COFAT: CORRELATION-AWARE QUERY DESIGN FOR CONTROLLABLE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generative models should not only produce realistic samples but also reliably follow the requests they are given. Evaluating this requires checking four things: whether the requested attribute combination is *feasible*, whether the generated sample is *obedient* to it, and whether it remains *realistic* and *novel*. Existing controllability evaluations score obedience, realism or novelty largely in isolation, and take feasibility for granted. The dominant query design, One-Factor-At-a-Time with non-swept attributes fixed at their marginal means (OFAT-μ), ignores dependence between attributes. On correlated attribute spaces it therefore issues requests that rarely or never occur in the data. Generators may then ignore the requested change, undermining the purpose of conditional generation. Our approach offers a new perspective by treating query design as part of controllability evaluation. We introduce a closed-form measure of query infeasibility and COFAT, a family of correlation-aware query designs that set non-swept attributes from their conditional rather than marginal distribution, within a framework that scores feasibility, obedience, realism and novelty jointly. For a synthetic generator that is perfectly obedient on feasible requests, switching from OFAT-μ to COFAT raises measured obedience from 0.15 to 0.99: the query, not the generator, can be the bottleneck. Across three tasks (lesion inpainting, lesion insertion, molecule generation) and against seven baseline designs, COFAT raises obedience (e.g., .07 $\rightarrow$ .66 for lesion intensity) and molecule validity ($0\% \rightarrow 67\%$), reduces copying of training samples, and keeps realism comparable. As the target range of the swept attribute is unchanged, these gains come from removing infeasible combinations, not from easing the target. Many reported controllability gaps may therefore be artifacts of query construction, which argues for treating query design as a first-class component of controllability evaluation.

1 INTRODUCTION

Conditional generative models are increasingly used to produce samples with specified properties: synthetic lesions of a given size, shape or texture for data augmentation and reader studies [56, 37, 29], or molecules with target physicochemical descriptors [6]. Such a model is only useful if it is *controllable*. When a user asks for a smaller, smoother, or brighter lesion, the generated sample should reflect the requested change, appear realistic, and remain novel without copying training data. The standard way to measure controllability is one-factor-at-a-time (OFAT): sweep one target attribute across a range of requested values, hold its companions (all other non-swept attributes) fixed, and check whether the generated samples follow the sweep [24, 13, 20]. A common instantiation fixes the companions at their training marginal mean [24] (OFAT-μ). This protocol is sensible when attributes are independent, but can fail if they are correlated [43, 13]. Such correlations are common in real-world data: lesion brightness and contrast tend to increase together; heavy-atom count, molecular weight, and molar refractivity are strongly correlated and often nearly redundant. Asking for a brighter lesion with average contrast, or a larger molecule with average weight, therefore request combinations that rarely or never occur in the training data [33].

This exposes two blind spots in current controllability evaluations. First, query feasibility has not been treated as a first-class, scored component of controllability evaluation alongside obedience and realism. Instead, feasibility been treated variously as an assumption [2], a limitation [47], a training objective [28], or a post-hoc, per-sample trust score that diagnoses outputs rather than the requests

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

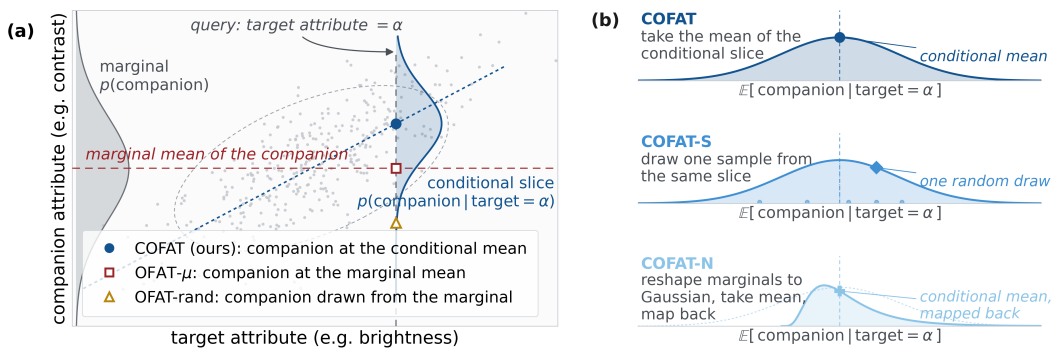

Figure 1: **COFAT method overview.** A query to a conditional generative model selects one target attribute (e.g., brightness $= \alpha$, vertical dashed line) but must also specify the remaining companion attributes (e.g., contrast). **(a)** OFAT-μ and OFAT-rand ignore the correlation and select unlikely attribute combinations, whereas COFAT selects the conditional mean. **(b)** COFAT variants operate on the same conditional slice: COFAT uses its Gaussian mean, COFAT-S samples from it, and COFAT-N accounts for non-Gaussian marginals before taking the mean.

that produced them [16]. Second, with the partial exception of [16], existing evaluations largely isolate obedience, realism and novelty rather than assessing them jointly. Recent studies quantify steerability [11] or LLM prompt sensitivity [19], measuring concordance between generated outputs and a target goal vector without verifying whether individual outputs are realistic. Dominant metrics such as Inception Score [42] and Fréchet Inception Distance [25] measure similarity in image embedding space rather than attribute space, without providing an obedience assessment. While other works assess novelty [4, 9, 40], they primarily focus on non-controllable generation settings. Consequently, neither line of evaluation answers the central question: ***did the model produce a realistically novel sample that satisfies a feasible request?***

We argue that a controllability evaluation should pose *feasible* requests, i.e., attribute combinations supported by the training distribution, and score each response jointly for *obedience* (does it satisfy the request?), *realism* (does it lie within the neighborhood of the real-data distribution?) and *novelty* (does it avoid copying training samples?). Building on this idea, our contribution is threefold. **A controllability framework.** For each attribute, we work from its empirical distribution in the training data. A percentile sweep fixes a set of L percentile levels of that attribute's training distribution and, at each level, queries the model for a response whose realized attribute value is intended to land at that percentile. We score every query–response pair on four properties: query feasibility, a closed-form measure of the queried percentile's distance from the training distribution; and response obedience, realism, and novelty, each computed from the same percentile-sweep protocol (Figure 2). **Correlation-aware queries.** COFAT sets companion attributes to their conditional mean given the swept-target; COFAT-S samples from this conditional to preserve diversity; COFAT-N extends both to non-Gaussian marginals (Figure 1). **Evidence that query design drives measured controllability.** With a synthetic generator of known obedience, changing only the query moves measured obedience from 0.15 to 0.99. Across three tasks and against seven baseline designs, COFAT raises obedience , improves novelty of training samples, and keeps realism comparable.

2 Related Work

Query Design for Controllability. One-Factor-at-a-Time (OFAT) evaluation varies a target attribute across a range of values while holding companion non-target attributes fixed, and checks that generated responses follow the sweep [24]. Despite well-known limitations under correlated inputs [43, 13], it has become a common evaluation strategy in conditional generative modeling [45, 32, 29], applied to latent semantic directions in image synthesis [45] and property-conditioned molecular generation [32] alike. As the target attribute is set, query design is determined by how it chooses the companion attributes. These designs fall into two groups. *Independent* designs (mean, median, random draws, factor fixing) set companions without regard to the target;

under correlation they produce queries outside the multivariate training envelope [43, 12]. This failure is also seen in deep generative models, where correlated factors cannot be manipulated independently [51]. *Data-anchored* designs (clustering [56], nearest-neighbour retrieval [8], or real training subjects [29]) are feasible by construction, but can only request already existing attribute combinations, restricting swept-able target values and inviting copying. COFAT combines the strengths of both: it conditions the companions on the target, but in closed form, so any target value can be requested. Conditioning on the target to avoid extrapolation is established in model explanation, for example in conditional alternatives to partial-dependence plots [5] and dependence-aware Shapley values [1]. We bring it to controllability evaluation and, in addition, score query feasibility explicitly.

Obedience. Prior work measures response to changed conditions via steerability [11], prompt sensitivity [19], or Pearson/Spearman correlation [29]. These metrics detect whether outputs change, but not whether they match the request. Mean Absolute Error (MAE), used in property-conditioned generation [27, 54], measures mismatch directly but collapses distinct failures into a single number. For requested values $c = [1, 3, 5]$, a generator that delivers half the requested change, $[0.5, 1.5, 2.5]$, and one with a constant offset, $[-0.5, 1.5, 3.5]$, both have MAE $= 1.5$ and a correlation of 1. In contrast, we propose to decompose obedience into slope, bias, and the concordance correlation coefficient (CCC) [34], separating such failures (Section 3.3).

Realism. Inception Score (IS) [42], Fréchet Inception Distance (FID) [25], and its conditional extension [17] summarize a whole sample set in one scalar and cannot be paired with per-sample obedience. Per-sample alternatives exist in image-feature space [30, 38], and Kim et al. [29] assess the image-level realism of individual lesions. All of these measure appearance; none checks whether a sample's attributes form a combination observed in real data. The Mahalanobis distance is a standard per-sample out-of-distribution score [31], and Demirel et al. [16] use it to assess the trustworthiness of generated samples. We apply it in attribute space, to both queries and responses as a measure of query feasibility and attribute-space realism.

Novelty. A realistic sample may simply be a copy: diffusion models memorize training data [10, 48, 49], including medical settings [3, 14, 15]. Molecular benchmarks such as GuacaMol [9] and MOSES [40] report novelty as the fraction of generated molecules absent from the training set, and [4] measure per-sample authenticity via nearest-neighbour distances. These works evaluate largely unconditional generation and treat copying as a property of the generator. We measure novelty per sample under conditional queries and show that copying also depends on the query design.

3 METHOD

A controllability evaluation provides a conditional generator g with an input $\mathbf{x}$ and a query $\mathbf{c} \in \mathbb{R}^d$ specifying the requested attribute values, and obtains a response $\mathbf{y} \sim g(\cdot \mid \mathbf{x}, \mathbf{c})$. We re-extract the same d attributes from $\mathbf{y}$, obtaining the realized attributes $\mathbf{a} \in \mathbb{R}^d$, which we compare directly with the requested attributes $\mathbf{c}$. We score the query for *feasibility* and the response for *obedience*, *realism*, and *novelty*. Section 3.1 shows that feasibility is decided by the query design and why OFAT-μ fails under correlated attributes, Section 3.2 introduces COFAT and its variants, and Section 3.3 defines the metrics.

3.1 QUERY DESIGN AND FEASIBILITY

Queries. $\mathbf{c} \in \mathbb{R}^d$ is the attribute vector of a *query*. A sweep varies one *target* attribute, indexed by s, at requested value $c_s = \alpha$, while the query design chooses the remaining *companions* $\mathbf{c}_{-s} \in \mathbb{R}^{d-1}$. All attributes are z-scored on the training set; $\mathbf{R} \succ 0$ is their correlation matrix. Ordering s first, $\mathbf{R} = \left[\begin{smallmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & \mathbf{Q} \end{smallmatrix}\right]$, with 1 as the self-correlation of c_s, vector $\mathbf{b} \in \mathbb{R}^{d-1}$ holds the correlations of c_s with its companions, and the matrix $\mathbf{Q} \in \mathbb{R}^{(d-1)\times(d-1)}$ the correlations among the companions. Analogous to [18], the *redundancy* of c_s is its variance explained by companion attributes:

$$R^2_{c_s} = 1 - \frac{\mathrm{Var}(c_s \mid \mathbf{c}_{-s})}{\mathrm{Var}(c_s)} = 1 - \frac{1}{[\mathbf{R}^{-1}]_{c_s c_s}}. \tag{1}$$

Infeasibility. We quantify a query's distance relative to the training distribution by its Mahalanobis radius, which we refer to as its *infeasibility*:

$$\rho(\mathbf{c}) = \sqrt{\mathbf{c}^\top \mathbf{R}^{-1} \mathbf{c}} \tag{2}$$

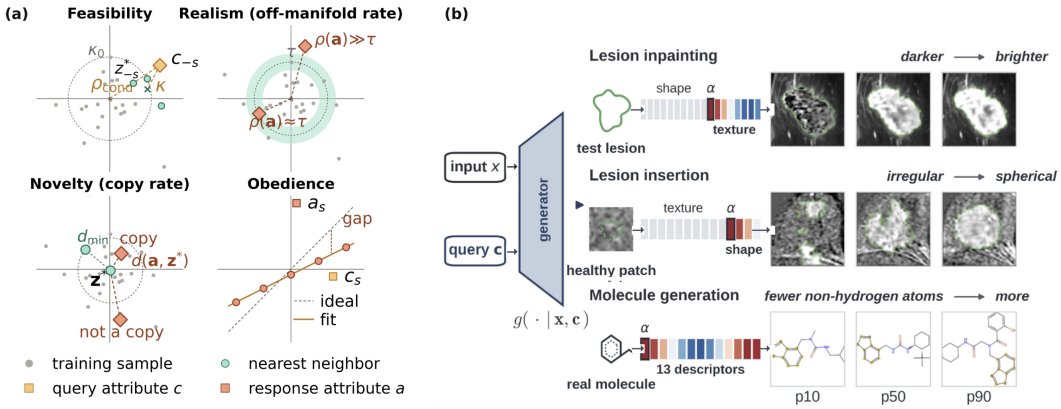

Figure 2: **Overview of controllability framework and experiments.** (a) Controllability framework: We evaluate query feasibility and response obedience, realism and novelty. (b) Experiments: we fix lesion shape and sweep texture (image *lesion inpainting*), fix texture and sweep shape (image *lesion insertion*), and sweep 13 molecule descriptors (*molecule generation*).

With the target fixed at $c_s = \alpha$, designs differ only in their companions $\mathbf{c}_{-s}$. We therefore separate the fixed target contribution from the additional infeasibility induced by the chosen companions.

Lemma 1 (Exact split). *Let $\mathbf{\Sigma} = \mathbf{Q} - \mathbf{b}\mathbf{b}^\top$. For any query with $c_s = \alpha$, squared radius splits as*

$$\rho^2(\mathbf{c}) = \underbrace{\alpha^2}_{\text{fixed request}} + \underbrace{(\mathbf{c}_{-s} - \alpha\mathbf{b})^\top \mathbf{\Sigma}^{-1} (\mathbf{c}_{-s} - \alpha\mathbf{b})}_{\rho^2_{\text{cond}}(\mathbf{c})\text{: added by the query design}}. \tag{3}$$

We define the *conditional infeasibility* of a query as $\rho_{\text{cond}}(\mathbf{c}) = \sqrt{\rho^2_{\text{cond}}(\mathbf{c})}$, the radius of how far its companions $\mathbf{c}_{-s}$ lie from the training distribution. See the proof in Appendix A.2.1.

OFAT-μ queries are infeasible under correlated attributes. Standard controllability evaluation (OFAT-μ) freezes companion attributes at their marginal means. After standardization, these means are zero, yielding the query vector $\mathbf{c}^{\text{OFAT}}$. Substituting $\mathbf{c}_{-s} = \mathbf{0}$ (all companions set to zero) into Eq. (3) gives the query and infeasibility in closed form:

$$\mathbf{c}^{\text{OFAT}} = (\alpha, \mathbf{0}) = \alpha \mathbf{e}_s, \qquad \rho_{\text{cond}}(\mathbf{c}^{\text{OFAT}}) = |\alpha| \frac{R_s}{\sqrt{1 - R_s^2}}. \tag{4}$$

where $\mathbf{e}_s$ is the unit vector for attribute s. This shows that OFAT-μ queries become infeasible when attributes are correlated ($R_s^2 \to 1$). See the proof in Appendix A.2.2.

3.2 COFAT: A CORRELATION-AWARE QUERY FAMILY

COFAT. Instead of the marginal mean, COFAT sets each companion to its conditional mean (Figure 1).

Proposition 1 (Least infeasible query). *For fixed $c_s = \alpha$, ρ_{cond} is uniquely minimized by*

$$\mathbf{c}^{\text{COFAT}} = \alpha\, \mathbf{R}_{\cdot s} = (\alpha,\ \alpha\mathbf{b}), \qquad \rho_{\text{cond}} = 0, \qquad \rho = |\alpha|. \tag{5}$$

The second term of Eq. (3) is a positive-definite quadratic form, which is zero if and only if $\mathbf{c}_{-s} = \alpha\mathbf{b}$. Under linear conditional expectations (e.g. Gaussian attributes), $\alpha\mathbf{b} = \mathbb{E}[\mathbf{c}_{-s} \mid c_s = \alpha]$, so the companions take the values they typically have when attribute s equals α. When s is independent of its companions, $\mathbf{R}_{\cdot s} = \mathbf{e}_s$ (the basis vector) thereby exactly reducing COFAT to OFAT.

COFAT-S (Stochastic). COFAT always asks for the most typical companions. COFAT-S instead samples them from the full conditional distribution $\mathbf{c}_{-s} = \alpha\mathbf{b} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

COFAT-N (Non-Gaussian Attributes). COFAT's guarantee (Prop. 1) holds for any distribution with respect to ρ (proof in Appendix A.2.3); its conditional-mean interpretation, however, requires linear conditional expectations, which follows from joint Gaussianity, not separate Gaussian marginals alone. Skewed or bounded attributes violate this. COFAT-N ($\rho_{\text{cond}} > 0$) relaxes it via a Gaussian copula (proof in Appendix A.2.4).

3.3 CONTROLLABILITY FRAMEWORK

Controllability is evaluated along four axes: *feasibility* depends on the queries alone and evaluates whether the requested configuration is compatible with the data distribution; *obedience* compares requested and realized attributes ($\mathbf{c}$ vs. $\mathbf{a}$); *realism* evaluates whether the sample remains on-manifold; and *novelty* assesses whether training samples are copied (Figure 2). For each test case, every target attribute s is swept over L training percentiles ($p_{10}, \ldots, p_{90}$), and each sweep is scored.

Feasibility metrics. We report *Infeasibility* $\bar{\rho}$ (Eq. 2) and *Conditional Infeasibility* $\bar{\rho}_c$, averaged over all sweep levels for query c. Although it measures deviation from the training distribution, it is circular for COFAT which minimizes ρ_{cond} by construction (Eq. 5).We therefore additionally report *conditional anomaly rate*, an independent check of empirical support. Let $\mathbf{z}$ denote the attributes of a real training sample. For each sweep level l, let κ be the average normalized distance between the query's companions c_{-s} and those of its k nearest training vectors z^*_{-s}:

$$\kappa^{(l)} = \frac{\|\mathbf{c}_{-s} - \mathbf{z}^*_{-s}\|}{\sigma}, \qquad \text{An.}\% = \frac{100}{L}\sum_{l=1}^{L} \mathbf{1}\Big[\kappa^{(l)} > \kappa_0\Big], \tag{6}$$

σ being the typical spread of the real training data and κ_0 a threshold. An.% (conditional anomaly rate) is the percentage of sweep levels whose companions have no nearby real counterpart (ideal 0).

Obedience metrics. Obedience is computed separately for each target attribute s, averaged across all levels, over all pairs $(\mathbf{c}_s, \mathbf{a}_s)$:

$$\text{slope}_\text{s} = \frac{\text{Cov}(c_s, a_s)}{\text{Var}(c_s)}, \quad \text{bias}_\text{s} = \bar{a}_s - \text{slope}\bar{c}_s, \quad \text{CCC}_\text{s} = \frac{2\,\text{Cov}(c_s, a_s)}{\text{Var}(c_s) + \text{Var}(a_s) + (\bar{c}_s - \bar{a}_s)^2}. \tag{7}$$

with bars denoting the mean across percentile levels. *Slope* is the fraction of the requested change that is delivered (ideal 1), *bias* captures constant offset (ideal 0), and *CCC* [34] the overall agreement in correlation, scale, and location (ideal 1). For the example in Section 2: slope=$0.5/1$, bias=$0/-1.5$, and CCC=$0.55/0.78$, separating the scale from the offset failure.

Companion drift is the mean absolute error on the companions: $\text{drift} = \frac{1}{(d-1)L}\sum_l |a^{(l)}_{-s} - c^{(l)}_{-s}|$ (ideal 0). It flags designs that deliver the target response by erroneously disturbing its companions.

Realism metrics. *Off-manifold rate.* A response $\mathbf{a}$ is off-manifold in attribute space if its Mahalanobis radius $\rho(\mathbf{a})$ (Eq. 2) deviates by more than δ from τ, a fixed high-percentile of the training radii:

$$\text{off-M.}\% = \frac{100}{N}\sum_{i=1}^{N} \mathbf{1}\Big[\,\big|\rho(\mathbf{a}^{(i)}) - \tau\big| > \delta\Big]. \tag{8}$$

Radii above $\tau + \delta$ indicate rare or implausible samples; radii below $\tau - \delta$, overly typical ones. For a 2-sided metric to remain valid at high-percentile, we choose sufficiently larger delta to reflect substantial deviations rather than small fluctuations.

Appearance realism (image space). Following [30], we embed each response image $f(y)$ and count it as realistic if it falls inside the union of spheres around real training embeddings, each reaching its k-th nearest real neighbour. P_{app} is the fraction of realistic responses (higher is better).

Validity (molecules). The percentage of generated molecules that are chemically well-formed; a molecule failing this check cannot be realistic. We report it at the extreme sweep levels, where designs are most likely to break. The validity of N generated samples y is defined as $\text{valid}\% = \frac{100}{N}\sum_{i=1}^{N} \mathbf{1}\big[y^i \in \mathcal{V}\big]$, with $\mathcal{V}$ the set of samples that satisfy a domain-specific check.

Novelty metric. We measure the *copy rate* [4] in attribute space. For a response attribute vector a, let $\mathbf{z}^*$ be its nearest training sample and $d(.,.)$ the distance between samples. The response is a likely copy if it lies closer to $\mathbf{z}^*$ than any other training sample does. $\mathrm{copy}\%$ is the fraction of such samples. The reference is the rate of held-out real data; rates above it indicate copying.

4 EXPERIMENTAL SETUP

Query designs. Each target attribute s is swept over L training percentiles, and a query design sets the companions $\mathbf{c}_{-s}$; all designs request $c_s = \alpha$ unless stated otherwise. We compare COFAT, COFAT-S, and COFAT-N (Sec. 3.2) against seven baseline designs in three groups (Details in Appendix A.3. *Independent designs* fix the companions at their marginal means (OFAT-μ [24]) or medians (OFAT-MED [46]), or draw them independently from $\mathcal{N}(0,1)$ (OFAT-RAND [36]). *Data-anchored designs* copy the k-means centroid (k=3) nearest to α (K-MEANS [56]) or average the five training samples nearest to α (KNN [8]), re-selected at every level. *Factor-fixing and negative-control designs* freeze the other evaluated attributes at 0 and set the remaining companions to their conditional mean (FREEZE [44]), or set the companions to their conditional mean $\alpha\mathbf{b}$ while muting the target, c_s=0 (DROP-c_s [35]); the latter tests specificity i.e, whether obedience stems from the target itself or from its companions.

Metric settings. For the anomaly rate we use k=50 neighbours and set κ_0 to the 95th percentile of κ over real training samples. The off-manifold rate uses τ=95, so real data scores 5%; rates within $[2.5, 7.5]\%$ pass the realism gate. P_{app} uses DINOv2 ViT-B/14 [39] embeddings at 224×224 with k=3; real test lesions score 0.68. The copy rate uses Euclidean distance on the 34 radiomic features for lesion textures (real-data reference 28.4%), 4 shape descriptors for lesion shapes (real-data reference 0%) and one minus the Tanimoto similarity of Morgan fingerprints (radius 2, 2048 bits) [41] for molecules (reference 46.7%).

Experiment A: appearance via lesion inpainting. *Can the model change a lesion's texture while keeping its shape?* MAMA-MIA [21] provides DCE-MRI breast cancer cases from four sites with patient-wise train/test splits ($1{,}080/300$), from which we extract 64×64×32 lesion patches, masks, and 34 radiomic features. A rectified-flow model based on the MAISI 3D backbone [23] is trained on the training split, conditioned via FiLM on all [d=34] features. At test time, each test mask (i.e. shape) is kept fixed while its interior is repainted, sweeping `fo_mean` over L=5 levels ($p_{10}, p_{25}, p_{50}, p_{75}, p_{90}$) with guidance γ=8 (300 lesions × 5 levels). Since `fo_mean` is nearly redundant with the other features (R^2_s=0.9998), this is the hardest setting for OFAT. Results for `glcm_contrast` and `glrlm_SRE` are in Appendix Table 7 and Table 8 respectively.

Experiment B: shape via lesion insertion. *Can the model alter a lesion's geometry while preserving its texture features?* A signed-distance-field model with the same conditioning is trained on MAMA-MIA with four shape features (d=4: `sphericity`, `elongation`, `flatness`, `irregularity`; R^2_s=0.55–0.88). At test time, we fix lesion volume, centroid and texture features, sweep each shape feature over L=5 levels with γ=3, and insert the lesions into lesion-free 64×64×32 patches from 100 healthy patients of an internal breast MRI cohort (100 hosts × 5 levels). We additionally report companion drift. Here 2, we report results for `sphericity`. Results for `elongation`,`flatness`,`irregularity` in Appendix Tables 12, 13 and 14 respectively.

Experiment C: molecular generation. *Can the model control physicochemical descriptors while producing valid, novel molecules?* We use ZINC-250k [22] (train/val $94{,}513/5{,}000$), representing each molecule by its SMILES [53] and 13 z-scored RDKit descriptors (d=13). A causal transformer [52] generates SMILES autoregressively, conditioned via FiLM on the descriptors. In the *free sweep*, `HeavyAtoms` is swept over L=5 levels with γ=1 and 500 molecules per level. In the *scaffold-anchored sweep*, the Bemis–Murcko scaffold [7] of a held-out molecule is fixed as prefix while `HeavyAtoms`, `BertzCT`, and `MolMR` are swept. Levels with fewer than 10 valid molecules are excluded from the obedience fit. *Free sweep* results reported on most redundant feature `HeavyAtoms` 3, while for BertzCT and MolMR in Appendix Tables 17 and 18 respectively. *Scaffoled sweep* in Appendix Tables 19.

Experiment D: synthetic sanity check. *Is low measured obedience a property of the generator or of the query?* We simulate a generator that obeys every request, but only along directions supported by the data: For a generator with perfect obedience ($g = I$) and realism strength η=0.05, this returns

Table 1: Quantitative results for appearance target `fo_mean` (Exp. A). **Bold** = best in column.

	Feasibility			Obedience			Realism		Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\to 1$	bias$\to 0$	CCC$\uparrow$	off-M.%$\to 5$	$P_{\text{app}} \to .68$	copy%$\downarrow$
OFAT-RAND[36]	190.1	190.1	80	-.66	+1.37	-.12	*63.5*	.28	7.0
OFAT-MED[46]	61.3	61.3	20	+.18	-.35	.27	3.3	**.68**	33.9
OFAT-μ[24]	62.8	62.8	20	+.09	+.62	.07	3.8	.63	33.1
DROP-c_s[35]	62.8	62.8	40	+.57	+.06	.52	4.1	.63	28.5
FREEZE[44]	**0.8**	0.10	**0**	+.81	-.07	.62	**5.6**	.56	27.8
K-MEANS[56]	32.2	32.2	**0**	+.53	+.11	.52	3.7	.65	29.9
KNN[8]	1.5	1.3	**0**	+.91	-.26	**.76**	4.0	.57	27.2
COFAT (ours)	**0.8**	**0.0**	**0**	+.94	**-.03**	.66	5.9	.57	26.9
COFAT-N (ours)	4.4	4.3	**0**	**+.99**	-.53	.74	7.2	.56	27.2
COFAT-S (ours)	5.2	5.1	80	-.39	+1.93	-.05	*74.7*	.28	**5.7**

$\mathbf{a} = \mathbf{M}\mathbf{c}$ with $\mathbf{M} = \mathbf{R}(\mathbf{R} + \eta\mathbf{I})^{-1}$, which reproduces requests along high-variance directions of $\mathbf{R}$ and shrinks them along unsupported ones. We build $\mathbf{R}(\omega)$ for d=20 attributes from a fixed loading matrix with three latent factors, interpolating from independent (ω=0) to nearly redundant attributes ($\omega \to 1$), and compare the slope under OFAT-μ and COFAT, averaged over all targets. Details in Appendix A.1.5.

5 Results and Discussion

A: Appearance. *Feasibility.* Because *fo_mean* is nearly redundant, OFAT-μ requests are $79\times$ farther from the data than COFAT's ($\bar{\rho}$=62.8 vs. 0.8; Table 1), and OFAT-RAND farther still (190.1). The model-free anomaly check agrees: COFAT's companions have a real counterpart at every level (An. 0%), whereas OFAT-μ and OFAT-MED lack one at 20% of levels and OFAT-RAND at 80%. COFAT-S is instructive: its draws are typical under the Gaussian model ($\bar{\rho}_c$=5.1, the 52nd percentile of real lesions), yet 80% lack empirical support, so sampling the Gaussian conditional is unreliable even where its mean is not. *Obedience and realism.* The gap carries over to generation. OFAT designs barely move or even reverse the target (slopes 0.09, 0.18, -0.66), whereas COFAT and COFAT-N track it (slopes 0.94/0.99, CCC 0.66/0.74) within the realism gate (off-M. 5.9/7.2%; Figure 3a). DROP-c_s keeps COFAT's companions but mutes the target and still reaches slope 0.57, so up to 0.37 of COFAT's slope stems from the target itself; this is an upper estimate, as the muted query is itself infeasible. The data-anchored KNN comes close (slope 0.91, highest CCC 0.76) but with a larger bias (-0.26 vs. -0.03). Appearance realism is lower for all designs that substantially change the target (P_{app}=0.56–0.57 vs. 0.63–0.68 for OFAT), consistent with extreme-intensity lesions being rarer. *Why joint evaluation matters.* Each metric alone would mislead. OFAT-μ and OFAT-MED pass the realism gate despite infeasible queries because they hardly change the target, and OFAT-MED even reaches the highest appearance realism (0.68) while copying more than real lesions (33.9% vs. 28.4%). Conversely, OFAT-RAND and COFAT-S copy least (7.0/5.7%) because they leave the manifold (off-M. 63.5/74.7%), and they fail obedience. Only COFAT, KNN, and FREEZE pass all criteria; COFAT does so with the most feasible queries and the lowest bias, without a nearest-neighbour search.

B: Shape. Shape features are less redundant (R_s^2=0.55–0.88), so the feasibility gap narrows ($\bar{\rho}$=2.47 for OFAT-μ vs. 0.87 for COFAT; Table 2). COFAT obeys best (CCC 0.90), ahead of KNN/K-MEANS (0.87/0.86) and OFAT-μ (0.58). However, on the skewed sphericity feature, its companions leave the real support (An. 100%), whereas COFAT-N stays within it (0%) at similar obedience (0.86) and lowest drift. As in Exp. A, OFAT-RAND and COFAT-S leave the manifold (off-M. 16.2/15.2%); all other designs score 0%, a generator property common to every query. Among them, COFAT copies least (45.4% vs. 47.0–52.8%). Figure 3b shows progressively rounder lesions under COFAT; other shape features are in Apppedix Tables 12, 13, 14

C: Molecules. *Free sweep.* Infeasible queries break generation outright: OFAT-μ, OFAT-MED, and DROP-c_s lack real support at 40% of levels, exactly the extremes where they produce no valid molecules, and OFAT-RAND fails at every level (Table 3). Their seemingly high CCCs (0.85–0.86) are thus fitted on the three intermediate levels only, where they still overshoot (slopes 1.57/1.74). COFAT tracks the target across all five levels (slope 0.96, CCC 0.99) with 82/67% validity at p_{10}/p_{90}, compared with 0% for OFAT-μ; COFAT-N, KNN, and K-MEANS perform similarly. FREEZE also lacks support at the extremes and nearly stops producing valid molecules there (5/4%), and

Table 2: Quantitative results for shape target `sphericity` (Exp. B). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\uparrow$	bias$\downarrow$	CCC$\uparrow$	drift$\downarrow$	off-M.%	copy%$\downarrow$
OFAT-RAND[36]	3.81	3.68	100	.75	.56	.61	0.69	16.2	42.2
OFAT-MED[46]	2.53	2.36	80	.39	.45	.52	**0.45**	0.0	52.8
OFAT-μ[24]	2.47	2.31	100	.38	**.21**	.58	0.50	0.0	50.8
DROP-c_s[35]	2.31	2.31	100	.59	.28	.77	0.54	0.0	50.4
K-MEANS[56]	1.62	1.31	100	.79	.30	.86	0.48	0.0	47.0
KNN[8]	1.15	0.68	**0**	.80	.32	.87	0.49	0.0	50.8
COFAT (ours)	**0.87**	**0.00**	100	.86	.29	**.90**	0.52	0.0	45.4
COFAT-N (ours)	1.15	0.67	**0**	.79	.33	.86	**0.45**	0.0	49.6
COFAT-S (ours)	1.79	*1.50*	100	**1.09**	.51	.81	1.04	15.2	**34.4**

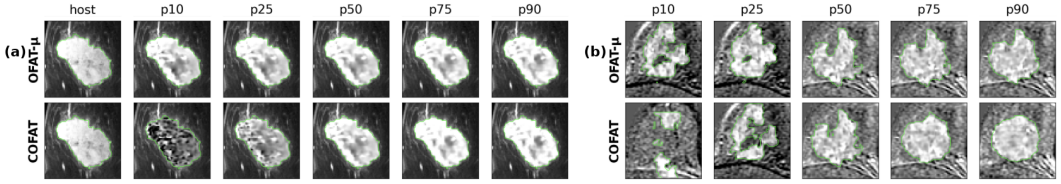

Figure 3: Qualitative results. (a) Appearance control: 3D-lesion inpainting within a fixed test mask, sweeping *fo_mean*. (b) Shape control: 3D-lesion insertion into a fixed healthy breast MRI patch, sweeping *sphericity*.

COFAT-S again trades realism ($39/48\%$) for lower copying (8.0%). All copy rates lie far below the real-data reference (46.7%). ***Scaffold-anchored sweep.*** With a fixed scaffold, COFAT and OFAT-μ reach comparable obedience, but COFAT yields far more valid molecules (29–79% vs. 8–32%; Figure 4).

D: Synthetic. With independent attributes, both designs measure the same slope (0.952, the generator's shrinkage $1/(1+\eta)$). As redundancy grows, the OFAT slope collapses to 0.149, while COFAT reaches 0.993. The generator never changes, only the query, so the apparent loss of obedience is an artifact of evaluating with increasingly infeasible queries.

Limitations. COFAT's companions equal the conditional mean only under linear conditional expectations; for skewed attributes they can leave the real support at extreme levels, as the anomaly check detects in Experiment B. COFAT-N handles these cases via a Gaussian copula, but neither captures multimodal dependence, and sampling the conditional (COFAT-S) proved unreliable. Moreover, OFAT and COFAT ask different questions: OFAT is interventional ("change s, hold the rest"), COFAT observational ("change s as the data does"). A model trained on observational data can only be expected to answer the latter; DROP-c_s bounds how much of its answer comes from the companions. Finally, we evaluate one model per task with a fixed guidance scale, coarse percentile sweeps, and fixed realism gates.

Conclusion. Controllability evaluations score whether a response is realistic and, sometimes, obedient, but not whether the request could be met at all. We make query feasibility a measured part of controllability, alongside obedience, realism, and novelty, and propose COFAT, which sets the unswept attributes to their conditional mean given the request and provably minimizes query infeasibility under the Mahalanobis measure. Across lesion inpainting, lesion insertion, and molecule generation, changing only the query turns weak or broken control into obedient, valid, and non-copied responses from the same model, and a synthetic generator isolates the same effect. Some reported controllability gaps are therefore artifacts of the query that no retraining will close. We recommend reporting every controllability result together with the feasibility of the queries behind it, so that failures of the model can be told apart from failures of the test.

Reproducibility Statement

All derivations underlying the method (the exact split of the Mahalanobis radius, the closed-form OFAT-μ penalty, the optimality of COFAT, the Gaussian-copula construction of COFAT-N, and the closed-form slopes of the synthetic experiment) are stated in Section 3 and proved in Appendix A.2

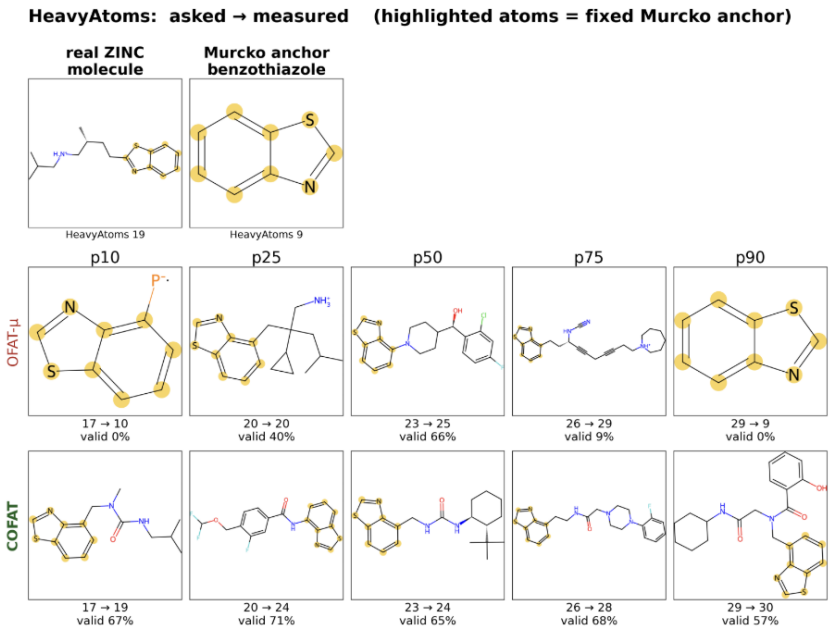

Figure 4: **Scaffold-Anchored Sweep.** A real ZINC molecule defines a fixed Bemis–Murcko scaffold, which is used as the prefix while HeavyAtoms is swept across five percentiles. COFAT remains obedient across the sweep, increasing the number of non-hydrogen atoms with the requested target, whereas OFAT-μ underestimates the target at the extremes (e.g., 10 vs. 17 at p_{10} and 9 vs. 29 at p_{90}). COFAT also maintains $> 50\%$ validity at every level, while OFAT-μ exceeds 50% only at the median level, demonstrating the advantage of COFAT in jointly preserving obedience and realism.

Table 3: Quantitative results for molecule target `HeavyAtoms` (Exp. C). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\to 1$	bias$\to 0$	CCC$\uparrow$	lv	valid% (p_{10}/p_{90})$\uparrow$	copy%$\downarrow$
OFAT-RAND[36]	14.2	14.1	100	—	—	—	*0/5*	*0/0*	—
OFAT-MED[46]	4.7	4.6	40	+1.74	+.05	.85	*3/5*	*0/0*	10.8
OFAT-μ[24]	4.4	4.3	40	+1.57	-.22	.86	*3/5*	*0/0*	11.6
DROP-c_s[35]	4.3	4.3	40	-.74	-.14	-.94	*3/5*	*0/0*	11.1
FREEZE[44]	3.3	3.15	40	+.72	-.14	.92	5/5	5/4	12.6
K-MEANS[56]	2.0	1.7	**0**	+.95	+.14	.98	5/5	76/65	11.3
KNN[8]	0.9	0.4	**0**	+.98	+.16	.98	5/5	80/60	13.1
COFAT (ours)	**0.8**	**0.0**	**0**	+.96	+.04	.99	5/5	82/**67**	14.0
COFAT-N (ours)	1.8	1.6	**0**	+1.03	+.07	.99	5/5	**87/67**	14.1
COFAT-S (ours)	3.2	3.1	**0**	**+.99**	**.00**	**1.00**	5/5	39/48	**8.0**

and A.1.5. All query designs, including the seven baselines, are specified in Section 4 and Appendix A.3, and all metric settings (neighbourhood sizes, thresholds, embeddings, fingerprints, sweep levels and guidance scales) are listed in the experimental setup. Per-level query-side values for every design and attribute are reported in the appendix, so feasibility results can be checked directly against the correlation matrices. Experiments A and C use the public MAMA-MIA and ZINC-250k datasets, for which we release the exact patient- and molecule-level split lists. We will release the complete code base: data preprocessing and radiomic feature extraction (including the extraction configuration), training and inference for all three generators, query construction for all designs, metric computation, the synthetic experiment, together with configuration files, random seeds, and the trained model weights subject to the dataset licenses. The only component that cannot be released is the internal healthy breast-MRI cohort used as insertion hosts in Experiment B; the code for Experiment B can run on any set of lesion-free breast MRI patches.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the research methodology and evaluation design, to check mathematical derivations, and to assist in interpreting results for the written text. We have not used generative AI tools to generate synthetic data or clean or reformat datasets. Additionally, we used generative AI tools to edit the manuscript for readability, suggest the structure of sections, identify relevant literature, and edit code for the experiments. All technical content, derivations, text, code, results and references were created or verified by the authors, who take full responsibility for the content of this work.

REFERENCES

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021. doi: 10.1016/j.artint.2021.103502.

[2] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations (ICLR)*, 2023.

[3] Muhammad Usman Akbar, Wuhao Wang, and Anders Eklund. Beware of diffusion models for synthesizing medical images – a comparison with gans in terms of memorizing brain mri and chest x-ray images. *arXiv preprint arXiv:2305.07644*, 2023.

[4] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning (ICML)*, PMLR, pp. 290–306, 2022.

[5] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020. doi: 10.1111/rssb.12377.

[6] Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2022.

[7] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

[8] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 15309–15324, 2022.

[9] Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019.

[10] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

[11] Trenton Chang, Tobias Schnabel, Adith Swaminathan, and Jenna Wiens. A course correction in steerability evaluation: Revealing miscalibration and side effects in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[12] Veronica Czitrom. One-factor-at-a-time versus designed experiments. *The American Statistician*, 53(2):126–131, 1999.

[13] Cuthbert Daniel. One-at-a-time plans. *Journal of the American Statistical Association*, 68 (342):353–360, 1973. doi: 10.1080/01621459.1973.10482433.

[14] Salman Ul Hassan Dar, Arman Ghanaat, Jannik Kahmann, Isabelle Ayx, Theano Papavassiliu, Stefan O. Schoenberg, and Sandy Engelhardt. Investigating data memorization in 3d latent diffusion models for medical image synthesis. In *Deep Generative Models (MICCAI 2023 Workshop)*, volume 14533 of *Lecture Notes in Computer Science*, pp. 56–65. Springer, 2024.

[15] Salman Ul Hassan Dar et al. Unconditional latent diffusion models memorize patient imaging data. *Nature Biomedical Engineering*, 2025. doi: 10.1038/s41551-025-01468-8.

[16] Berker Demirel, Valentino Maiorca, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Assessing sample quality in conditional generation under compositional shift. *arXiv preprint arXiv:2606.09601*, 2026.

[17] Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[18] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Wiley, New York, 3rd edition, 1998.

[19] Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. What did i do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1543–1558, 2025.

[20] Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

[21] Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, Maria-Laura Cosaka, Pasant M Abo-Elhoda, Sara W Tantawy, Shorouq S Sakrana, Norhan O Shawky-Abdelfatah, Amr Muhammad Abdo Salem, Androniki Kozana, Eugen Divjak, Gordana Ivanac, Katerina Nikiforaki, Michail E Klontzas, Rosa García-Dosdá, Meltem Gulsun-Akpinar, Oğuz Lafcı, Ritse Mann, Carlos Martín-Isla, Fred Prior, Kostas Marias, Martijn P A Starmans, Fredrik Strand, Oliver Díaz, Laura Igual, and Karim Lekadir. A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. *Scientific Data*, 12(1):453, 2025. doi: 10.1038/s41597-025-04707-4.

[22] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572.

[23] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi: Medical ai for synthetic imaging. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441, 2025.

[24] D. M. Hamby. A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, 32(2):135–154, 1994.

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[27] Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

[28] Eui-Jin Kim and Prateek Bansal. A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, 148:104053, 2023.

[29] Jonghun Kim, Inye Na, Eun Sook Ko, and Hyunjin Park. Tumor synthesis conditioned on radiomics. *arXiv preprint arXiv:2509.24182*, 2025.

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[31] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7167–7177, 2018.

[32] Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, 10:31, 2018.

[33] Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical Science*, 11:1153–1164, 2020.

[34] Lawrence I.-K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989. doi: 10.2307/2532051.

[35] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

[36] Max D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991. doi: 10.1080/00401706.1991.10484804.

[37] Inye Na, Jonghun Kim, Eun Sook Ko, and Hyunjin Park. Radiomicsfill-mammo: Synthetic mammogram mass manipulation with radiomics features. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15003 of *LNCS*, pp. 723–733, 2024.

[38] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[40] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, et al. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020.

[41] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[43] Andrea Saltelli and Paola Annoni. How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, 25(12):1508–1517, 2010.

[44] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, 2008. doi: 10.1002/9780470725184.

[45] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(4):2004–2018, 2022.

[46] Christopher G. Small. A survey of multidimensional medians. *International Statistical Review*, 58(3):263–277, 1990. doi: 10.2307/1403809.

[47] Duncan Soiffer, Chandler Squires, Yuan Guan, Jason Hartford, and Pradeep Ravikumar. Catastrophic compositional generation: Why vanilla diffusion models fail to extrapolate. *arXiv preprint arXiv:2606.23920*, 2026.

[48] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023.

[49] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[50] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics–Doklady*, 4:1035–1038, 1963.

[51] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 10401–10412, 2021.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[53] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

[54] Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

[55] Fuzhen Zhang. *The Schur Complement and Its Applications*. Springer, New York, 2006.

[56] Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

A Appendix

A.1 Additional Analysis

A.1.1 Authenticity experiment: is the generator copying?

Setup. We test whether generated lesions and molecules are genuinely novel or close copies of the training set, using held-out real samples as the non-copying baseline. **Lesions**: 34 radiomic features; 1,080 training / 300 held-out MAMA-MIA lesions; generated lesions from the rectified-flow inpainting model (`fo_mean` sweep, $\gamma = 8$, 300 subjects $\times$ 5 levels, 1,500/design, 15,000 total). **Molecules**: 94,513 ZINC training molecules, 2,000 held-out as baseline; 2048-bit Morgan fingerprints (radius 2); generated via `HeavyAtoms` sweep ($\gamma = 1$, 500/level $\times$ 5 levels, 2,500/design, 25,000 total); metrics computed only on RDKit-valid SMILES.

Table 4: Copying test: generated samples against held-out real data. NN: distance to the nearest training sample (lesions: Mahalanobis over principal components carrying 99% of variance, 34 features; molecules: 1−Tanimoto). Copy: % of samples closer to their nearest training sample $\mathbf{z}^*$ than $\mathbf{z}^*$ is to its own nearest neighbour [4]; a design at or below the *real test* row is not copying. Lesion intervals: 95% subject bootstrap. OFAT-RAND produced one valid molecule, so its molecule columns are omitted.

	Lesions (`fo_mean`)		Molecules (`HeavyAtoms`)		
Design	NN ↓	Copy% ↓	Novel%	NN ↓	Copy% ↓
real test	2.17	28.4 [23.5,33.3]	100.0	0.46	46.7
OFAT-RAND[36]	6.17	7.0 [5.9,8.3]	—	—	—
OFAT-MED[46]	2.01	33.9 [30.1,38.0]	99.9	0.60	10.8
OFAT-μ[24]	2.11	33.1 [29.4,36.8]	100.0	0.61	11.6
DROP-c_s[35]	2.16	28.5 [25.7,31.3]	99.9	0.60	11.1
FREEZE[44]	2.23	27.8 [25.0,30.5]	100.0	0.59	12.6
K-MEANS[56]	2.10	29.9 [26.9,33.0]	99.9	0.58	11.3
KNN[8]	2.23	27.2 [24.7,29.7]	100.0	0.56	13.1
COFAT (ours)	2.24	26.9 [24.2,29.8]	100.0	0.56	14.0
COFAT-N (ours)	2.25	27.2 [24.5,29.9]	99.9	0.55	14.1
COFAT-S (ours)	9.42	5.7 [4.7,6.7]	100.0	0.58	8.0

Metrics. For a response attribute vector $\mathbf{a}$ (real held-out or generated), let $\mathbf{z}^* = \arg\min_{\mathbf{z}\in\mathcal{T}} d(\mathbf{a}, \mathbf{z})$ be its nearest training neighbour; $d(\mathbf{a}, \mathbf{z}^*)$ is the **NN distance** (Euclidean in whitened PCA space for lesions; 1−Tanimoto for molecules), reported as the median. NN distance alone cannot signal copying — small values can reflect a dense training region, large values can reflect novelty *or* unrealistic generation.

To test copying directly, we use the authenticity test of [4]: letting $d_{\text{ref}} = \min_{\mathbf{z}'\neq\mathbf{z}^*} d(\mathbf{z}^*, \mathbf{z}')$ be $\mathbf{z}^*$'s own nearest-neighbour distance, $\mathbf{a}$ is flagged as a likely copy when $d(\mathbf{a}, \mathbf{z}^*) < d_{\text{ref}}$, i.e. when it sits closer to a training point than that point normally sits to its own neighbours. The **copy rate** is the % of samples flagged, consistent with the novelty metric in Section 3.3. We compute both metrics for generated and held-out real samples; the real-test copy rate is the natural baseline. Desired behaviour: NN distance comparable to the real baseline, and copy rate not substantially above it. Neither metric alone can distinguish unrealistic generation from rare-but-valid generation. For molecules we additionally report exact-match novelty (% of valid molecules whose SMILES are absent from training).

Results. Table 4 shows little evidence of copying for the main designs. COFAT and COFAT-N match the real-test NN distance (2.24, 2.25 vs. 2.17) with copy rates (26.9%, 27.2%) at or below the 28.4% baseline — as close to the training distribution as unseen real data, with no elevated risk of memorisation. FREEZE, KNN, K-MEANS, and DROP-c_s behave similarly. OFAT-MED and OFAT-μ show higher copy rates (33.9%, 33.1%), suggesting a stronger tendency toward suspiciously close neighbours. COFAT-S and OFAT-RAND have very low copy rates but much larger NN distances (9.42, 6.17) — not better generation, just samples far from the training distribution, illustrating why the two metrics must be read jointly.

For molecules, all designs exceed the real-test NN distance (0.46) while copy rates stay well below the 46.7% baseline (COFAT: 0.56 NN, 14.0% copy; COFAT-N: 0.55 NN, 14.1% copy) — close to the training distribution but not memorising individual molecules. Exact-match novelty is ≈100% across the board. Overall, the combined analysis gives no evidence that COFAT memorises the training set.

A.1.2 IS REALISM CIRCULAR?

One might worry that this comparison is rigged: realism is measured against the same reference distribution used to construct the COFAT query in the first place, so a query that starts closer to the manifold should trivially yield a "more realistic" output. If that were the whole story, realism would simply be an echo of query feasibility rather than a genuine signal.

Here is the test: if the model carried the query's feasibility straight through into its output, realism would decrease *monotonically* as $\bar{\rho}_c$ increases, with no exceptions. Figure 5 shows this is not the case — realism has no consistent relationship to query infeasibility across designs.

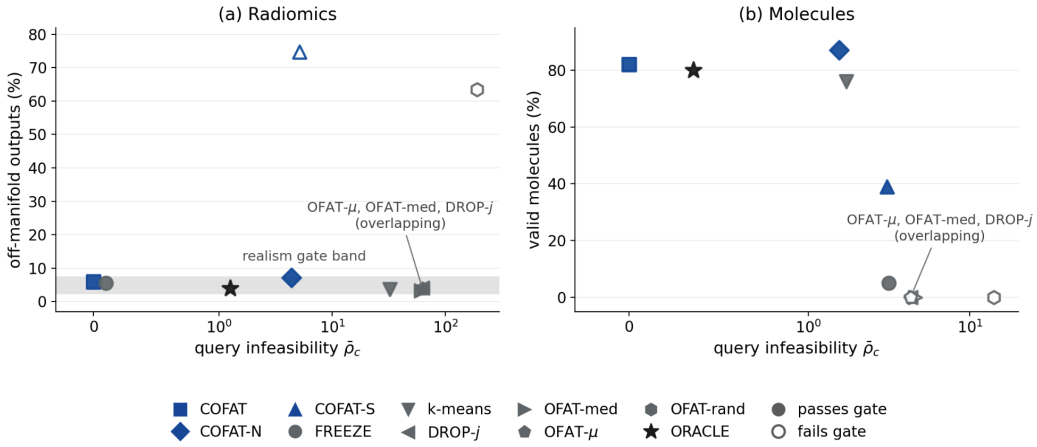

Figure 5: Realism against query infeasibility $\bar{\rho}_c$, one point per design, read directly from Table 1 and Table 3. If realism merely inherited the query's feasibility, points would fall on a decreasing curve; instead the most infeasible baselines (OFAT-μ/med, DROP-c_s, K-MEANS) sit at or below COFAT's off-manifold rate by failing obedience instead, and COFAT-S/COFAT-N show realism can fail or exceed a neighbour's at matched $\bar{\rho}_c$.

What the figure actually reveals are two distinct failure modes hiding behind a passing realism score, plus one clean success case:

1. **A hard, infeasible query can still "pass" realism, for the wrong reason.** In radiomics (Table 1), OFAT-μ[24], OFAT-MED[46], DROP-c_s[35], and K-MEANS[56] operate at $\bar{\rho}_c \in [32, 63]$ — an order of magnitude beyond COFAT's 0 — yet their off-manifold rates (3.3–4.1%) are *lower* than COFAT's (5.9%). This is not because they handle hard queries well: their obedience is low (slope 0.09–0.57), so they barely move the output away from its typical, unconditional value regardless of what was asked. They pass the realism gate by effectively ignoring the query — the same "passing by under-moving the dial" failure noted for OFAT-RAND[36].
2. **An easy, feasible query can still fail realism.** COFAT-S has $\bar{\rho}_c = 5.14$, close to COFAT-N's 4.34, yet fails the gate outright at 74.7%. Feasibility of the query in expectation says nothing about the variance of the realized sample, so an easy query can still land off-manifold once drawn. This may reflect impossible attribute combinations or rare combinations absent from the training distribution; ρ_{cond} alone cannot distinguish between the two cases.
3. **COFAT achieves both at once.** In molecules (Table 3), COFAT-N ($\bar{\rho}_c = 1.55$) is more feasible than the unconditioned baseline but less feasible than COFAT (0), and still reaches higher validity (87% vs. 82%).

Taken together, infeasibility is not proportional to realism in either direction: a hard query does not guarantee an unrealistic output — models can dodge the difficulty by simply not obeying it (case 1) — and an easy query does not guarantee a realistic one, since a feasible mean can still yield an infeasible draw (case 2). This is precisely the paper's central point: realism alone is a hollow metric, since a model can satisfy it just by ignoring the request. What *is* true by construction is only that COFAT pays no infeasibility penalty for asking; whether the model then returns a realistic, obedient response is an empirical property of the model itself. That is why we report obedience and realism *jointly* and gate on both rather than reading either in isolation — a design only earns credit if it is both listening to the query *and* remaining on-manifold, not one or the other.

A.1.3 RARE-BUT-POSSIBLE COMBINATIONS: A STRESS TEST FOR THE REALISM GATE

The objection. COFAT always sets companion attributes to their conditional mean, so every query is, by construction, *typical*. A generator that merely retrieves training samples satisfies such queries perfectly, while a generator that can compose genuinely novel — rare but valid — combinations gets no credit, since COFAT never asks for one. This objection hinges on a distinction ρ_{cond} cannot

make: is an off-conditional combination *impossible* (outside the support) or merely *rare* (inside it, at low density)? Since ρ_{cond} is a density-based measure, it cannot tell. This experiment tests that gap directly.

Setup. For each swept attribute s, we displace the companions by a controlled offset:

$$\mathbf{c}(\alpha, t) = \alpha\, \mathbf{R}_{\cdot s} + \boldsymbol{\delta}_t, \qquad \boldsymbol{\delta}_t = (0,\ t\, \mathbf{L}\mathbf{u}), \tag{9}$$

where $\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}$ is the Cholesky factorisation of the conditional covariance of the companions ($\boldsymbol{\Sigma} = \mathbf{Q} - \mathbf{b}\mathbf{b}^\top$) and $\mathbf{u}$ is a random unit vector, so that $\rho_{\text{cond}}(\mathbf{c}) = t$ exactly. At $t = 0$ the query reduces to COFAT; $t > 0$ requests a combination t conditional standard deviations away from what the swept value predicts. OFAT-μ corresponds to a single point on this axis at $t = |\alpha|\, R_s / \sqrt{1 - R_s^2}$ (Eq. 4). We sweep $t \in \{0.25\} \cup \{1, \ldots, 20\}$, averaging over 9 levels of α (i.e., $L = 9$ percentiles), 20 random directions per swept attribute, and all $d = 20$ attributes.

Generators. Four *perfectly obedient*, noiseless generators differ only in how they resolve an off-conditional request. Three are Tikhonov responders $\mathbf{a} = \mathbf{M}_\kappa\, \mathbf{c}$, with $\mathbf{M}_\kappa = \mathbf{R}(\mathbf{R} + \kappa \mathbf{I})^{-1}$: **compositional** ($\kappa = 10^{-5}$, near-identity), **default** ($\kappa = 0.05$), and **anchored** ($\kappa = 1$, strongly pulled back toward the mean). The fourth, **copier**, returns the nearest of $n = 50{,}000$ training vectors $\mathbf{z}$ — pure retrieval, no composition.

Metrics. **Delivered fraction** measures how much of the requested offset appears in the output:

$$\text{delivered}(t) = \frac{\langle\, \mathbf{a}(\mathbf{c}) - \mathbf{a}(\mathbf{c}_0),\ \boldsymbol{\delta}_t \,\rangle}{\|\boldsymbol{\delta}_t\|^2}, \tag{10}$$

the OLS slope of the achieved shift onto the requested shift, where $\mathbf{c}_0 = \alpha\, \mathbf{R}_{\cdot s}$ is the COFAT baseline query. The dot product credits movement only in the requested *direction*, separating obedience from mere displacement: 1 means the offset is fully delivered, 0 means it was ignored. Numerator and denominator are pooled across s, directions, and α before dividing, giving one stable regression slope per t. **Off-manifold rate** is the standard realism gate (Eq. 8): the fraction of outputs whose Mahalanobis radius exceeds the training set's τ-percentile radius by more than δ.

Results. **Compositional** and **default** deliver $\gtrsim 0.95$–1.0 across the whole dose range — fully obedient regardless of rarity. **Anchored** plateaus at ≈ 0.83, obedient but damped. **Copier** degrades steadily with t, since the nearest training vector $\mathbf{z}^*$ falls further behind an increasingly rare target. Under the realism gate, **compositional** jumps to 100% off-manifold almost immediately past $t \approx 5$ — it obeys so exactly that the gate flags it the moment the request itself leaves the training density's high-mass region. **Default** and **copier** cross the gate band later, more gradually; **anchored** stays inside it longest, simply by moving the least.

Interpretation. The generator best at composing genuinely novel, valid combinations (**compositional**) fails the realism gate soonest, precisely for doing what was asked. The generator that looks most "realistic" under the gate alone (**anchored**) is not more realistic in any generative sense — it is the least obedient. This is the same "passing by under-moving the dial" failure seen in the OFAT baselines, now reproduced under fully controlled conditions where ρ_{cond}, obedience, and realism sit on one continuous axis.

Takeaways.

- ρ_{cond} cannot distinguish an impossible combination from a rare-but-valid one — both register as "far from the conditional mean," since a density-based measure has no notion of a hard support boundary.
- A realism gate used in isolation rewards the wrong behaviour: it favours generators that ignore atypical requests (**anchored**, **copier**) over generators that faithfully compose rare, legitimate combinations (**compositional**), inverting the intended incentive.
- Obedience and realism must therefore be reported *jointly*: a model deserves credit only when it both follows the query and remains on-manifold, not for satisfying either criterion alone.

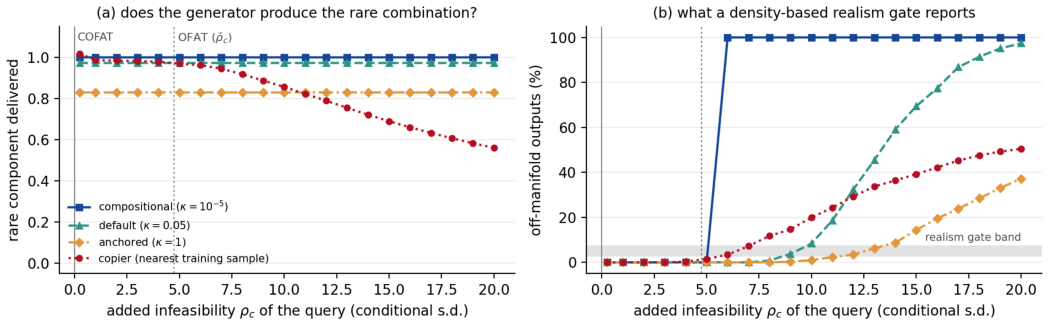

Figure 6: Rare-combination sweep with perfectly obedient generators ($\mathbf{G} = \mathbf{I}$, $\overline{R_s^2} = 0.975$). **(a)** Fraction of the requested offset delivered (Eq. 10). **(b)** Off-manifold rate (grey band: realism gate).

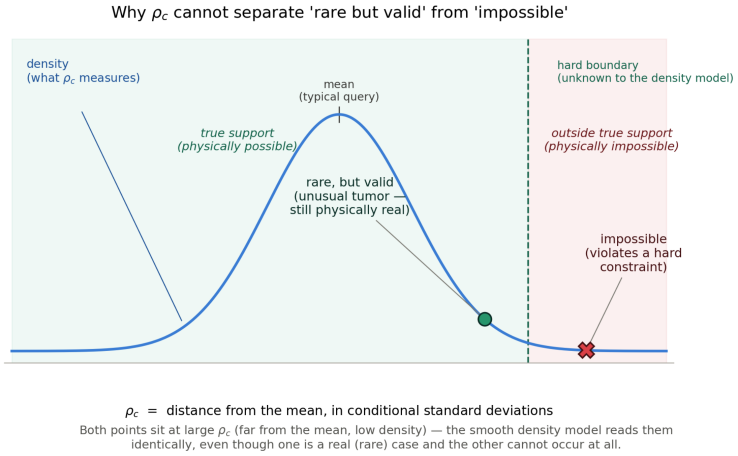

Figure 7: The Mahalanobis-distance dilemma: ρ_{cond} is a density-based measure, so it assigns the same large value to an impossible attribute combination (outside the support) and to a rare-but-valid one (inside the support at low density). The two cases require different responses from the generator but are indistinguishable from ρ_{cond} alone.

- Even under a highly infeasible query, an obedient generator with moderate anchoring (**default**) can recover much of the requested rare component while staying near the realism gate. A high off-manifold rate therefore does not necessarily imply that the generated sample is unrealistic — it may instead reflect a rare but genuine attribute combination that the density-based gate cannot distinguish from an impossible one.

A.1.4 Response-Side Theory

The response-side analysis first formalizes the response achieved and the obedience–realism tradeoff, then identifies the necessary and sufficient conditions required for achieving both obedience and realism.

Response Formalization We formalize the response of a conditional generator using a minimal Tikhonov-regularized abstraction [50, 26] demonstrating an obedience-realism tradeoff. Let $\mathbf{c}$ denote the requested change, $\mathbf{a}$ the achieved response, $\mathbf{G}$ the ideal request-to-response map of the model's generator, and $\mathbf{R} \succ 0$ the training-distribution correlation matrix. We define

$$\mathbf{a}(\mathbf{c}) = \arg\min_{\mathbf{a}} \underbrace{\|\mathbf{a} - \mathbf{G}\mathbf{c}\|^2}_{\text{obedience}} + \underbrace{\kappa\, \mathbf{a}^\top \mathbf{R}^{-1} \mathbf{a}}_{\text{realism}} = \underbrace{\mathbf{R}(\mathbf{R} + \kappa I)^{-1}}_{=: \mathbf{M}} \mathbf{G}\mathbf{c}, \tag{11}$$

where $\mathbf{M} := \mathbf{R}(\mathbf{R} + \kappa I)^{-1}$ and $\kappa \geq 0$ controls the strength of the distributional anchor. The matrix $\mathbf{G}$ captures the model's inherent obedience, with $\mathbf{G} = I$ denoting perfect obedience. Total

obedience combines the model's intrinsic obedience G with query dependence R_s^2, where $R_s^2 = 0$ indicates independent attributes and $R_s^2 > 0$ indicates dependence.

Proof. The objective is strictly convex in $\mathbf{a}$ since $\mathbf{R} \succ 0$ and $\kappa \geq 0$. Setting its gradient to zero gives $(I + \kappa \mathbf{R}^{-1})\mathbf{a} = \mathbf{G}\mathbf{c}$, and multiplying by $\mathbf{R}$ yields $(\mathbf{R} + \kappa I)\mathbf{a} = \mathbf{R}\mathbf{G}\mathbf{c}$. Since $\mathbf{R} + \kappa I$ is invertible, we obtain $\mathbf{a} = \mathbf{R}(\mathbf{R} + \kappa I)^{-1}\mathbf{G}\mathbf{c} = \mathbf{M}\mathbf{G}\mathbf{c}$. □

Necessary and Sufficient Conditions for Joint Obedience–Realism We study when a controllable model produces responses that are both *obedient* to the query and *realistic* under the training distribution. The limiting cases illustrate the failure modes.

Condition 1: Nontrivial attribute dependence. A model's obedience depends on its intrinsic, unknown obedience G and the query $\mathbf{c}$. COFAT exploits conditional dependencies in the training distribution R_s^2 by fixing non-swept attributes at their conditional means. If the swept attribute j is independent of the others, $R_s^2 = 0$, then $\mathbf{R}_{\cdot j} = \mathbf{e}_j$, and COFAT reduces to OFAT:

$$R_s^2 = 0 \implies \mathbf{c}_{\text{COFAT}} = \mathbf{c}_{\text{OFAT}} = \alpha\, \mathbf{e}_j, \qquad \forall\, \kappa, G. \tag{12}$$

Thus, with no attribute dependence, COFAT cannot exploit the training distribution to improve obedience, regardless of G or κ. Here, G is unknown and is estimated only from the training data.

Condition 2: Positive realism regularization. Conversely, suppose the model is moderately obedient. The parameter κ controls the pull toward the real distribution. When $\kappa = 0$, this regularization vanishes and

$$\mathbf{M} = \mathbf{R}(\mathbf{R} + \kappa I)^{-1}\Big|_{\kappa=0} = \mathbf{R}\mathbf{R}^{-1} = I. \tag{13}$$

Hence, the model may remain obedient but is no longer constrained toward realism.

Together, these cases imply that joint obedience–realism requires both nontrivial attribute dependence and positive realism regularization, i.e., $R_s^2 \neq 0$ and $\kappa > 0$. If $R_s^2 = 0$, COFAT collapses to OFAT; if $\kappa = 0$, the response is obedient but non-realistic.

A.1.5 Synthetic Experiment D: Mathematical Derivation

The synthetic experiment isolates the effect of query design from generator quality by pairing a near-perfect generator with attribute spaces of increasing correlation. We derive the measured obedience (slope) in closed form for both OFAT-μ and COFAT, showing that the gap is fully explained by the correlation structure.

Correlation matrix construction. Sample a fixed $d \times m$ loading matrix Λ ($d = 20$, $m = 3$) with entries drawn i.i.d. from $\mathcal{N}(0, 1)$. For a scalar $\omega \in [0, 1)$, define the raw covariance

$$\mathbf{\Sigma}(\omega) = \omega^2\, \Lambda\Lambda^\top + \mathbf{I}_d, \tag{14}$$

and normalize to unit diagonal:

$$\mathbf{R}(\omega) = \mathbf{D}(\omega)^{-1/2}\, \mathbf{\Sigma}(\omega)\, \mathbf{D}(\omega)^{-1/2}, \qquad \mathbf{D}(\omega) = \operatorname{diag}\big(\mathbf{\Sigma}(\omega)\big). \tag{15}$$

At $\omega = 0$, $\mathbf{R} = \mathbf{I}$ (independent attributes). As $\omega \to 1$, the off-diagonal entries grow and the average redundancy $\overline{R_s^2}$ approaches 1.

Response model. We model a near-perfect generator via the linear response

$$\mathbf{a} = \mathbf{M}\,\mathbf{c}, \qquad \mathbf{M} = \mathbf{R}\big(\mathbf{R} + \kappa\mathbf{I}\big)^{-1}, \quad \kappa = 0.05. \tag{16}$$

This has a Bayesian interpretation: if the latent attribute vector has prior $\mathcal{N}(\mathbf{0}, \mathbf{R})$ and the query $\mathbf{c}$ is observed with isotropic noise $\kappa\mathbf{I}$, then $\mathbf{M}\mathbf{c}$ is the posterior mean. The generator attempts to deliver $\mathbf{c}$ exactly ($\mathbf{G} = \mathbf{I}$), but the correlation structure pulls the response toward the data manifold by an amount controlled by κ. When $\kappa \to 0$, $\mathbf{M} \to \mathbf{I}$ and the generator is perfectly obedient on *every* query. For small $\kappa > 0$, obedience depends on the spectral structure of $\mathbf{R}$ and, crucially, on the query $\mathbf{c}$.

Spectral form. Let $\mathbf{R} = \mathbf{U}\mathbf{\Delta}\mathbf{U}^\top$ be the eigendecomposition, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d > 0$ and orthonormal eigenvectors. Then

$$\mathbf{M} = \mathbf{U}\ \mathrm{diag}\Big(\frac{\lambda_j}{\lambda_j + \kappa}\Big)\mathbf{U}^\top. \tag{17}$$

The diagonal entry and column product needed below are:

$$M_{ss} = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \kappa}\, U_{sj}^2, \qquad [\mathbf{MR}]_{ss} = \sum_{j=1}^{d} \frac{\lambda_j^2}{\lambda_j + \kappa}\, U_{sj}^2, \tag{18}$$

using $\sum_j U_{sj}^2 = 1$ (orthonormality) and $R_{ss} = \sum_j \lambda_j U_{sj}^2 = 1$ (unit diagonal).

OFAT-μ slope. The OFAT query sets $c_s = \alpha$ and $\mathbf{c}_{-s} = \mathbf{0}$ (Eq. 4), i.e. $\mathbf{c}^{\mathrm{OFAT}} = \alpha\mathbf{e}_s$. The realized value of attribute s is

$$a_s^{\mathrm{OFAT}} = [\mathbf{M}\,\alpha\mathbf{e}_s]_s = \alpha\, M_{ss} = \alpha \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \kappa}\, U_{sj}^2. \tag{19}$$

Since the sweep varies α and the response is linear in α, the measured slope is

$$\boxed{\mathrm{slope}_s^{\mathrm{OFAT}} = M_{ss} = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \kappa}\, U_{sj}^2.} \tag{20}$$

Independent case ($\omega = 0$). $\mathbf{R} = \mathbf{I}$, so all $\lambda_j = 1$ and $\mathrm{slope}_s^{\mathrm{OFAT}} = 1/(1+\kappa) = 1/1.05 \approx 0.952$.

High-redundancy case ($\omega \to 1$). The factor model concentrates most of the variance into $m = 3$ large eigenvalues; the remaining $d - m = 17$ eigenvalues shrink toward zero. For any small eigenvalue $\lambda_j \ll \kappa$, $\lambda_j/(\lambda_j + \kappa) \approx \lambda_j/\kappa \approx 0$. Because $\sum_j U_{sj}^2 = 1$, eigenvectors associated with small eigenvalues generically carry non-negligible weight U_{sj}^2, and their near-zero factors drag M_{ss} well below 1.

COFAT slope. The COFAT query sets $\mathbf{c}^{\mathrm{COFAT}} = \alpha\mathbf{R}_{\cdot s}$ (Eq. 5), so $c_s = \alpha R_{ss} = \alpha$ and $\mathbf{c}_{-s} = \alpha\mathbf{b}$. The realized value of attribute s is

$$a_s^{\mathrm{COFAT}} = [\mathbf{M}\,\alpha\mathbf{R}_{\cdot s}]_s = \alpha\,[\mathbf{MR}]_{ss} = \alpha \sum_{j=1}^{d} \frac{\lambda_j^2}{\lambda_j + \kappa}\, U_{sj}^2, \tag{21}$$

where we used $\mathbf{R}_{\cdot s} = \mathbf{U}\mathbf{\Delta}\mathbf{U}^\top\mathbf{e}_s$ and Eq. (18). The measured slope is

$$\mathrm{slope}_s^{\mathrm{COFAT}} = [\mathbf{MR}]_{ss} = \sum_{j=1}^{d} \frac{\lambda_j^2}{\lambda_j + \kappa}\, U_{sj}^2. \tag{22}$$

Key identity linking the two slopes. We decompose each eigenvalue term:

$$\frac{\lambda_j^2}{\lambda_j + \kappa} = \frac{\lambda_j(\lambda_j + \kappa) - \kappa\lambda_j}{\lambda_j + \kappa} = \lambda_j - \kappa\,\frac{\lambda_j}{\lambda_j + \kappa}.$$

Substituting into Eq. (22) and using $\sum_j \lambda_j U_{sj}^2 = R_{ss} = 1$:

$$\boxed{\mathrm{slope}_s^{\mathrm{COFAT}} = \underbrace{\sum_j \lambda_j\, U_{sj}^2}_{R_{ss}=1} - \kappa \underbrace{\sum_j \frac{\lambda_j}{\lambda_j + \kappa}\, U_{sj}^2}_{M_{ss} = \mathrm{slope}_s^{\mathrm{OFAT}}} = 1 - \kappa\,\mathrm{slope}_s^{\mathrm{OFAT}}.} \tag{23}$$

Equivalently, in matrix form, the identity $\lambda_j^2/(\lambda_j+\kappa) = \lambda_j - \kappa\lambda_j/(\lambda_j+\kappa)$ applied to all eigenvalues simultaneously gives

$$\mathbf{MR} = \mathbf{R}^2(\mathbf{R} + \kappa\mathbf{I})^{-1} = \mathbf{R} - \kappa\mathbf{M}. \tag{24}$$

Analysis of the identity. Equation (23) reveals three properties:

1. **Independent attributes** ($\omega = 0$, all $\lambda_j = 1$). Both slopes are $1/(1+\kappa) = 0.952$, and the identity confirms: $0.952 = 1 - 0.05 \times 0.952$ ✓.
2. **High redundancy** ($\omega \to 1$). The OFAT slope M_{ss} drops (e.g., to 0.149). The COFAT slope $1 - \kappa\, M_{ss}$ remains near 1:
$$\text{slope}^{\text{COFAT}} = 1 - 0.05 \times 0.149 = 0.993.$$
3. **Bounds.** Since $0 < M_{ss} \le 1/(1+\kappa)$, the COFAT slope satisfies
$$\frac{1}{1+\kappa} \;\le\; \text{slope}^{\text{COFAT}} \;\le\; 1. \tag{25}$$
For $\kappa = 0.05$, the COFAT slope is always in $[0.952,\, 1]$. In contrast, the OFAT slope M_{ss} can fall to arbitrarily small values as the smallest eigenvalue of $\mathbf{R}$ shrinks relative to κ.

Why COFAT is immune to redundancy. The OFAT query $\alpha \mathbf{e}_s$ projects onto *all* eigenvectors of $\mathbf{R}$, including those with $\lambda_j \ll \kappa$ whose response is heavily attenuated ($\lambda_j/(\lambda_j + \kappa) \approx 0$). Each such direction contributes U_{sj}^2 weight to the slope sum, dragging it down.

The COFAT query $\alpha \mathbf{R}_{\cdot s}$ pre-multiplies by $\mathbf{R}$, which scales each eigenvector by its eigenvalue λ_j. Small-eigenvalue directions receive weight proportional to $\lambda_j^2 U_{sj}^2$ instead of $\lambda_j U_{sj}^2$, suppressing their contribution quadratically. The surviving response is dominated by the large-eigenvalue (data-supported) directions, which the generator handles well. This is precisely why conditioning companions on the swept target (Proposition 1) removes the query-induced obedience penalty: the query no longer asks the generator to produce attribute combinations that lie outside the correlation structure of the training data.

Average slope across attributes. The reported slopes are averaged over all d swept attributes:

$$\overline{\text{slope}}^{\text{OFAT}} = \frac{1}{d}\sum_{s=1}^{d} M_{ss} = \frac{1}{d}\,\text{tr}(\mathbf{M}) = \frac{1}{d}\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \kappa}, \tag{26}$$

$$\overline{\text{slope}}^{\text{COFAT}} = 1 - \kappa\,\overline{\text{slope}}^{\text{OFAT}} = 1 - \frac{\kappa}{d}\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \kappa}. \tag{27}$$

The OFAT average slope is the mean eigenvalue shrinkage factor. Under high redundancy, m eigenvalues are large ($\lambda_j/(\lambda_j + \kappa) \approx 1$) and the remaining $d - m$ are small (≈ 0), giving $\overline{\text{slope}}^{\text{OFAT}} \approx m/d = 3/20 = 0.15$, matching the reported 0.149. The COFAT average slope is $1 - 0.05 \times 0.15 = 0.993$, matching the reported 0.993.

Panel (a): Varying redundancy at fixed κ. Figure 8(a) fixes $\kappa = 0.05$ and sweeps ω from 0 (independent) to near 1 (highly redundant). The x-axis plots $1 - \overline{R_s^2}$ on a log scale (decreasing = more dependent).

At high redundancy the factor model yields m large eigenvalues $\lambda_j \gg \kappa$ and $d - m$ small ones $\lambda_j \ll \kappa$. From Eq. (26):

$$\overline{\text{slope}}^{\text{OFAT}} \;\approx\; \frac{1}{d}\left(\underbrace{\sum_{j=1}^{m} 1}_{m \text{ large } \lambda_j} \;+\; \underbrace{\sum_{j=m+1}^{d} 0}_{(d-m) \text{ small } \lambda_j} \right) = \frac{m}{d}, \tag{28}$$

which for m=3, d=20 gives the ≈ 0.15 plateau visible in the red curve. COFAT (Eq. (27)) stays near 1 throughout because $\kappa \cdot m/d$ is small.

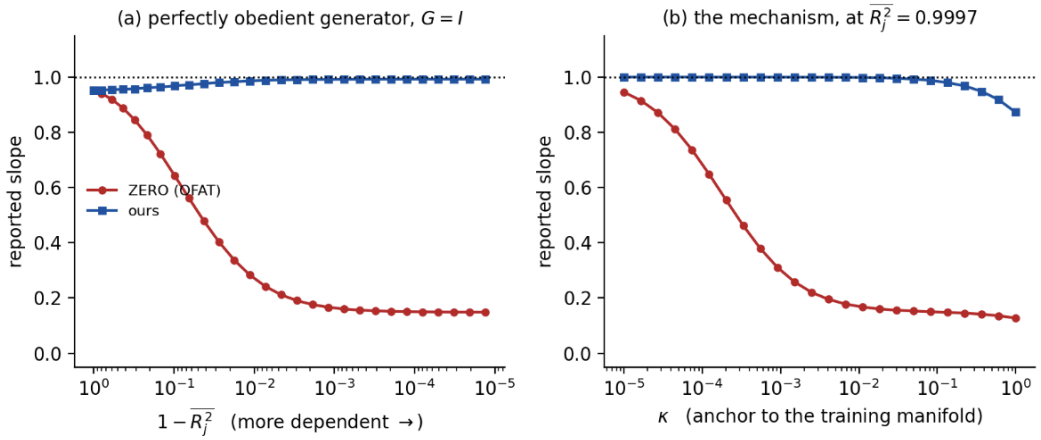

Figure 8: Results of a synthetic generator g that is perfectly obedient on feasible request. (a) Obedience decreases for OFAT as attribute redundancy increases, while COFAT remains near-perfect. (b) At fixed high redundancy, increasing realism strength κ widens the gap between OFAT and COFAT

Panel (b): Varying κ at fixed high redundancy. Figure 8(b) fixes $\overline{R_s^2} = 0.9997$ (so $1-\overline{R_s^2} \approx 3\times10^{-4}$) and sweeps κ from 10^{-5} to 10^{0}. The eigenvalue spectrum is now fixed: $m = 3$ eigenvalues of order $\lambda_{\max}$ and $d-m = 17$ eigenvalues of order $\lambda_{\min} \ll 1$. Both slopes are functions of κ alone.

Three regimes emerge:

Regime I: $\kappa \ll \lambda_{\min}$. All shrinkage factors $\lambda_j/(\lambda_j + \kappa) \approx 1$, so $\overline{\text{slope}}^{\text{OFAT}} \approx 1$ and $\overline{\text{slope}}^{\text{COFAT}} \approx 1 - \kappa \approx 1$. Both curves hug the dotted line at 1.

Regime II: $\lambda_{\min} \ll \kappa \ll \lambda_{\max}$. The $d-m$ small eigenvalues are attenuated ($\lambda_j/(\lambda_j + \kappa) \approx 0$) while the m large ones remain unaffected (≈ 1). Substituting into Eqs. (26)–(27):

$$\overline{\text{slope}}^{\text{OFAT}} \approx \frac{m}{d} = 0.15, \tag{29}$$

$$\overline{\text{slope}}^{\text{COFAT}} \approx 1 - \kappa\,\frac{m}{d} \approx 1, \qquad (\text{since } \kappa \cdot m/d \ll 1). \tag{30}$$

This is the wide plateau in the red curve where OFAT has already collapsed but COFAT remains near-perfect.

Regime III: $\kappa \gtrsim \lambda_{\max}$. Now even the large eigenvalues are attenuated: $\lambda_j/(\lambda_j + \kappa) \approx \lambda_j/\kappa$. Both slopes decrease:

$$\overline{\text{slope}}^{\text{OFAT}} \approx \frac{1}{d\kappa}\sum_j \lambda_j = \frac{\text{tr}(\mathbf{R})}{d\kappa} = \frac{1}{\kappa}, \tag{31}$$

$$\overline{\text{slope}}^{\text{COFAT}} = 1 - \kappa\,\overline{\text{slope}}^{\text{OFAT}} \approx 1 - \frac{\kappa}{\kappa} \;\to\; 0. \tag{32}$$

The last line uses $\text{tr}(\mathbf{R}) = d$ (correlation matrix), so $\overline{\text{slope}}^{\text{OFAT}} \to 1/\kappa$ and $\overline{\text{slope}}^{\text{COFAT}} \to 1 - 1 = 0$ in the extreme limit. In practice $\kappa = 1$ is the largest value tested, the large eigenvalues satisfy $\lambda_{\max} \gg 1$ (e.g. $\lambda_{\max} \sim 10$–15 for d=20, m=3), so Regime III is only partially entered and the COFAT slope drops to ≈ 0.88 rather than zero, as visible in the figure.

Summary. The κ-sweep confirms that COFAT's advantage comes from suppressing the query's projection onto small-eigenvalue directions (Regime II). Only when κ overwhelms even the dominant eigenvalues (Regime III) does COFAT begin to degrade—but it still maintains a strictly higher slope than OFAT at every κ.

A.2 PROOFS

A.2.1 PROOF OF LEMMA 1 (EXACT SPLIT OF SQUARED MAHALANOBIS RADIUS)

[Restatement] Let $\mathbf{R} = \left[\begin{smallmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & \mathbf{Q} \end{smallmatrix}\right] \succ 0$ be the correlation matrix of the z-scored attributes with swept attribute s ordered first, and let $\mathbf{\Sigma} = \mathbf{Q} - \mathbf{b}\mathbf{b}^\top$ be the Schur complement of the $(1,1)$ block. For any query $\mathbf{c} = (\alpha,\, \mathbf{c}_{-s})$,

$$\rho^2(\mathbf{c}) \;=\; \mathbf{c}^\top \mathbf{R}^{-1}\mathbf{c} \;=\; \underbrace{\alpha^2}_{\text{fixed request}} \;+\; \underbrace{(\mathbf{c}_{-s} - \alpha\mathbf{b})^\top \mathbf{\Sigma}^{-1}(\mathbf{c}_{-s} - \alpha\mathbf{b})}_{\rho^2_{\text{cond}}(\mathbf{c})}.$$

Proof. Because $\mathbf{R} \succ 0$ and $[\mathbf{R}]_{11} = 1 > 0$, the Schur complement $\mathbf{\Sigma} = \mathbf{Q} - \mathbf{b}\mathbf{b}^\top$ is positive definite and the block inversion formula [55, Thm. 2.1] gives

$$\mathbf{R}^{-1} = \begin{bmatrix} 1 + \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b} & -\mathbf{b}^\top \mathbf{\Sigma}^{-1} \\ -\mathbf{\Sigma}^{-1}\mathbf{b} & \mathbf{\Sigma}^{-1} \end{bmatrix}. \tag{33}$$

Write $\mathbf{c} = (\alpha,\, \mathbf{c}_{-s})^\top$ and expand the quadratic form:

$$\begin{aligned} \mathbf{c}^\top \mathbf{R}^{-1}\mathbf{c} &= \alpha^2\big(1 + \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b}\big) - 2\alpha\, \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{c}_{-s} + \mathbf{c}_{-s}^\top \mathbf{\Sigma}^{-1}\mathbf{c}_{-s} \\ &= \alpha^2 + \big(\alpha^2 \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b} - 2\alpha\, \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{c}_{-s} + \mathbf{c}_{-s}^\top \mathbf{\Sigma}^{-1}\mathbf{c}_{-s}\big) \\ &= \alpha^2 + (\mathbf{c}_{-s} - \alpha\mathbf{b})^\top \mathbf{\Sigma}^{-1}(\mathbf{c}_{-s} - \alpha\mathbf{b}), \end{aligned} \tag{34}$$

where the last step completes the square. Since $\mathbf{\Sigma}^{-1} \succ 0$, the second term is non-negative and equals zero if and only if $\mathbf{c}_{-s} = \alpha\mathbf{b}$. □

A.2.2 PROOF OF OFAT-μ INFEASIBILITY PENALTY (EQ. 4)

[OFAT-μ conditional infeasibility] Setting companions to their marginal means, $\mathbf{c}_{-s} = \mathbf{0}$, gives

$$\rho_{\text{cond}}\big(\mathbf{c}^{\text{OFAT}}\big) = |\alpha|\, \frac{R_s}{\sqrt{1 - R_s^2}},$$

where $R_s^2 = \mathbf{b}^\top \mathbf{Q}^{-1}\mathbf{b}$ is the redundancy of attribute s.

Proof. **Step 1.** Substitute $\mathbf{c}_{-s} = \mathbf{0}$ into the split (Lemma 1):

$$\rho^2_{\text{cond}} = (\mathbf{0} - \alpha\mathbf{b})^\top \mathbf{\Sigma}^{-1}(\mathbf{0} - \alpha\mathbf{b}) = \alpha^2\, \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b}. \tag{35}$$

It remains to evaluate $\mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b}$ in terms of R_s^2.

Step 2. Define $q\mathbf{b}^\top \mathbf{Q}^{-1}\mathbf{b}$. Because $\mathbf{\Sigma} = \mathbf{Q} - \mathbf{b}\mathbf{b}^\top$ is a rank-one perturbation of $\mathbf{Q}$, the Sherman–Morrison formula yields

$$\mathbf{\Sigma}^{-1} = \big(\mathbf{Q} - \mathbf{b}\mathbf{b}^\top\big)^{-1} = \mathbf{Q}^{-1} + \frac{\mathbf{Q}^{-1}\mathbf{b}\mathbf{b}^\top \mathbf{Q}^{-1}}{1 - q}.$$

Multiplying on both sides by $\mathbf{b}^\top$ and $\mathbf{b}$:

$$\mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b} = q + \frac{q^2}{1 - q} = \frac{q}{1 - q}. \tag{36}$$

Step 3. Identify $q = R_s^2$. From the block inverse (33), the $(1,1)$ entry of $\mathbf{R}^{-1}$ is

$$[\mathbf{R}^{-1}]_{ss} = 1 + \mathbf{b}^\top \mathbf{\Sigma}^{-1}\mathbf{b} = 1 + \frac{q}{1 - q} = \frac{1}{1 - q}.$$

The redundancy definition (Eq. 1) states $R_s^2 = 1 - 1/[\mathbf{R}^{-1}]_{ss} = 1 - (1 - q) = q$.

Step 4. Combine (35) and (36):

$$\rho^2_{\text{cond}} = \alpha^2 \cdot \frac{R_s^2}{1 - R_s^2}, \qquad \rho_{\text{cond}} = |\alpha|\, \frac{R_s}{\sqrt{1 - R_s^2}}.$$ □

Interpretation. When the swept attribute is independent of its companions, $R_s^2 = 0$ and $\rho_{\text{cond}} = 0$: OFAT-μ is perfectly feasible. As $R_s^2 \to 1$ (full redundancy), $\rho_{\text{cond}} \to \infty$: the query requests a combination that diverges from the training distribution proportionally to the sweep magnitude $|\alpha|$.

A.2.3 DISTRIBUTION-FREE OPTIMALITY OF COFAT (PROPOSITION 1)

[Restatement] For any fixed $c_s = \alpha$, the conditional infeasibility ρ_{cond} is uniquely minimized by $\mathbf{c}_{-s} = \alpha\mathbf{b}$, giving $\rho_{\text{cond}} = 0$ and $\rho = |\alpha|$. This holds for *any* distribution of the attributes.

Proof. The conditional infeasibility is defined purely in terms of the correlation matrix:

$$\rho^2_{\text{cond}}(\mathbf{c}) = (\mathbf{c}_{-s} - \alpha\mathbf{b})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{c}_{-s} - \alpha\mathbf{b}), \qquad \boldsymbol{\Sigma} = \mathbf{Q} - \mathbf{b}\mathbf{b}^\top.$$

The matrix $\boldsymbol{\Sigma}$ is positive definite (it is the Schur complement of a positive-definite matrix $\mathbf{R}$). A positive-definite quadratic form $\mathbf{u}^\top\boldsymbol{\Sigma}^{-1}\mathbf{u} \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.

Setting $\mathbf{u} = \mathbf{c}_{-s} - \alpha\mathbf{b}$, the unique minimizer is $\mathbf{c}^\star_{-s} = \alpha\mathbf{b}$, at which $\rho^2_{\text{cond}} = 0$ and, from Lemma 1, $\rho^2 = \alpha^2 + 0 = \alpha^2$.

No distributional assumption is used. The quantities $\mathbf{R}$, $\mathbf{b}$, $\mathbf{Q}$, and $\boldsymbol{\Sigma}$ are computed from second-order statistics (the sample correlation matrix) of the training data. The proof uses only the algebraic fact that $\boldsymbol{\Sigma}^{-1} \succ 0$, which follows from $\mathbf{R} \succ 0$ via Schur complement theory and holds regardless of whether the underlying data distribution is Gaussian, skewed, heavy-tailed, or discrete.

Conditional-mean interpretation. While the optimality of $\mathbf{c}_{-s} = \alpha\mathbf{b}$ is distribution-free, its interpretation as the conditional expectation $\mathbb{E}[\mathbf{c}_{-s} \mid c_s = \alpha] = \alpha\mathbf{b}$ requires additional structure. For a multivariate Gaussian, the conditional expectation is linear and equals $\alpha\mathbf{b}$ exactly. More generally, if the conditional expectation is linear in c_s (as guaranteed, e.g., by elliptical distributions), the same identification holds. For distributions with nonlinear conditional expectations, COFAT still minimizes ρ_{cond} to zero but the companions $\alpha\mathbf{b}$ may differ from the true conditional mean; COFAT-N (Appendix A.2.4) addresses this case. □

A.2.4 COFAT-N: DERIVATION VIA THE GAUSSIAN COPULA

When marginals are non-Gaussian (skewed, bounded, discrete), the linear conditional mean $\alpha\mathbf{b}$ may not match the true conditional expectation. COFAT-N corrects for this by working in a Gaussianized space via the probability-integral transform and a Gaussian copula.

Setup. Let $z_1, \ldots, z_d$ be the z-scored training attributes (zero mean, unit variance) with possibly non-Gaussian marginals. Denote the empirical CDF of attribute k by $\hat{F}_k$.

Step 1: Transform to Gaussian marginals. Apply the probability-integral transform followed by the inverse standard-normal CDF:

$$\tilde{z}_k = \Phi^{-1}\big(\hat{F}_k(z_k)\big), \qquad k = 1, \ldots, d. \tag{37}$$

Each $\tilde{z}_k$ has, by construction, a standard-normal marginal distribution. The dependence structure among $\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_d)$ is captured by their correlation matrix $\tilde{\mathbf{R}}$, which defines the Gaussian copula of the original data [**?**].

Step 2: Apply COFAT in the Gaussian space. Partition $\tilde{\mathbf{R}}$ with the swept attribute first: $\tilde{\mathbf{R}} = \left[\begin{smallmatrix} 1 & \tilde{\mathbf{b}}^\top \\ \tilde{\mathbf{b}} & \tilde{\mathbf{Q}} \end{smallmatrix}\right]$. Transform the target value to the Gaussian space:

$$\tilde{\alpha} = \Phi^{-1}\big(\hat{F}_s(\alpha)\big). \tag{38}$$

In this space, the attributes are jointly Gaussian (by the copula assumption), so the conditional mean is exact:

$$\mathbb{E}[\tilde{\mathbf{c}}_{-s} \mid \tilde{c}_s = \tilde{\alpha}] = \tilde{\alpha}\,\tilde{\mathbf{b}}. \tag{39}$$

By Proposition 1 applied to $\tilde{\mathbf{R}}$, this choice minimizes ρ_{cond} in the transformed space to zero.

Step 3: Map back to the original space. Each companion value $\tilde{c}_k = \tilde{\alpha}\,\tilde{b}_k$ (for $k \neq s$) is a standard-normal quantile. Convert it to a probability and invert through the original empirical CDF:

$$c_k = \hat{F}_k^{-1}\Big(\Phi\big(\tilde{\alpha}\,\tilde{b}_k\big)\Big), \qquad k \neq s. \tag{40}$$

The complete COFAT-N query is therefore $\mathbf{c}^{\text{COFAT-N}} = \big(\alpha,\ c_1, \ldots, c_{s-1}, c_{s+1}, \ldots, c_d\big)$ with each companion given by (40).

Properties.

1. **Reduces to COFAT under Gaussian marginals.** If every marginal is already Gaussian, $\hat{F}_k = \Phi$ and the transform (37) is the identity, so $\tilde{\alpha} = \alpha$, $\tilde{\mathbf{b}} = \mathbf{b}$, and $c_k = \Phi^{-1}(\Phi(\alpha b_k)) = \alpha b_k$, recovering Eq. 5.
2. **Respects marginal support.** Because each companion is produced by inverting through $\hat{F}_k^{-1}$, COFAT-N values always lie within the observed range of attribute k, even for bounded or discrete attributes.
3. **Conditional infeasibility.** In the Gaussian copula space, $\rho_{\text{cond}} = 0$ (by Proposition 1). In the original space, the back-transformation is nonlinear, so $\rho_{\text{cond}} > 0$ in general. However, COFAT-N inherits the copula's conditional dependence structure, which typically yields substantially lower ρ_{cond} than OFAT-μ or other designs that ignore correlations. The residual infeasibility reflects the gap between the Gaussian copula model and the true dependence structure.
4. **COFAT-NS (stochastic variant).** Analogously to COFAT-S, one may sample from the full conditional in the Gaussian space, $\tilde{\mathbf{c}}_{-s} = \tilde{\alpha}\,\tilde{\mathbf{b}} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\, \tilde{\mathbf{Q}} - \tilde{\mathbf{b}}\tilde{\mathbf{b}}^\top)$, and map each draw back via (40). This preserves diversity while respecting the copula dependence.

Summary of the COFAT-N procedure.

1. Estimate empirical CDFs $\hat{F}_1, \ldots, \hat{F}_d$ from training data.
2. Transform all training samples to Gaussian scores via (37).
3. Estimate the correlation matrix $\tilde{\mathbf{R}}$ of the transformed data.
4. For a sweep of attribute s at value α: transform $\tilde{\alpha} = \Phi^{-1}(\hat{F}_s(\alpha))$, compute conditional mean $\tilde{c}_k = \tilde{\alpha}\,\tilde{b}_k$, and invert $c_k = \hat{F}_k^{-1}(\Phi(\tilde{c}_k))$ for all $k \neq s$.

A.3 Baselines

We evaluate COFAT, COFAT-S and COFAT-N against seven baselines. **OFAT-μ[13]**: $c_s = \alpha$. Fix other $d-1$ attributes at their population means (0 in z-units). **OFAT-RAND[36]**: $c_s = \alpha$. Independently sample other $d-1$ attributes from $\mathcal{N}(0,1)$, ignoring correlation structure R. **OFAT-MED[46]**: $c_s = \alpha$. Fix other $d-1$ attributes at their population medians. **FREEZE[44]**: $c_s = \alpha$. Freeze other attributes (except c_s) in S_r, S_s, S_m at baseline (0). Companion attributes are obtained via COFAT 3.2. **K-MEANS[56]**: $c_s = \alpha$. Select the k-means centroid (default $k = 3$) closest to α in coordinate s, copy its vector, and overwrite $c_s = \alpha$. Re-select the centroid for each target percentile at every sweep. **DROP-S[35]**: $c_s = 0$. Set the remaining attributes as $\alpha R_{:,s}$. This tests whether obedience depends on swept attribute alone or also arises from the companions. **KNN[8]**: Select $N = 5$ real training samples $\mathbf{z}^*$ closest to c_s and average their vectors. Perform a fresh lookup at each sweep percentile.

A.4 Additional Figures

We provide some additional figures for reference 9 and 10.

A.5 Additional tables

Overview of appendix tables. The appendix tables are grouped by experiment: Exp. A (radiomic appearance, Tables 5, 6, 7, 8), Exp. B (shape, Tables 9, 10, 11, 12, 13, 14), and Exp. C (molecules, Tables 15, 16, 17, 18, 19). Each group follows the same order. (i) A *redundancy table* lists R_s^2 for every attribute and checks the closed forms of Sec. 3 against the queries actually passed to the generator: $\rho/|\alpha|$ equals 1 for COFAT and $1/\sqrt{1-R_s^2}$ for OFAT-μ, computed from the training correlation matrix alone, before any sample is generated. (ii) A *query-side table* reports ρ and ρ_{cond} of every design at each sweep level. (iii) *Full result tables*, in the format of Tables 1–3, cover the targets not shown in the main text: `glcm_contrast` and `glrlm_SRE` (Exp. A); `elongation`, `flatness`, and `irregularity` (Exp. B); and `BertzCT` and `MolMR` (Exp. C). For Exp. C,

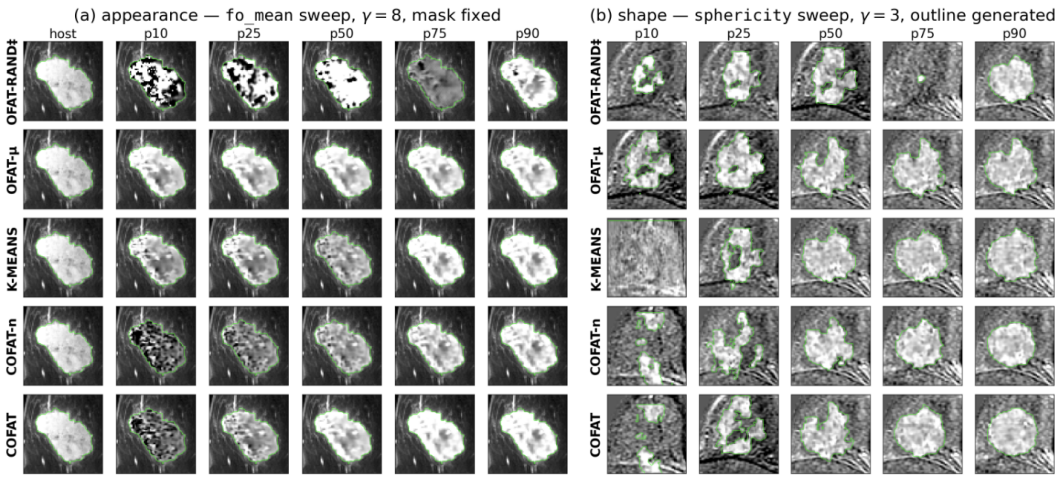

Figure 9: Qualitative results across all designs. (a) Appearance control: 3D-lesion inpainting within a fixed test mask, sweeping *fo_mean*. (b) Shape control: 3D-lesion insertion into a fixed healthy breast MRI patch, sweeping *sphericity*.

Table 5: **Radiomics: redundancy and both closed forms.** For each of the three swept attributes: how redundant it is with its 33 companions (R_s^2) *closed form* evaluates the proposition 1 ; *observed* builds the actual 34-vector at each of the five percentile levels, measures $\rho = \sqrt{\mathbf{c}^\top \mathbf{R}^{-1}\mathbf{c}}$, divides by the demand $|\alpha|$, and averages the five ratios — exactly the code path the generator's own queries go through, clipping to the model's ± 5 input range included. Agreement between the two columns of each pair is the check. The first pair is the floor: the best possible request is exactly as implausible as the demand itself, $\rho(\mathbf{c}^\star)/|\alpha| = 1$, and stays there no matter how redundant the attribute becomes. The second pair is the penalty for freezing the companions instead, $1/\sqrt{1-R_s^2}$, which is the whole effect size of the correction — from $12\times$ on `glcm_contrast` to $79\times$ on `fo_mean`. Nothing here is measured off a generated volume; it is arithmetic on the correlation matrix of the 1080 real training lesions. R_s^2 is given to six decimals.

		$\rho(\mathbf{c}^\star)/\lvert\alpha\rvert$		$\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert$	
attribute	R_s^2	closed form	observed	closed form	observed
`fo_mean`	0.999839	1.00000	1.00000	78.90919	78.90919
`glcm_contrast`	0.993096	1.00000	1.00000	12.03537	12.03537
`glrlm_SRE`	0.996069	1.00000	1.00000	15.94984	15.94984

Table 19 additionally reports the scaffold-anchored sweep per scaffold and descriptor, including how often the scaffold is kept and extended.

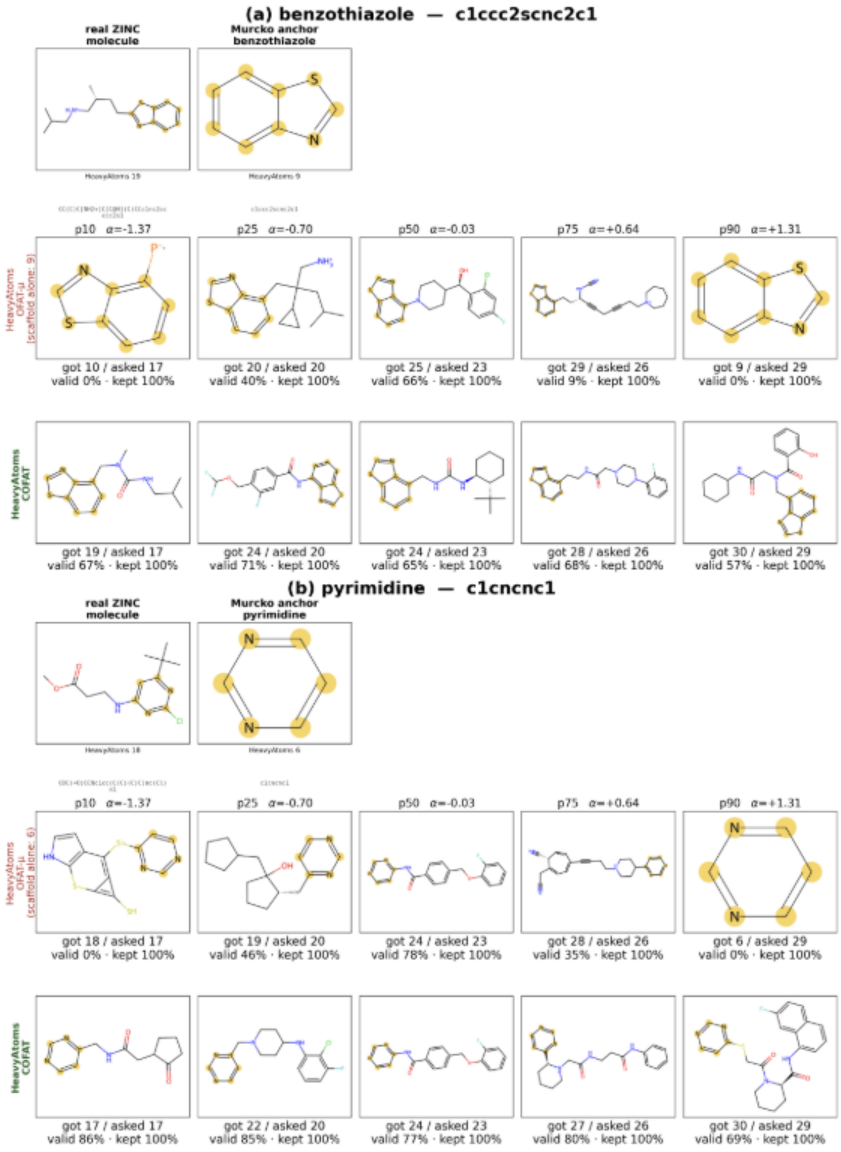

Figure 10: **Scaffold-Anchored Sweep.** A real ZINC molecule defines a fixed Bemis–Murcko scaffold, which is used as the prefix while HeavyAtoms is swept across five percentiles for 2 scaffolds. COFAT remains obedient across the sweep, increasing the number of non-hydrogen atoms with the requested target, whereas OFAT-μ underestimates the target at the extremes (e.g., 10 vs. 17 at p_{10} and 9 vs. 29 at p_{90}). COFAT also maintains $> 50\%$ validity at every level, while OFAT-μ exceeds 50% only at the median level, demonstrating the advantage of COFAT in jointly preserving obedience and realism.

Table 6: Radiomics, query side: infeasibility of each design at each sweep level. Lower ρ indicates lower request cost, while lower ρ_{cond} indicates less infeasibility added by the design. The row order follows the main results table.

	$\rho \downarrow$ *request cost*						$\rho_{\text{cond}} \downarrow$ *design contribution*					
Design	p10	p25	p50	p75	p90	mean	p10	p25	p50	p75	p90	mean
fo_mean $R_s^2 = 0.999839$, $\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert = 78.91$												
requested α	-1.16	-0.68	-0.14	$+0.58$	$+1.43$	0.80	—	—	—	—	—	—
OFAT-RAND[36]	182.0	314.7	169.5	124.3	159.9	190.06	182.0	314.7	169.5	124.3	159.9	190.05
OFAT-MED[46]	79.8	41.8	3.21	57.3	124.6	61.33	79.8	41.8	3.21	57.3	124.6	61.32
OFAT-μ[24]	91.3	53.3	10.8	45.6	112.9	62.78	91.3	53.3	10.8	45.6	112.9	62.77
DROP-c_s[35]	91.3	53.3	10.8	45.6	112.9	62.77	91.3	53.3	10.8	45.6	112.9	62.77
FREEZE[44]	1.17	**0.68**	**0.14**	**0.58**	1.44	**0.80**	0.15	0.09	0.02	0.07	0.18	0.10
K-MEANS[56]	56.6	18.7	18.1	65.6	2.29	32.26	56.6	18.6	18.1	65.6	1.79	32.16
KNN[8]	2.02	1.29	1.01	1.17	2.06	1.51	1.65	1.10	1.00	1.02	1.48	1.25
COFAT (ours)	**1.16**	**0.68**	**0.14**	**0.58**	**1.43**	**0.80**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N (ours)	3.15	3.57	3.21	3.37	8.84	4.43	2.93	3.51	3.21	3.32	8.72	4.34
COFAT-S (ours)	4.79	5.54	5.07	5.82	4.89	5.22	4.65	5.50	5.06	5.79	4.68	5.14
glcm_contrast $R_s^2 = 0.993096$, $\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert = 12.04$												
requested α	-1.16	-0.68	-0.11	$+0.52$	$+1.27$	0.75	—	—	—	—	—	—
OFAT-RAND[36]	310.5	152.2	164.8	113.7	127.3	173.70	310.5	152.2	164.8	113.7	127.3	173.69
OFAT-MED[46]	12.9	7.50	3.21	8.35	17.0	9.79	12.9	7.47	3.21	8.33	16.9	9.76
OFAT-μ[24]	13.9	8.20	1.34	6.29	15.2	9.00	13.9	8.17	1.33	6.27	15.2	8.96
DROP-c_s[35]	13.9	8.17	1.33	6.27	15.2	8.96	13.9	8.17	1.33	6.27	15.2	8.96
FREEZE[44]	**1.16**	**0.68**	**0.11**	**0.52**	**1.27**	**0.75**	0.11	0.06	0.01	0.05	0.12	0.07
K-MEANS[56]	10.2	4.60	2.11	3.43	6.59	5.38	10.1	4.55	2.11	3.39	6.47	5.32
KNN[8]	1.47	1.20	1.26	1.19	1.82	1.39	0.92	0.99	1.26	1.07	1.31	1.11
COFAT (ours)	**1.16**	**0.68**	**0.11**	**0.52**	**1.27**	**0.75**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N (ours)	4.19	3.53	3.21	6.24	2.87	4.01	4.03	3.47	3.21	6.22	2.58	3.90
COFAT-S (ours)	5.56	6.03	5.48	5.93	6.76	5.95	5.44	5.99	5.48	5.90	6.64	5.89
glrlm_SRE $R_s^2 = 0.996069$, $\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert = 15.95$												
requested α	-1.19	-0.55	$+0.09$	$+0.65$	$+1.20$	0.74	—	—	—	—	—	—
OFAT-RAND[36]	451.7	258.4	310.7	160.0	386.3	313.40	451.7	258.4	310.7	160.0	386.3	313.40
OFAT-MED[46]	20.7	10.9	3.21	9.37	17.8	12.40	20.7	10.8	3.21	9.34	17.8	12.38
OFAT-μ[24]	19.0	8.85	1.47	10.3	19.1	11.74	18.9	8.83	1.47	10.3	19.0	11.71
DROP-c_s[35]	18.9	8.83	1.47	10.3	19.0	11.71	18.9	8.83	1.47	10.3	19.0	11.71
FREEZE[44]	1.20	0.56	**0.09**	**0.65**	1.20	**0.74**	0.14	0.07	0.01	0.08	0.14	0.09
K-MEANS[56]	8.38	2.72	1.58	5.41	14.1	6.44	8.30	2.66	1.58	5.37	14.1	6.39
KNN[8]	3.58	2.02	1.31	1.69	3.50	2.42	3.38	1.94	1.31	1.57	3.28	2.30
COFAT (ours)	**1.19**	**0.55**	**0.09**	**0.65**	**1.20**	**0.74**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N (ours)	6.77	3.15	3.21	4.52	5.18	4.57	6.66	3.10	3.21	4.48	5.04	4.50
COFAT-S (ours)	6.43	5.47	5.14	6.04	5.44	5.71	6.32	5.44	5.14	6.01	5.31	5.64

Table 7: Quantitative results for appearance target `glcm_contrast` (Exp. A). **Bold** = best in column.

	Feasibility			Obedience			Realism		Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%↓	slope$\to 1$	bias$\to 0$	CCC↑	off-M.%$\to 5$	$P_{\text{app}} \to .68$	copy%↓
OFAT-RAND[36]	173.7	173.7	60	+.65	+.48	.22	*79.1*	.15	**0.9**
OFAT-MED[46]	9.8	9.8	**0**	+.29	+.42	.43	3.3	.69	34.9
OFAT-μ[24]	9.0	9.0	**0**	+.15	-.33	.25	4.5	.61	29.5
DROP-c_s[35]	9.0	9.0	**0**	+.42	-.20	.68	4.5	.62	29.2
FREEZE[44]	**0.8**	0.07	**0**	**+.69**	-.20	**.90**	*8.4*	.61	28.6
K-MEANS[56]	5.4	5.3	**0**	+.34	**+.01**	.49	4.0	.65	30.1
KNN[8]	1.4	1.1	**0**	+.46	+.16	.68	3.0	.66	33.2
COFAT (ours)	**0.8**	**0.0**	**0**	+.60	-.20	.85	**4.9**	.62	28.1
COFAT-N (ours)	4.0	3.9	**0**	+.44	+.45	.59	3.4	**.69**	31.3
COFAT-S (ours)	6.0	5.9	100	+1.72	-.10	.63	*88.3*	.12	**0.9**

Table 8: Quantitative results for appearance target `glrlm_SRE` (Exp. A). **Bold** = best in column.

	Feasibility			Obedience			Realism		Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%↓	slope→1	bias→0	CCC↑	off-M.%→5	$P_{\text{app}}\to.68$	copy%↓
OFAT-RAND[36]	313.4	313.4	100	-.60	-.40	-.41	*96.6*	.13	**1.6**
OFAT-MED[46]	12.4	12.4	**0**	-.70	+1.03	-.49	*16.5*	.46	20.4
OFAT-μ[24]	11.7	11.7	**0**	-.11	+.35	-.19	*25.9*	.47	23.3
DROP-c_s[35]	11.7	11.7	80	+1.32	+.87	.64	*30.1*	.41	16.5
FREEZE[44]	**0.7**	0.09	80	+.88	+.35	.89	*20.1*	.53	24.7
K-MEANS[56]	6.4	6.4	**0**	+.16	+.31	.28	*9.5*	.56	22.6
KNN[8]	2.4	2.3	**0**	+.61	+.32	.78	**3.8**	.60	30.9
COFAT (ours)	**0.7**	**0.0**	80	+.78	**+.05**	**.96**	**3.8**	.62	32.7
COFAT-N (ours)	4.6	4.5	**0**	+.40	+.54	.52	3.5	**.68**	33.7
COFAT-S (ours)	5.7	5.6	80	**+.93**	-.08	.64	*57.3*	.25	5.6

Table 9: **Shape: redundancy and both closed forms.** The four dials the shape flow is conditioned on. `volume` is omitted: it is not a conditioning input — it is supplied at each query and is fixed at all percentiles. *closed form* evaluates the proposition 1; *observed* builds the actual five-dial vector at each of the five percentile levels the generator sweeps, measures $\rho = \sqrt{\mathbf{c}^\top \mathbf{R}^{-1}\mathbf{c}}$, divides by the demand $|\alpha|$, and averages the five ratios. The first pair is the floor, $\rho(\mathbf{c}^\star)/|\alpha| = 1$ however redundant the dial; the second is the penalty for freezing the other four, $1/\sqrt{1-R_s^2}$, here from $1.49\times$ on `elongation` to $2.85\times$ on `sphericity`. These are not the 34-attribute numbers: scored against the full radiomics vocabulary the ratios are far larger, because the redundancy is carried by texture companions this five-dial space does not contain. Nothing here is measured off a generated volume.

		$\rho(\mathbf{c}^\star)/\lvert\alpha\rvert$		$\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert$	
attribute	R_s^2	closed form	observed	closed form	observed
`sphericity`	0.877251	1.00000	1.00000	2.85424	2.85424
`irregularity`	0.856781	1.00000	1.00000	2.64241	2.64241
`flatness`	0.665493	1.00000	1.00000	1.72901	1.72901
`elongation`	0.551095	1.00000	1.00000	1.49253	1.49253

Table 10: **Shape, query side: infeasibility across sweep levels.** For `sphericity` and `elongation`, ρ is total request cost and ρ_{cond} is the design contribution in Mahalanobis units.

	$\rho \downarrow$ — the whole request						$\rho_{\text{cond}} \downarrow$ — what the design added					
design	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}	mean	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}	mean
swept: `sphericity` $R_s^2 = 0.8773$ *(4-dial)*, $\alpha \in [-1.49, 1.19]$												
OFAT-RAND[36]	3.95	2.70	3.71	5.07	3.60	3.81	3.65	2.61	3.71	5.01	3.39	3.68
OFAT-MED[46]	5.17	2.93	0.63	1.39	2.56	2.53	4.95	2.84	0.60	1.16	2.26	2.36
OFAT-μ[24]	4.25	1.99	0.51	2.19	3.40	2.47	3.98	1.87	0.48	2.05	3.19	2.31
DROP-c_s[35]	3.98	1.87	0.48	2.05	3.19	2.31	3.98	1.87	0.48	2.05	3.19	2.31
FREEZE[44]	1.49	0.70	0.18	0.77	1.19	0.87	0.00	0.00	0.00	0.00	0.00	0.00
K-MEANS[56]	2.22	1.70	1.33	0.92	1.92	1.62	1.64	1.55	1.32	0.51	1.51	1.31
KNN[8]	1.61	1.33	0.57	0.81	1.46	1.15	0.62	1.13	0.54	0.25	0.84	0.68
COFAT	**1.49**	**0.70**	**0.18**	**0.77**	**1.19**	**0.87**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.56	1.25	0.63	0.86	1.46	1.15	0.47	1.04	0.60	0.39	0.84	0.67
COFAT-S	1.79	1.81	1.16	1.64	2.54	1.79	1.00	1.67	1.15	1.45	2.24	1.50
swept: `elongation` $R_s^2 = 0.5511$ *(4-dial)*, $\alpha \in [-1.42, 1.17]$												
OFAT-RAND[36]	4.96	2.37	3.93	3.35	2.20	3.36	4.75	2.29	3.92	3.26	1.87	3.22
OFAT-MED[46]	2.24	1.11	0.63	1.21	1.80	1.40	1.73	0.93	0.60	0.96	1.37	1.12
OFAT-μ[24]	2.12	0.91	0.28	1.10	1.75	1.23	1.58	0.67	0.21	0.82	1.30	0.91
DROP-c_s[35]	1.58	0.67	0.21	0.82	1.30	0.91	1.58	0.67	0.21	0.82	1.30	0.91
FREEZE[44]	1.42	0.61	0.19	0.74	1.17	0.83	0.00	0.00	0.00	0.00	0.00	0.00
K-MEANS[56]	1.48	1.12	2.15	0.83	1.27	1.37	0.39	0.94	2.14	0.37	0.49	0.87
KNN[8]	1.46	0.66	0.29	0.92	1.19	0.90	0.33	0.26	0.23	0.54	0.18	0.31
COFAT	**1.42**	**0.61**	**0.19**	**0.74**	**1.17**	**0.83**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.55	0.89	0.63	0.90	1.25	1.04	0.61	0.65	0.60	0.51	0.43	0.56
COFAT-S	1.80	1.77	1.74	0.83	2.93	1.81	1.10	1.66	1.73	0.38	2.69	1.51

Table 11: **Shape, query side: infeasibility across sweep levels.** For `flatness` and `irregularity`, ρ is total request cost and ρ_{cond} is the design contribution in Mahalanobis units.

	$\rho\ \downarrow$ — the whole request						$\rho_{\text{cond}}\ \downarrow$ — what the design added					
design	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}	mean	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}	mean
swept: `flatness` $R_s^2 = 0.6655$ *(4-dial)*, $\alpha \in [-1.38, 1.23]$												
OFAT-RAND[36]	3.83	2.71	6.04	6.29	3.88	4.55	3.57	2.61	6.04	6.25	3.68	4.43
OFAT-MED[46]	2.62	1.55	0.63	1.33	2.07	1.64	2.22	1.36	0.62	1.09	1.67	1.39
OFAT-μ[24]	2.39	1.26	0.14	1.32	2.13	1.45	1.95	1.03	0.11	1.08	1.73	1.18
DROP-c_s[35]	1.95	1.03	0.11	1.08	1.73	1.18	1.95	1.03	0.11	1.08	1.73	1.18
FREEZE[44]	1.38	0.73	0.08	0.76	1.23	0.84	0.00	0.00	0.00	0.00	0.00	0.00
K-MEANS[56]	1.50	2.08	1.10	0.80	1.36	1.37	0.58	1.94	1.09	0.24	0.59	0.89
KNN[8]	1.41	0.86	0.26	0.86	1.31	0.94	0.28	0.46	0.24	0.40	0.46	0.37
COFAT	**1.38**	**0.73**	**0.08**	**0.76**	**1.23**	**0.84**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.54	1.06	0.63	0.87	1.26	1.07	0.69	0.77	0.62	0.42	0.27	0.56
COFAT-S	2.99	2.23	2.95	3.70	1.62	2.70	2.65	2.11	2.95	3.62	1.06	2.48
swept: `irregularity` $R_s^2 = 0.8568$ *(4-dial)*, $\alpha \in [-0.80, 1.27]$												
OFAT-RAND[36]	3.37	2.34	1.68	4.05	3.24	2.93	3.27	2.25	1.64	4.04	2.98	2.84
OFAT-MED[46]	1.72	1.30	0.63	1.24	3.80	1.74	1.52	1.14	0.52	1.21	3.59	1.60
OFAT-μ[24]	2.11	1.68	0.92	0.74	3.35	1.76	1.95	1.55	0.86	0.68	3.10	1.63
DROP-c_s[35]	1.95	1.55	0.86	0.68	3.10	1.63	1.95	1.55	0.86	0.68	3.10	1.63
FREEZE[44]	0.80	0.63	0.35	0.28	1.27	0.67	0.00	0.00	0.00	0.00	0.00	0.00
K-MEANS[56]	1.04	0.80	0.87	1.11	2.13	1.19	0.67	0.48	0.79	1.07	1.71	0.95
KNN[8]	1.46	0.79	0.58	1.36	1.59	1.16	1.22	0.48	0.46	1.33	0.96	0.89
COFAT	**0.80**	**0.63**	**0.35**	**0.28**	**1.27**	**0.67**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.46	0.86	0.63	1.25	1.56	1.15	1.22	0.58	0.52	1.22	0.91	0.89
COFAT-S	1.78	1.96	1.11	1.76	1.88	1.70	1.59	1.85	1.05	1.74	1.39	1.52

Table 12: Quantitative results for shape target `elongation` (Exp. B). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\uparrow$	bias$\downarrow$	CCC$\uparrow$	drift$\downarrow$	off-M.%	copy%$\downarrow$
OFAT-RAND[36]	3.36	3.22	20	.00	.50	.00	0.66	21.2	**18.2**
OFAT-MED[46]	1.40	1.12	0	.11	.72	.14	0.39	0.0	59.0
OFAT-μ[24]	1.23	0.91	0	.12	.61	.16	0.40	0.0	50.8
DROP-c_s[35]	0.91	0.91	0	.30	.60	.40	0.39	0.0	48.0
K-MEANS[56]	1.37	0.87	0	.42	.46	.58	0.45	0.0	54.8
KNN[8]	0.90	0.31	0	.42	.58	.54	**0.38**	0.0	57.2
COFAT (ours)	**0.83**	**0.00**	0	.41	.58	.53	0.39	0.0	47.2
COFAT-N (ours)	1.04	0.56	0	.43	.64	.52	**0.38**	0.0	56.4
COFAT-S (ours)	1.81	1.51	0	**.54**	**.43**	**.72**	0.53	0.0	55.6

Table 13: Quantitative results for shape target `flatness` (Exp. B). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\uparrow$	bias$\downarrow$	CCC$\uparrow$	drift$\downarrow$	off-M.%	copy%$\downarrow$
OFAT-RAND[36]	4.55	4.43	0	.40	**.78**	.41	0.74	0.0	36.4
OFAT-MED[46]	1.64	1.39	0	.40	.97	.36	0.32	0.0	55.8
OFAT-μ[24]	1.45	1.18	0	.41	.82	.43	0.36	0.0	50.4
DROP-c_s[35]	1.18	1.18	0	.30	.88	.31	0.37	0.0	56.0
K-MEANS[56]	1.37	0.89	0	.67	.85	.59	0.36	0.0	52.0
KNN[8]	0.94	0.37	0	**.68**	.85	.60	0.32	0.0	53.6
COFAT (ours)	**0.84**	**0.00**	0	**.68**	.80	**.62**	0.35	0.0	54.8
COFAT-N (ours)	1.07	0.56	0	.66	.88	.57	**0.31**	0.0	55.4
COFAT-S (ours)	2.70	2.48	0	.63	.96	.50	0.65	10.6	**31.4**

Table 14: Quantitative results for shape target `irregularity` (Exp. B). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\uparrow$	bias$\downarrow$	CCC$\uparrow$	drift$\downarrow$	off-M.%	copy%$\downarrow$
OFAT-RAND[36]	2.93	2.84	80	.45	-.34	.65	0.64	0.0	58.6
OFAT-MED[46]	1.74	1.60	80	.36	-.39	.53	0.52	0.0	53.6
OFAT-μ[24]	1.76	1.63	100	.37	-.35	.56	0.56	0.0	52.4
DROP-c_s[35]	1.63	1.63	100	.22	-.34	.37	0.50	0.0	54.6
K-MEANS[56]	1.19	0.95	100	**.67**	-.29	**.84**	0.48	0.0	47.2
KNN[8]	1.16	0.89	**0**	.64	**-.28**	**.84**	0.51	0.0	50.2
COFAT (ours)	**0.67**	**0.00**	100	.56	-.32	.76	0.51	0.0	49.0
COFAT-N (ours)	1.15	0.89	**0**	.64	-.29	.83	**0.47**	0.0	49.6
COFAT-S (ours)	1.70	1.52	80	.54	-.38	.71	0.83	0.0	**46.6**

Table 15: **Molecules: redundancy and both closed forms.** All thirteen ZINC descriptors the decoder is conditioned on, ordered by redundancy. *closed* evaluates the proposition 1; *observed* builds the actual 13-vector at each of the five percentile levels, measures $\rho = \sqrt{\mathbf{c}^\top \mathbf{R}^{-1}\mathbf{c}}$, divides by the demand $|\alpha|$, and averages the five ratios. The first pair is the floor, $\rho(\mathbf{c}^\star)/|\alpha| = 1$ for every descriptor however redundant; the second is the penalty for freezing the companions, $1/\sqrt{1-R_s^2}$. This is the domain where the law makes a *within-domain, quantitative* prediction rather than a demonstration: molecular descriptors are not uniformly entangled the way radiomics features are — redundancy here ranges from 0.97 on `HeavyAtoms` down to 0.25 on `QED` — so the ordering of this column predicts, from the correlation matrix alone and before a single molecule is generated, which dials the frozen query will break. The three swept dials (`HeavyAtoms`, `BertzCT`, `MolMR`) are the three most redundant. Nothing here is measured off a generated molecule; it is arithmetic on the correlation matrix of the $94{,}513$ ZINC training molecules.

		$\rho(\mathbf{c}^\star)/\|\alpha\|$		$\rho(\mathbf{c}^{\text{OFAT-}\mu})/\|\alpha\|$	
attribute	R_s^2	closed form	observed	closed form	observed
`HeavyAtoms`	0.9656	1.0000	1.0000	5.3912	5.3912
`BertzCT`	0.9641	1.0000	1.0000	5.2756	5.2756
`MolMR`	0.9516	1.0000	1.0000	4.5460	4.5460
`MolWt`	0.9129	1.0000	1.0000	3.3884	3.3884
`NumAromRings`	0.9116	1.0000	1.0000	3.3640	3.3640
`TPSA`	0.8239	1.0000	1.0000	2.3830	2.3830
`NumHAcceptors`	0.7605	1.0000	1.0000	2.0433	2.0433
`FracCSP3`	0.7538	1.0000	1.0000	2.0154	2.0154
`NumAliphRings`	0.6952	1.0000	1.0000	1.8113	1.8113
`MolLogP`	0.5821	1.0000	1.0000	1.5469	1.5469
`NumRotBonds`	0.5549	1.0000	1.0000	1.4988	1.4988
`NumHDonors`	0.5394	1.0000	1.0000	1.4734	1.4734
`QED`	0.2517	1.0000	1.0000	1.1560	1.1560

Table 16: **Molecules, query side: infeasibility across sweep levels.** For `HeavyAtoms`, `BertzCT`, and `MolMR`, ρ is total request cost and ρ_{cond} is the design contribution in Mahalanobis units.

	$\rho\ \downarrow$ — the whole request						$\rho_{\text{cond}}\ \downarrow$ — what the design added					
design	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}	mean	p_{10}	p_{25}	p_{50}	p_{75}	p_{90}	mean
HeavyAtoms $R_s^2 = 0.9656$, $\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert = 5.39$, $\alpha \in [-1.37, 1.31]$												
OFAT-RAND[36]	8.90	17.73	16.78	17.88	9.50	14.16	8.80	17.72	16.78	17.87	9.41	14.11
OFAT-MED[46]	7.60	4.06	1.15	3.50	7.02	4.67	7.47	4.00	1.15	3.44	6.90	4.59
OFAT-μ[24]	7.36	3.75	0.14	3.48	7.09	4.36	7.23	3.68	0.13	3.42	6.97	4.29
DROP-c_s[35]	7.23	3.68	0.13	3.42	6.97	4.29	7.23	3.68	0.13	3.42	6.97	4.29
FREEZE[44]	5.49	2.80	0.10	2.59	5.29	3.26	5.32	2.71	0.10	2.51	5.13	3.15
K-MEANS[56]	2.14	2.26	0.67	3.26	1.62	1.99	1.65	2.15	0.67	3.20	0.94	1.72
KNN[8]	1.41	0.75	0.44	0.70	1.39	0.94	0.37	0.27	0.43	0.27	0.45	0.36
COFAT	**1.37**	**0.70**	**0.03**	**0.64**	**1.31**	**0.81**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.84	2.69	1.18	1.49	1.90	1.82	1.23	2.60	1.18	1.35	1.37	1.55
COFAT-S	4.65	2.96	2.30	2.55	3.51	3.20	4.45	2.88	2.30	2.47	3.25	3.07
BertzCT $R_s^2 = 0.9641$, $\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert = 5.28$, $\alpha \in [-1.25, 1.34]$												
OFAT-RAND[36]	13.40	14.39	7.42	7.23	12.24	10.93	13.34	14.37	7.42	7.20	12.16	10.90
OFAT-MED[46]	6.93	4.14	1.15	3.22	6.88	4.46	6.81	4.08	1.15	3.16	6.75	4.39
OFAT-μ[24]	6.60	3.76	0.32	3.32	7.07	4.21	6.48	3.69	0.31	3.26	6.94	4.14
DROP-c_s[35]	6.48	3.69	0.31	3.26	6.94	4.14	6.48	3.69	0.31	3.26	6.94	4.14
FREEZE[44]	2.22	1.26	0.11	1.12	2.38	1.42	1.83	1.04	0.09	0.92	1.96	1.17
K-MEANS[56]	1.43	2.49	0.78	3.13	1.48	1.86	0.69	2.39	0.78	3.07	0.63	1.51
KNN[8]	1.34	0.77	0.63	0.71	1.41	0.97	0.48	0.28	0.63	0.32	0.43	0.43
COFAT	**1.25**	**0.71**	**0.06**	**0.63**	**1.34**	**0.80**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.46	1.29	1.15	1.51	1.61	1.40	0.76	1.08	1.15	1.37	0.89	1.05
COFAT-S	3.22	4.20	2.86	3.37	3.37	3.40	2.97	4.14	2.86	3.31	3.09	3.27
MolMR $R_s^2 = 0.9516$, $\rho(\mathbf{c}^{\text{OFAT-}\mu})/\lvert\alpha\rvert = 4.55$, $\alpha \in [-1.28, 1.28]$												
OFAT-RAND[36]	12.83	9.79	8.71	13.35	12.21	11.38	12.77	9.77	8.71	13.33	12.14	11.34
OFAT-MED[46]	5.80	3.13	1.15	3.27	6.02	3.87	5.66	3.06	1.15	3.21	5.89	3.79
OFAT-μ[24]	5.80	3.02	0.01	2.95	5.80	3.52	5.66	2.95	0.01	2.88	5.66	3.43
DROP-c_s[35]	5.66	2.95	0.01	2.88	5.66	3.43	5.66	2.95	0.01	2.88	5.66	3.43
FREEZE[44]	2.26	1.18	0.01	1.15	2.26	1.37	1.87	0.97	0.00	0.95	1.87	1.13
K-MEANS[56]	1.76	1.99	0.59	2.64	1.60	1.71	1.21	1.87	0.59	2.56	0.96	1.44
KNN[8]	1.31	0.72	0.43	0.70	1.33	0.89	0.29	0.26	0.43	0.25	0.36	0.32
COFAT	**1.28**	**0.67**	**0.00**	**0.65**	**1.28**	**0.77**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
COFAT-N	1.55	1.57	1.15	1.11	2.62	1.60	0.88	1.42	1.15	0.89	2.28	1.32
COFAT-S	4.26	2.36	3.96	2.51	3.55	3.33	4.07	2.26	3.96	2.42	3.31	3.20

Table 17: Quantitative results for molecule target `BertzCT` (Exp. C). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%$\downarrow$	slope$\to 1$	bias$\to 0$	CCC$\uparrow$	lv	valid% (p_{10}/p_{90})$\uparrow$	copy%$\downarrow$
OFAT-RAND[36]	10.9	10.9	80	—	—	—	*0/5*	*0/0*	—
OFAT-MED[46]	4.5	4.4	0	+.10	-.10	.16	*4/5*	*2/6*	11.2
OFAT-μ[24]	4.2	4.1	0	-.00	-.47	-.00	5/5	5/8	11.2
DROP-c_s[35]	4.1	4.1	0	+.65	-.24	.87	5/5	9/6	13.3
FREEZE[44]	1.7	1.5	0	+.97	**-.00**	**.99**	5/5	91/32	14.4
K-MEANS[56]	1.9	1.5	0	+.90	**+.00**	.97	5/5	89/58	13.7
KNN[8]	1.0	0.4	0	+.94	+.13	**.99**	5/5	88/61	13.8
COFAT (ours)	**0.8**	**0.0**	0	+.93	+.06	**.99**	5/5	**93**/63	15.5
COFAT-N (ours)	1.4	1.1	0	**+.98**	+.09	**.99**	5/5	90/**64**	14.1
COFAT-S (ours)	3.4	3.3	0	+.94	-.10	.98	5/5	65/24	**8.9**

Table 18: Quantitative results for molecule target `MolMR` (Exp. C). **Bold** = best in column.

	Feasibility			Obedience				Realism	Novelty
Design	$\bar{\rho}\downarrow$	$\bar{\rho}_c\downarrow$	An.%↓	slope$\to 1$	bias$\to 0$	CCC↑	lv	valid% (p_{10}/p_{90})↑	copy%↓
OFAT-RAND[36]	11.4	11.3	100	—	—	—	*0/5*	*0/0*	—
OFAT-MED[46]	3.9	3.8	40	+.56	-.13	.81	*4/5*	*46/0*	8.1
OFAT-μ[24]	3.5	3.4	40	+.38	-.33	.64	*4/5*	*41/0*	8.4
DROP-c_s[35]	3.4	3.4	40	+.25	-.13	.37	*4/5*	*0/39*	8.2
FREEZE[44]	2.7	2.6	40	+.81	**+.02**	.97	5/5	44/6	**9.2**
K-MEANS[56]	1.7	1.4	**0**	+.94	+.12	**.99**	5/5	80/60	10.3
KNN[8]	0.9	0.3	**0**	+.95	+.10	**.99**	5/5	82/63	12.5
COFAT (ours)	**0.8**	**0.0**	**0**	**+.98**	+.07	**.99**	5/5	82/70	14.4
COFAT-N (ours)	1.6	1.3	**0**	+.95	+.13	**.99**	5/5	**91/76**	15.2
COFAT-S (ours)	3.3	3.2	**0**	+.85	+.05	.94	5/5	49/51	11.0

Table 19: **Molecules, Arm 2: scaffold-anchored controllability across three descriptors.** Obedience, validity, and anchor retention are reported for `HeavyAtoms`, `BertzCT`, and `MolMR`.

	obedience			realism	anchor honoured?	
design	slope $\to 1$	bias $\to 0$	CCC $\uparrow$	valid % $\uparrow$	kept %	grew %
HeavyAtoms						
anchor `C1CCNCC1`						
COFAT	+0.98	+0.07	**0.99** [0.97,0.99]	**29**	100	98
OFAT-μ[24]	+1.98	-0.26	0.76 [0.66,0.84]	8	80	56
anchor `c1cncnc1` *(CCC indistinguishable)*						
COFAT	+1.00	+0.15	0.98 [0.97,0.99]	**79**	100	100
OFAT-μ[24]	+1.21	+0.06	0.98 [0.94,1.00]	32	100	79
anchor `C1COCCN1` *(CCC indistinguishable)*						
COFAT	+0.91	+0.21	0.96 [0.95,0.97]	**60**	100	100
OFAT-μ[24]	+1.40	+0.07	0.94 [0.90,0.98]	17	80	66
anchor `c1ccccc1` *(CCC indistinguishable)*						
COFAT	+0.97	+0.25	0.95 [0.93,0.96]	**71**	100	100
OFAT-μ[24]	+1.38	+0.19	0.92 [0.88,0.94]	26	80	60
anchor `c1ccc2scnc2c1` *(CCC indistinguishable)*						
COFAT	+0.85	+0.28	0.92 [0.90,0.93]	**65**	100	95
OFAT-μ[24]	+0.67	-0.07	0.91 [0.30,0.99]	23	100	73
BertzCT						
anchor `C1CCNCC1`						
COFAT	+0.82	-0.16	**0.92** [0.88,0.95]	**26**	100	99
OFAT-μ[24]	-0.05	-0.88	-0.05 [-0.21,0.12]	16	100	93
anchor `c1cncnc1`						
COFAT	+0.89	+0.01	**0.98** [0.98,0.99]	**84**	100	100
OFAT-μ[24]	+0.17	-0.55	0.20 [0.06,0.35]	45	100	95
anchor `C1COCCN1`						
COFAT	+0.90	-0.02	**0.98** [0.97,0.99]	**62**	100	100
OFAT-μ[24]	-0.07	-0.69	-0.08 [-0.30,0.17]	25	100	98
anchor `c1ccccc1`						
COFAT	+0.90	+0.08	**0.98** [0.97,0.99]	**76**	100	100
OFAT-μ[24]	-0.04	-0.60	-0.05 [-0.15,0.08]	41	100	99
anchor `c1ccc2scnc2c1`						
COFAT	+0.68	+0.52	**0.75** [0.73,0.77]	**58**	100	96
OFAT-μ[24]	-0.03	-0.25	-0.05 [-0.17,0.10]	35	100	87
MolMR						
anchor `C1CCNCC1`						
COFAT	+0.97	+0.03	**0.99** [0.97,0.99]	**25**	100	98
OFAT-μ[24]	+0.40	-0.32	0.58 [0.35,0.75]	16	80	78
anchor `c1cncnc1`						
COFAT	+1.00	+0.03	**0.99** [0.99,1.00]	**80**	100	100
OFAT-μ[24]	+0.33	-0.24	0.58 [0.46,0.68]	49	100	100
anchor `C1COCCN1`						
COFAT	+0.93	+0.08	**0.98** [0.97,0.99]	**59**	100	100
OFAT-μ[24]	+0.36	-0.28	0.61 [0.44,0.75]	30	100	100
anchor `c1ccccc1`						
COFAT	+0.96	+0.21	**0.96** [0.94,0.97]	**71**	100	100
OFAT-μ[24]	+0.47	-0.02	0.73 [0.62,0.81]	40	80	80
anchor `c1ccc2scnc2c1`						
COFAT	+0.87	+0.43	**0.86** [0.84,0.88]	**62**	100	95
OFAT-μ[24]	+0.35	+0.06	0.55 [0.41,0.67]	39	100	95