k^{\mathrm{DP}}) + \varepsilon_{\mathrm{gen}}(N,\eta) && \text{(Lem. 2 for } L_{\mathrm{anchor}})\\
&\le \widehat{L}_{\mathrm{anchor}}(\mathcal{D}^{\mathrm{pop}}) + \varepsilon_{\mathrm{gen}}(N,\eta) && \text{(Thm. 4; } \mathcal{D}^{\mathrm{pop}}\in\mathcal{H}_k)\\
&\le \widehat{L}_{\mathrm{true}}(\mathcal{D}^{\mathrm{pop}}) + \varepsilon_{\mathrm{sur}} + \varepsilon_{\mathrm{gen}}(N,\eta) && \text{(Def. 5 at } \mathcal{D}^{\mathrm{pop}}\in\mathcal{H}_k)\\
&\le L^{\mathrm{pop}}_{\mathrm{true}}(\mathcal{D}^{\mathrm{pop}}) + \varepsilon_{\mathrm{sur}} + 2\varepsilon_{\mathrm{gen}}(N,\eta) && \text{(Lem. 2 for } L_{\mathrm{true}}).
\end{aligned}$$

The second and fifth steps hold simultaneously for all $\mathcal{D}_k \in \mathcal{H}_k$ and both losses on the single high-probability event of Lemma 2, whose union bound already covers both, so no extra union bound is needed. The fourth step uses that $\varepsilon_{\mathrm{sur}}$ is a supremum over the whole class $\mathcal{H}_k$ and so covers $\mathcal{D}^{\mathrm{pop}}$; it never compares a population quantity to an empirical one, avoiding the error of bounding $L^{\mathrm{pop}}_{\mathrm{anchor}} - L^{\mathrm{pop}}_{\mathrm{true}}$ by the empirical $\varepsilon_{\mathrm{sur}}$. The same chain holds verbatim with $\varepsilon_{\mathrm{gen}}$ replaced by $\sup_{\mathcal{D}_k\in\mathcal{H}_k}\widehat{\varepsilon_{\mathrm{gen}}}$ from Proposition 6. Thus the population excess consists of the surrogate error $\varepsilon_{\mathrm{sur}}$ and the finite-sample deviation $2\varepsilon_{\mathrm{gen}}$; the former is controlled by the frontier geometry and spacing constraint, while the latter decreases with the number of requests N. □

D Stability of the Deployed System Under Request Shift and Measurement Error

D.1 Lipschitz Stability of the Matching Cost and Request-Distribution Shift

A request is an *observed* quantity, subject to quantization, measurement error, or elicitation noise, and the true request distribution need not be Gaussian. The deployed loss should not be brittle to either.

Lemma 3 (Nearest-atom distance is 1-Lipschitz in the request)**.** *For any fixed $\mathcal{D}_k$ and requests Q,Q', $\big|\sqrt{e(Q,\mathcal{D}_k)} - \sqrt{e(Q',\mathcal{D}_k)}\big| \le d_W(Q,Q')$.*

Proof. Let $\Gamma(\mathcal{D}_k) = \{z(P) : P \in \mathcal{D}_k\}$, finite hence closed and nonempty, so $\sqrt{e(Q, \mathcal{D}_k)} = \operatorname{dist}(z(Q), \Gamma(\mathcal{D}_k))$. Distance to a fixed nonempty closed set is 1-Lipschitz: for any $\zeta \in \Gamma$, $\operatorname{dist}(z, \Gamma) \le \|z - \zeta\| \le \|z - z'\| + \|z' - \zeta\|$; take the infimum over ζ and symmetrize. Finally $\|z(Q) - z(Q')\|_2 = d_W(Q, Q')$. □

Corollary 6 (Additive stability). *If $e(Q, \mathcal{D}_k), e(Q', \mathcal{D}_k) \le B^2$ then $|e(Q, \mathcal{D}_k) - e(Q', \mathcal{D}_k)| \le 2B\, d_W(Q, Q')$.*

Proof. With $u, v \in [0, B]$ the two square roots, $|u^2 - v^2| = |u - v|(u + v) \le 2B|u - v|$; apply Lemma 3. □

Theorem 9 (Wasserstein stability of the population risk). *Let $\mathcal{Q}, \mathcal{Q}'$ be request distributions supported on the compact set of Assumption 1 , so that $e(\cdot, \mathcal{D}_k) \le B^2$ $\mathcal{Q}$- and $\mathcal{Q}'$-a.s. for every $\mathcal{D}_k$. Then for every $\mathcal{D}_k \subseteq \mathcal{F}$,*

$$\big|L_{\mathcal{Q}}(\mathcal{D}_k) - L_{\mathcal{Q}'}(\mathcal{D}_k)\big| \;\le\; 2B\, W_1(\mathcal{Q}, \mathcal{Q}') \tag{43}$$

where W_1 is the 1-Wasserstein distance with respect to d_W and $L_{\mathcal{Q}}(\mathcal{D}_k) = \mathbb{E}_{Q \sim \mathcal{Q}}[e(Q, \mathcal{D}_k)]$. Consequently the ordering of any two dictionaries by population risk can be reversed by a shift only if $W_1(\mathcal{Q}, \mathcal{Q}') \ge |L_{\mathcal{Q}}(\mathcal{D}_1) - L_{\mathcal{Q}}(\mathcal{D}_2)|/(4B)$.

Proof. By Corollary 6, $Q \mapsto e(Q, \mathcal{D}_k)$ is $2B$-Lipschitz with respect to d_W on the common support. Kantorovich–Rubinstein duality (14) states $W_1(\mathcal{Q}, \mathcal{Q}') = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \big[\mathbb{E}_{\mathcal{Q}} f - \mathbb{E}_{\mathcal{Q}'} f\big]$, the supremum over 1-Lipschitz f with respect to d_W. Apply it to $f = e(\cdot, \mathcal{D}_k)/(2B)$:

$$\big|L_{\mathcal{Q}}(\mathcal{D}_k) - L_{\mathcal{Q}'}(\mathcal{D}_k)\big| = 2B\big|\mathbb{E}_{\mathcal{Q}} f - \mathbb{E}_{\mathcal{Q}'} f\big| \le 2B\, W_1(\mathcal{Q}, \mathcal{Q}').$$

For the ordering claim, if $L_{\mathcal{Q}}(\mathcal{D}_1) < L_{\mathcal{Q}}(\mathcal{D}_2)$, reversing the order under $\mathcal{Q}'$ requires each risk to move by at least half their gap, hence $4B\, W_1 \ge |L_{\mathcal{Q}}(\mathcal{D}_1) - L_{\mathcal{Q}}(\mathcal{D}_2)|$. □

Concentration of the Physically Deployed Operating Point. At deployment a request is routed to a single frozen atom $m^\star(Q)$ and a batch of n real test images is passed through it. Let $d^{(i)}, \ell^{(i)}$ be per-image distortion and KL values with means $D_{m^\star}, \bar{R}_{m^\star}$ (the atom's *nominal* point, estimated on the validation split) and finite variances $\sigma_D^2(m^\star), \sigma_R^2(m^\star)$.

Theorem 10 (Finite-batch concentration). *With $\hat{D}_n = \frac{1}{n}\sum_i d^{(i)}$, $\hat{R}_n = \frac{1}{n}\sum_i \ell^{(i)}$ over an i.i.d. batch, for any $t > 0$*

$$\Pr\big[|\hat{D}_n - D_{m^\star}| \ge t\big] \le \frac{\sigma_D^2(m^\star)}{nt^2}, \qquad \Pr\big[|\hat{R}_n - \bar{R}_{m^\star}| \ge t\big] \le \frac{\sigma_R^2(m^\star)}{nt^2},$$

and since the pixel-space error on $[0,1]$-valued images with a sigmoid output satisfies $d^{(i)} \in [0, d_{\max}]$ with $d_{\max} = 1$, $\Pr[|\hat{D}_n - D_{m^\star}| \ge t] \le 2\exp(-2nt^2/d_{\max}^2)$. In either case the deviation is $O_P(n^{-1/2})$ and $\hat{D}_n \to D_{m^\star}$ almost surely.

Proof. Chebyshev: for i.i.d. $d^{(i)}$ with mean $D_{m^\star}$ and variance σ_D^2, $\operatorname{Var}(\hat{D}_n) = \sigma_D^2/n$, so $\Pr[|\hat{D}_n - D_{m^\star}| \ge t] \le \sigma_D^2/(nt^2)$; identically for $\hat{R}_n$. For the sharper bound, $d^{(i)} = \|x - \hat{x}\|_2^2/(CHW) \in [0, 1]$ because each pixel error is at most 1, and Hoeffding for the mean of n i.i.d. variables in $[0, d_{\max}]$ gives $2\exp(-2nt^2/d_{\max}^2)$. Almost-sure convergence is the strong law, valid since the $d^{(i)}$ are i.i.d. conditional on the fixed frozen model with finite mean. The $O_P(n^{-1/2})$ rate follows by inverting either bound. The rate $\ell^{(i)}$ is not bounded a priori, so only the Chebyshev form is claimed for $\hat{R}_n$. □

D.2 Stability of the Deployed Loss Under Measurement Error

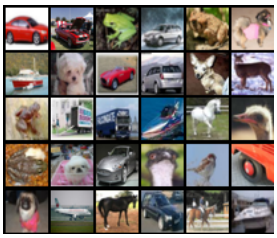
Proposition 7 (Stability of the true objective under atom-coordinate error). *Suppose each measured atom satisfies $\|z(P_i) - z(\widetilde{P}_i)\|_2 \le \varepsilon$ for all i, where $\widetilde{P}_i$ are the (unknown) exact atoms. For a fixed dictionary $\mathcal{D}_k$ and request Q write $d^\star_{\mathcal{D}_k}(Q) = \min_{P \in \mathcal{D}_k} d_W(Q, P)$ and $\widetilde{d}^\star_{\mathcal{D}_k}(Q) = \min_{P \in \mathcal{D}_k} d_W(Q, \widetilde{P})$. Then*

$$\big|d^\star_{\mathcal{D}_k}(Q) - \widetilde{d}^\star_{\mathcal{D}_k}(Q)\big| \le \varepsilon, \qquad \big|e(Q, \mathcal{D}_k) - \widetilde{e}(Q, \mathcal{D}_k)\big| \le 2\varepsilon\, d^\star_{\mathcal{D}_k}(Q) + \varepsilon^2, \tag{44}$$

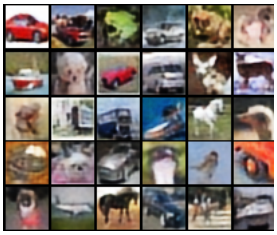
and hence for any weighted sample, (so that $d^\star_{\mathcal{D}_k}(Q) \le B$),

$$\big|L_{\mathrm{true}}(\mathcal{D}_k) - \widetilde{L}_{\mathrm{true}}(\mathcal{D}_k)\big| \le 2\varepsilon \sum_n w_n\, d^\star_{\mathcal{D}_k}(Q_n) + \varepsilon^2 \le 2\varepsilon B + \varepsilon^2. \tag{45}$$

Moreover $|\Delta_{ij} - \widetilde{\Delta}_{ij}| \le 2\varepsilon$, so a spacing constraint imposed with slack 2ε is satisfied by the exact atoms as well.



(a) Original Cifar-10 Images



(b) Reconstructed Cifar-10 Images



(c) Original vs. Reconstructed Cifar-10 Images

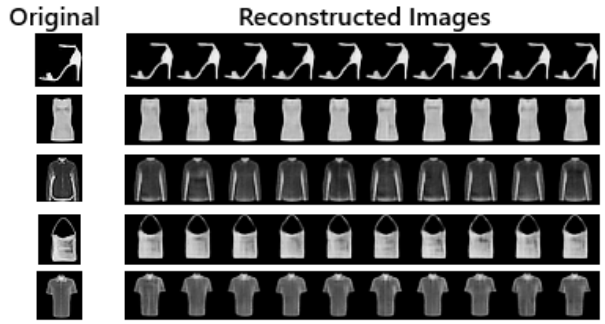
Figure 8: Reconstructed Cifar-10 Images by dictionaries in fig. 4a



(a) Original FashionMnist Images



(b) Reconstructed FashionMnist Images



(c) Original vs. Reconstructed FashionMnist Images

Figure 9: Reconstructed FashionMNIST Images by dictionaries in fig. 4c

Proof. Each $|d_W(Q,P_i)-d_W(Q,\widetilde{P}_i)| \le \varepsilon$ by the triangle inequality, and a minimum of finitely many ε-close numbers moves by at most ε, giving the first claim. Then $|(d^\star)^2-(\widetilde{d}^\star)^2| = |d^\star-\widetilde{d}^\star|(d^\star+\widetilde{d}^\star) \le \varepsilon(2d^\star+\varepsilon)$. Weight and sum. There is a finite B with $d_W^2(Q,P) \le B^2$, define $B := \max_{Q\in\mathrm{supp}(\mathcal{Q}),,P\in\mathcal{F}} d_W(Q,P) < \infty$, for every $Q \in \mathrm{supp}(\mathcal{Q})$ and every $P \in \mathcal{F}$, so in particular $d^\star_{\mathcal{D}_k}(Q_n) \le B$ for every $\mathcal{D}_k \in \mathcal{H}_k$. For the gaps, apply the triangle inequality twice. □

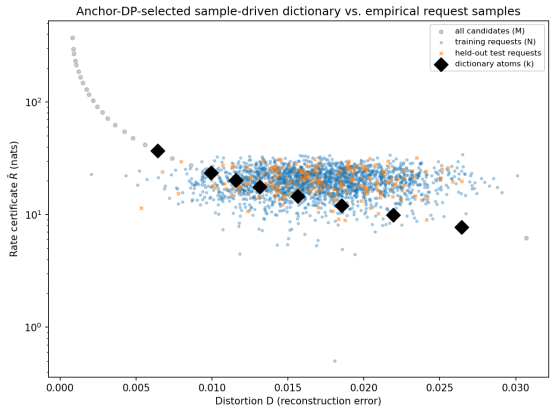
E All Experiments

E.1 Request Distribution

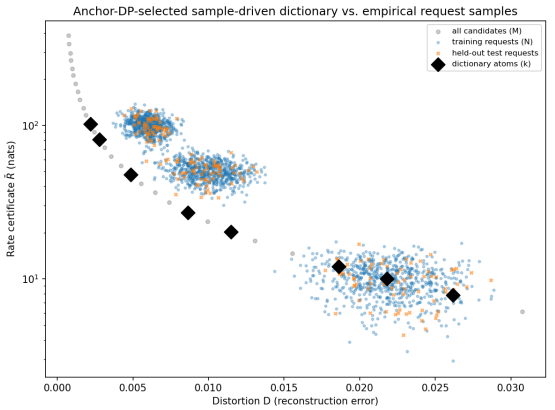
Requests are drawn from a bivariate Gaussian, truncated to a compact region,

$$(D,R) \sim \mathcal{N}\left(\begin{bmatrix}\mu_D\\ \mu_R\end{bmatrix}, \begin{bmatrix}\sigma_D^2 & \rho\sigma_D\sigma_R\\ \rho\sigma_D\sigma_R & \sigma_R^2\end{bmatrix}\right) \Big|\ (D,R)\in\mathcal{K}, \tag{46}$$

where $\rho \in (-1,1)$ is the correlation coefficient and $\mathcal{K} = \{D,R \ge 0\} \cap \{\|z(Q)-\bar{z}_{\mathcal{F}}\|_2 \le \varrho\}$ is bounded (nonnegativity plus a radius ϱ in normalized coordinates around the frontier centroid), enforced by rejection sampling. Compactness of $\mathcal{K}$ and finiteness of $\mathcal{F}$ make Assumption 1 hold exactly, with B computed as the maximum z-distance from any point of $\mathcal{K}$ to any atom. We report $(\mu_D,\mu_R,\sigma_D,\sigma_R,\rho,\varrho)$, the value of B, and the fraction of requests outside $[D_{\min},D_{\max}]$.

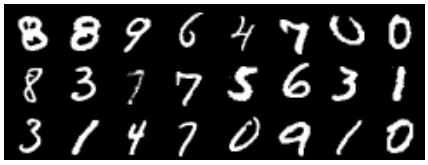


(a) Dictionaries selected by the proposed DP method for MNIST

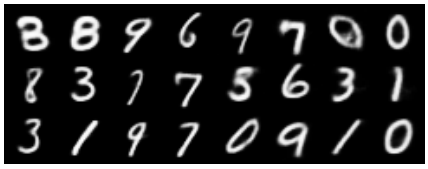


(b) Dictionaries selected by the proposed DP method for MNIST with Multi-Query

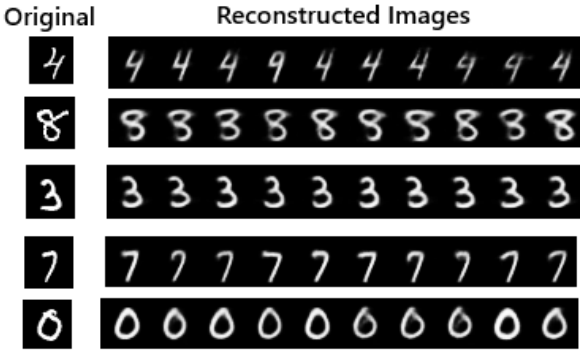
Figure 10: Dictionaries selected by the proposed DP method for MNIST with different Query Methods



(a) Original Mnist Images

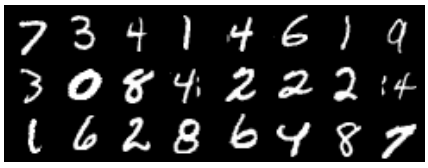


(b) Reconstructed Mnist Images

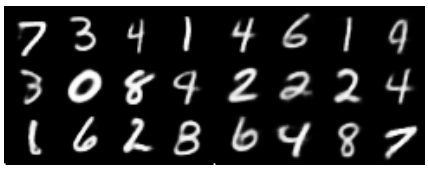


(c) Original vs. Reconstructed Mnist Images

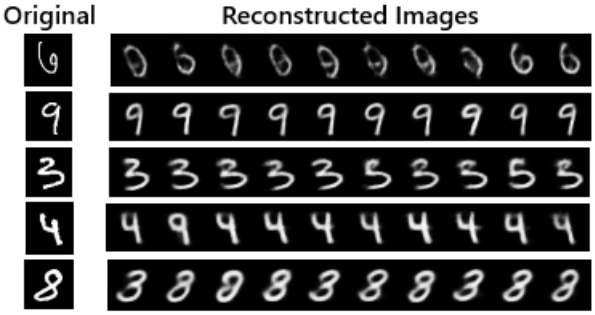
Figure 11: Reconstructed MNIST Images by dictionaries in fig. 10a



(a) Original Mnist Images



(b) Reconstructed 3-Mnist Images



(c) Original Mnist vs. Reconstructed 3-Mnist Images

Figure 12: Reconstructed MNIST Images by dictionaries in fig. 10b

Table 2: Request distribution parameters and coverage of the feasible frontier region. For mixture distributions, each component is listed on its own row.

Dataset	π	μ_D	μ_R	σ_D	σ_R	ρ	N_{train}	N_{test}
MNIST (single Gaussian)	—	0.017	20.0	0.004	5.0	0.05	2000	200
MNIST (3-mixture)	0.3333	0.022	10.0	0.0025	2.5	-0.2	2000	200
	0.3333	0.01	50.0	0.0015	7.0	-0.25		
	0.3333	0.006	100.0	0.0008	12.0	-0.25		
CIFAR10	—	0.0028	630.0	0.001	20.0	0.05	2000	200
Fashion-MNIST	—	0.007	100.0	0.002	20.0	0.05	2000	200

Table 3: Regret of the proposed DP dictionary relative to the full-frontier reference (M=30 atoms).

Dataset	k	L_{proposed}	L_{full}	Regret (%)
MNIST (single Gaussian)	4	0.09723	0.02050	374.3
	6	0.03683	0.02050	79.6
	8	0.02482	0.02050	21.1
	10	0.02093	0.02050	2.1
	15	0.02050	0.02050	0.0
	20	0.02050	0.02050	0.0
	25	0.02050	0.02050	0.0
MNIST (3-mixture)	4	0.24805	0.08858	180.0
	6	0.11861	0.08858	33.9
	8	0.10098	0.08858	14.0
	10	0.09409	0.08858	6.2
	15	0.08933	0.08858	0.9
	20	0.08858	0.08858	0.0
	25	0.08858	0.08858	0.0
CIFAR10	4	0.07334	0.07334	0.0
	6	0.07334	0.07334	0.0
	8	0.07334	0.07334	0.0
	10	0.07334	0.07334	0.0
	15	0.07334	0.07334	0.0
	20	0.07334	0.07334	0.0
	25	0.07334	0.07334	0.0
Fashion-MNIST	4	0.10612	0.08482	25.1
	6	0.09058	0.08482	6.8
	8	0.08711	0.08482	2.7
	10	0.08640	0.08482	1.9
	15	0.08482	0.08482	0.0
	20	0.08482	0.08482	0.0
	25	0.08482	0.08482	0.0

We make no claim that this reflects real traffic, and no algorithm sees a closed-form density. **Re-centering:** for each dataset we report a run with shared parameters (method ordering) and one with μ_D, σ_D re-centered on that dataset's D-range (absolute error and certificate tightness). $N_{\text{tr}} = 2000$, $N_{\text{te}} = 200$.

E.2 Formal Definitions of the Baseline Selection Rules

Every baseline below operates on the same sorted frontier $\mathcal{F} = (P_0, \ldots, P_{M-1})$ and, once constructed, is scored by the identical rule equation 11 on the identical held-out sample, construction differs across methods, evaluation does not. We give each construction rule precisely so that every number in Table 1 is traceable to a formula. Throughout, $\Delta_{ij} := d_W(P_i, P_j)$ denotes the normalized gap between two frontier candidates, computed with the *linear*-scale $\bar{R}$ term unless stated otherwise.

Table 4: Deployed dictionary summary (M=30, k=10). $L^{\text{atom}}_{\text{anchor}}$ and $L^{\text{atom}}_{\text{true}}$ are the anchor-routed surrogate and the deployed nearest-atom loss on the N_{train}=2000 training requests. Test error is the mean squared matching distance d^2_W between each of the N_{test}=200 held-out requests and its single served atom, measured in the normalized $(D, \bar{R})$ plane; it is not a pixel-space reconstruction error. Values are computed with dataset-specific normalization, so they are comparable across methods within a dataset but not across datasets.

Dataset	M	k	Reduction (%)	$L^{\text{atom}}_{\text{anchor}}$	$L^{\text{atom}}_{\text{true}}$	Test error
MNIST (single Gaussian)	30	10	66.7 (3.00×)	0.0186	0.0186	0.0209
MNIST (3-mixture)	30	10	66.7 (3.00×)	0.0975	0.0975	0.0941
CIFAR10	30	10	66.7 (3.00×)	0.0684	0.0684	0.0733
Fashion-MNIST	30	10	66.7 (3.00×)	0.0799	0.0799	0.0864

Definition 6 (Interior random-k)**.** *Draw k interior atoms uniformly without replacement from $\mathcal{F}^\circ$,*

$$\mathcal{D}_{\text{rand}} = \text{SRS}\big(\mathcal{F}^\circ,\, k\big). \tag{47}$$

*We draw $S = 30$ independent realizations $\mathcal{D}^{(1)}_{\text{rand}}, \ldots, \mathcal{D}^{(S)}_{\text{rand}}$ and report the mean and population standard deviation (*ddof $= 0$*) of $\{L_{\text{true}}(\mathcal{D}^{(s)}_{\text{rand}})\}_{s=1}^{S}$, matching the estimator used by the released implementation.*

Definition 7 (Uniform-rank)**.** *Select atoms at evenly spaced* ranks *along the D-sorted frontier,*

$$i_\ell = \text{round}\left(\frac{\ell(M-1)}{k+1}\right), \qquad \ell = 1, \ldots, k, \qquad \mathcal{D}_{\text{rank}} = \{P_{i_\ell}\}_{\ell=1}^{k}, \tag{48}$$

with round$(\cdot)$ *using round-half-to-even (banker's rounding), matching Python's built-in* `round`*. Because candidate D-spacing is not uniform (the β-grid is log-spaced,* **??***), this is uniform in* index*, not in the value D. This selects exactly k interior candidates; the two frontier endpoints correspond to the omitted boundary positions $\ell = 0$ and $\ell = k+1$; duplicate i_ℓ among $\ell = 1, \ldots, k$ are removed, so $|\mathcal{D}_{\text{rank}}| \le k$ in general, with equality whenever $M \ge k+2$, as holds throughout the reported k-sweep.*

Definition 8 (Uniform-arc-length)**.** *Let $A_0 = 0$, $A_i = \sum_{j<i} \Delta_{j,j+1}$ be the cumulative normalized arc length of the frontier polyline (linear-scale $\bar{R}$, distinct from the log-scale spacing metric used by the DP,* **??***). For k interior evenly spaced targets $t_\ell = \frac{\ell}{k+1} A_{M-1}$, $\ell = 1, \ldots, k$,*

$$i_\ell = \operatorname*{arg\,min}_{i \in \{1, \ldots, M-2\}} |A_i - t_\ell|, \qquad \mathcal{D}_{\text{arc}} = \{P_{i_\ell}\}_{\ell=1}^{k}. \tag{49}$$

Definition 9 (Farthest-point sampling)**.** *Initialize the geometric reference set $S_0 = \{0, M-1\}$ and greedily add k interior candidates, farthest in minimum normalized gap from the current selection,*

$$S_{t+1} = S_t \cup \Big\{ \operatorname*{arg\,max}_{i \in \{1, \ldots, M-2\} \setminus S_t} \min_{j \in S_t} \Delta_{ij} \Big\}, \qquad t = 0, \ldots, k-1, \tag{50}$$

with Δ_{ij} the linear-scale normalized gap. The deployed dictionary is

$$\mathcal{D}_{\text{fps}} = \{P_i : i \in S_k \setminus \{0, M-1\}\}, \tag{51}$$

so $|S_k| = k+2$ but only its k interior atoms are deployed. This targets geometric coverage of $(D, \bar{R})$-space, not L_{true} directly.

Definition 10 (Greedy-nearest-atom)**.** *Initialize $S_0 = \{0, M-1\}$. For $t = 0, \ldots, k-1$, let the currently deployed set be $\mathcal{D}_t = \{P_j : j \in S_t \setminus \{0, M-1\}\}$ and add the interior candidate that most reduces the* true *deployed loss L_{true} itself (not a surrogate),*

$$S_{t+1} = S_t \cup \Big\{ \operatorname*{arg\,min}_{i \in \{1, \ldots, M-2\} \setminus S_t} L_{\text{true}}\big(\mathcal{D}_t \cup \{P_i\}; \mathcal{Q}_{\text{tr}}\big) \Big\}. \tag{52}$$

The final deployed dictionary is $\mathcal{D}_{\text{greedy}} = \{P_i : i \in S_k \setminus \{0, M-1\}\}$, at cost $O(kM)$ evaluations of L_{true}, each $O(Nk)$, for a total of $O(k^2MN)$. By comparison, the anchor-routed DP runs in $O\big(M^2(N+k)\big)$ (Proposition 2), so Greedy is cheaper only when $k^2 < M$. Because it directly optimizes the deployed objective rather than the anchor-routed surrogate L_{anchor} (Definition 3), it is the most direct baseline for isolating surrogate slack (the price of exactness on L_{anchor} rather than L_{true}) from ordinary search suboptimality, Table 6.

Remark 2 (Shared evaluation, divergent construction)**.** *Every $\mathcal{D} \in \{\mathcal{D}_{\text{rand}}, \mathcal{D}_{\text{rank}}, \mathcal{D}_{\text{arc}}, \mathcal{D}_{\text{fps}}, \mathcal{D}_{\text{km}}, \mathcal{D}_{\text{greedy}}, \mathcal{D}^{\text{DP}}_k\}$ is scored by the identical rule equation 11; only $\mathcal{D}^{\text{DP}}_k$ (Theorem 4) and $\mathcal{D}_{\text{greedy}}$ (Definition 10) are* constructed *to target an objective related to L_{true} — the anchor-routed surrogate L_{anchor} and L_{true} itself, respectively. The remaining five ($\mathcal{D}_{\text{rand}}, \mathcal{D}_{\text{rank}}, \mathcal{D}_{\text{arc}}, \mathcal{D}_{\text{fps}}, \mathcal{D}_{\text{km}}$) are constructed by geometric or coverage criteria unrelated to either objective. We report this distinction explicitly rather than implying a uniform construction standard across all seven methods.*

Table 5: Selected dictionary atoms $(D, \bar{R}, \beta)$ for the deployed dictionary (k=10).

Dataset	Atom	D	$\bar{R}$	β
MNIST (single Gaussian)	1	0.00073	377.3449	1.000×10^{-6}
	2	0.00642	36.7587	1.897×10^{-4}
	3	0.00992	23.5292	4.343×10^{-4}
	4	0.01157	20.1410	5.724×10^{-4}
	5	0.01314	17.6718	7.544×10^{-4}
	6	0.01563	14.5895	9.943×10^{-4}
	7	0.01854	12.0655	1.310×10^{-3}
	8	0.02194	9.9507	1.727×10^{-3}
	9	0.02644	7.7584	2.276×10^{-3}
	10	0.03066	6.1810	3.000×10^{-3}
MNIST (3-mixture)	1	0.00070	378.9683	1.000×10^{-6}
	2	0.00243	91.4445	2.747×10^{-5}
	3	0.00424	54.5538	8.287×10^{-5}
	4	0.00735	31.6833	2.500×10^{-4}
	5	0.00989	23.5743	4.343×10^{-4}
	6	0.01141	20.2554	5.724×10^{-4}
	7	0.01842	12.1311	1.310×10^{-3}
	8	0.02168	10.0129	1.727×10^{-3}
	9	0.02649	7.7866	2.276×10^{-3}
	10	0.03106	6.1097	3.000×10^{-3}
CIFAR10	1	0.00250	590.3171	1.000×10^{-6}
	2	0.00259	551.9289	1.318×10^{-6}
	3	0.00262	511.4361	1.737×10^{-6}
	4	0.00265	463.3455	2.289×10^{-6}
	5	0.00274	422.4189	3.017×10^{-6}
	6	0.00288	378.2544	3.977×10^{-6}
	7	0.00300	342.1454	5.241×10^{-6}
	8	0.00323	302.1741	6.907×10^{-6}
	9	0.00350	265.2086	9.103×10^{-6}
	10	0.03477	2.9677	3.000×10^{-3}
Fashion-MNIST	1	0.00179	558.5290	1.000×10^{-6}
	2	0.00445	117.1783	2.084×10^{-5}
	3	0.00490	100.1639	2.747×10^{-5}
	4	0.00559	81.0500	3.620×10^{-5}
	5	0.00631	65.3533	4.771×10^{-5}
	6	0.00766	45.6673	8.287×10^{-5}
	7	0.00911	31.8573	1.440×10^{-4}
	8	0.01087	22.5834	2.500×10^{-4}
	9	0.01454	13.3784	5.724×10^{-4}
	10	0.02575	4.7628	3.000×10^{-3}

Table 6: Time complexity of the selection methods (M candidates, N requests, deployed dictionary size k; the two frontier endpoints are DP boundary guards and are not counted toward k).

Method	Time	Guarantee
Greedy-nearest-atom	$O(k^2MN)$	none (heuristic)
Proposed DP	$O\big(M^2(N+k)\big)$	exact for L_{anchor} over $\mathcal{H}_k$
Exhaustive search	$O\big(\binom{M-2}{k}Nk\big)$	exact for L_{true}