STEERED NOT STIRRED: EFFECTIVE REWARD TILTING USING SURROGATE REWARD MIXING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous-time generative models, from diffusion models, flow matching, and now flow maps, owe their practical value to high-fidelity generation and steerability. In many applications, a ground-truth latent "oracle" reward R^* is unavailable or expensive to evaluate, e.g., human judgments or experimental measurements. Although the oracle reward is not directly observable, it can often be approximated through multiple cheaper surrogates that, individually, provide incomplete characterizations but collectively reveal substantial information about the downstream task. In this paper, we present a method for composing such surrogates to steer a pretrained model toward the distribution tilted by an unavailable R^*. Given a small feedback set of oracle rankings or numerical rewards, we introduce **M**ixing **A**nd **R**eward **T**ilting for **IN**ference-time **I**mprovement (MARTINI), which learns a surrogate composition for plug-and-play use with standard inference-time steering methods. Our theory characterizes surrogate information limits and relates population recovery error to ideal-target KL at small tilt strength. For preferences, the bound additionally accounts for an irreducible mismatch between the Plackett–Luce population target and the conditional mean. Experiments demonstrate improved oracle-tilt fidelity in language generation and ImageNet alignment, and improved joint pose accuracy and validity in protein–ligand cofolding, without inference-time oracle queries.

1 INTRODUCTION

In many real-world deployments, generative models are judged not only by sample fidelity to the pretraining distribution, but by whether their generations satisfy downstream task-specific objectives (Zhang et al., 2026). Consequently, it remains imperative to align generative models to desired rewards through algorithmic frameworks such as inference-time steering or post-training (Uehara et al., 2025) that, in effect, sharpen the model distribution on samples with high downstream utility, where that utility is too costly or inaccessible to supervise at pretraining scale. The central discrepancy between pre-training and downstream task alignment, therefore, requires high-quality post-training data under an oracle reward, whose acquisition is often bottlenecked due to high collection costs (Gao et al., 2023), from capturing the nature of human preferences (Rafailov et al., 2023; Ethayarajh et al., 2024) to experimental measurements needed to evaluate biomolecular design (Pathak et al., 2025).

In lieu of carte blanche access to a reward oracle that is perfectly predictive of a desired downstream task, practitioners often have access to multiple weakly informative signals in the form of surrogate or proxy reward models. In text-to-image generation, for example, human preferences are approximated by aesthetic, alignment, and learned preference scores such as, ImageReward (Xu et al., 2023), CLIP (Radford et al., 2021), and PickScore (Kirstain et al., 2023). In molecular and protein design, physical plausibility checks (Buttenschoen et al., 2024) and machine-learned interatomic potentials (Wood et al., 2025) similarly serve as proxies for physical assays or high-fidelity simulations. In each example, cheap proxy reward signals are individually imperfect and only weakly correlated with the true downstream task, yet collectively they contain substantial information and often complementary evidence for successful candidates. Despite this promise—including the potential to mitigate reward hacking (Coste et al., 2024; Eisenstein et al., 2024)—existing inference-time steering methods for diffusion and flow-matching models are largely formulated around a single surrogate reward. As a result, a key missing development is a principled framework for composing multiple

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.

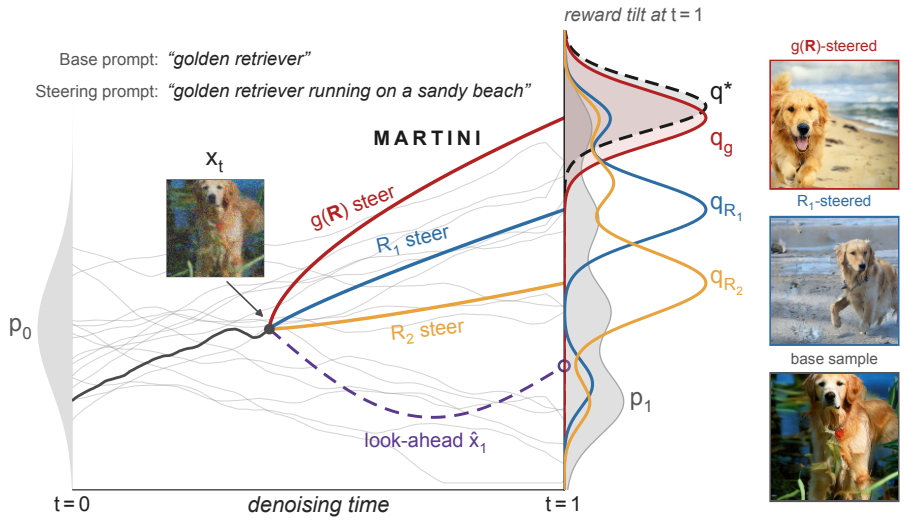


Figure 1: MARTINI learns a surrogate composition $g \circ \mathbf{R}$ from limited offline oracle feedback and uses it to steer toward q_g, aiming to approximate the gold reward tilt q^*. At inference, MARTINI steers without querying R^*.

proxies to recover the latent oracle objective, rather than standard multi-objective optimization that treats surrogates as separate objectives. This motivates a fundamental question: *How can multiple weak reward surrogates be composed to steer a pretrained diffusion or flow model towards the reward-tilted distribution favored by an unavailable oracle?*

Present work. We study steering toward the reward-tilted distribution of a latent oracle reward, R^*, available only on a small annotated feedback set, using multiple cheap surrogates and no query access to R^* at inference. More precisely, we formulate two settings representative of real-world use, in which the feedback set supplies either preference rankings over candidate samples or direct scalar reward values. We introduce MARTINI (**M**ixing **A**nd **R**eward **T**ilting for **IN**ference-time **I**mprovement, Figure 1), which learns a composition of the surrogates from that feedback and allows plug-and-play deployment in existing reward steering frameworks. MARTINI exploits complementary surrogate information to improve steering toward R^*, targeting a different solution concept from Pareto optimization. We summarize our main contributions as follows:

1. **Methods**. We develop a novel algorithm in MARTINI that exploits the problem structure of steering towards a gold reward R^*, witnessed only on a feedback set, by learning a composition function over cheap reward surrogates that can readily interface with existing steering pipelines.
2. **Theory**. We characterize the information limit of the surrogates and identify the exact KL-optimal composition. Conditional-mean regression minimizes recovery error and the leading small-guidance KL coefficient. We also bound reward gain and gap, and relate preference fitting to these quantities through the discrepancy between its population target and the conditional mean.
3. **Results**. We evaluate MARTINI on language generation, ImageNet text alignment, and protein–ligand cofolding, as well as toy experiments, spanning FKS (Singhal et al., 2025), FMRG (Huang et al., 2026), MFM (Potaptchik et al., 2026), and Itô steering (Pan et al., 2026) methods.

2 Background and Preliminaries

Notation. We denote by $\mathcal{X} = \mathbb{R}^d$ the state space of the generative dynamics and by p_1 the data distribution on this space, using embedded representations for discrete data. A reward is a scalar function $R : \mathcal{X} \to \mathbb{R}$. A superscript asterisk marks the gold reward R^*, unqueryable at inference. We use $\mathbf{R}$ to denote the vector of k surrogate rewards $\mathbf{R} = (R_1, \dots, R_k) : \mathcal{X} \to \mathbb{R}^k$, and $\mathbf{r} = (r_1, \dots, r_k) \in \mathbb{R}^k$ values it takes. A composition function $g : \mathbb{R}^k \to \mathbb{R}$ turns the surrogates into a scalar reward on $\mathcal{X}$, i.e., $g \circ \mathbf{R} : x \mapsto g(\mathbf{R}(x))$ while $\mathcal{G}$ denotes the function class used to fit the composition. For a reward R and inverse temperature $\beta > 0$, we denote $q_R \propto p_1 \exp(\beta R)$ as its induced tilt. All such tilts assign zero probability to sets of zero p_1-probability. In particular, we define $q^* := q_{R^*}$ as the gold tilt and correspondingly $q_g := q_{g \circ \mathbf{R}}$ the g-composition tilt. $\mathrm{KL}(p \,\|\, q)$ is the Kullback–Leibler divergence, while conditional moments of R^* given $\mathbf{R}$ are denoted by:

$\mu := \mathbb{E}_{p_1}[R^* \mid \mathbf{R}]$, $v := \mathrm{Var}_{p_1}(R^* \mid \mathbf{R})$, and $\bar{v} := \mathbb{E}_{p_1}[v]$, so that μ is the best surrogate estimate of R^*, v is the variance of R^* that remains once the surrogates are known, and $\bar{v}$ the expected variance. Specifically, $\bar{v}$ quantifies the additional information about R^* that the surrogates do not contain. Finally, SCC and PCC denote the Spearman and Pearson correlation coefficients, respectively.

2.1 Stochastic Interpolants and Inference Steering

Stochastic interpolants, such as flow matching (Lipman et al., 2022; Liu, 2022; Albergo and Vanden-Eijnden, 2023) and diffusion models (Ho et al., 2020; Song et al., 2021), transport samples from a tractable prior p_0, e.g., $\mathcal{N}(0, I)$, to a data distribution $p_1 \in \mathbb{P}(\mathcal{X})$, typically represented by a dataset $\mathcal{D}_{\text{train}} = \{x^i\}_{i=1}^N$ of N samples. These models then construct an interpolant $I_t := \alpha_t x_1 + \sigma_t x_0$, that connects independent samples $x_0 \sim p_0$, $x_1 \sim p_1$ and induces the marginal law $p_t := \mathrm{Law}(I_t)$. This interpolant can be constructed in different ways: Flow matching regresses a velocity network $u_\phi : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ against the interpolant velocity $\dot{I}_t = \dot{\alpha}_t x_1 + \dot{\sigma}_t x_0$. Diffusion models learn endpoint denoisers $\hat{x}_1(x_t, t) := D_\phi(x_t, t) \approx \mathbb{E}[x_1 \mid I_t = x_t]$, or implicitly parametrize the Stein score $\nabla \log p_t(x_t)$. Sampling can then be performed by integrating the probability flow ODE $\dot{x}_t = u_\phi(x_t, t)$, or by simulating an SDE with the same marginals and terminal law $p_{1,\phi}$, respectively:

$$\mathrm{d}x_t = \left[u_\phi(x_t, t) + \tfrac{c_t^2}{2} \nabla \log p_t(x_t) \right] \mathrm{d}t + c_t \, \mathrm{d}W_t, \quad x_0 \sim p_0, \tag{1}$$

where $c_t > 0$ is the diffusion coefficient and W_t is the standard Wiener process on $\mathbb{R}^d$.

Steering. Given a perfectly trained diffusion model with terminal law $p_{1,\phi} \overset{d}{=} p_1$, inference-time steering seeks to target the reward-tilted q_R without further retraining. SDE methods achieve this by adding a Doob h-transform drift $\beta c_t^2 \nabla_{x_t} V_t$ with $V_t(x_t) = \beta^{-1} \log \mathbb{E}[\exp(\beta R(x_1)) \mid x_t]$, while sequential Monte-Carlo schemes (Doucet et al., 2001) propagate particles and resample by look-ahead reward potentials $R(\hat{x}_1(x_t))$. Finally, few-step optimal-control methods treat the velocity as a control and modify the velocity by pushing samples along ∇R, also using a look-ahead reward estimate.

Flow maps. Rather than integrating the probability-flow ODE step-by-step, one can train a flow map, $X : [0,1]^2 \times \mathcal{X} \to \mathcal{X}$, $X_{t,s}(x) := X(t, s, x)$. A flow map jumps a state from time t to s along the ODE trajectory, $X_{t,s}(x_t) = x_s$, and is naturally parametrized via the average velocity $u_{t,s}$ for $s \neq t$, $X_{t,s}(x_t) = x_t + (s-t)\, u_{t,s}(x_t)$, with $u_{t,s}$ given by: $u_{t,s}(x_t) = (s-t)^{-1} \int_t^s u_\phi(x_\tau, \tau) \, \mathrm{d}\tau$. As $s \to t$ the average velocity recovers the instantaneous one, $\lim_{s \to t} u_{t,s}(x_t) = u_\phi(x_t, t)$. We can distill flow maps from a pretrained model, or directly from data, using Lagrangian, Eulerian, or Progressive distillation or losses. The one-step map $\hat{x}_1(x_t) = X_{t,1}(x_t)$ provides a fast look-ahead to the endpoint, so we can evaluate a reward on a predicted sample without integrating the remaining trajectory.

Continuous treatment of discrete data. Let $\mathcal{V}$ be a vocabulary and L a sequence length, such that token sequences satisfy $y \in \mathcal{V}^L$. Here, $\mathcal{X} = \mathbb{R}^{L \times |\mathcal{V}|}$, where $d = L|\mathcal{V}|$, and we use tokenwise one-hot encoder $\mathrm{Enc} : \mathcal{V}^L \to \mathcal{X}$ with the tokenwise argmax decoder $\mathrm{Dec} : \mathcal{X} \to \mathcal{V}^L$ (Lee et al., 2026a), such that p_1 is the distribution of one-hot embeddings $\mathrm{Enc}(\mathcal{V}^L) \subset \mathcal{X}$. Since $\mathrm{Dec}(\mathrm{Enc}(y)) = y$, this representation preserves the text distribution exactly. We apply the preceding interpolant and flow-map construction to p_1, with $x_1 = \mathrm{Enc}(y)$ and $x_0 \sim \mathcal{N}(0, I)$. The intermediate marginals p_t have smooth densities and converge in distribution to p_1 as $t \to 1$. Learned continuous terminal predictions are then decoded by tokenwise argmax. Rewards on one-hot encodings are defined by the corresponding text rewards.

2.2 Preference Modelling

Preference learning aims to fit a latent utility h from relative rankings rather than scalar-valued scores (as in, e.g., regression tasks), by minimizing a negative log-likelihood (NLL). The Bradley–Terry (BT) model (Bradley and Terry, 1952) models pairwise rankings $p(x_i \succ x_j) = \sigma(h(x_i) - h(x_j))$, with σ the logistic sigmoid function. The Plackett–Luce (PL) model (Plackett, 1975; Luce, 1959) extends this to a full ranking π of a candidate set S of size m, decomposed into a sequence of softmax choices over the items not yet selected, $p_{\mathrm{PL}}(\pi \mid h, S) = \prod_{i=1}^{m-1} \exp(h\left(x_{\pi(i)}\right)) / \sum_{i'=i}^{m} \exp\left(h\left(x_{\pi(i')}\right)\right)$, recovering BT at $m = 2$ and is invariant to additive constants. Such models are pervasive in LLM post-training: annotators rank responses to a prompt, and the goal is to align the model to human judgements of quality using RL or direct preference optimization (Schulman et al., 2017; Rafailov et al., 2023).

3 METHOD

MARTINI learns a composition of fixed surrogates from an offline feedback set $\mathcal{D}_{\text{feedback}}$, distinct from the generator's training data $\mathcal{D}_{\text{train}}$. Feedback consists of numerical oracle rewards or rankings; the fitted composition drives inference-time steering without oracle queries.

3.1 PROBLEM SETUP

At a given tilt strength $\beta > 0$, we seek a surrogate tilt q_g approximating $q^* \propto p_1 \exp(\beta R^*)$, using k cheap surrogates stacked together as $\mathbf{R}$. We reiterate that R^* is not available at inference; instead, we only have knowledge of the gold reward through a feedback set $\mathcal{D}_{\text{feedback}}$. We seek $g^{\text{opt}} \in \arg\min_g \mathrm{KL}(q^* \| q_g)$ over measurable compositions with finite, positive tilt normalizers. In practice, we fit $g \in \mathcal{G}$ from $\mathcal{D}_{\text{feedback}}$. Performance depends on the surrogate information, composition class, and feedback budget. We consider two feedback regimes.

(i) Value feedback. R^* returns scores: $\mathcal{D}_{\text{feedback}} = \{(x_i, R^*(x_i))\}_{i=1}^n$, where $\mathcal{D}_{\text{feedback}}$ represents a budget of expensive evaluations. The feedback is cardinal, namely the scale of R^* is known.

(ii) Preference feedback. R^* is seen only through order: $\mathcal{D}_{\text{feedback}} = \{(S^{(j)}, \pi^{(j)})\}_{j=1}^M$ comprises M query groups of $m \geq 2$ samples, $S^{(j)} \sim p_1^{\otimes m}$, and $\pi^{(j)}$ is the preferred order. Rankings are drawn independently across groups from the PL model of §2.2 at the gold reward score:

$$\pi^{(j)} \sim p^*_{\text{PL}}(\cdot \mid S^{(j)}), \qquad p^*_{\text{PL}}(\pi \mid S) := p_{\text{PL}}(\pi \mid R^*, S) = \prod_{i=1}^{m-1} \frac{\exp(R^*(x_{\pi(i)}))}{\sum_{i'=i}^{m} \exp\left(R^*\left(x_{\pi(i')}\right)\right)}. \quad (2)$$

Equivalently, by the Gumbel-sort construction (Yellott, 1977), $\pi^{(j)}$ is obtained by perturbing each candidate's oracle score with independent standard Gumbel noise and sorting in descending order. By the shift invariance of p_{PL}, preferences constrain R^* only up to an additive constant.

Relation to multi-objective optimization. In multi-objective optimization (MOO), the R_j are distinct objectives. For maximized rewards, x dominates x' if $R_j(x) \geq R_j(x')$ for every j, with at least one strict inequality. Undominated feasible candidates form the Pareto set, where their reward vectors form the Pareto front, and scalarizations express trade-offs among the objectives.

Our surrogates instead provide imperfect information about one oracle R^*. A candidate can dominate another on every surrogate yet receive a lower oracle reward, so the surrogate Pareto set can exclude the oracle optimum (lemmas 4 and 5 and Figure 2). We therefore allow nonmonotone compositions, although no composition can recover information absent from its inputs. Further discussions are in §K.

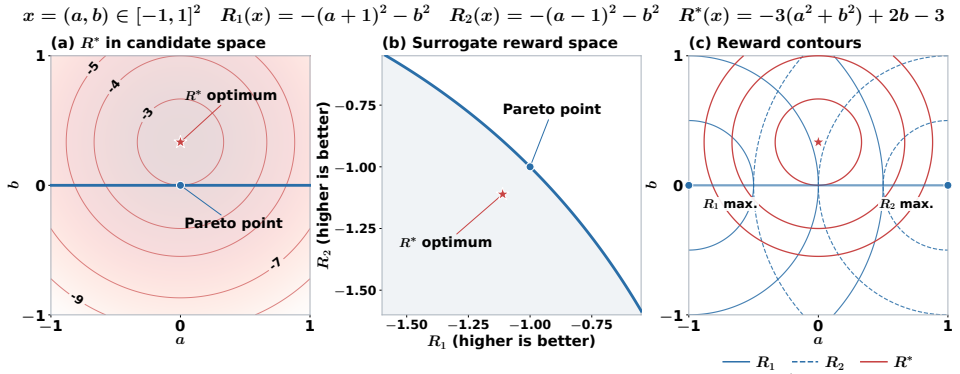


Figure 2: **The oracle optimum can be dominated by the surrogates. (a)** Contours of R^* in candidate space. Its maximizer lies outside the Pareto set. **(b)** Visualization in surrogate reward space, Pareto front in blue. The Pareto point scores higher on both surrogates but lower on R^*. **(c)** Visualization of surrogate reward and R^* contours.

3.2 TRAINING MARTINI

Given $\mathcal{D}_{\text{feedback}}$ drawn from p_1, MARTINI learns a composition g offline using a value or preference loss, with the map $g \circ \mathbf{R}$ compatible with any inference steering approach without querying R^*.

Value feedback. For numerical evaluations of R^*, the fitting objective is the mean squared error:

$$\mathcal{L}_{\text{MSE}}(g) = \frac{1}{n}\sum_{i=1}^{n}\left(g(\mathbf{R}(x_i)) - R^*(x_i)\right)^2 \quad x_i \sim \mathcal{D}_{\text{feedback}}. \tag{3}$$

Over unrestricted measurable compositions, the population counterpart of Equation (3) is minimized by $g \circ \mathbf{R} = \mu$, a quantity whose steering advantage to q^* we theoretically establish in §3.3.

Preference feedback. For rankings, we fit the composition by minimizing the PL NLL. Let $x_{(i)}^{(j)}$ denote the candidate ranked ith in group j, with the most preferred candidate first. The objective is

$$\mathcal{L}_{\text{PL}}(g) = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{m-1}\left[g\Big(\mathbf{R}(x_{(i)}^{(j)})\Big) - \log\sum_{l=i}^{m}\exp\Big(g\Big(\mathbf{R}(x_{(l)}^{(j)})\Big)\Big)\right]. \tag{4}$$

Equation (4) fits the observed rankings without requiring numerical gold rewards. Its population target need not equal μ, whose implications for recovery and steering are discussed in §3.3.

Constructing the composition. We consider three ways to construct the composition: correlation weighting, fitted linear models, and MLPs. Linear and MLP compositions are trained using the appropriate empirical loss with family-specific regularization, while correlation weights are computed directly from Pearson correlations with observed oracle values or Spearman correlations with observed oracle rankings. Correlation and linear models describe a weighted sum of surrogate scores, while MLPs can capture nonlinear relationships. Derivations and implementation details are provided in §A.

Inference-time steering. After fitting, we hold g fixed and use $g \circ \mathbf{R}$ to steer toward q_g. Adding the Doob drift correction to (1) gives:

$$\mathrm{d}x_t = \left[u_\phi(x_t, t) + \frac{c_t^2}{2}\nabla \log p_t(x_t) + \beta c_t^2 \nabla_{x_t} V_t^g(x_t)\right]\mathrm{d}t + c_t\,\mathrm{d}W_t, \tag{5}$$

where $V_t^g(x_t) = \frac{1}{\beta}\log \mathbb{E}[\exp(\beta g(\mathbf{R}(x_1))) \mid x_t]$. In practice, an endpoint prediction $\hat{x}_1 := D_\phi(x_t, t)$ gives the look-ahead approximation. In practice, we use the look-ahead reward $g(\mathbf{R}(\hat{x}_1(x_t)))$. When the composition, surrogates, and endpoint predictor are differentiable, its gradient follows by the chain rule (c.f, algorithm 1 for details).

3.3 Theoretical Results

Our setting raises two questions: ***Q1.*** *when does learning to predict R^* from surrogates lead to useful steering*, and ***Q2.*** *how well can any composition of the surrogates do?* The objectives in Equations (3) and (4) use oracle feedback, whereas our goal is to approximate q^*. Our results connect these questions and identify the limits imposed by information about R^* missing from the surrogates.

We examine four quantities: *Steering divergence*, $\mathrm{KL}(q^* \| q_g)$, measures fidelity to the oracle tilt, and *reward gain* and *reward gap* compare the induced distribution's expected R^* reward with that of p_1 and q^* respectively. Finally, *recovery error* measures how accurately $g \circ \mathbf{R}$ predicts R^* under p_1.

We analyze population quantities and ideal tilts under the assumptions stated in §G.1. The KL and reward expansions in the theorems hold for fixed compositions as $\beta \downarrow 0$, while the recovery identities and bounds are exact. For both theorems, we use mean-aligned representatives (Assumption G.3), which preserve the induced tilt. For conditional generation, i.e. $p_1(\cdot|c)$ we fix the context c in the analysis and drop the notation for brevity.

Value feedback. We write $\mathrm{MSE}(g) := \mathbb{E}_{p_1}[(g \circ \mathbf{R} - R^*)^2]$ for population recovery error and $\omega_R := \exp(\beta R)/\mathbb{E}_{p_1}[\exp(\beta R)]$ for a tilt's relative weight.

Theorem 1 (Value Regime)**.** *Suppose R^* and $g \circ \mathbf{R}$ satisfy Assumptions G.1–G.3. Then:*

(i) (Steering.) *As $\beta \downarrow 0$, with Assumption G.3 used for the second equality,*

$$\mathrm{KL}(q^* \| q_g) = \frac{\beta^2}{2}\mathrm{Var}_{p_1}(R^* - g \circ \mathbf{R}) + O(\beta^3) = \frac{\beta^2}{2}\mathrm{MSE}(g) + O(\beta^3).$$

(ii) (Reward gain.) *For any* $R \in L^\infty(p_1)$ *and* $\beta > 0$: $\mathbb{E}_{q_R}[R^*] - \mathbb{E}_{p_1}[R^*] = \mathrm{Cov}_{p_1}(R^*, \omega_R)$. *Furthermore, as* $\beta \downarrow 0$:

$$\mathbb{E}_{q_g}[R^*] - \mathbb{E}_{p_1}[R^*] \geq \beta \left[\mathrm{Var}_{p_1}(\mu) - \sqrt{\mathrm{Var}_{p_1}(\mu)(\mathrm{MSE}(g) - \bar{v})} \right] + O(\beta^2).$$

(iii) (Reward gap.) *As* $\beta \downarrow 0$, $\mathbb{E}_{q^*}[R^*] - \mathbb{E}_{q_g}[R^*] \leq \beta[\bar{v} + \sqrt{\mathrm{Var}_{p_1}(\mu)(\mathrm{MSE}(g) - \bar{v})}] + O(\beta^2)$.

(iv) (Best recovery.) *Under Assumption G.1 alone,*

$$\mathrm{MSE}(g) = \bar{v} + \|g \circ \mathbf{R} - \mu\|^2_{L^2(p_1)} \geq \bar{v}, \qquad \mathrm{Var}_{p_1}(R^*) = \mathrm{Var}_{p_1}(\mu) + \bar{v}.$$

Part (i) answers **Q1** by connecting the value-fitting objective directly to steering: reducing population MSE reduces the leading KL error. This also explains why the unrestricted regression target μ is useful even though the exact KL-optimal composition g^{opt} can differ from it. Proposition 1 and Corollary 1 show that $\mathrm{KL}(q^* \| q_\mu) - \mathrm{KL}(q^* \| q_{g^{\mathrm{opt}}}) = O(\beta^4)$ as $\beta \downarrow 0$. Thus, using μ instead of g^{opt} incurs only fourth-order excess KL, providing a distributional justification for MSE fitting in this regime.

Part (iv) answers **Q2**: fitting can reduce the discrepancy from μ, but recovery error remains at least $\bar{v}$, the variation in R^* unexplained by the surrogates. Improving the fit of g therefore addresses only one source of error, and reducing this floor requires more informative surrogate inputs (Corollary 5).

Parts (ii)–(iii) relate recovery error to expected reward after steering under R^*. At leading order, reducing recovery error raises the gain lower bound and lowers the gap upper bound. When $\mathrm{Var}_{p_1}(\mu) > 0$, sufficiently accurate approximation of μ guarantees positive gain for small enough $\beta > 0$, even if $\bar{v} > 0$ (Corollary 2). Useful steering therefore does not require perfect oracle recovery. The exact gain identity in part (ii) appears concurrently in Paes et al. (2026), while here we relate it to learned surrogate compositions through asymptotic gain and gap bounds.

Preference feedback. When numerical oracle rewards are unavailable, we instead ask whether fitting oracle rankings controls the same recovery and steering quantities.

Theorem 2 (Preference Regime). *Under Assumptions G.1, G.3, and G.4, define the population loss* $\mathrm{PL}(g) := \mathbb{E}_{S,\pi}[-\log p_{\mathrm{PL}}(\pi \mid g \circ \mathbf{R}, S)]$ *and let* $g_{\mathrm{PL}} \in \arg\min_{g \in \mathcal{G}} \mathrm{PL}(g)$. *With the score-range bound* Δ *from G.4 and* $c_\Delta = (1 - 1/m)e^{-\Delta}$, *every* $g \in \mathcal{G}$ *satisfies the bound*

$$\mathrm{MSE}(g) \leq B(g) := \bar{v} + \left(\sqrt{\mathrm{Var}_{p_1}(g_{\mathrm{PL}} \circ \mathbf{R} - \mu)} + \sqrt{\frac{2}{c_\Delta}\big(\mathrm{PL}(g) - \mathrm{PL}(g_{\mathrm{PL}})\big)} \right)^2.$$

Parts (i)–(iii) further assume the oracle bound in Assumption G.2 and hold for fixed g *as* $\beta \downarrow 0$:

(i) (Steering.) $\mathrm{KL}(q^* \| q_g) \leq \frac{\beta^2}{2} B(g) + O(\beta^3)$.

(ii) (Reward gain.) $\mathbb{E}_{q_g}[R^*] - \mathbb{E}_{p_1}[R^*] \geq \beta \left[\mathrm{Var}_{p_1}(\mu) - \sqrt{\mathrm{Var}_{p_1}(\mu)(B(g) - \bar{v})} \right] + O(\beta^2)$.

(iii) (Reward gap.) $\mathbb{E}_{q^*}[R^*] - \mathbb{E}_{q_g}[R^*] \leq \beta \left[\bar{v} + \sqrt{\mathrm{Var}_{p_1}(\mu)(B(g) - \bar{v})} \right] + O(\beta^2)$.

(iv) (Minimum bound.) $B(g) \geq B(g_{\mathrm{PL}}) = \mathrm{MSE}(g_{\mathrm{PL}})$.

The empirical objective (4) estimates $\mathrm{PL}(g)$, whose excess controls recovery through $B(g)$. This bound combines the information floor $\bar{v}$, the discrepancy between $g_{\mathrm{PL}} \circ \mathbf{R}$ and μ, and excess PL loss. For a fixed problem and class, reducing excess loss tightens the leading KL, gain, and gap, bounds in parts (i)–(iii), justifying preference fitting without numerical oracle rewards. At zero excess loss, the discrepancy remains unless $g_{\mathrm{PL}} \circ \mathbf{R} = \mu$, in which case $B(g_{\mathrm{PL}}) = \bar{v}$ (iv). Thus g_{PL} minimizes the bound, without necessarily minimizing recovery error or exact steering divergence. The guarantee requires a convex class, a uniform score-range bound Δ, and an attained population minimum (Assumption G.4); nonconvex classes require separate analysis. Proofs and supporting results appear in §G–H.

4 EXPERIMENTS

We evaluate MARTINI on language generation, text-to-image alignment, and protein–ligand cofolding, fitting $\hat{g}$ offline and steering without inference-time oracle queries, with synthetic experiments in §I.

Evaluation. For language and images, we assess recovery using held-out PCC and MSE for values, or SCC and PL loss for preferences. Experimental tilts use the sampler law $p_{1,\phi}$ in place of p_1. We evaluate fixed finite target tilts $\beta = 1$ and sweep sampler guidance separately, probing behavior beyond our small-β guarantees. Composition fidelity is measured by $\mathrm{KL}(q^* \| q_g)$, estimated from unsteered samples. We assess generated samples using embedding-space energy distance and distances on the R^* marginal, including Wasserstein distances and Gaussian and KDE estimates of KL. We also report the absolute reward gap and a diversity-adjusted cost Q (prior: 1 and lower is better), which evaluates reward gains accompanied by diversity loss. Definitions and estimators appear in §C. Specific protein cofolding metrics are described in §4.3.

Language and Image Baselines. Best-of-N (BoN), reported in Appendix ablation tables, selects the highest-R^* candidate, requiring N oracle evaluations per sample. Deployable steering baselines are the best single surrogate (BSS), uniform weighting, and ParetoFlow (Yuan et al., 2025), plus ReNO's fixed reward weights for images (Eyring et al., 2024). BSS is selected on training feedback using either PCC or SCC depending on regime. We also evaluate SLOP's calibrated selection protocol (Wang et al., 2026). SLOP requires R^* values, so its preference results serve only as a value-calibrated reference.

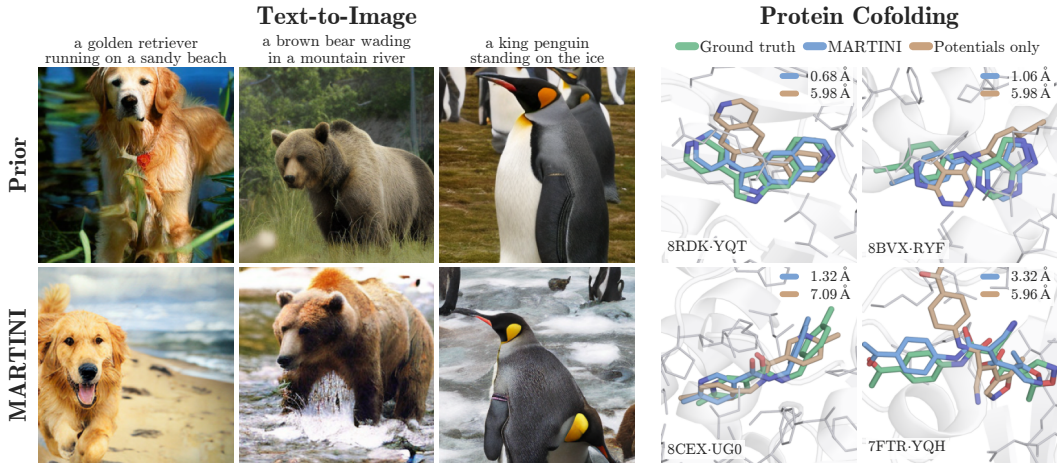


Figure 3: **MARTINI steering.** Left: ImageNet prior samples (top) and steered samples (bottom). Right: Predicted and reference ligand poses. Annotations give ligand RMSD in Å.

4.1 LANGUAGE GENERATION

We use FMRG to steer the released flow map language model (FMLM) (Lee et al., 2026a), trained on One Billion Words (Chelba et al., 2013). For fitting and evaluation, we generate unconditional sequences of 128 tokens in 8 steps using the `bert-base-uncased` vocabulary.

Setup. We use the negative generative perplexity under Qwen2.5-1.5B (Qwen et al., 2025) as the oracle, $R^* = -\mathrm{gPPL}$. Our four differentiable surrogates are BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) scores, CoLA acceptability (Warstadt et al., 2019), and empirical unigram entropy H_{tok}. Entropy supplies a diversity signal to address repetitive outputs under gPPL minimization. Feedback comprises $n = 256$ numerical rewards or $M = 256$ rankings of $m = 8$ candidates. We generate 256 samples per method, regime, and seed, and average metrics over five seeds.

Results. Held-out loss selects an MLP for values and a linear composition for preferences (Figure 5), and both attain the lowest reported composition KL (Figure 6). At steering strength 0.03, both outperform the displayed deployable baselines in energy distance and Q (Figure 4). Increasing BSS guidance (the same BSS was chosen in both regimes) worsens reward-marginal KL and Q, with both exceeding the prior. Stronger MARTINI guidance improves reward-marginal KL but

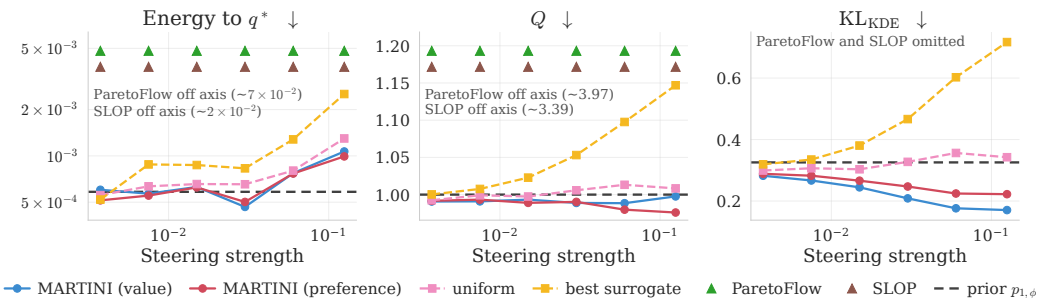


Figure 4: **Language steering.** Lower is better for all metrics. Triangles indicate values beyond the plotted range.

Table 1: **ImageNet steering.** Bold/underline mark best/second-best deployable results within each regime. On shared rows, $*/\dagger$ denote values/preferences, with superscripts/subscripts indicating first/second.

Method	KL_g	Energy ↓	Sl. W_1 ↓	$\mathrm{KL}_{\mathcal{N}}$ ↓	$\mathrm{KL}_{\mathrm{KDE}}$ ↓	|gap| ↓	Q ↓
Prior $p_{1,\phi}$	–	0.0683	0.0080	0.752	0.658	7.51	1.000
ParetoFlow	–	0.0885	0.0088	0.825	0.937	5.35	0.844
SLOP	0.520	0.1200	0.0098	<u>0.168</u>$_*$	0.170	<u>1.98</u>$_{*\dagger}$	0.364
ReNO	0.553	0.0503	0.0069	0.197	<u>0.166</u>$_*$	2.10	<u>0.244</u>$_{*\dagger}$
Uniform	**0.436**$^*_\dagger$	0.0603	0.0074	0.605	0.635	2.73	**0.238***†
BSS (Val.)	0.514	<u>0.0498</u>	<u>0.0068</u>	0.217	0.183	2.57	0.354
MARTINI (Val.)	<u>0.466</u>	**0.0321**	**0.0058**	**0.034**	**0.049**	**1.27**	0.359
BSS (Pref.)	0.440	<u>0.0423</u>	<u>0.0065</u>	<u>0.135</u>	<u>0.133</u>	2.04	0.316
MARTINI (Pref.)	**0.423**	**0.0372**	**0.0061**	**0.093**	**0.121**	**1.71**	0.297

eventually raises energy distance above the prior: closer reward-marginal matching need not improve embedding-space fidelity. Additional results and generated examples appear in Tables 3 and 14.

4.2 TEXT-TO-IMAGE

We use FMRG (Huang et al., 2026) to steer a distilled, class-conditioned ImageNet flow map toward text prompts, fitting a separate composition for each class–prompt pair, as shown in Figure 3 (left).

Setup. Our surrogates are HPSv2 (Wu et al., 2023), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and CLIP (Radford et al., 2021), matching ReNO's reward set (Eyring et al., 2024). The oracle R^* measures image–prompt alignment using Skywork-VL-Reward 7B (Wang et al., 2025a). Feedback comprises $n = 32$ scalar values or $M = 128$ rankings. Recovery uses all 19 class–prompt pairs while the main steering comparison uses seven with reference ESS > 1000 in both regimes. For each pair and method, sample metrics use the lowest-energy evaluated strength. All-pair results and cutoff sensitivity appear in §E, and implementation details in §B.

Results. Held-out loss selects a linear composition in both regimes. On these seven pairs, MARTINI achieves the lowest embedding-space and reward-marginal distances and absolute reward gap in both regimes (Table 1). With values, $\mathrm{KL}_{\mathcal{N}}$ and $\mathrm{KL}_{\mathrm{KDE}}$ are reduced by factors of 4.9 and 3.4, respectively, relative to the best baselines. Only uniform weighting achieves lower KL_g (0.436 versus 0.466). With preferences, MARTINI also achieves the lowest KL_g (0.423). Both selected compositions also outperform BSS, uniform weighting, and ReNO in mean energy at shared strengths $\lambda \in \{2, 5, 8\}$ (Table 5). Full sweeps and additional feedback budget comparisons appear in §E. Qualitative results are in §J.2.

4.3 PROTEIN-LIGAND CO-FOLDING

Protein–ligand cofolding provides a setting with experimental scalar feedback. We ask whether a surrogate composition learned from binding-affinity measurements can improve the accuracy and physical validity of the jointly generated poses. Qualitative results shown in Figure 3 (right) and Figure 21.

Table 2: **Protein–ligand steering on post-2023 Runs N' Poses**. 15 particles, all rewards normalized, best@1.

Method	Success Rate ↑	RMSD< 2 Å ↑	PB Valid ↑	lDDT-PLI ↑	Diversity ↑
Boltz1 potentials	59.5	59.5	**99.9**	65.2	0.0108
DeCAF potentials	61.8	61.8	<u>98.3</u>	65.7	0.0132
smina + potentials	61.8	61.8	98.2	65.4	0.0129
MARTINI (smina, potentials)	<u>62.7</u>	<u>62.8</u>	93.3	<u>66.5</u>	**0.0139**
MARTINI (5 surrogates)	**64.5**	**64.5**	91.8	**67.6**	<u>0.0133</u>

Setup. Using the value regime, we fit a composition of five surrogates to experimental $\mathrm{p}K_D$ measurements from HiQBind (Wang et al., 2025b), restricting the feedback set to PoseBusters-valid monomers and multimers (Buttenschoen et al., 2024). The fitted composition $\hat{g} \circ \mathbf{R}$ then steers the DeCAF-Boltz flow map (Scarpellini et al., 2026), distilled from Boltz-1 (Wohlwend et al., 2025), without further experimental feedback. We evaluate on 186 complexes from the post-2023 subset of Runs N' Poses (Škrinjar et al., 2026), beyond Boltz-1's training cutoff. As these evaluation complexes lack measured affinities, we assess transfer through pose accuracy and physical validity, reporting best@1 metrics averaged over five seeds. Dataset filtering, surrogate definitions, and steering details are in §F.

Metrics. Our *success rate* is the percentage of predictions with RMSD below 2 Å that pass PoseBusters validity checks (Buttenschoen et al., 2024). We also report accuracy and validity rates alongside *lDDT-PLI*, which measures local protein–ligand distance agreement within a 6 Å pocket (Mariani et al., 2013), as well as structural diversity detailed in §F.3.

Results. The BSS is smina (Figure 8). MARTINI (correlation g) achieves the highest reported best@1 (§F.3) success rate (64.5%) over both potentials-only steering and BSS + potentials by 2.7 percentage points (Table 2). Relative to potentials-only, *lDDT-PLI* improves by 1.9 points, indicating better local geometry, with similar diversity (0.0133 versus 0.0132). PB-validity falls by 6.5 points, even though MARTINI trades validity for accuracy, all the accurate predictions also satisfy validity criteria.

5 Related work

Reward-based control. Generative models can be aligned through reinforcement learning (Black et al., 2024) or preference optimisation (Rafailov et al., 2023; Wallace et al., 2024; Bai et al., 2026), and steered at inference through guidance (Bansal et al., 2023; Singhal et al., 2025), posterior sampling (Chung et al., 2023), or flow-map optimal control (Huang et al., 2026). MARTINI complements these methods by learning the steering reward from offline oracle feedback.

Reward compositions and limited feedback. Reward ensembles can mitigate over-optimisation (Coste et al., 2024). ReNO uses fixed reward weights (Eyring et al., 2024), while SLOP calibrates pooled scores to maximize estimated gold reward (Wang et al., 2026). We study approximation of a fixed oracle tilt from values or rankings, characterizing the information limit imposed by the surrogates. Multi-objective methods balance distinct rewards (Ramé et al., 2023; Yuan et al., 2025). Limited oracle access also motivates query-efficient fine-tuning (Uehara et al., 2024), protein steering (Yang et al., 2025), and optimisation over generated samples (Willis et al., 2025). We use feedback to learn how fixed surrogates predict a single oracle, keeping the generator unchanged. Complementing analyses of proxy error and ensembling (Paes et al., 2026), we relate fitting loss and surrogate information limits to steering fidelity and reward bounds.

6 Conclusion

MARTINI learns surrogate compositions offline from oracle values or rankings for reuse across compatible steerers without inference-time oracle queries. Our theory links fitting to ideal-tilt fidelity and reward improvement, identifying surrogate information limits and conditions for useful steering despite imperfect recovery. Experiments demonstrate benefits across language, images, and protein–ligand cofolding. KL and reward expansions are local in β, with larger tilts assessed empirically. Preference guarantees require convex, range-bounded classes, and their tightness remains unassessed. Robustness beyond synthetic PL rankings remains untested. ImageNet guidance uses per-pair reference energy, leaving selection without oracle-scored references an open task.

AI USE STATEMENT

We used generative AI tools to develop and review mathematical statements and proofs and to support research execution, including implementation and feedback on experimental methodology, and interpretation of results. Additionally, we used these tools to review, draft, and improve the clarity of the manuscript. The authors take responsibility for the final content of this work, including all text, mathematical claims, results, and artifacts produced with AI assistance.

ETHICS STATEMENT

Steering pretrained generative models, including language models, can amplify biases or errors in the base model, oracle, or surrogate rewards, and can generate harmful content. Improvements in measured reward do not establish safety, fairness, or factual correctness. Likewise, computational protein–ligand metrics do not replace experimental validation. Applications require task-specific evaluation and appropriate safeguards.

REPRODUCIBILITY STATEMENT

The fitting objectives and steering procedure are described in §3.2, with composition-fitting details in §A. Experimental implementation and metric estimation are described in §B and §C. Formal assumptions and proofs are provided in §G.1 and §H. We commit to publicly releasing the code and supporting materials needed to reproduce our methods and experimental results upon acceptance.

REFERENCES

M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

J. Bai, X. Yu, M. Xu, W. Lu, X. Pan, K. Maeng, D. Kifer, J. Wang, and Y. Wang. Towards better optimization for listwise preference in diffusion models. In *International Conference on Learning Representations*, volume 2026, pages 133254–133282, 2026.

A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal guidance for diffusion models, 2023. URL https://arxiv.org/abs/2302.07121.

M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D*, 69(5):701–709, May 2013. doi: 10.1107/S0907444913007051. URL https://doi.org/10.1107/S0907444913007051.

K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

M. Buttenschoen, G. M. Morris, and C. M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9): 3130–3139, 2024. doi: 10.1039/D3SC04185A.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010. doi: 10.1107/S0907444909042073.

H. Chung, J. Kim, M. T. McCann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14687.

T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential monte carlo methods. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, pages 3–14. Springer New York, New York, NY, 2001. ISBN 978-1-4757-3437-9. doi: 10.1007/978-1-4757-3437-9_1. URL https://doi.org/10.1007/978-1-4757-3437-9_1.

J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. N. D'Amour, K. D. Dvijotham, A. Fisch, K. A. Heller, S. R. Pfohl, D. Ramachandran, P. Shaw, and J. Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5u1GpUkKtG.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

L. Eyring, S. Karthik, K. Roth, A. Dosovitskiy, and Z. Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization, 2024. URL https://arxiv.org/abs/2406.04312.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

J. Y. Huang, J. Lin, S. Shah, K. Nair, and N. M. Boffi. How to guide your flow: Few-step alignment via flow map reward guidance, 2026. URL https://arxiv.org/abs/2604.27147.

Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

D. Koes, M. Baumgartner, and C. Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8): 1893–1904, 2013. doi: 10.1021/ci300604z.

C. Lee, J. Yoo, M. Agarwal, S. Shah, J. Huang, A. Raghunathan, S. Hong, N. M. Boffi, and J. Kim. Flow map language models: One-step language modeling via continuous denoising, 2026a. URL https://arxiv.org/abs/2602.16813.

K. Lee, S. Yu, and J. Shin. Decoupled meanflow: Turning flow models into flow maps for accelerated sampling. In *International Conference on Learning Representations*, volume 2026, pages 64555–64583, 2026b.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Q. Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. doi: 10.48550/arXiv.2209.14577.

R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

V. Mariani, M. Biasini, A. Barbato, and T. Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21): 2722–2728, 2013. doi: 10.1093/bioinformatics/btt473.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

L. M. Paes, N. Mackraz, B.-J. Theobald, and F. Danieli. Theoretical limits of language model alignment. *arXiv preprint arXiv:2605.07105*, 2026.

Z. Pan, P. Potaptchik, W. Yao, M. S. Albergo, and J. Pidstrigach. Itô maps for any-step sdes, 2026. URL https://arxiv.org/abs/2606.11156.

S. Passaro, G. Corso, J. Wohlwend, M. Reveiz, S. Thaler, V. R. Somnath, N. Getz, T. Portnoi, J. Roy, H. Stark, D. Kwabi-Addo, D. Beaini, T. Jaakkola, and R. Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/10.1101/2025.06.14.659707v1.

A. Pathak, R. Theagarajan, M. M. Rizqi, A. S. Nugraha, T. Boruah, Pratiksha, H. Kumar, B. Naik, S. Yadav, A. K. Jha, et al. Ai-enabled drug and molecular discovery: computational methods, platforms, and translational horizons. *Discover Molecules*, 2(1):32, 2025.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

P. Potaptchik, A. Saravanan, A. Mammadov, A. Prat, M. S. Albergo, and Y. W. Teh. Meta flow maps enable scalable reward alignment, 2026. URL https://arxiv.org/abs/2601.14430.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

A. Ramé, G. Couairon, M. Shukor, C. Dancette, J.-B. Gaya, L. Soulier, and M. Cord. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.04488.

A. O. Rausch, M. I. Freiberger, C. O. Leonetti, D. M. Luna, L. G. Radusky, P. G. Wolynes, D. U. Ferreiro, and R. G. Parra. Frustratometer: an r-package to compute local frustration in protein structures, point mutants and md simulations. *Bioinformatics*, 37(18):3038–3040, 09 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab176. URL https://doi.org/10.1093/bioinformatics/btab176.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

G. Scarpellini, R. Shprints, P. Holderrieth, J. Nam, P. Murugan, R. Gómez-Bombarelli, T. Jaakkola, M. Al-Shedivat, N. M. Boffi, and A. J. Bose. Few-step cofolding with all-atom flow maps, 2026. URL https://arxiv.org/abs/2606.08375.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

R. Singhal, Z. Horvitz, R. Teehan, M. Ren, Z. Yu, K. McKeown, and R. Ranganath. A General Framework for Inference-time Scaling and Steering of Diffusion Models, 2025. URL https://arxiv.org/abs/2501.06848.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

M. Uehara, Y. Zhao, K. Black, E. Hajiramezanali, G. Scalia, N. L. Diamant, A. M. Tseng, S. Levine, and T. Biancalani. Feedback efficient online fine-tuning of diffusion models. In *International Conference on Machine Learning*, 2024.

M. Uehara, Y. Zhao, C. Wang, X. Li, A. Regev, S. Levine, and T. Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

P. Škrinjar, J. Eberhardt, G. Studer, G. Tauriello, T. Schwede, and J. Durairaj. Evaluating generalization in protein–ligand cofolding methods. *Nature Structural & Molecular Biology*, 33: 782–794, 2026. doi: 10.1038/s41594-026-01797-5. URL https://doi.org/10.1038/s41594-026-01797-5.

B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

X. Wang, P. Wang, J. Pei, W. Shen, Y. Peng, Y. Hao, W. Qiu, A. Jian, T. Xie, X. Song, et al. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*, 2025a.

Y. Wang, K. Sun, J. Li, X. Guan, O. Zhang, D. Bagni, Y. Zhang, H. A. Carlson, and T. Head-Gordon. A workflow to create a high-quality protein–ligand binding dataset for training, validation, and prediction tasks. *Digital Discovery*, 4(5):1209–1220, 2025b.

Y. Wang, J. Liu, and T. Koike-Akino. Temper and tilt lead to slop: Reward hacking mitigation with inference-time alignment. *arXiv preprint arXiv:2605.13537*, 2026.

A. Warstadt, A. Singh, and S. R. Bowman. Cola: The corpus of linguistic acceptability (with added annotations). 2019.

S. Willis, P. Duckworth, J. Simons, A. Kalisz, K. Sinkovics, N. Ghenassia, S. Surana, H. T. Oldroyd, A. I. Stere, D. D. Margineantu, C. H. Ek, H. Moss, and E. Bodin. Sample-efficient optimisation over the outputs of generative models, 2025. URL https://arxiv.org/abs/2509.23800.

J. Wohlwend, G. Corso, S. Passaro, N. Getz, M. Reveiz, K. Leidal, W. Swiderski, L. Atkinson, T. Portnoi, I. Chinn, J. Silterra, T. Jaakkola, and R. Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2025. doi: 10.1101/2024.11.19.624167. URL https://www.biorxiv.org/content/early/2025/05/06/2024.11.19.624167.

B. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. Kitchin, D. S. Levine, et al. Uma: A family of universal models for atoms. *Advances in Neural Information Processing Systems*, 38:129391–129427, 2025.

X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.05977.

J. Yang, W. Chu, D. Khalil, R. Astudillo, B. Wittmann, F. Arnold, and Y. Yue. Steering generative models with experimental data for protein fitness optimization. *Advances in Neural Information Processing Systems*, 38:145383–145421, 2025.

J. I. Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977. ISSN 0022-2496. doi: https://doi.org/10.1016/0022-2496(77)90026-8. URL https://www.sciencedirect.com/science/article/pii/0022249677900268.

Y. Yuan, C. Chen, C. Pal, and X. Liu. Paretoflow: Guided flows in multi-objective optimization. In *International Conference on Learning Representations*, 2025. arXiv:2412.03718.

S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, G. Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: https://doi.org/10.1002/prot.20264. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20264.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

APPENDIX CONTENTS

A Learning the composition g: derivations . . . 15
B Implementation details for image and language generation . . . 17
C Metric estimation for image and language generation . . . 20
D Composition selection and recovery . . . 22
E Additional results and sensitivity analyses . . . 24
F Protein–ligand cofolding . . . 30
G Additional theoretical results . . . 33
H Proofs . . . 37
I Toy Examples . . . 42
J Qualitative examples . . . 46
K Relation to multi-objective optimisation . . . 53

A LEARNING THE COMPOSITION g: DERIVATIONS

We consider correlation weighting, linear compositions, and MLPs. Linear and MLP compositions minimize the value or preference loss in Equations (3) and (4), with regularization. Correlation weights are computed directly from the feedback. The fitted composition is then held fixed during steering.

A.1 FITTING OBJECTIVES

For value feedback, write $y_i = R^*(x_i)$ and $f_i = g(\mathbf{R}(x_i))$. The objective is

$$\mathcal{L}_{\mathrm{MSE}}(g) = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2.$$

For preferences, let $x_{(i)}^{(j)}$ be the candidate ranked ith in query j, with the most preferred first, and write $f_{(i)}^{(j)} = g(\mathbf{R}(x_{(i)}^{(j)}))$. The PL loss is

$$\mathcal{L}_{\mathrm{PL}}(g) = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{m-1}\left[f_{(i)}^{(j)} - \log\sum_{l=i}^{m} e^{f_{(l)}^{(j)}}\right].$$

It is convex in the candidate scores and invariant to a common additive shift. Proposition 2 gives its derivatives.

A.2 CORRELATION WEIGHTING

The composition is $\hat{g}(\mathbf{R}) = \sum_{j=1}^{k} w_j R_j$. For value feedback, ρ_j is the empirical Pearson correlation between R_j and the oracle values. For preferences, it is the Spearman correlation with the observed rankings, assigning larger ordinal scores to more preferred candidates. We retain signed correlations and use

$$w_j = \begin{cases} \rho_j / \sum_{l=1}^{k} \rho_l, & \sum_{l=1}^{k} \rho_l > 0, \\ 1/k, & \text{otherwise.} \end{cases}$$

This is a weighting heuristic rather than, in general, a minimizer of either fitting loss.

A.3 LINEAR COMPOSITIONS

Let $g_{\mathbf{w}}(\mathbf{R}) = \mathbf{w}^\top \mathbf{R}$ and let Φ have rows $\mathbf{R}(x_i)^\top$. For values, ridge regression minimizes

$$\frac{1}{2}\|\Phi\mathbf{w} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \qquad \lambda > 0,$$

with solution

$$\hat{\mathbf{w}} = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top \mathbf{y}.$$

For preferences, we minimize

$$\mathcal{L}_{\mathrm{PL}}(g_{\mathbf{w}}) + \frac{\lambda_{\mathrm{PL}}}{2} \|\mathbf{w}\|_2^2, \qquad \lambda_{\mathrm{PL}} > 0.$$

Both objectives are strictly convex and coercive, so each has a unique minimizer. The fitted score is $\hat{g}(\mathbf{R}) = \hat{\mathbf{w}}^\top \mathbf{R}$.

A.4 MLP COMPOSITIONS

We fit a neural composition g_θ using the appropriate value or preference objective with weight decay, and set $\hat{g} = g_{\hat{\theta}}$ after training. Unlike the linear family, this parameterization can represent nonlinear relationships between the surrogates and the oracle.

In practice, we use the same small network in all image and language experiments. The MLP family maps the four surrogate scores, standardised over the pool, through two hidden layers of 64 units with $\tanh$ activations to a scalar ($4 \to 64 \to 64 \to 1$, no dropout). It is trained full-batch from a single initialisation using Adam with learning rate 3×10^{-3} and weight decay 10^{-3}.

B IMPLEMENTATION DETAILS FOR IMAGE AND LANGUAGE GENERATION

Algorithm 1 MARTINI: fit a surrogate composition and steer

Require: pretrained base model with terminal law $p_{1,\phi}$; surrogates $\mathbf{R}$; feedback $\mathcal{D}_{\text{feedback}}$; composition family $\mathcal{G}$; tilt strength β; steerer STEER

1: **Offline fitting:** Evaluate $\mathbf{R}$ on the feedback candidates in $\mathcal{D}_{\text{feedback}}$
2: **if** correlation weighting is selected **then**
3: | Construct $\hat{g}$ from empirical correlations (§A)
4: **else**
5: | Set $\ell \leftarrow \mathcal{L}_{\text{MSE}}$ for values (3) or $\ell \leftarrow \mathcal{L}_{\text{PL}}$ for preferences (4)
6: | Fit $\hat{g} \in \mathcal{G}$ using ℓ and family-specific regularization
7: **Inference:**
8: **return** STEER(base model, $\hat{g} \circ \mathbf{R}, \beta$)

We give implementation details for the experiments in §4. Additional results appear in §E.

Compute. Experiments were run on NVIDIA H100-80GB GPUs, with independent experiments distributed across devices.

B.1 IMAGENET

We use the released model of Potaptchik et al. (2026) (`mfm-xl2`), a SiT-XL/2 transformer distilled from the Decoupled MeanFlow ImageNet model (Lee et al., 2026b). It generates class-conditional ImageNet images at 256×256 in the latent space of the Stable Diffusion VAE (`sd-vae-ft-mse`; latents of shape $4 \times 32 \times 32$, scaled by 0.18215), with noise at $t = 0$, data at $t = 1$ and the linear interpolant $x_t = (1-t)\varepsilon + t\,x_1$. The model is conditioned only on the ImageNet class label: the text prompt never enters the model, and acts only through the rewards — R^* and the surrogates score the decoded image against it.

Reward models. The oracle uses `Skywork/Skywork-VL-Reward-7B`. The four surrogates are `RE-N-Y/hpsv21`, `RE-N-Y/ImageReward`, `openai/clip-vit-base-patch16`, and `yuvalkirstain/PickScore_v1`. For each image and steering prompt, the oracle score is the Yes/No logit margin,

$$R^*(x) = \text{logit}_{\text{Yes}}(x) - \text{logit}_{\text{No}}(x).$$

Target. $q^* \propto p_{1,\phi}\exp(\beta z(R^*))$, $\beta = 1$, is estimated by self-normalised reweighting of 8192 independent draws from the model's reverse SDE (32 steps); the effective sample size of these weights is reported per class in Table 4.

Feedback and fitting. ImageNet uses n=32 scalar R^* evaluations (value) or M=128 ranking queries of m=8 candidates each (preference), drawn from the class's pool. We fit separate composition coefficients for each class–prompt pair. A fifth of the value labels (6) and a quarter of the preference queries (32, built from images disjoint from the fitting queries) are held out; every family is fit on the remainder, and the family is selected by held-out MSE (value) or Plackett–Luce loss (preference), averaged over all 19 pairs. The linear family is selected in both regimes. The held-out part is used only for this selection and is not refit. loss.

Preference utilities. The synthetic annotator sorts $z(R^*) + \sigma\epsilon$ with independent standard Gumbel noise ϵ and $\sigma = 0.5$, which samples Plackett–Luce rankings with utilities $z(R^*)/\sigma$. We standardise R^* once over the reference pool, fixing its scale across queries. In our eight-candidate calibration (a standard-normal oracle), $\sigma = 0.5$ gives 82.7% pairwise agreement with the unperturbed oracle order. This noise level is loosely informed by the approximately 81% non-tie human agreement reported on MT-Bench (Zheng et al., 2023); agreement with an oracle order and agreement between annotators are different quantities. A fitted utility f is expressed in PL units, so the tilt it implies on the reference scale is $\exp(\beta\sigma f)$. In practice we standardise the fitted composition over the pool before use, so the steering reward and KL_g both use $\exp(\beta z(f))$, which coincides with $\exp(\beta\sigma f)$ when f recovers the utility exactly.

Pools and data splits. For each class–prompt pair and feedback regime, we draw an independent pool of 8192 samples and score the oracle and surrogates. Within each regime, the pool supplies

feedback candidates, normalization statistics, and the importance-weighted reference §C. Held-out recovery is measured on pool rows that no fit or selector touched, subsampled to at most 4000. Oracle and surrogate scores are computed once per pool sample and cached. The prior $p_{1,\phi}$ is 64 fresh draws from the FMRG sampler at zero strength, and BoN, sBoN and SLOP select from an independent pool of 4096 such draws, never from the reference they are scored against; these rows come from a separate run with its own 8192-draw reference.

Steering and strength selection. We use FMRG for ImageNet and sweep its guidance strength $\lambda \in \{2, 3, 5, 8\}$ separately from the target tilt strength $\beta = 1$. Each run takes 32 kernel transitions with a reward-gradient step after each, the reward evaluated on one look-ahead endpoint (one noise draw, one flow-map jump), and each setting is a single run of 64 samples, scored afresh. For each method and class–prompt pair, all generated-sample metrics are reported at the evaluated strength with the lowest embedding-space energy distance to q^*. BoN is reported at fixed rungs (n=8, 32), and SLOP at 16 candidates per sample, the rung matched to MARTINI's inference cost, calibrated on the same n=32 labels. This selection uses the oracle-scored reference and describes performance within the evaluated sweep.

Baselines. BSS is selected from training feedback via either PCC or SCC for value and preference regimes respectively. Uniform weighting combines the normalized surrogate scores with equal weights. For images, ReNO uses its published fixed reward weights within the same FMRG pipeline (Eyring et al., 2024). This compares reward mixtures rather than reproducing ReNO's original noise-optimization procedure. We also compare ParetoFlow (Yuan et al., 2025) and SLOP's calibrated selection protocol (Wang et al., 2026). SLOP is calibrated using numerical oracle values; its preference-regime results are therefore a value-calibrated reference. BoN selects the highest-R^* candidate among N candidates, requiring N oracle evaluations per output.

B.2 LANGUAGE GENERATION

We use the released FMLM checkpoint (Lee et al., 2026a) (`david3684/FMLM-B-LM1B`), trained on One Billion Words (Chelba et al., 2013). Generation is unconditional, with 128 tokens from the 30,522-token `bert-base-uncased` vocabulary generated in 8 flow-map steps, at softmax temperature 1 and without per-step noise injection ($\gamma = 0$). We use FMRG steering and feedback budgets of $n = 256$ numerical rewards or $M = 256$ rankings of $m = 8$ candidates, labelled by the same Plackett–Luce annotator as ImageNet ($\sigma = 0.5$). Each evaluated setting is run with 5 seeds of 256 generated samples; metrics are computed per seed and averaged, and $|\text{gap}|$ is taken from the seed-averaged mean.

The oracle is negative per-sequence generative perplexity under Qwen2.5-1.5B (Qwen et al., 2025), scored in fp32 after removing the `[CLS]` separators that LM1B's sentence packing leaves inside each sample. The surrogates are unmasked pseudo-log-likelihoods from BERT and DistilBERT (`bert-base-uncased`, `distilbert-base-uncased`; the mean log-probability the MLM head assigns to each token actually present, in a single forward pass without masking), a CoLA acceptability score (the acceptable-minus-unacceptable logit margin of `textattack/bert-base-uncased-CoLA`), and the empirical unigram entropy of the sample. All four share the generator's vocabulary, so their differentiable extensions read a distribution over tokens directly: it is multiplied into each scorer's word-embedding matrix (between the `[CLS]` and `[SEP]` embeddings), and the entropy is taken of the position-averaged distribution. On one-hot inputs these reduce exactly to the discrete scores. Held-out loss selects an MLP for values and a linear composition for preferences.

Pools and steering. For each run, 2048 unsteered samples are scored by the oracle and all surrogates: 1536 rows supply the feedback and 512 are held out for recovery, and the whole pool, importance-weighted, forms the q^* reference (ESS 1155 of 2048). FMRG follows each flow-map step with two unit-normalised reward-gradient steps, $x \leftarrow x + \lambda \nabla_x r / \|\nabla_x r\|$, where r is evaluated on the endpoint prediction $X_{t,1}(x)$ — one call of the two-time denoiser, on the vocabulary simplex — at every step. The strength grid is $\lambda \in \{0.00375, 0.0075, 0.015, 0.03, 0.06, 0.125, 0.25, 0.5, 1, 2, 4\}$; figures and tables show $\lambda \leq 0.125$, where the methods separate. Candidates may recur across ranking queries, so the $M = 256$ groups contain 2048 candidate occurrences rather than 2048 distinct samples.

B.3 LICENSES AND ATTRIBUTION

The MFM and DMF checkpoints, SiT, SD-VAE-ft-MSE, SD-VAE-ft-EMA, Skywork-VL-Reward-7B, and CLIP are released under MIT licenses; the PickScore project repository also uses MIT. The Qwen2.5-VL-7B-Instruct backbone and the upstream HPSv2.1 and ImageReward models use Apache-2.0; the upstream reward-model licenses apply to the `RE-N-Y` repackagings identified above.

For the underlying image dataset, ImageNet is governed by its research and educational-use agreement. Dataset terms are distinct from the licenses of the pretrained models used to generate the experimental pools.

C METRIC ESTIMATION FOR IMAGE AND LANGUAGE GENERATION

We distinguish the tilt induced by a fitted composition from the samples produced by a steerer. Image and language references use the sampler law $p_{1,\phi}$ and target strength $\beta = 1$ on standardized oracle rewards. The sampler's guidance parameter is swept separately and does not redefine this target.

Reference distribution and effective sample size. Let $x_i \sim p_{1,\phi}$, $i = 1, \dots, N_{\mathrm{ref}}$, be reference samples. With oracle mean a_* and standard deviation $s_* > 0$ estimated on this pool, define

$$z_*(x) = \frac{R^*(x) - a_*}{s_*}, \qquad w_i^* = \frac{e^{\beta z_*(x_i)}}{\sum_{j=1}^{N_{\mathrm{ref}}} e^{\beta z_*(x_j)}}, \qquad \widehat{q}^* = \sum_i w_i^* \delta_{x_i}.$$

The same standardization is used for all compared methods within a task. The effective sample size is

$$\mathrm{ESS} = \frac{1}{\sum_i (w_i^*)^2}.$$

Small ESS indicates that a few reference samples carry most of the weight. The main ImageNet comparison uses the seven class–prompt pairs with reference ESS above 1000 in both feedback regimes, limiting sensitivity to concentrated weights. Sensitivity to this restriction is reported in §E.

Composition KL. Evaluate the fitted score on the same reference samples and standardize it there, $z_{\hat{g}}(x) = (\hat{g}(\mathbf{R}(x)) - a_{\hat{g}})/s_{\hat{g}}$, with $a_{\hat{g}}$ and $s_{\hat{g}}$ its mean and standard deviation over the pool; this is the same standardization applied to the steering reward. Define

$$w_i^{\hat{g}} = \frac{e^{\beta z_{\hat{g}}(x_i)}}{\sum_j e^{\beta z_{\hat{g}}(x_j)}}, \qquad \widehat{\mathrm{KL}}_g = \sum_i w_i^* \log \frac{w_i^*}{w_i^{\hat{g}}}.$$

Because $\widehat{\mathrm{KL}}_g$ depends on $\hat{g}$ only through $z_{\hat{g}}$, it is unchanged by positive affine rescaling of the fitted score. This is the exact KL between the two empirical tilts on the common pool and estimates the corresponding population KL. It requires no steered samples and remains fixed across the sampler-guidance sweep when the score and β are fixed. Standardization changes the induced tilt. With population normalization, conditional-mean regression nevertheless minimizes its leading KL coefficient among standardized surrogate scores (corollary 4); the fourth-order excess-KL result applies to the unscaled conditional mean. Our implementation estimates normalization moments from the reference pool.

Embedding-space energy distance. Let P be the generated distribution and ψ the evaluation embedding. For independent $X, X' \sim P$ and $Y, Y' \sim q^*$, define

$$\mathcal{E}_\psi(P, q^*) = 2\mathbb{E}\|\psi(X) - \psi(Y)\|_2 - \mathbb{E}\|\psi(X) - \psi(X')\|_2 - \mathbb{E}\|\psi(Y) - \psi(Y')\|_2.$$

This compares the embedding distributions: it vanishes when those distributions agree, but need not identify the complete image or text distribution. It requires no bandwidth and includes a within-sample diversity term. We use DINOv2-base (Oquab et al., 2023) for images and token-mean-pooled `all-mpnet-base-v2` (Reimers and Gurevych, 2019) for text, with L_2-normalized embeddings. Reference expectations use self-normalized importance weights. The resulting finite-sample estimate is not generally unbiased and can be negative when self-pairs are excluded.

Sliced Wasserstein distance. For $B = 256$ unit directions θ_b, sampled once using a fixed seed shared across methods, we report

$$\widehat{\mathrm{SW}}_1 = \frac{1}{B} \sum_{b=1}^{B} W_1\Big((\theta_b^\top \psi)_\# \widehat{P}, (\theta_b^\top \psi)_\# \widehat{q}^* \Big),$$

where $T_\# P$ denotes the distribution of $T(X)$ for $X \sim P$. Each projected distance is computed between the unweighted generated samples and weighted reference samples.

Oracle-reward marginal KL. Let $\widehat{f}_P$ and $\widehat{f}_*$ be Gaussian KDEs of the generated and reference oracle rewards, with the reference KDE using weights w_i^*. We estimate

$$\mathrm{KL}_{\mathrm{KDE}} = \int \widehat{f}_*(r) \log \frac{\widehat{f}_*(r)}{\widehat{f}_P(r)} \, \mathrm{d}r.$$

The numerical grid contains 200 points spanning the combined observed reward range with 5% padding. This is a truncated numerical approximation to KL on the oracle-reward marginal, rather than on the full generated distribution.

Gaussian approximation to reward-marginal KL. Let a_P, s_P^2 and a_q, s_q^2 be the means and variances of R^* under P and q^*, respectively. For positive variances, the Gaussian approximation in the same direction is

$$\mathrm{KL}_{\mathcal{N}} = \log \frac{s_P}{s_q} + \frac{s_q^2 + (a_q - a_P)^2}{2s_P^2} - \frac{1}{2}.$$

Reference moments are estimated using w_i^*.

Standardized reward Wasserstein distances. Let F_P and F_q be the CDFs of $z_*(X)$ under P and q^*. For $p \in \{1, 2\}$, define

$$W_p^z = \left[\int_0^1 |F_q^{-1}(u) - F_P^{-1}(u)|^p \, \mathrm{d}u \right]^{1/p}.$$

The same oracle standardization is applied to both distributions. These compare reward marginals, not distributions in the generation space.

Reward gain and gap. For a generated law P, define

$$\mathrm{gain}(P) = \mathbb{E}_P[R^*] - \mathbb{E}_{p_{1,\phi}}[R^*], \qquad \mathrm{gap}(P) = \mathbb{E}_{q^*}[R^*] - \mathbb{E}_P[R^*].$$

The absolute gap is $|\mathrm{gap}(P)|$; a negative signed gap indicates a mean reward above that of the reference tilt. Matching the mean reward alone does not establish distributional fidelity.

Language quality and diversity. Let NLL_i be the total negative log-likelihood of sequence i under the evaluation language model and T_i its number of scored tokens. The per-sequence oracle uses $R^*(x_i) = -\exp(\mathrm{NLL}_i/T_i)$, whereas the aggregate quality metric is corpus perplexity,

$$\mathrm{gPPL}(P) = \exp\left(\frac{\sum_i \mathrm{NLL}_i}{\sum_i T_i} \right).$$

Self-BLEU is the mean BLEU score of each generated sample against the others, computed on the generator's token IDs. Higher Self-BLEU indicates greater overlap across samples. We use *diversity collapse* to describe outputs dominated by repeated tokens or near-duplicate sequences, assessed through within-sequence entropy and Self-BLEU. We report the diversity-adjusted cost

$$Q(P) = \frac{\mathrm{gPPL}(P)}{\mathrm{gPPL}(p_{1,\phi})} \max\left(1, \frac{\mathrm{sBLEU}(P)}{\mathrm{sBLEU}(p_{1,\phi})}\right).$$

It equals 1 for the prior, and lower is better. The one-sided factor penalizes increased repetition without rewarding reduced repetition; it does not ensure fluency or rule out all forms of reward hacking.

Image quality and diversity. Let $\mathbb{E}_P[R^*]$ be the mean oracle reward over the samples of method P, and $\mathbb{E}_{q^*}[R^*]$ the target's, estimated by self-normalised importance weighting of the reference pool. We report the reward error $|\mathrm{gap}|(P) = |\mathbb{E}_{q^*}[R^*] - \mathbb{E}_P[R^*]|$, which penalises falling short of q^* and overshooting it alike. Diversity $\mathrm{Div}(P)$ is the mean pairwise cosine distance, $1 - \cos(f_i, f_j)$ averaged over all pairs $i < j$, between the DINOv2 embeddings (`dinov2-base`, images resized to 224×224) of the decoded samples; higher indicates greater spread. For images, we combine standardized oracle-reward gain with an analogous one-sided diversity penalty:

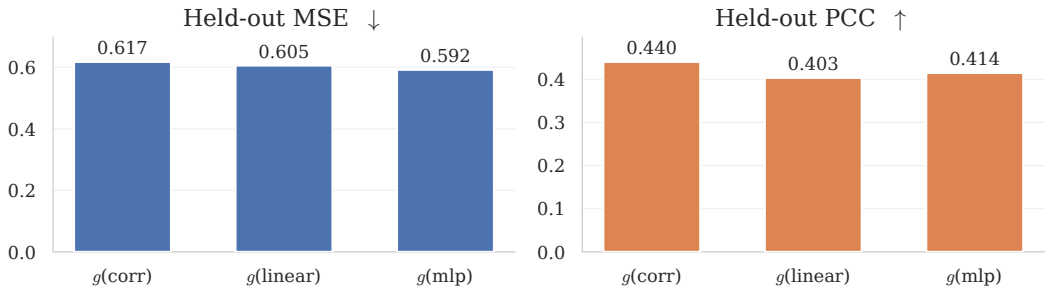
$$Q(P) = \exp\left(-\frac{\mathbb{E}_P[R^*] - \mathbb{E}_{p_{1,\phi}}[R^*]}{s_*} \right) \max\left(1, \frac{\mathrm{Div}(p_{1,\phi})}{\mathrm{Div}(P)}\right).$$

where s_* is the standard deviation of R^* over the reference pool, the scale on which the tilt $z(R^*)$ is defined. Q equals 1 for the prior and lower is better. Because it is always positive, the one-sided factor always penalises lost diversity without rewarding added diversity. As for language, Q rewards reward gain without limit, including gain beyond $\mathbb{E}_{q^*}[R^*]$, so fidelity to the target is read from $|\mathrm{gap}|$ and the distributional metrics. Q is computed per class–prompt pair and then averaged.

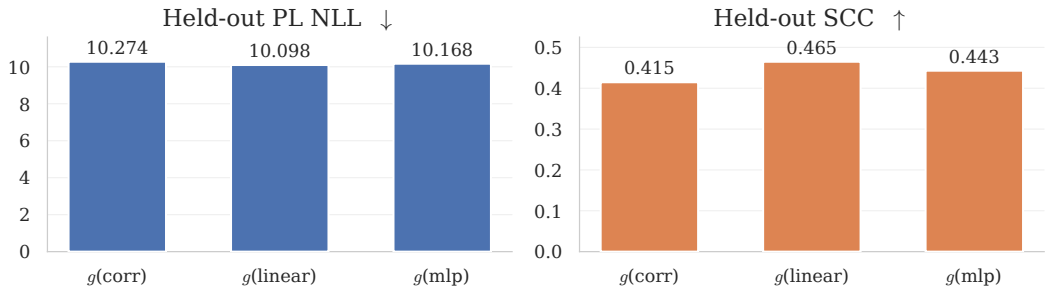
D COMPOSITION SELECTION AND RECOVERY

D.1 LANGUAGE GENERATION

We compare correlation, linear, and MLP compositions using held-out fitting loss. MSE selects an MLP for value feedback, and PL loss selects a linear composition for preferences. The recovery and composition-KL comparisons appear in Figures 5 and 6.

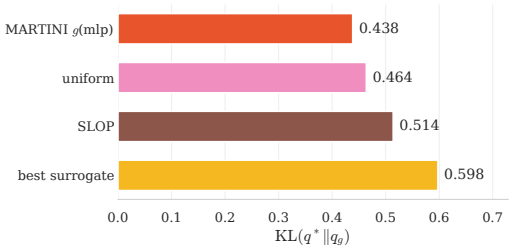


(a) Value feedback.

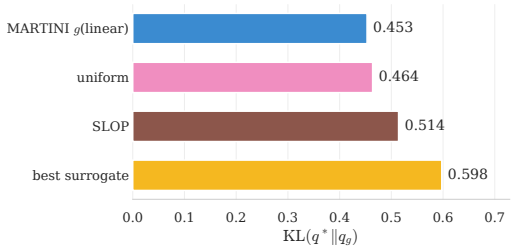


(b) Preference feedback.

Figure 5: **Language reward recovery.** Held-out losses and correlations for the three composition families. Lower loss and higher correlation are better.



(a) Value feedback.

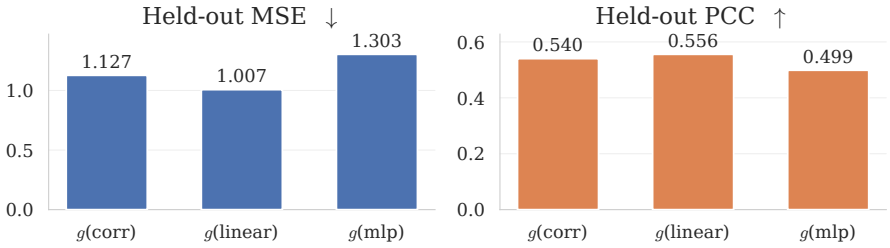


(b) Preference feedback.

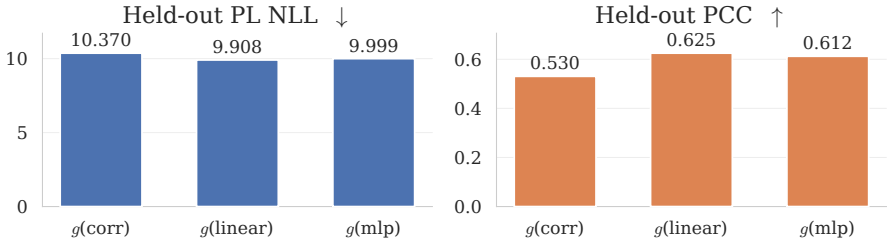
Figure 6: **Language composition fidelity.** Estimated $\mathrm{KL}(q^* \| q_g)$ for the selected compositions and displayed baselines. Lower is better.

D.2 IMAGENET

We average held-out fitting loss across all 19 class–prompt pairs in Table 4. The linear family has the lowest loss in both regimes and is used for steering. Coefficients are fitted separately for each pair.



(a) Value regime, n=32.



(b) Preference regime, M=128.

Figure 7: **ImageNet reward recovery.** Left: held-out MSE for values and PL loss for preferences (lower is better). Right: held-out correlation between predicted and oracle rewards (higher is better).

E ADDITIONAL RESULTS AND SENSITIVITY ANALYSES

E.1 LANGUAGE GENERATION

Table 3 reports the complete language results, including oracle baselines and displayed guidance strengths.

Table 3: **Language generation: complete results.** SB denotes Self-BLEU; Un. counts distinct sequences among 256 outputs. Orange rows use inference-time oracle evaluations. Blue rows show MARTINI compositions.

Method	Strength	Energy ↓	gPPL ↓	SB ↓	Un. ↑	$\mathrm{KL}_{\mathcal{N}}$ ↓	$\mathrm{KL}_{\mathrm{KDE}}$ ↓	W_1^z ↓	W_2^z ↓	|gap|↓
Prior $p_{1,\phi}$	—	5.84×10^{-4}	94.5	0.232	256	0.380	0.326	0.790	0.858	20.33
BoN ($N=8$)	—	1.63×10^{-2}	64.6	0.909	82	1.119	1.517	0.481	0.582	12.20
BoN ($N=64$)	—	1.33×10^{-1}	49.4	0.987	17	20.317	14.635	1.095	1.219	28.02
sBoN ($N=8$)	—	5.40×10^{-3}	79.3	0.773	141	0.024	0.037	0.162	0.184	3.87
sBoN ($N=64$)	—	7.32×10^{-3}	76.0	0.763	137	0.009	0.034	0.091	0.125	0.62
Uniform	0.00375	5.56×10^{-4}	93.4	0.233	256	0.350	0.299	0.743	0.808	19.12
Uniform	0.015	6.56×10^{-4}	93.6	0.234	256	0.351	0.303	0.747	0.810	19.24
Uniform	0.03	6.54×10^{-4}	94.5	0.233	256	0.376	0.327	0.784	0.839	20.18
Uniform	0.06	8.02×10^{-4}	95.8	0.228	256	0.415	0.356	0.841	0.906	21.66
Uniform	0.125	1.30×10^{-3}	95.3	0.225	256	0.403	0.342	0.828	0.902	21.31
BSS	0.00375	5.20×10^{-4}	94.3	0.233	256	0.374	0.319	0.779	0.848	20.06
BSS	0.015	8.72×10^{-4}	96.5	0.232	256	0.437	0.380	0.873	0.952	22.49
BSS	0.03	8.30×10^{-4}	99.6	0.230	256	0.523	0.466	0.995	1.072	25.61
BSS	0.06	1.28×10^{-3}	103.8	0.224	256	0.650	0.602	1.164	1.247	29.98
BSS	0.125	2.52×10^{-3}	108.4	0.216	256	0.782	0.717	1.367	1.476	35.20
ParetoFlow	0.00375	7.65×10^{-2}	92.0	0.975	140	0.340	0.368	0.707	0.855	18.01
ParetoFlow	0.015	6.92×10^{-2}	93.1	0.973	148	0.380	0.337	0.771	0.955	19.70
ParetoFlow	0.03	5.81×10^{-2}	87.9	0.972	154	0.258	0.247	0.575	0.719	14.02
ParetoFlow	0.06	6.15×10^{-2}	90.4	0.969	155	0.335	0.388	0.652	0.779	16.41
ParetoFlow	0.125	8.38×10^{-2}	84.6	0.970	133	0.272	0.422	0.479	0.551	9.76
SLOP	—	1.77×10^{-2}	81.4	0.915	75	0.068	0.069	0.245	0.326	6.09
MARTINI: *values*										
g(corr)	0.00375	5.56×10^{-4}	93.0	0.234	256	0.337	0.286	0.726	0.793	18.69
g(corr)	0.015	5.56×10^{-4}	91.0	0.239	256	0.281	0.239	0.642	0.704	16.52
g(corr)	0.03	4.78×10^{-4}	89.8	0.242	256	0.250	0.212	0.593	0.649	15.26
g(corr)	0.06	6.80×10^{-4}	88.8	0.243	256	0.226	0.185	0.555	0.610	14.28
g(corr)	0.125	9.72×10^{-4}	88.0	0.246	256	0.200	0.167	0.517	0.557	13.29
g(linear)	0.00375	5.02×10^{-4}	92.6	0.235	256	0.323	0.275	0.707	0.771	18.21
g(linear)	0.015	3.82×10^{-4}	89.9	0.239	256	0.249	0.214	0.596	0.644	15.33
g(linear)	0.03	4.66×10^{-4}	87.8	0.245	256	0.196	0.162	0.508	0.558	13.09
g(linear)	0.06	9.68×10^{-4}	86.3	0.250	256	0.160	0.133	0.447	0.491	11.50
g(linear)	0.125	1.69×10^{-3}	84.8	0.256	256	0.124	0.109	0.385	0.417	9.90
g(mlp)	0.00375	6.02×10^{-4}	92.9	0.234	256	0.333	0.282	0.721	0.784	18.56
g(mlp)	0.015	6.28×10^{-4}	91.4	0.239	256	0.292	0.245	0.661	0.718	17.01
g(mlp)	0.03	4.66×10^{-4}	90.2	0.241	256	0.265	0.209	0.615	0.681	15.84
g(mlp)	0.06	7.72×10^{-4}	88.7	0.245	256	0.225	0.176	0.552	0.610	14.21
g(mlp)	0.125	1.07×10^{-3}	88.1	0.249	256	0.213	0.171	0.529	0.590	13.60
MARTINI: *preferences*										
g(corr)	0.00375	5.14×10^{-4}	93.6	0.234	256	0.356	0.302	0.753	0.820	19.37
g(corr)	0.015	5.58×10^{-4}	93.9	0.235	256	0.362	0.311	0.763	0.831	19.63
g(corr)	0.03	5.88×10^{-4}	94.9	0.235	256	0.390	0.335	0.806	0.877	20.74
g(corr)	0.06	8.28×10^{-4}	95.9	0.233	256	0.420	0.370	0.843	0.904	21.71
g(corr)	0.125	1.00×10^{-3}	97.6	0.228	256	0.469	0.406	0.920	0.994	23.69

Continued on next page

Table 3 (continued)

Method	Strength	Energy ↓	gPPL ↓	SB ↓	Un. ↑	$\mathrm{KL}_{\mathcal{N}}$ ↓	$\mathrm{KL}_{\mathrm{KDE}}$ ↓	W_1^z ↓	W_2^z ↓	|gap|↓
g(linear)	0.00375	5.14×10^{-4}	93.2	0.234	256	0.343	0.289	0.736	0.804	18.94
g(linear)	0.015	6.24×10^{-4}	92.0	0.236	256	0.306	0.266	0.682	0.737	17.55
g(linear)	0.03	5.02×10^{-4}	91.5	0.238	256	0.296	0.247	0.666	0.727	17.15
g(linear)	0.06	7.68×10^{-4}	90.5	0.238	256	0.269	0.224	0.625	0.681	16.09
g(linear)	0.125	9.92×10^{-4}	90.1	0.238	256	0.258	0.222	0.609	0.660	15.66
g(mlp)	0.00375	4.56×10^{-4}	93.4	0.234	256	0.353	0.295	0.749	0.826	19.28
g(mlp)	0.015	5.88×10^{-4}	92.2	0.237	256	0.313	0.267	0.692	0.752	17.80
g(mlp)	0.03	3.66×10^{-4}	91.8	0.239	256	0.306	0.250	0.681	0.749	17.54
g(mlp)	0.06	7.14×10^{-4}	90.7	0.244	256	0.276	0.229	0.636	0.696	16.36
g(mlp)	0.125	1.11×10^{-3}	91.1	0.245	256	0.284	0.240	0.649	0.703	16.71

E.2 IMAGENET

Table 4 lists the class–prompt pairs and reference ESS values. Full guidance sweeps and per-pair MARTINI results appear in Tables 5 and 6.

Table 4: **ImageNet class–prompt pairs.** ESS is computed from each regime's independent 8192-sample reference pool; displayed values are rounded. † marks the seven pairs with reference ESS above 1000 in both regimes.

ID	Class	Steering prompt	ESS (Val.)	ESS (Pref.)
555	fire engine†	a fire engine on a busy city street	4043	4081
294	brown bear†	a brown bear wading in a mountain river	3903	3826
84	peacock†	a peacock displaying in a lush garden	3728	3757
145	king penguin†	a king penguin standing on the ice	2451	2418
347	bison†	a bison grazing on the open prairie	2199	2291
386	African elephant†	an African elephant at a watering hole	2031	2001
779	school bus†	a school bus on a tree-lined suburban street	1250	1304
483	castle	a castle on a hilltop at dawn	993	828
250	Siberian husky	a Siberian husky in deep snow	975	975
130	flamingo	a flamingo wading in shallow water at sunset	935	908
717	pickup truck	a pickup truck on a dirt road in the countryside	596	640
144	pelican	a pelican perched on a wooden pier by the sea	500	349
340	zebra	a zebra grazing on the open savanna	351	374
9	ostrich	an ostrich running across a dusty plain	111	85
291	lion	a lion resting on the savanna at sunset	90	44
817	sports car	a sports car parked on a city street at night	48	34
281	tabby cat	a tabby cat sitting in the snow	11	13
354	Arabian camel	an Arabian camel crossing desert dunes at sunset	3	2
207	golden retriever	a golden retriever running on a sandy beach	1	1

Table 5: **ImageNet steering-strength sweep.** Means over the seven marked pairs in Table 4. Setting gives guidance strength, except for SLOP, where it gives candidate count N. Composition KL is shown once per fitted composition. Lower is better for all metrics; SLOP is value-calibrated.

Method	Setting	KL_g	Energy	Sl. W_1	$\mathrm{KL}_{\mathcal{N}}$	$\mathrm{KL}_{\mathrm{KDE}}$	W_1^z	W_2^z	|gap|	Q
Prior $p_{1,\phi}$	—	—	0.0683	0.0080	0.752	0.658	1.253	1.380	7.51	1.000
BoN ($N=8$)	—	—	0.0654	0.0078	0.501	0.671	0.546	0.658	3.17	0.193
BoN ($N=16$)	—	—	0.0795	0.0084	2.918	3.121	0.951	1.120	5.53	0.142
BoN ($N=32$)	—	—	0.0844	0.0085	9.964	7.721	1.271	1.473	7.44	0.106
Uniform	2	0.436	0.0832	0.0086	0.044	0.089	0.311	0.374	1.29	0.418
Uniform	3		0.0688	0.0078	0.069	0.091	0.291	0.332	1.47	0.291

Continued on next page

Table 5 (continued)

Method	Setting	KL_g	Energy	Sl. W_1	$\mathrm{KL}_{\mathcal{N}}$	$\mathrm{KL}_{\mathrm{KDE}}$	W_1^z	W_2^z	|gap|	Q
Uniform	5		0.0648	0.0076	0.696	0.646	0.544	0.627	3.19	0.222
Uniform	8		0.0813	0.0081	1.415	1.123	0.631	0.725	3.55	0.207
ReNO	2	0.553	0.0749	0.0082	0.051	0.090	0.309	0.379	1.36	0.400
ReNO	3		0.0578	0.0074	0.039	0.080	0.248	0.289	0.90	0.314
ReNO	5		0.0527	0.0070	0.238	0.196	0.411	0.465	2.34	0.235
ReNO	8		0.0812	0.0081	1.419	0.893	0.552	0.634	3.17	0.218
ParetoFlow	2 ($t = 0$)	—	0.1696	0.0116	0.490	0.528	0.932	1.086	5.19	0.944
ParetoFlow	3 ($t = 0$)		0.1422	0.0109	0.404	0.438	0.917	1.079	5.08	0.801
ParetoFlow	5 ($t = 0$)		0.1405	0.0108	0.550	0.498	0.885	1.012	4.96	0.811
ParetoFlow	8 ($t = 0$)		0.1282	0.0101	0.145	0.225	0.465	0.541	1.92	0.447
SLOP	$N = 8$	0.520	0.1020	0.0091	0.077	0.091	0.328	0.388	1.38	0.359
SLOP	$N = 16$		0.1200	0.0098	0.168	0.170	0.365	0.446	1.98	0.364
SLOP	$N = 32$		0.1147	0.0095	0.420	0.327	0.551	0.643	3.08	0.273
Values ($n = 32$)										
BSS	2	0.514	0.0674	0.0078	0.131	0.177	0.539	0.635	2.79	0.532
BSS	3		0.0535	0.0070	0.113	0.142	0.455	0.548	2.23	0.468
BSS	5		0.0529	0.0069	0.162	0.144	0.385	0.432	1.94	0.340
BSS	8		0.0960	0.0088	0.157	0.161	0.399	0.456	2.03	0.330
g(corr)	2	0.425	0.0673	0.0078	0.059	0.097	0.348	0.431	1.83	0.434
g(corr)	3		0.0618	0.0075	0.062	0.084	0.300	0.339	1.61	0.314
g(corr)	5		0.0596	0.0074	0.774	0.625	0.531	0.605	2.98	0.247
g(corr)	8		0.0778	0.0080	1.317	0.991	0.698	0.800	4.02	0.241
g(linear)	2	0.466	0.0559	0.0073	0.104	0.124	0.471	0.556	2.69	0.495
g(linear)	3		0.0409	0.0064	0.039	0.069	0.288	0.352	1.40	0.388
g(linear)	5		0.0339	0.0059	0.061	0.069	0.319	0.377	1.64	0.360
g(linear)	8		0.1010	0.0091	0.246	0.312	0.400	0.475	1.88	0.431
g(mlp)	2	0.488	0.0525	0.0071	0.186	0.204	0.653	0.744	3.80	0.594
g(mlp)	3		0.0355	0.0060	0.068	0.091	0.359	0.433	1.97	0.433
g(mlp)	5		0.0319	0.0058	0.165	0.168	0.410	0.492	2.18	0.436
g(mlp)	8		0.0956	0.0090	0.277	0.316	0.514	0.600	2.80	0.434
Preferences ($M = 128$)										
BSS	2	0.440	0.0628	0.0077	0.066	0.102	0.368	0.463	1.79	0.429
BSS	3		0.0552	0.0073	0.035	0.080	0.258	0.306	0.90	0.359
BSS	5		0.0425	0.0065	0.133	0.134	0.388	0.433	2.00	0.312
BSS	8		0.0861	0.0084	0.264	0.254	0.427	0.494	2.08	0.323
g(corr)	2	0.417	0.0752	0.0082	0.045	0.089	0.314	0.366	1.10	0.385
g(corr)	3		0.0716	0.0080	0.048	0.079	0.314	0.348	1.57	0.284
g(corr)	5		0.0638	0.0075	1.013	0.866	0.578	0.648	3.50	0.199
g(corr)	8		0.0786	0.0080	3.831	1.905	0.670	0.777	4.02	0.181
g(linear)	2	0.423	0.0580	0.0074	0.095	0.132	0.403	0.490	2.11	0.441
g(linear)	3		0.0484	0.0068	0.053	0.106	0.345	0.398	1.54	0.384
g(linear)	5		0.0380	0.0061	0.097	0.123	0.365	0.423	1.93	0.287
g(linear)	8		0.0780	0.0081	0.400	0.327	0.389	0.454	2.13	0.303
g(mlp)	2	0.426	0.0620	0.0076	0.081	0.107	0.408	0.486	2.11	0.432
g(mlp)	3		0.0516	0.0070	0.070	0.096	0.313	0.365	1.51	0.350
g(mlp)	5		0.0330	0.0059	0.525	0.458	0.384	0.458	1.86	0.310
g(mlp)	8		0.0783	0.0081	0.263	0.248	0.503	0.576	2.56	0.360

Table 6: **MARTINI by ImageNet class–prompt pair.** Linear compositions use each pair's lowest-energy evaluated strength. ESS is rounded and refers to the regime-specific 8192-sample pool. † marks pairs in the main comparison. Dashes indicate unreported estimates.

Class	ESS	Str.	KL_g	Energy	Sl. W_1	$\mathrm{KL}_{\mathcal{N}}$	$\mathrm{KL}_{\mathrm{KDE}}$	W_1^z	W_2^z	|gap|	Q
Values ($n = 32$)											
fire engine†	4043	5	0.360	0.0351	0.0063	0.028	0.029	0.187	0.211	1.10	0.277
brown bear†	3903	5	0.257	0.0289	0.0056	0.035	0.054	0.140	0.234	0.94	0.241
peacock†	3728	5	0.416	0.0229	0.0052	0.033	0.068	0.194	0.229	0.51	0.449
king penguin†	2451	3	0.652	0.0352	0.0055	0.011	0.034	0.200	0.228	0.81	0.272
bison†	2199	3	0.462	0.0238	0.0052	0.031	0.042	0.296	0.363	1.71	0.546
African elephant†	2031	5	0.399	0.0635	0.0075	0.012	0.025	0.172	0.240	0.95	0.179
school bus†	1250	2	0.718	0.0152	0.0051	0.086	0.091	0.593	0.670	2.88	0.547
castle	993	5	0.720	0.0721	0.0095	0.013	0.031	0.282	0.312	0.78	0.175
Siberian husky	975	5	0.525	0.0571	0.0081	2.349	1.117	2.203	2.545	12.48	0.715
flamingo	935	3	0.443	0.0429	0.0058	0.021	0.043	0.332	0.372	1.00	0.294
pickup truck	596	3	0.959	0.0162	0.0063	0.024	0.077	0.470	0.567	1.60	0.072
pelican	500	3	1.010	0.0127	0.0046	0.466	0.364	1.682	1.917	7.27	0.415
zebra	351	5	0.916	0.0308	0.0058	0.234	0.174	1.030	1.134	3.95	0.366
ostrich	111	2	1.278	0.0302	0.0060	0.213	0.159	1.590	1.803	5.01	0.208
lion	90	8	1.828	0.0583	0.0077	0.333	0.285	1.969	2.081	4.46	0.003
sports car	48	5	2.317	0.0417	0.0082	0.528	0.436	2.982	3.396	7.54	0.075
tabby cat	11	2	1.363	0.0773	0.0130	8.625	2.676	8.810	9.189	22.16	0.526
Arabian camel	3	2	3.820	0.0532	0.0140	8.717	3.435	8.824	9.073	12.13	0.717
golden retriever	1	5	4.779	0.0924	0.0271	1.815	1.470	12.072	13.751	13.90	0.000
Mean† (7 pairs)			0.466	0.0321	0.0058	0.034	0.049	0.255	0.310	1.27	0.359
Preferences ($M = 128$)											
fire engine†	4081	5	0.317	0.0508	0.0072	0.390	0.330	0.615	0.669	3.90	0.195
brown bear†	3826	5	0.266	0.0409	0.0064	0.053	0.130	0.282	0.305	1.96	0.209
peacock†	3757	5	0.286	0.0208	0.0050	0.092	0.168	0.438	0.500	2.18	0.282
king penguin†	2418	5	0.639	0.0352	0.0053	0.068	0.067	0.260	0.300	0.68	0.217
bison†	2291	5	0.423	0.0287	0.0056	0.011	0.030	0.184	0.220	0.98	0.492
African elephant†	2001	3	0.433	0.0717	0.0083	0.001	0.034	0.131	0.171	0.33	0.224
school bus†	1304	5	0.596	0.0123	0.0048	0.034	0.086	0.455	0.602	1.96	0.459
Siberian husky	975	5	0.479	0.0515	0.0071	1.233	0.809	1.835	2.147	10.39	0.498
flamingo	908	3	0.418	0.0200	0.0045	0.021	0.047	0.370	0.385	1.05	0.286
castle	828	3	0.630	0.0524	0.0085	0.008	0.031	0.278	0.335	0.56	0.167
pickup truck	640	5	0.797	0.0181	0.0064	0.172	0.173	0.800	0.885	3.30	0.025
zebra	374	3	0.885	0.0326	0.0055	0.500	0.318	1.470	1.622	5.51	0.641
pelican	349	3	0.720	0.0156	0.0048	0.217	0.162	1.226	1.352	5.35	0.194
ostrich	85	2	1.174	0.0167	0.0059	0.102	0.074	1.159	1.333	3.60	0.113
lion	44	5	2.179	0.0333	0.0068	0.161	0.224	1.602	1.818	3.21	0.043
sports car	34	5	2.648	0.0591	0.0095	0.904	0.629	3.627	3.970	8.75	0.090
tabby cat	13	3	2.062	0.0566	0.0124	8.150	3.535	8.472	8.796	20.94	0.339
Arabian camel	2	2	4.655	—	0.0155	12.306	10.363	12.002	12.207	16.79	0.502
golden retriever	1	3	3.261	—	0.0280	6.451	2.279	17.331	17.984	22.68	0.064
Mean† (7 pairs)			0.423	0.0372	0.0061	0.093	0.121	0.338	0.395	1.71	0.297

Table 7 assesses sensitivity to the reference ESS cutoff. For each cutoff, we aggregate each method's metrics across retained pairs using the mean or median, then rank the methods. The comparison includes MARTINI, BSS, uniform weighting, ReNO, ParetoFlow, and SLOP; composition KL excludes ParetoFlow. SLOP remains a value-calibrated reference in the preference regime. Steering methods use their per-pair lowest-energy evaluated strength. The Siberian husky pair is retained whenever it meets the cutoff.

Table 7: **Sensitivity to reference ESS.** Each entry gives MARTINI 's rank using mean/median metrics across retained pairs; bold 1 marks first place. Cutoffs apply separately to each regime's 8192- sample reference pool.

ESS cutoff	Pairs	KL_g	Energy	Sl. W_1	$\mathrm{KL}_{\mathcal{N}}$	$\mathrm{KL}_{\mathrm{KDE}}$	W_1^z	W_2^z	$\lvert\mathrm{gap}\rvert$	Q
Values ($n = 32$)										
≥ 0	19	**1**/2	**1/1**	**1/1**	4/**1**	2/**1**	3/**1**	3/**1**	4/2	5/4
≥ 100	14	2/3	**1/1**	**1/1**	3/**1**	2/**1**	4/**1**	4/**1**	4/**1**	4/4
≥ 350	13	2/3	**1/1**	**1/1**	3/**1**	2/**1**	3/**1**	2/**1**	3/**1**	4/4
≥ 500	11	2/**1**	**1/1**	**1/1**	4/**1**	**1/1**	2/**1**	2/**1**	2/**1**	5/3
≥ 850	10	2/**1**	**1/1**	**1/1**	4/**1**	2/**1**	3/**1**	3/**1**	3/**1**	5/3
≥ 950	9	2/**1**	**1/1**	**1/1**	4/**1**	3/**1**	3/**1**	3/**1**	3/**1**	5/3
≥ 1000	7	2/2	**1/1**	**1/1**	**1/1**	**1/1**	**1/1**	**1/1**	**1/1**	4/3
Preferences ($M = 128$)										
≥ 0	19	**1/1**	**1/1**	**1/1**	3/3	3/4	4/4	4/4	4/4	3/3
≥ 100	13	**1/1**	1/2	**1/1**	4/2	3/4	4/**1**	4/**1**	5/**1**	4/3
≥ 350	12	**1/1**	**1/1**	**1/1**	4/2	3/2	3/2	3/**1**	4/**1**	4/3
≥ 500	11	**1/1**	**1/1**	**1/1**	3/2	3/2	3/2	3/**1**	3/2	3/3
≥ 850	9	**1/1**	**1/1**	**1/1**	4/2	4/2	4/2	4/**1**	4/3	4/3
≥ 950	8	**1/1**	**1/1**	**1/1**	4/3	4/4	5/2	5/2	5/3	4/3
≥ 1000	7	1/2	**1/1**	**1/1**	1/2	1/3	**1/1**	**1/1**	1/3	3/**1**

Across all cutoffs, MARTINI ranks first on mean energy and sliced Wasserstein distance. Reward-marginal ranks depend on the cutoff and aggregation, with the Siberian husky pair substantially affecting the means.

E.3 BUDGET ANALYSIS

We compare $n \in \{32, 256\}$ numerical rewards and $M \in \{8, 128\}$ rankings of $m = 8$ candidates. Increasing feedback improves held-out loss, PCC, and KL_g for all three families (Table 8). Linear compositions remain selected at every budget, although the MLP's loss approaches that of the linear family at the larger budgets. For the selected compositions, KL_g decreases from 0.466 to 0.415 with values and from 0.475 to 0.423 with preferences.

Table 8: **ImageNet composition fitting across feedback budgets.** Loss is MSE on $z(R^*)$ for values and PL NLL per group for preferences. Bold marks the lowest mean selection loss. BSS uses the same validation split to select its surrogate and family; only its composition KL is reported.

		Held-out loss ↓			PCC ↑			KL_g ↓			
Regime	Budget	Corr.	Linear	MLP	Corr.	Linear	MLP	Corr.	Linear	MLP	BSS
Value	$n = 32$	1.127	**1.007**	1.303	0.540	0.556	0.499	0.425	0.466	0.488	0.514
	$n = 256$	0.898	**0.706**	0.726	0.549	0.629	0.621	0.416	0.415	0.372	0.460
Pref.	$M = 8$	10.774	**10.652**	11.692	0.453	0.532	0.426	0.420	0.475	0.519	0.858
	$M = 128$	10.370	**9.908**	9.999	0.530	0.625	0.612	0.417	0.423	0.426	0.440

For each steering method and pair, all generated-sample metrics use the strength with the lowest observed energy distance to q^*. Improved fitting does not uniformly improve realized steering (Table 9). MARTINI achieves lower energy than BSS at every budget, but energy does not improve with additional feedback. Reward-marginal fidelity improves with more preferences, whereas it worsens at $n = 256$: several pairs overshoot the target mean reward at the selected strengths, despite improved composition KL. This separates improvements in fitting from the effects of guidance and sampling.

Table 9: **ImageNet steering across feedback budgets.** Means over the seven evaluation pairs. MARTINI uses the linear composition at every budget. Lower is better; bold marks the better of each method's two budgets. Fit-free baselines are shared.

			Embedding space		R^*-marginal				Reward	
Method	Budget	KL_g	Energy	Sl. W_1	$\mathrm{KL}_{\mathcal{N}}$	$\mathrm{KL}_{\mathrm{KDE}}$	W_1^z	W_2^z	$\lvert\mathrm{gap}\rvert$	Q
Fit-free baselines										
Prior $p_{1,\phi}$	–	–	0.0683	0.0080	0.752	0.658	1.253	1.380	7.51	1.000
ParetoFlow	–	–	0.0885	0.0088	0.825	0.937	0.948	1.068	5.35	0.844
ReNO	–	0.553	0.0503	0.0069	0.197	0.166	0.380	0.430	2.10	0.244
Uniform	–	0.436	0.0603	0.0074	0.605	0.635	0.475	0.552	2.73	0.238
Value feedback										
SLOP	$n = 32$	0.520	0.1200	0.0098	**0.168**	**0.170**	**0.365**	**0.446**	**1.98**	0.364
SLOP	$n = 256$	**0.423**	**0.0867**	**0.0086**	0.376	0.265	0.532	0.623	2.99	**0.250**
BSS	$n = 32$	0.514	0.0498	0.0068	0.217	0.183	0.478	0.547	2.57	0.354
BSS	$n = 256$	**0.460**	**0.0389**	**0.0063**	**0.093**	**0.108**	**0.291**	**0.348**	**1.35**	**0.285**
MARTINI	$n = 32$	0.466	**0.0321**	**0.0058**	**0.034**	**0.049**	**0.255**	**0.310**	**1.27**	0.359
MARTINI	$n = 256$	**0.415**	0.0377	0.0060	0.549	0.764	0.454	0.527	2.72	**0.279**
Preference feedback										
BSS	$M = 8$	0.858	**0.0412**	**0.0063**	0.323	0.262	0.570	0.637	2.82	0.585
BSS	$M = 128$	**0.440**	0.0423	0.0065	**0.135**	**0.133**	**0.392**	**0.442**	**2.04**	**0.316**
MARTINI	$M = 8$	0.475	**0.0337**	**0.0059**	0.544	0.629	0.606	0.679	3.49	0.386
MARTINI	$M = 128$	**0.423**	0.0372	0.0061	**0.093**	**0.121**	**0.338**	**0.395**	**1.71**	**0.297**

F PROTEIN–LIGAND COFOLDING

F.1 SURROGATES

We fit the surrogate composition to experimental binding-affinity feedback and evaluate its transfer to generated structures. The five surrogates provide information about structural quality, binding-site energetics, and predicted affinity.

SMINA. SMINA is a widely used docking and scoring tool forked from AutoDock Vina and optimised for high-throughput scoring and for user-specified scoring functions (Koes et al., 2013).

Frustration. The FrustratometeR R package Rausch et al. (2021) is used to calculate local energetic frustration in protein structures, where high frustration relates to unstable local regions and correlates with active sites.

MolProbity. MolProbity (Chen et al., 2010) assesses the physical quality of molecular structures through all-atom steric contact analysis with explicit hydrogens, together with covalent-geometry and backbone and side-chain torsion-angle criteria.

Boltz Affinity. The boltz affinity prediction head introduced in Passaro et al. (2025) is used here as an affinity surrogate. The Boltz-2 weights are used to featurise the complex, but the pose itself is taken as atomic coordinates from the steering run (e.g. with DeCAF) and then used to constrain the affinity prediction, so that the prediction head uses the steered poses. Default Boltz-2 settings were used for this surrogate.

Boltz-1x physical potentials. We include the physical potential terms used by Boltz-1x (Wohlwend et al., 2025) to penalize stereochemical violations, clashes, and inconsistent chain geometry. The resampling coefficients are

`symmetric_chain_com`	0.5
`connections`	1.0
`posebusters`	0.1
`chiral`	1.0
`stereo`	1.0
`planar`	1.0
`vdw_overlap`	$\begin{cases} 0.01, & t < 0.6, \\ 0, & t \geq 0.6. \end{cases}$

Here t is the resampling-time variable used by the implementation. These seven terms include both within-chain and between-chain contributions.

We normalize the surrogate reward values both for fitting g and during steering. Figure 8 displays the correlation of the surrogate rewards with the experimental HiQBind $\mathrm{p}K_D$ as well as the theoretical recovery ceiling when using these surrogate rewards. SMINA is the best single surrogate based on the Pearson correlation. Boltz energy potentials are instrumental in ensuring physical validity of the generated poses; therefore, each baseline includes normalized energy potentials.

The theory ceiling in Figure 8 is $\rho_{\mathrm{info}} = \sqrt{1 - \bar{v}/\mathrm{Var}(R^*)}$, the highest PCC any function of the scores can reach, where $\bar{v} = \mathbb{E}[\mathrm{Var}(R^* \mid \mathbf{R})]$ is the variance of the measured affinity that the HiQBind scores leave unexplained, estimated by leave-one-out local-linear k-nearest-neighbour regression of R^* on $\mathbf{R}$.

F.2 FEEDBACK AND EVALUATION DATASETS

HiQBind. HiQBind is a curated protein–ligand binding dataset constructed by matching experimentally determined binding free energies from BioLiP, Binding MOAD, and BindingDB to co-crystallized PDB complexes and reprocessing each entry through the open-source HiQBind-WF workflow (Wang et al., 2025b). The release comprises 18,160 PDB entries with binding-affinity annotations and 32,275 refined non-covalent complexes, annotated with affinity values, release year, resolution, ligand and protein identifiers, and derived ligand properties. We subset this data set to monomers and multimers that have experimental $\mathrm{p}K_D$, leaving us with N=7095 for the feedback set.

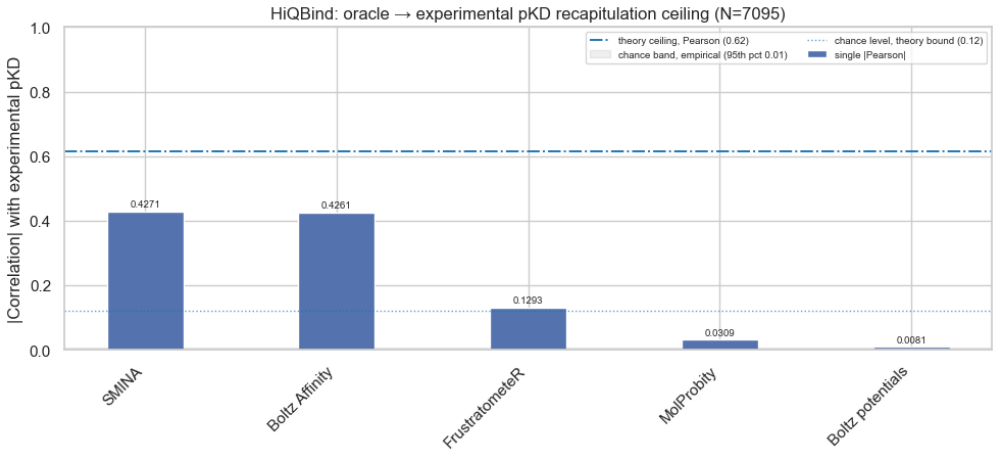


Figure 8: Correlation of HiQBind scores with the experimentally measured binding affinity values

Runs N' Poses. We evaluate on 186 complexes from the post-2023 subset of Runs N' Poses (Škrinjar et al., 2026), beyond Boltz-1's training cutoff. These evaluation complexes lack experimental affinity labels, so we measure pose accuracy and physical validity. This tests whether a composition fitted to affinity feedback transfers to structure generation; it does not directly measure recovery of the experimental affinity tilt. Figure 9 compares the surrogate marginals in the fitting and evaluation datasets.

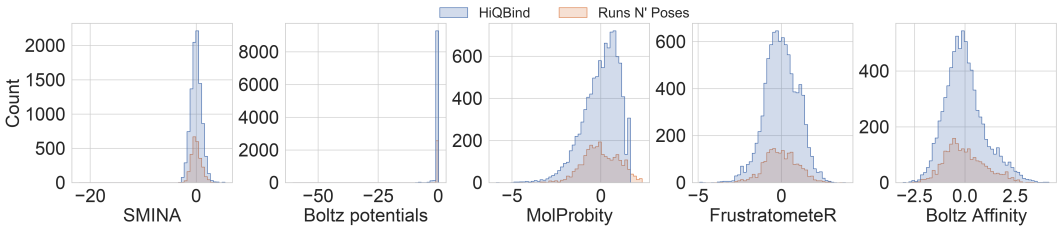


Figure 9: **Surrogate-score distributions.** Normalized marginals on HiQBind and Runs N' Poses.

F.3 Protein success metrics

We use the OpenStructure (`ost`) package for calculating both the RMSD and *lDDT-PLI* Biasini et al. (2013), with the same defaults as in Scarpellini et al. (2026).

Diversity. We report structural diversity, defined as one minus the similarity, where similarity is the mean pairwise TM-score Zhang and Skolnick (2004) across the top n_{seed} best poses (best within one seed across particles) where n_{seed} is the number of seeds.

We report best@1 metrics as defined in DeCAFScarpellini et al. (2026) averaged over 5 seeds.

F.4 Extended results and ablations

We compare FK steering with 10 and 15 particles in Table 11 and vary the relative weights of smina and the energy potentials in Table 10 to choose the strongest baseline combination. We report joint success, physical validity, and lDDT-PLI to assess sensitivity to these choices.

Ablating between the mixtures of fixed-weights shows that $[0.1, 0.9]$ gives the highest success rate (61.8%), while $[0.9, 0.1]$ gives the highest lDDT-PLI and lower PB-validity. We pick the mixture with the highest success rate as the BSS baseline reported in other tables, and use this mixture for BSS baseline experiment with 15 particles. The learned linear mixture reaches 62.4% success, which beats all of the ablated mixtures, demonstrating utility of MARTINI even with fewer surrogates.

Table 10: **Steering with smina and energy potentials, weight ablation.** 10 particles. Best@1 averaged over 5 seeds. **Bold** and underline mark the best and runner-up value per column.

$[w_{\text{smina}}, w_{\text{potentials}}]$	Success Rate ↑	RMSD < 2 Å ↑	PB Valid ↑	lDDT-PLI ↑
$[0.1, 0.9]$	<u>61.8</u>	<u>61.8</u>	<u>97.6</u>	65.1
$[0.3, 0.7]$	61.2	61.2	**97.8**	65.1
$[0.5, 0.5]$	61.1	61.1	97.4	65.0
$[0.7, 0.3]$	61.3	61.3	96.5	65.1
$[0.9, 0.1]$	61.5	61.6	94.4	<u>65.4</u>
g(linear)	**62.4**	**62.7**	93.5	**66.3**

Table 11: **Numer of particles ablation, FK-Steering.** All surrogate rewards are normalized. Best@1 averaged over 5 seeds. **Bold** and underline mark the best and runner-up value per column.

	Num Particles $= 10$			Num Particles $= 15$		
Method	Success Rate ↑	PB Valid ↑	lDDT-PLI ↑	Success Rate ↑	PB Valid ↑	lDDT-PLI ↑
Baselines						
DeCAF potentials	61.8	**97.9**	65.2	61.8	**98.3**	65.7
smina + potentials	61.8	<u>97.6</u>	65.1	61.8	<u>98.2</u>	65.4
MARTINI						
g(lin: smina, poten.)	62.4	93.5	<u>66.3</u>	62.7	93.3	66.5
g(corr: all 5)	**63.3**	91.9	**66.4**	**64.5**	91.8	**67.6**
g(lin: all 5)	<u>62.8</u>	95.1	65.6	63.4	94.7	66.5
g(mlp: all 5)	62.2	83.0	65.5	<u>63.7</u>	82.7	<u>67.3</u>

G ADDITIONAL THEORETICAL RESULTS

This section gives the assumptions and supporting results for the two main theorems in §3.3. Proofs appear in §H.

G.1 ASSUMPTIONS

We analyze population quantities and ideal tilted distributions, with $p_{1,\phi} = p_1$. The space $\mathcal{X}$ carries its Borel σ-algebra, p_1 is a probability law, and the oracle, surrogates, and compositions are measurable. Conditional statements hold almost surely under the specified law. Compositions in the main theorems are fixed as $\beta \downarrow 0$; the explicitly defined g^{opt} in Proposition 1 may depend on β.

Assumption G.1 (Finite second moments)**.** The oracle and every compared score satisfy

$$\mathbb{E}_{p_1}[(R^*)^2] < \infty, \qquad \mathbb{E}_{p_1}[(g \circ \mathbf{R})^2] < \infty.$$

Assumption G.2 (Essential boundedness)**.** There exist constants $M_*, M_g < \infty$, independent of β, such that

$$p_1(|R^*| \le M_*) = 1, \qquad p_1(|g \circ \mathbf{R}| \le M_g) = 1.$$

The score bound is required for each composition used in an expansion; M_g need not be uniform over all compositions.

Assumption G.2 implies Assumption G.1 and ensures finite, positive tilt normalizers. It concerns the oracle and composed scores p_1-almost surely, rather than bounds on every raw surrogate globally on $\mathcal{X}$.

Assumption G.3 (Mean-aligned representatives)**.** Every compared composition is represented so that

$$\mathbb{E}_{p_1}[g \circ \mathbf{R}] = \mathbb{E}_{p_1}[R^*].$$

In the preference analysis, this holds for every $g \in \mathcal{G}$, including the population minimizer g_{PL}.

This is a normalization convention for analysis. For an unshifted composition g, define

$$\bar{g}(\mathbf{r}) := g(\mathbf{r}) + \mathbb{E}_{p_1}[R^*] - \mathbb{E}_{p_1}[g \circ \mathbf{R}].$$

Then $q_{\bar{g}} = q_g$ and $\mathrm{PL}(\bar{g}) = \mathrm{PL}(g)$, whereas

$$\mathrm{MSE}(\bar{g}) = \mathrm{MSE}(g) - \big(\mathbb{E}_{p_1}[g \circ \mathbf{R}] - \mathbb{E}_{p_1}[R^*]\big)^2.$$

Thus mean alignment preserves tilts and PL probabilities, but can reduce MSE. In the preference analysis, $\mathcal{G}$ denotes the mean-aligned image of the original class; this affine transformation preserves convexity, score-range widths, PL risks, and induced tilts. The theoretical bounds use these representatives and require no additional fitting step or oracle access. The aligned MSE is no larger than the unshifted population MSE estimated by (3).

Assumption G.4 (Preference observations and composition class)**.** (i) Each query contains $m \ge 2$ independent candidates $S \sim p_1^{\otimes m}$, with ranking

$$\pi \mid S \sim p^*_{\mathrm{PL}}(\cdot \mid S) := p_{\mathrm{PL}}(\cdot \mid R^*, S).$$

Observed queries are independent.

(ii) The class $\mathcal{G}$ is nonempty and convex, and there is a common $\Delta < \infty$ such that, for every $g \in \mathcal{G}$,

$$\operatorname*{ess\,sup}_{p_1}(g \circ \mathbf{R}) - \operatorname*{ess\,inf}_{p_1}(g \circ \mathbf{R}) \le \Delta.$$

(iii) The population loss

$$\mathrm{PL}(g) := \mathbb{E}_{S,\pi}[-\log p_{\mathrm{PL}}(\pi \mid g \circ \mathbf{R}, S)]$$

attains its minimum over $\mathcal{G}$ at some $g_{\mathrm{PL}} \in \arg\min_{g \in \mathcal{G}} \mathrm{PL}(g)$.

Existence of g_{PL} is assumed; convexity and a range bound alone do not establish it. Together with Assumption G.3, the common width implies

$$|g \circ \mathbf{R} - \mathbb{E}_{p_1}[R^*]| \le \Delta \quad p_1\text{-almost surely}, \qquad g \in \mathcal{G}.$$

Hence the preference scores are already bounded. Boundedness of R^* is additionally imposed for the KL and reward expansions, but is not needed for the exact recovery bound when $R^* \in L^2(p_1)$.

Relation to fitted models. The common range bound applies to the entire admissible score class, not merely to observed scores. For example, if $\|\mathbf{R}(x)\|_2 \le B$ p_1-almost surely, the convex linear class $\{g_{\mathbf{w}} : \|\mathbf{w}\|_2 \le W\}$ has score-range width at most $2BW$. Ridge regularization and finite empirical ranges alone do not establish this condition for an unrestricted class. Theorem 2 remains conditional on the stated class assumptions.

Which assumptions each conclusion uses. The following conditions suffice for the individual conclusions; the main statements collect them for convenience.

Conclusion	Assumptions used
Theorem 1(i)	G.2; also G.3 for the equality with MSE.
Theorem 1(ii), exact identity	G.1 for R^* and $R \in L^\infty(p_1)$; no small-β limit.
Theorem 1(ii)–(iii), asymptotic bounds	G.2, with g fixed.
Theorem 1(iv)	G.1 only.
Theorem 2, $\mathrm{MSE}(g) \le B(g)$ and (iv)	G.1, G.3, and G.4.
Theorem 2(i)–(iii)	The preceding conditions and the oracle bound in G.2, with g fixed.

Mean alignment is unnecessary for the value gain and gap bounds: $\mathrm{Var}_{p_1}(g \circ \mathbf{R} - \mu) \le \mathrm{MSE}(g) - \bar{v}$ suffices. The exact recovery identities and the preference recovery bound hold without taking $\beta \downarrow 0$. For fixed-score expansions, the $O(\beta^j)$ constants may depend on the problem and score and need not be uniform over the fitted class. The expansions involving g^{opt} account explicitly for its dependence on β.

Lemma 1 (Likelihood controls recovery)**.** *Under Assumptions G.1, G.3, and G.4, for every $g \in \mathcal{G}$, with $c_\Delta = (1 - \frac{1}{m})e^{-\Delta}$,*

$$\mathrm{Var}_{p_1}(g \circ \mathbf{R} - g_{\mathrm{PL}} \circ \mathbf{R}) \le \frac{2}{c_\Delta}\big(\mathrm{PL}(g) - \mathrm{PL}(g_{\mathrm{PL}})\big).$$

The variance inequality is invariant to additive shifts.

G.2 Theoretical floor

The exact recovery floor is $\bar{v}$. The exact KL floor is $\mathcal{K}_{\min}$, whose leading term is $\beta^2\bar{v}/2$. For a law q on $\mathcal{X}$, write $q(\mathbf{R})$ for the distribution of $\mathbf{R}(X)$ when $X \sim q$.

Lemma 2 (Level-set invariance)**.** *For every measurable composition g with $0 < \mathbb{E}_{p_1}[e^{\beta g \circ \mathbf{R}}] < \infty$,*

$$q_g(\cdot \mid \mathbf{R} = \mathbf{r}) = p_1(\cdot \mid \mathbf{R} = \mathbf{r}) \quad \textit{for } p_1(\mathbf{R})\textit{-almost every } \mathbf{r}.$$

The tilt reweights surrogate level sets without changing their conditional laws. This statement also applies when individual level sets have zero mass.

Proposition 1 (Exact KL floor and optimal composition)**.** *Fix $\beta > 0$. Under Assumption G.2, for every compared composition g,*

$$\mathrm{KL}(q^* \| q_g) = \mathcal{K}_{\min} + \mathrm{KL}(q^*(\mathbf{R}) \| q_g(\mathbf{R})) \ge \mathcal{K}_{\min},$$

where

$$\mathcal{K}_{\min} := \mathbb{E}_{\mathbf{r} \sim q^*(\mathbf{R})}\left[\mathrm{KL}\big(q^*(\cdot \mid \mathbf{r}) \| p_1(\cdot \mid \mathbf{r})\big)\right]$$

does not depend on g. Over all admissible measurable compositions, the minimum is attained at

$$g^{\mathrm{opt}}(\mathbf{r}) = \frac{1}{\beta} \log \mathbb{E}_{p_1}[e^{\beta R^*} \mid \mathbf{R} = \mathbf{r}],$$

uniquely up to an additive constant, $p_1(\mathbf{R})$-almost surely; it need not belong to a restricted fitted family.

Moreover, $\mathcal{K}_{\min} = 0$ if and only if $\bar{v} = 0$. As $\beta \downarrow 0$, with g fixed in the second expansion,

$$\mathcal{K}_{\min} = \frac{\beta^2}{2}\bar{v} + O(\beta^3),$$

$$\mathrm{KL}(q^* \| q_g) = \mathcal{K}_{\min} + \frac{\beta^2}{2}\mathrm{Var}_{p_1}(g \circ \mathbf{R} - \mu) + O(\beta^3).$$

Corollary 1 (The regression solution is leading-order optimal)**.** *Under Assumption G.2, for the displayed g^{opt} in Proposition 1,*

$$g^{\mathrm{opt}} \circ \mathbf{R} = \mu + \tfrac{\beta}{2} v + O(\beta^2)$$

uniformly almost surely. As $\beta \downarrow 0$,

$$\mathrm{KL}(q^* \| q_\mu) - \mathcal{K}_{\min} = \frac{\beta^4}{8}\mathrm{Var}_{p_1}(v) + O(\beta^5).$$

Here $\mathrm{Var}_{p_1}(v)$ measures variation in conditional oracle variance across surrogate values. Thus the conditional mean induces a tilt whose excess KL above the exact optimum is fourth order in β.

G.3 Reward error controls steering

Lemma 3 (Reward error controls steering)**.** *Under Assumption G.2, with g fixed,*

$$\mathrm{KL}(q^* \| q_g) = \frac{\beta^2}{2}\mathrm{Var}_{p_1}(R^* - g \circ \mathbf{R}) + O(\beta^3) \quad (\beta \downarrow 0).$$

No mean or variance matching is required for this expansion.

The orthogonal decomposition $R^* - g \circ \mathbf{R} = (R^* - \mu) + (\mu - g \circ \mathbf{R})$ gives the recovery floor plus approximation error in Theorem 1.

Corollary 2 (Positive gain from imperfect recovery)**.** *Under the assumptions of Theorem 1, suppose*

$$\mathrm{MSE}(g) < \mathrm{Var}_{p_1}(R^*),$$

or equivalently, $\|g \circ \mathbf{R} - \mu\|^2_{L^2(p_1)} < \mathrm{Var}_{p_1}(\mu)$. Then there exists $\beta_0 > 0$ such that

$$\mathbb{E}_{q_g}[R^*] > \mathbb{E}_{p_1}[R^*] \qquad \text{for all } 0 < \beta < \beta_0.$$

Thus, predicting the oracle more accurately than the best constant predictor suffices for positive reward gain at sufficiently small guidance strength.

G.4 Correlation ceiling

Corollary 3 (Correlation ceiling)**.** *Under Assumption G.1, if $\mathrm{Var}_{p_1}(R^*) > 0$, every nonconstant score $g \circ \mathbf{R}$ satisfies*

$$|\mathrm{PCC}_{p_1}(g \circ \mathbf{R}, R^*)| \le \sqrt{1 - \bar{v}/\mathrm{Var}_{p_1}(R^*)}.$$

If $\mathrm{Var}_{p_1}(\mu) > 0$, equality is attained at $g \circ \mathbf{R} = \mu$. If $\mathrm{Var}_{p_1}(\mu) = 0$, the correlations are zero and PCC with the constant μ is undefined.

The identity $\bar{v} = \mathrm{Var}_{p_1}(R^*)(1 - \rho^2_{\mathrm{ceil}})$ involves the population ceiling. An arbitrary predictor's held-out correlation need not attain it or determine $\bar{v}$. A fitted linear predictor estimates performance

within the linear class and need not attain this unrestricted ceiling. The best nonlinear function of a single surrogate R_j instead has Pearson ceiling

$$\sqrt{\frac{\mathrm{Var}_{p_1}(\mathbb{E}_{p_1}[R^* \mid R_j])}{\mathrm{Var}_{p_1}(R^*)}}.$$

Corollary 4 (Standardized surrogate tilts). *Under Assumptions G.2 and G.3, let $f = g \circ \mathbf{R}$, with $s_*^2 = \mathrm{Var}_{p_1}(R^*) > 0$ and $s_f^2 = \mathrm{Var}_{p_1}(f) > 0$. For a nonconstant score F, define*

$$z_F := \frac{F - \mathbb{E}_{p_1}[F]}{\sqrt{\mathrm{Var}_{p_1}(F)}}, \qquad q_F^{\mathrm{std}} \propto p_1 e^{\beta z_F}.$$

For fixed g, as $\beta \downarrow 0$,

$$\mathrm{KL}(q_{R^*}^{\mathrm{std}} \| q_f^{\mathrm{std}}) = \beta^2 \big[1 - \mathrm{PCC}_{p_1}(R^*, f)\big] + O(\beta^3).$$

If $s_\mu^2 := \mathrm{Var}_{p_1}(\mu) > 0$, the leading coefficient is minimized over bounded, nonconstant surrogate scores by $f = \mu$. Moreover, with $\rho_{\mathrm{ceil}} = s_\mu / s_*$,

$$0 \le \rho_{\mathrm{ceil}} - \mathrm{PCC}_{p_1}(R^*, f) \le \frac{2\big(\mathrm{MSE}(g) - \bar{v}\big)}{s_* s_\mu}.$$

G.5 Surrogates and floors

Corollary 5 (More surrogates can improve the optimal floors). *Let $\mathbf{R}_S$ be any sub-vector of $\mathbf{R}$, and let $\bar{v}_S$ and $\mathcal{K}_{\min}(S)$ be its optimal recovery and KL floors. Then*

$$\bar{v} \le \bar{v}_S, \qquad \mathcal{K}_{\min} \le \mathcal{K}_{\min}(S).$$

The recovery comparison uses Assumption G.1; the KL comparison holds for every $\beta > 0$ under Assumption G.2. If $\mathrm{Var}_{p_1}(R^*) > 0$, the Pearson ceiling weakly increases. Under Assumption G.2, the conditional-mean leading reward-gain coefficient $\mathrm{Var}_{p_1}(R^*) - \bar{v}$ weakly increases, and its leading mean-gap coefficient $\bar{v}$ weakly decreases.

The unrestricted best composition of the full vector can match or improve on that of any sub-vector. Redundant surrogates need not give strict improvement. These comparisons do not guarantee that a finite-data fit improves, or impose an unrestricted reward ceiling.

G.6 Plackett–Luce curvature

The following derivatives give the curvature used in the preference recovery bound.

Proposition 2 (Plackett–Luce curvature). *For a query with $m \ge 2$ candidates, finite score vector $f \in \mathbb{R}^m$, and observed order π, let $A_i = \{\pi(i), \dots, \pi(m)\}$ and let $p^{(i)}$ be the softmax on A_i, zero elsewhere. The query NLL satisfies*

$$\nabla_f \mathcal{L}_S = \sum_{i=1}^{m-1} (p^{(i)} - e_{\pi(i)}), \qquad \nabla_f^2 \mathcal{L}_S = \sum_{i=1}^{m-1} \Big[\mathrm{diag}(p^{(i)}) - p^{(i)} p^{(i)\top}\Big] \succeq 0.$$

Its Hessian has null space $\mathrm{span}\{\mathbf{1}\}$ and rank $m - 1$, where e_j is the jth standard basis vector.

H PROOFS

Proofs are ordered by dependency. All expectations below are under the stated base or tilted laws, and all assertions about conditional moments hold almost surely.

Lemma 2 (level-set invariance). For a measurable event A, conditional change of measure gives

$$\mathbb{E}_{q_g}[\mathbf{1}_A \mid \mathbf{R}] = \frac{\mathbb{E}_{p_1}[\omega_{g\circ\mathbf{R}}\mathbf{1}_A \mid \mathbf{R}]}{\mathbb{E}_{p_1}[\omega_{g\circ\mathbf{R}} \mid \mathbf{R}]} = \mathbb{E}_{p_1}[\mathbf{1}_A \mid \mathbf{R}].$$

The density ratio is positive and $\mathbf{R}$-measurable, so it cancels. The base and tilted surrogate marginals have the same null sets. No division by the probability of a level set is used. □

Lemma 3 (reward error controls steering). Let $Z_g = \mathbb{E}_{p_1}[e^{\beta g\circ\mathbf{R}}]$ and $Z^* = \mathbb{E}_{p_1}[e^{\beta R^*}]$. The density ratio gives

$$\mathrm{KL}(q^*\|q_g) = \beta\mathbb{E}_{q^*}[R^* - g\circ\mathbf{R}] + \log Z_g - \log Z^*.$$

Assumption G.2 justifies the Taylor expansions:

$$\begin{aligned}
\log Z_g &= \beta\mathbb{E}_{p_1}[g\circ\mathbf{R}] + \tfrac{\beta^2}{2}\mathrm{Var}_{p_1}(g\circ\mathbf{R}) + O(\beta^3),\\
\log Z^* &= \beta\mathbb{E}_{p_1}[R^*] + \tfrac{\beta^2}{2}\mathrm{Var}_{p_1}(R^*) + O(\beta^3),\\
\mathbb{E}_{q^*}[R^* - g\circ\mathbf{R}] &= \mathbb{E}_{p_1}[R^* - g\circ\mathbf{R}] + \beta\mathrm{Cov}_{p_1}(R^*, R^* - g\circ\mathbf{R}) + O(\beta^2).
\end{aligned}$$

The linear terms cancel, leaving

$$\tfrac{\beta^2}{2}\left[\mathrm{Var}_{p_1}(R^*) - 2\mathrm{Cov}_{p_1}(R^*, g\circ\mathbf{R}) + \mathrm{Var}_{p_1}(g\circ\mathbf{R})\right] + O(\beta^3),$$

which is the claimed variance expansion. □

Theorem 1 (Value Regime). *Suppose R^* and $g\circ\mathbf{R}$ satisfy Assumptions G.1–G.3. Then:*

(i) (Steering.) *As $\beta \downarrow 0$, with Assumption G.3 used for the second equality,*

$$\mathrm{KL}(q^*\|q_g) = \frac{\beta^2}{2}\mathrm{Var}_{p_1}(R^* - g\circ\mathbf{R}) + O(\beta^3) = \frac{\beta^2}{2}\mathrm{MSE}(g) + O(\beta^3).$$

(ii) (Reward gain.) *For any $R \in L^\infty(p_1)$ and $\beta > 0$: $\mathbb{E}_{q_R}[R^*] - \mathbb{E}_{p_1}[R^*] = \mathrm{Cov}_{p_1}(R^*, \omega_R)$. Furthermore, as $\beta \downarrow 0$:*

$$\mathbb{E}_{q_g}[R^*] - \mathbb{E}_{p_1}[R^*] \geq \beta\left[\mathrm{Var}_{p_1}(\mu) - \sqrt{\mathrm{Var}_{p_1}(\mu)(\mathrm{MSE}(g) - \bar{v})}\right] + O(\beta^2).$$

(iii) (Reward gap.) *As $\beta \downarrow 0$, $\mathbb{E}_{q^*}[R^*] - \mathbb{E}_{q_g}[R^*] \leq \beta[\bar{v} + \sqrt{\mathrm{Var}_{p_1}(\mu)(\mathrm{MSE}(g) - \bar{v})}] + O(\beta^2)$.*

(iv) (Best recovery.) *Under Assumption G.1 alone,*

$$\mathrm{MSE}(g) = \bar{v} + \|g\circ\mathbf{R} - \mu\|^2_{L^2(p_1)} \geq \bar{v}, \qquad \mathrm{Var}_{p_1}(R^*) = \mathrm{Var}_{p_1}(\mu) + \bar{v}.$$

Proof of Theorem 1. (iv), using Assumption G.1. Write $\varepsilon = R^* - \mu$. Then $\mathbb{E}_{p_1}[\varepsilon \mid \mathbf{R}] = 0$ and $\mathbb{E}_{p_1}[\varepsilon^2] = \bar{v}$. Thus ε is orthogonal in $L^2(p_1)$ to every square-integrable function of $\mathbf{R}$. Expanding the square gives

$$\mathrm{MSE}(g) = \mathbb{E}_{p_1}[(\varepsilon + \mu - g\circ\mathbf{R})^2] = \bar{v} + \|g\circ\mathbf{R} - \mu\|^2_{L^2(p_1)}.$$

Equality with $\bar{v}$ holds exactly when $g\circ\mathbf{R} = \mu$. Taking the constant score $\mathbb{E}_{p_1}[R^*]$ proves $\mathrm{Var}_{p_1}(R^*) = \mathrm{Var}_{p_1}(\mu) + \bar{v}$.

(i), using Assumptions G.2 and G.3. The same orthogonality yields

$$\mathrm{Var}_{p_1}(R^* - g\circ\mathbf{R}) = \bar{v} + \mathrm{Var}_{p_1}(g\circ\mathbf{R} - \mu).$$

By mean alignment the error has mean zero, so this equals $\mathrm{MSE}(g)$. Lemma 3 proves (i).

(ii). For $R \in L^\infty(p_1)$ and $R^* \in L^2(p_1)$, the density ratio satisfies $\mathbb{E}_{p_1}[\omega_R] = 1$, giving the exact identity

$$\mathbb{E}_{q_R}[R^*] - \mathbb{E}_{p_1}[R^*] = \mathbb{E}_{p_1}[R^*\omega_R] - \mathbb{E}_{p_1}[R^*]\mathbb{E}_{p_1}[\omega_R] = \mathrm{Cov}_{p_1}(R^*, \omega_R).$$

This identity also appears in Paes et al. (2026). For a composition, $\omega_{g\circ\mathbf{R}}$ is $\mathbf{R}$-measurable, so the covariance equals

$$\mathrm{Cov}_{p_1}(\mu, \omega_{g\circ\mathbf{R}}).$$

Under Assumption G.2, expanding the density ratio gives

$$\mathbb{E}_{q_g}[R^*] - \mathbb{E}_{p_1}[R^*] = \beta \mathrm{Cov}_{p_1}(\mu, g \circ \mathbf{R}) + O(\beta^2).$$

Part (iv) implies

$$\mathrm{Var}_{p_1}(g \circ \mathbf{R} - \mu) \le \mathrm{MSE}(g) - \bar{v},$$

with equality under Assumption G.3. Consequently,

$$|\mathrm{Cov}_{p_1}(\mu, \mu - g \circ \mathbf{R})| \le \sqrt{\mathrm{Var}_{p_1}(\mu)(\mathrm{MSE}(g) - \bar{v})}$$

by Cauchy–Schwarz. Substituting

$$\mathrm{Cov}_{p_1}(\mu, g \circ \mathbf{R}) = \mathrm{Var}_{p_1}(\mu) - \mathrm{Cov}_{p_1}(\mu, \mu - g \circ \mathbf{R})$$

proves (ii).

(iii), using Assumption G.2. The oracle tilt's gain is $\beta \mathrm{Var}_{p_1}(R^*) + O(\beta^2)$, so

$$\mathbb{E}_{q^*}[R^*] - \mathbb{E}_{q_g}[R^*] = \beta\left[\bar{v} + \mathrm{Cov}_{p_1}(\mu, \mu - g \circ \mathbf{R})\right] + O(\beta^2). \tag{6}$$

The same Cauchy–Schwarz bound proves (iii). □

The conditional mean minimizes recovery error and the leading KL coefficient. It need not maximize expected reward over arbitrary score scalings: replacing μ by $a\mu$ changes the leading gain to $a\mathrm{Var}_{p_1}(\mu)$.

Proof of Corollary 2. By Theorem 1(iv),

$$\mathrm{MSE}(g) - \bar{v} < \mathrm{Var}_{p_1}(\mu).$$

Consequently, the coefficient of β in the gain lower bound of Theorem 1(ii) is strictly positive. For fixed g, this term dominates the $O(\beta^2)$ remainder for all sufficiently small positive β. □

Proposition 2 (PL curvature). For a fixed query and order,

$$\mathcal{L}_S(f) = \sum_{i=1}^{m-1}\left[\log \sum_{l \in A_i} e^{f_l} - f_{\pi(i)}\right].$$

Differentiating log-sum-exp gives the stated gradient and Hessian. For any $w \in \mathbb{R}^m$,

$$w^\top\left[\mathrm{diag}(p^{(i)}) - p^{(i)}p^{(i)\top}\right] w = \mathrm{Var}_{p^{(i)}}(w) \ge 0.$$

The first stage assigns positive probability to every candidate, so this quadratic form vanishes only for constant w. Every stage vanishes on constant vectors; hence the summed Hessian has the stated null space and rank. □

Lemma 1 (likelihood controls recovery). Under Assumptions G.1, G.3, and G.4, set $w = (g - g_{\mathrm{PL}}) \circ \mathbf{R}$ and $g_t = g_{\mathrm{PL}} + t(g - g_{\mathrm{PL}})$, $0 \le t \le 1$. The segment is feasible by convexity of $\mathcal{G}$. Minimality gives

$$\frac{d}{dt}\mathrm{PL}(g_t)\Big|_{t=0} \ge 0,$$

not necessarily equality at a constrained boundary. For $S = (x_1, \ldots, x_m)$ let $w_{S,j} = w(x_j)$ and $\bar{w}_S = m^{-1}\sum_j w_{S,j}$.

Proposition 2 bounds the likelihood's second directional derivative below by its first-stage variance. The range bound gives $p_j^{(1)} \geq e^{-\Delta}/m$, and hence

$$\mathrm{Var}_{p^{(1)}}(w_S) = \min_c \sum_j p_j^{(1)}(w_{S,j} - c)^2 \geq \frac{e^{-\Delta}}{m} \sum_j (w_{S,j} - \bar{w}_S)^2.$$

For independent items,

$$\mathbb{E}_S \left[m^{-1} \sum_j (w_{S,j} - \bar{w}_S)^2 \right] = (1 - 1/m)\mathrm{Var}_{p_1}(w).$$

Taking expectations over both S and π gives

$$\frac{d^2}{dt^2}\mathrm{PL}(g_t) \geq c_\Delta \mathrm{Var}_{p_1}(w).$$

The bounded scores and directions justify differentiating under these expectations. Taylor's integral formula then yields

$$\begin{aligned}\mathrm{PL}(g) - \mathrm{PL}(g_{\mathrm{PL}}) &= \frac{d}{dt}\mathrm{PL}(g_t)\Big|_{t=0} + \int_0^1 (1-t)\frac{d^2}{dt^2}\mathrm{PL}(g_t)\,dt \\ &\geq \tfrac{c_\Delta}{2}\mathrm{Var}_{p_1}(w),\end{aligned}$$

proving the claim. □

Theorem 2 (Preference Regime). *Under Assumptions G.1, G.3, and G.4, define the population loss* $\mathrm{PL}(g) := \mathbb{E}_{S,\pi}[-\log p_{\mathrm{PL}}(\pi \mid g \circ \mathbf{R}, S)]$ *and let* $g_{\mathrm{PL}} \in \arg\min_{g \in \mathcal{G}} \mathrm{PL}(g)$. *With the score-range bound* Δ *from G.4 and* $c_\Delta = (1 - 1/m)e^{-\Delta}$, *every* $g \in \mathcal{G}$ *satisfies the bound*

$$\mathrm{MSE}(g) \leq B(g) := \bar{v} + \left(\sqrt{\mathrm{Var}_{p_1}(g_{\mathrm{PL}} \circ \mathbf{R} - \mu)} + \sqrt{\frac{2}{c_\Delta}\big(\mathrm{PL}(g) - \mathrm{PL}(g_{\mathrm{PL}})\big)} \right)^2.$$

Parts (i)–(iii) further assume the oracle bound in Assumption G.2 and hold for fixed g *as* $\beta \downarrow 0$:

(i) (Steering.) $\mathrm{KL}(q^* \| q_g) \leq \frac{\beta^2}{2} B(g) + O(\beta^3)$.

(ii) (Reward gain.) $\mathbb{E}_{q_g}[R^*] - \mathbb{E}_{p_1}[R^*] \geq \beta \left[\mathrm{Var}_{p_1}(\mu) - \sqrt{\mathrm{Var}_{p_1}(\mu)(B(g) - \bar{v})} \right] + O(\beta^2)$.

(iii) (Reward gap.) $\mathbb{E}_{q^*}[R^*] - \mathbb{E}_{q_g}[R^*] \leq \beta \left[\bar{v} + \sqrt{\mathrm{Var}_{p_1}(\mu)(B(g) - \bar{v})} \right] + O(\beta^2)$.

(iv) (Minimum bound.) $B(g) \geq B(g_{\mathrm{PL}}) = \mathrm{MSE}(g_{\mathrm{PL}})$.

Proof of Theorem 2. First use Assumptions G.1, G.3, and G.4. Theorem 1(iv), the L^2 triangle inequality, mean alignment, and Lemma 1 give

$$\begin{aligned}\sqrt{\mathrm{MSE}(g) - \bar{v}} &= \|g \circ \mathbf{R} - \mu\|_{L^2(p_1)} \\ &\leq \|g \circ \mathbf{R} - g_{\mathrm{PL}} \circ \mathbf{R}\|_{L^2(p_1)} + \|g_{\mathrm{PL}} \circ \mathbf{R} - \mu\|_{L^2(p_1)} \\ &\leq \sqrt{\frac{2}{c_\Delta}\big(\mathrm{PL}(g) - \mathrm{PL}(g_{\mathrm{PL}})\big)} + \sqrt{\mathrm{Var}_{p_1}(g_{\mathrm{PL}} \circ \mathbf{R} - \mu)}.\end{aligned}$$

Squaring and adding $\bar{v}$ proves $\mathrm{MSE}(g) \leq B(g)$. The nonnegative excess PL risk is zero at g_{PL}, so B is minimized there, with

$$B(g_{\mathrm{PL}}) = \bar{v} + \mathrm{Var}_{p_1}(g_{\mathrm{PL}} \circ \mathbf{R} - \mu) = \mathrm{MSE}(g_{\mathrm{PL}}).$$

This proves (iv), not minimization of the actual quantities bounded by B.

For (i)–(iii), additionally assume the oracle bound in Assumption G.2. The common finite mean and range width already imply

$$|g \circ \mathbf{R}| \leq |\mathbb{E}_{p_1}[R^*]| + \Delta$$

almost surely for every $g \in \mathcal{G}$. Theorem 1 therefore applies to each fixed g. Substituting $\mathrm{MSE}(g) \leq B(g)$ into its bounds proves (i)–(iii): the gain lower bound decreases, and the KL and gap upper bounds increase, with recovery error. □

The discrepancy need not vanish under conditional oracle variation: mixing PL laws over oracle values need not produce PL probabilities with utility μ. A sufficient realizable case is $\bar{v} = 0$ and a conditional-mean composition in $\mathcal{G}$. Then $R^* = \mu$ almost surely, and the excess conditional cross-entropy over that score is a KL divergence between ranking laws. Thus μ minimizes PL risk; Lemma 1 and mean alignment give $g_{\mathrm{PL}} \circ \mathbf{R} = \mu$.

Proposition 1 (exact KL floor). The KL chain rule and Lemma 2 give

$$\mathrm{KL}(q^* \| q_g) = \mathrm{KL}(q^*(\mathbf{R}) \| q_g(\mathbf{R})) + \mathbb{E}_{\mathbf{r} \sim q^*(\mathbf{R})} \left[\mathrm{KL}\big(q^*(\cdot \mid \mathbf{r}) \| p_1(\cdot \mid \mathbf{r})\big) \right].$$

The second term is $\mathcal{K}_{\min}$. The surrogate marginals have density ratios relative to $p_1(\mathbf{R})$ proportional to $\mathbb{E}_{p_1}[e^{\beta R^*} \mid \mathbf{R}]$ and $e^{\beta g \circ \mathbf{R}}$, respectively. They agree exactly for the displayed g^{opt}, up to an additive constant, proving optimality.

The conditional density ratio of q^* relative to p_1 is

$$\frac{e^{\beta R^*}}{\mathbb{E}_{p_1}[e^{\beta R^*} \mid \mathbf{R}]}.$$

Therefore $\mathcal{K}_{\min} = 0$ if and only if this ratio is one almost surely, equivalently R^* is $\mathbf{R}$-measurable, or $\bar{v} = 0$.

Assumption G.2 gives the uniform conditional cumulant expansion

$$\log \mathbb{E}_{p_1}[e^{\beta R^*} \mid \mathbf{R}] = \beta\mu + \tfrac{\beta^2}{2} v + O(\beta^3).$$

Thus $g^{\mathrm{opt}} \circ \mathbf{R} = \mu + O(\beta)$. Since $|g^{\mathrm{opt}} \circ \mathbf{R}| \le \|R^*\|_\infty$, the constants in the Taylor remainder bounds of Lemma 3 are independent of β, so that expansion also applies to g^{opt}. Orthogonality gives

$$\mathrm{Var}_{p_1}(R^* - g^{\mathrm{opt}} \circ \mathbf{R}) = \bar{v} + O(\beta^2),$$

so $\mathcal{K}_{\min} = \beta^2 \bar{v}/2 + O(\beta^3)$. For a fixed g, the same orthogonal decomposition and Lemma 3 give the final expansion. □

Corollary 1 (fourth-order excess). Dividing the conditional cumulant expansion by β proves

$$g^{\mathrm{opt}} \circ \mathbf{R} = \mu + \beta v/2 + O(\beta^2)$$

uniformly. The exact decomposition, equality of the surrogate marginals of q^* and $q_{g^{\mathrm{opt}}}$, and level-set invariance for $q_{g^{\mathrm{opt}}}$ and q_μ give

$$\mathrm{KL}(q^* \| q_\mu) - \mathcal{K}_{\min} = \mathrm{KL}(q_{g^{\mathrm{opt}}} \| q_\mu).$$

Twice differentiating the log normalizer along the interpolating score, and integrating, yields the exact identity

$$\mathrm{KL}(q_{g^{\mathrm{opt}}} \| q_\mu) = \beta^2 \int_0^1 t \, \mathrm{Var}_{q_{(1-t)\mu + t(g^{\mathrm{opt}} \circ \mathbf{R})}} (g^{\mathrm{opt}} \circ \mathbf{R} - \mu) \, dt.$$

For completeness, if the log normalizer along this path is $F(t)$, the KL is

$$F'(1) - F(1) + F(0) = \int_0^1 t F''(t) \, dt,$$

and F'' is β^2 times the displayed variance. The tilted density is $1 + O(\beta)$ relative to p_1 uniformly in t. Since $g^{\mathrm{opt}} \circ \mathbf{R} - \mu = \beta v/2 + O(\beta^2)$ uniformly, the integrand's variance equals

$$\beta^2 \mathrm{Var}_{p_1}(v)/4 + O(\beta^3).$$

Integration gives $\beta^4 \mathrm{Var}_{p_1}(v)/8 + O(\beta^5)$. This argument controls the remainder for the shrinking score difference; the generic $O(\beta^3)$ remainder of Lemma 3 alone would not suffice. □

Corollary 3 (correlation ceiling). Conditional expectation gives

$$\mathrm{Cov}_{p_1}(g \circ \mathbf{R}, R^*) = \mathrm{Cov}_{p_1}(g \circ \mathbf{R}, \mu).$$

Cauchy–Schwarz and total variance imply

$$|\mathrm{PCC}_{p_1}(g \circ \mathbf{R}, R^*)| \le \sqrt{\frac{\mathrm{Var}_{p_1}(\mu)}{\mathrm{Var}_{p_1}(R^*)}} = \sqrt{1 - \frac{\bar{v}}{\mathrm{Var}_{p_1}(R^*)}}.$$

If μ is nonconstant, its own score attains equality. If it is constant, the covariance vanishes for every admissible nonconstant score. The single-surrogate ceiling follows by the same argument conditioning on R_j. □

Corollary 4. Lemma 3 applied to the standardized scores gives the expansion, since $\mathrm{Var}_{p_1}(z_{R^*} - z_f) = 2[1 - \mathrm{PCC}_{p_1}(R^*, f)]$. Optimality follows from Corollary 3. Conditional expectation gives

$$\rho_{\mathrm{ceil}} - \mathrm{PCC}_{p_1}(R^*, f) = \frac{\rho_{\mathrm{ceil}}}{2} \|z_f - z_\mu\|^2_{L^2(p_1)}.$$

By mean alignment and the reverse triangle inequality,

$$\|z_f - z_\mu\|_{L^2(p_1)} \leq \frac{2}{s_\mu} \|f - \mu\|_{L^2(p_1)}.$$

The recovery decomposition in Theorem 1 completes the bound.

Corollary 5. Every composition of $\mathbf{R}_S$ is also a composition of $\mathbf{R}$. The full-vector population MSE infimum therefore cannot be larger; the two infima are $\bar{v}$ and $\bar{v}_S$ by Theorem 1(iv). The same inclusion for the exact KL infima in Proposition 1 proves $\mathcal{K}_{\min} \leq \mathcal{K}_{\min}(S)$ for every $\beta > 0$. The remaining weak comparisons follow by substitution in the correlation ceiling and the conditional-mean reward coefficients. □

I TOY EXAMPLES

I.1 VALUE FEEDBACK

Setup. We use a two-dimensional example to isolate the benefit of a nonlinear composition. Let $x = (u, v)$, $p_1 = \mathcal{N}(0, 4I)$, $p_0 = \mathcal{N}(0, I)$, and $R_1(x) = u$, $R_2(x) = v$. The oracle is $R^*(x) = |u| - |v|$, giving

$$q^*(x) \propto p_1(x) e^{\beta R^*(x)}, \qquad \beta = 1. \tag{7}$$

The target is bimodal along the u-axis, and no affine composition of the surrogates represents R^* exactly. Although the oracle is unbounded, the Gaussian prior gives a finite tilt normalizer. This example therefore extends beyond the bounded-reward assumptions of the main results.

Fitting. We use $n = 400$ prior samples to fit linear and MLP compositions with the value loss in Equation (3). The linear composition uses ridge regression on standardized surrogates. The MLP has two hidden layers of 64 GELU units and is trained with Adam, using a held-out quarter of the feedback set for early stopping. The predicted reward fields are compared with the oracle in Figure 10.

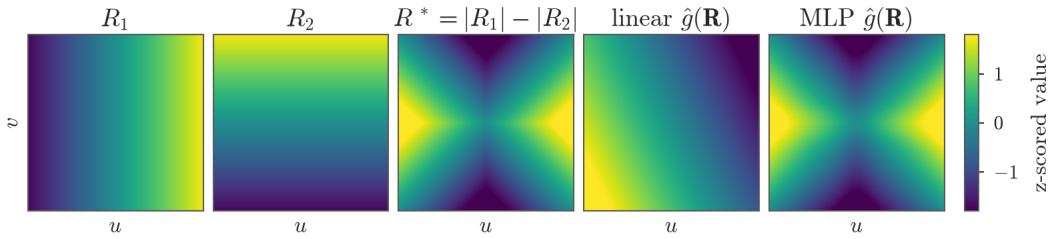


Figure 10: **Value-toy reward recovery.** Surrogate fields, oracle reward, and fitted linear and MLP compositions.

Steering. We steer the distilled meta flow map with MFM using each fitted composition or an individual surrogate. The MLP composition gives lower estimated W_1 and KL to q^* (0.51 and 0.05) than the linear composition (1.49 and 0.60), R_1 (1.85 and 1.06), or R_2 (2.32 and 0.91), consistent with the distributions in Figure 11.

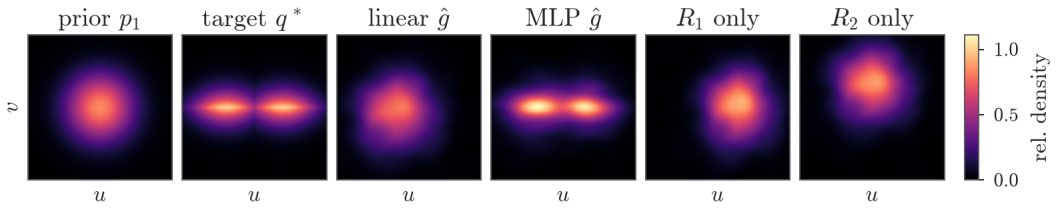


Figure 11: **Value-toy steering with MFM.** Prior and target distributions, followed by samples generated using the linear composition, MLP composition, and individual surrogates.

I.2 PREFERENCE FEEDBACK

We next learn from PL rankings for a multimodal oracle and compare FKS, FMRG, MFM, and Itô steering.

Setup. On $\mathcal{X} = \mathbb{R}^2$, with coordinates $x = (u, v)$ and prior $p_1 = \mathcal{N}(0, 4I)$, the oracle combines three Gaussian bumps:

$$R^*(x) = \sum_{c=1}^{3} h_c \, \phi(x; \mu_c, w),$$

where μ_c are the centers, h_c control their amplitudes, and w is the width parameter. The resulting target $q^*(x) \propto p_1(x) \exp(\beta R^*(x))$ at $\beta = 1$ is multimodal (Figure 12). Reference samples are obtained by rejection sampling. The three surrogates R_1, R_2, R_3 are corrupted views of the oracle, with shifted centers, rescaled heights, and noise localized around the modes.

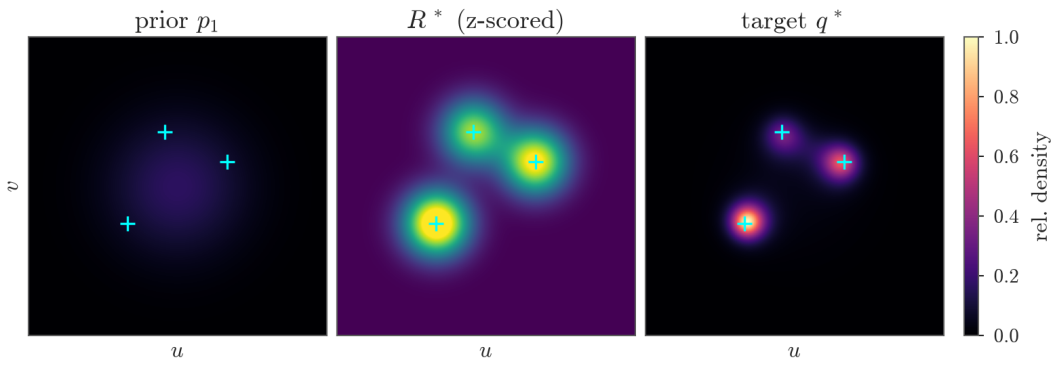


Figure 12: **Preference-toy setup.** Prior, oracle reward, and target tilt. Reward-mode centers are marked "+". Density panels here and in Figures 14 and 15 use a common scale relative to the peak of q^*.

Fitting. Preference feedback consists of a PL ranking of $m = 50$ independent prior samples, generated with utilities $\lambda R^*(x)$ and $\lambda = 10$. Linear and MLP compositions are fitted using the PL loss in Equation (4); correlation weights are computed from the observed ranking. Fitting uses these rankings rather than numerical oracle rewards. The fitted fields are shown in Figure 13.

The fitted PL utility is on the λR^* scale. To express the corresponding composition in oracle-reward units, it is divided by λ; we denote this calibrated composition by $\hat{g}$. We repeat the experiment over five ranking draws.

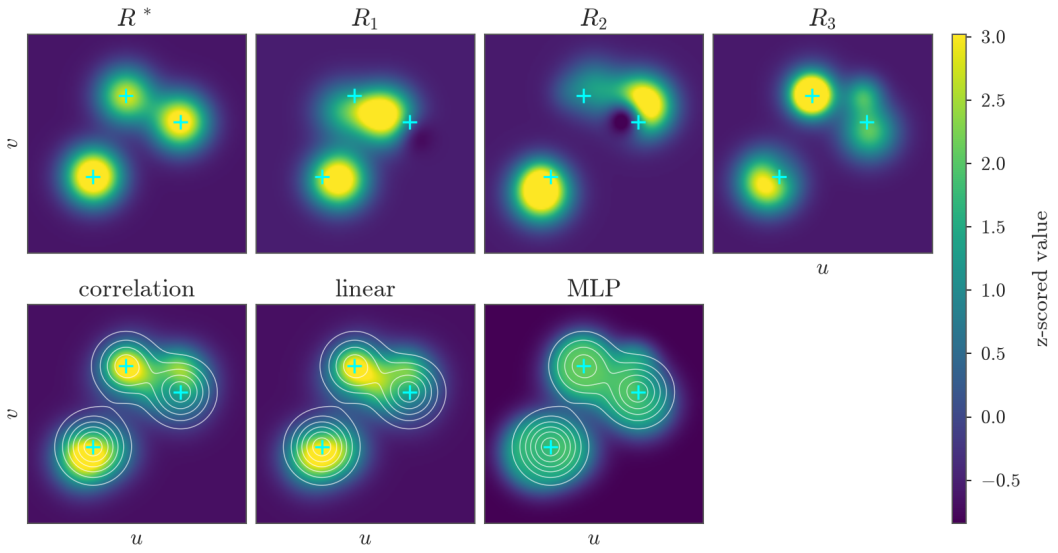


Figure 13: **Preference-toy reward recovery.** Oracle and surrogate fields, together with correlation, linear, and MLP compositions. Oracle contours are overlaid on the fitted fields.

Steering. We supply the same fitted composition to FKS, FMRG, MFM, and Itô steering (Figure 14). Its ideal tilt $q_{\hat{g}}$ can differ from q^*, while look-ahead, discretization, and particle approximations introduce additional sampling error. FMRG, MFM, and Itô are reported at the evaluated setting with the lowest reference-based KL; this is retrospective sweep selection. FKS uses a fixed configuration. The target remains fixed at $\beta = 1$ throughout the guidance sweeps.

The metrics rank the steerers differently in Table 12. FMRG and achieves the lowest KL (0.21), while FKS achieves the lowest W_1 (0.55). FMRG also gives the largest mean oracle reward (0.55), which alone does not establish closer agreement with the full target distribution.

Comparison with individual surrogates. Using FMRG, the fitted composition achieves lower W_1 and KL than any individual surrogate in Figure 15 and Table 13. Each reward uses its evaluated

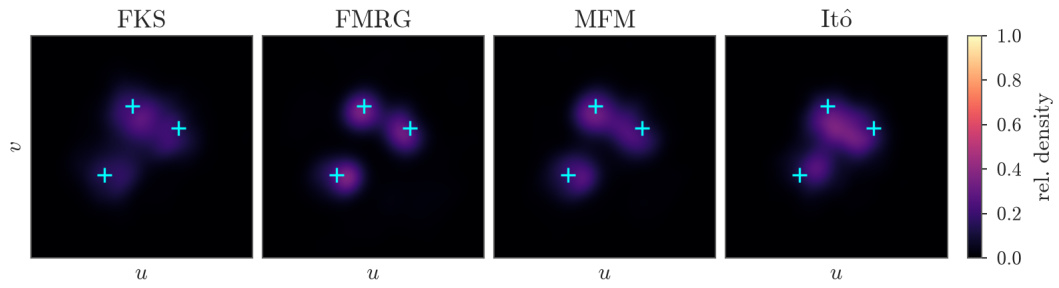


Figure 14: **Preference-toy steering.** FKS, FMRG, MFM, and Itô outputs under the fitted composition, using the density scale of Figure 12.

Table 12: **Preference-toy steering.** Estimated distances to q^* and mean oracle reward, averaged over five ranking draws. Bold marks the best value in each column.

Steerer	$W_1 \downarrow$	$\mathrm{KL}(q^* \Vert \cdot) \downarrow$	Mean $R^* \uparrow$
FKS	**0.55**	0.32	0.44
FMRG	0.69	**0.21**	**0.55**
MFM	0.57	0.26	0.49
Itô	0.68	0.43	0.48

KL-minimizing guidance strength, with the oracle target held fixed. This comparison demonstrates the benefit of composition within the evaluated sweeps.

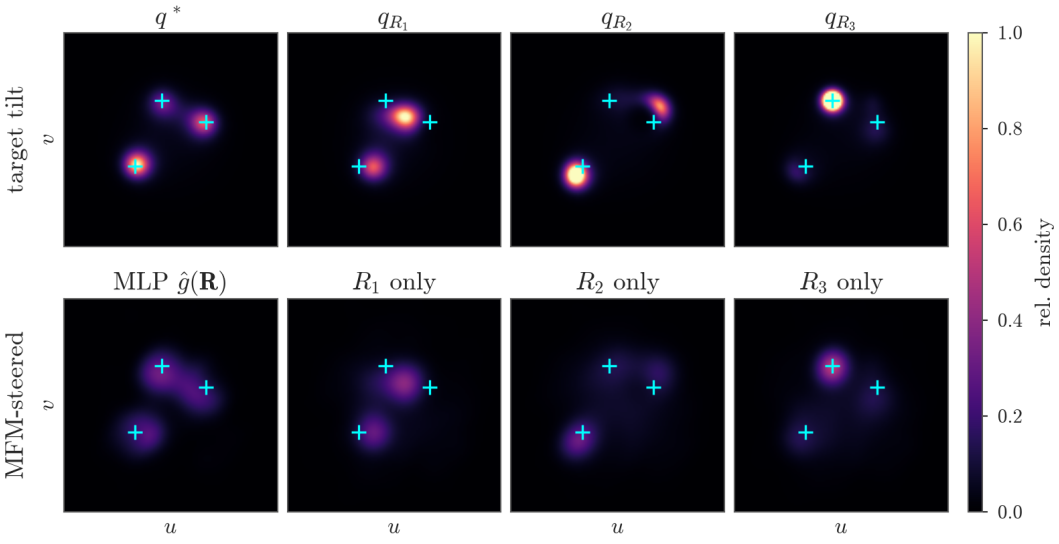


Figure 15: **Composition and individual surrogates.** Top: oracle and single-surrogate tilts. Bottom: FMRG outputs using the fitted composition and each surrogate separately. Densities use the scale of Figure 12.

Table 13: **Preference-toy reward comparison.** FMRG using the fitted composition or an individual surrogate, averaged over five ranking draws. Lower is better; bold marks the best result.

Steering reward	$W_1 \downarrow$	$\mathrm{KL}(q^* \Vert \cdot) \downarrow$
$\hat{g}(\mathbf{R})$	$\mathbf{0.69}$	$\mathbf{0.21}$
R_1	0.79	0.52
R_2	0.73	0.47
R_3	0.99	0.46

J QUALITATIVE EXAMPLES

J.1 LANGUAGE GENERATION

Method	Generated text
prior $p_{1,\phi}$	[CLS] june. [CLS] for the moment, some muslims in america will be told to thinker the united states, the united states also of course for the same for being 4, 000. [CLS] set a hands - on date for one local resident after this three of each college that, in spades, she is the only way when i move it to show…
BoN	[CLS] [CLS] nine officers involved in the trade said they expected 7 online, up from their peak at £25, 500 for up to 1, 000 people to be arrested in 2006. [CLS] in the same, it was spent about 10 minutes in the previous week…
sBoN	[CLS] are going though not to get " women to help out " is always just thousands of childcare - - and questions about whether orwell should not have set up but if there are simply no social issues, we co - authors of thousands…
uniform	[CLS] to [CLS] [CLS] this [CLS] ' s better earnings [CLS], or $ 1, 000 [CLS]. to match [CLS]. [CLS] and will he be [CLS] to [CLS] once this weekend, which will [CLS] that it along further. [CLS] japan ' s idea [CLS] than [CLS] in the [CLS] [CLS] city withinus…
BSS	[CLS] leaders. [CLS] in the senate, the 2010 parliament road will be political by long, the united states, the gas and also of announced for the public - private distribution of food. [CLS] i last emailed on thursday on new zealand for being " six of report " that, in bads, she is the only survivor when i move it to li…
SLOP *(draw 1)*	[CLS] himself up at the masters again to show a time and can work on training this week. [CLS] the end of the recession, he signed up to cutting 2, 000 jobs from 2030 by $ 2. 1. 5. [CLS] so might actually see your loan based on interest rates…
SLOP *(draw 2)*	[CLS] himself up at the masters again to show a time and can work on training this week. [CLS] the end of the recession, he signed up to cutting 2, 000 jobs from 2030 by $ 2. 1. 5. [CLS] so might actually see your loan based on interest rates…
ParetoFlow	[CLS] a year. [CLS] " there is none of the very how they, " she will. [CLS] there is none of this one, but the fact that that 100 lobby phone ' s office is not going back to war, what did, have that of the president? [CLS] about 19s out of the then back of two - time champion…
MARTINI	[CLS] parents. [CLS] for the times, some parents in staff will be told to do, the has investment, the act and also of out for the world for last 4 of england. [CLS] (did this blacks) this one is a almost and industry of the college that, in spades, she is the only business when i move it to off…

Table 14: **Language-generation examples.** Red highlights repetitions and malformed spans in the displayed excerpts. The two SLOP outputs are identical; ParetoFlow repeats a phrase within a sample.

J.2 IMAGENET

Figures 16 to 20 show five of the seven class–prompt pairs in the main table. Rows compare the unsteered prior with FMRG steering using MARTINI's linear composition, ReNO weights, and BSS in the value regime at $n = 32$. Each row contains that method's highest-oracle-reward samples. This oracle-based selection is used only for illustration; the grids do not describe typical output quality or an oracle-free deployment selection rule.

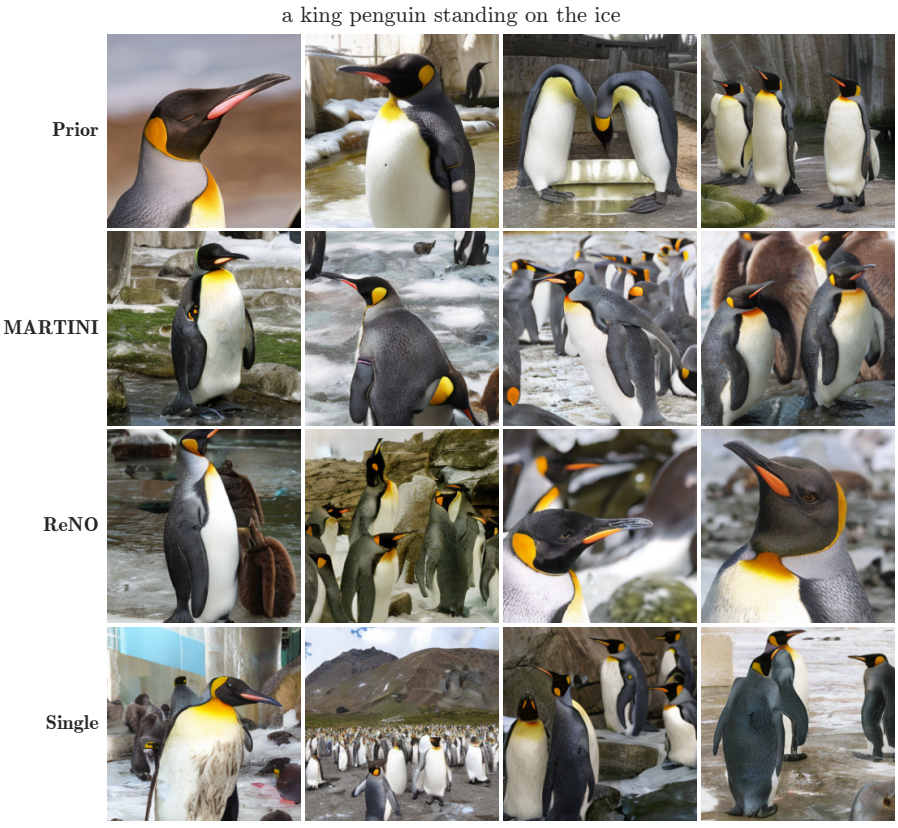


Figure 16: **ImageNet: king penguin.** Prompt: "a king penguin standing on the ice."

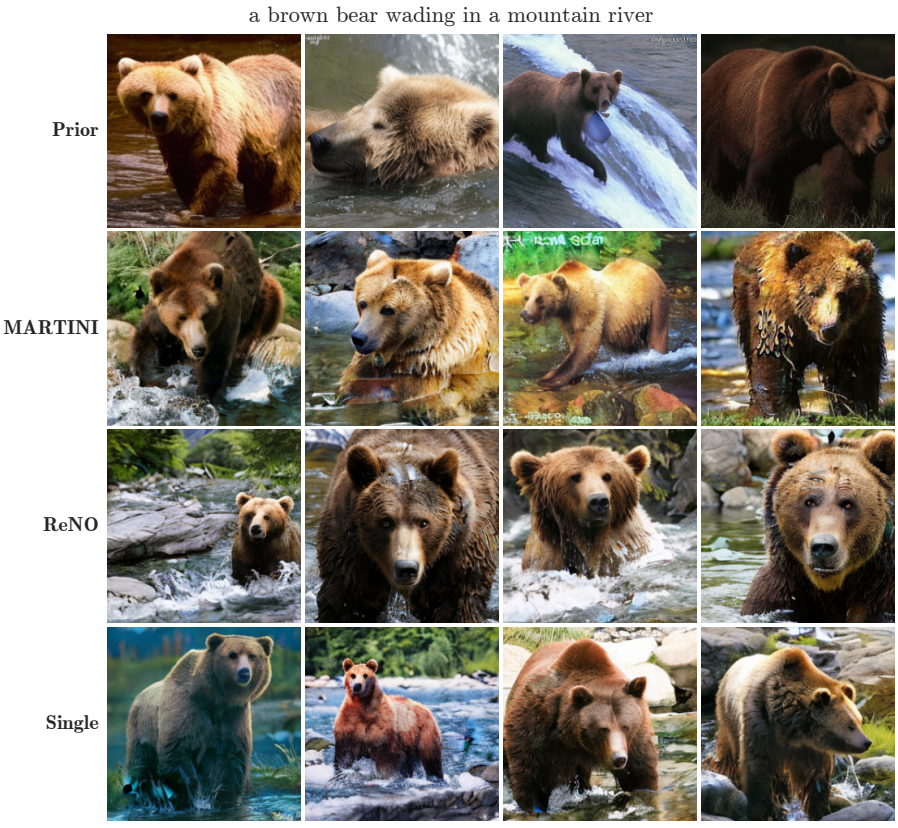


Figure 17: **ImageNet: brown bear.** Prompt: "a brown bear wading in a mountain river."

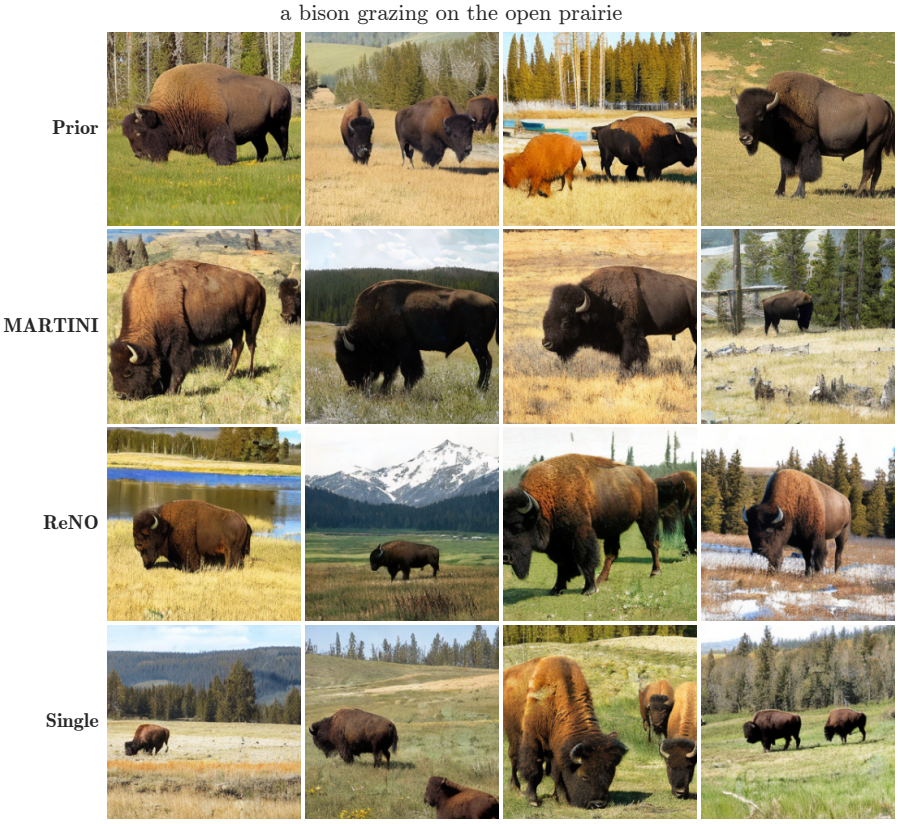


Figure 18: **ImageNet: bison.** Prompt: "a bison grazing on the open prairie."



Figure 19: **ImageNet: fire engine.** Prompt: "a fire engine on a busy city street."



Figure 20: **ImageNet: school bus.** Prompt: "a school bus on a tree-lined suburban street."

J.3 PROTEIN COFOLDING

Figure 21 shows 20 complexes, comparing ground truth, ligand crystal, the baseline DeCAF, and the MARTINI-best [gCorrelation].

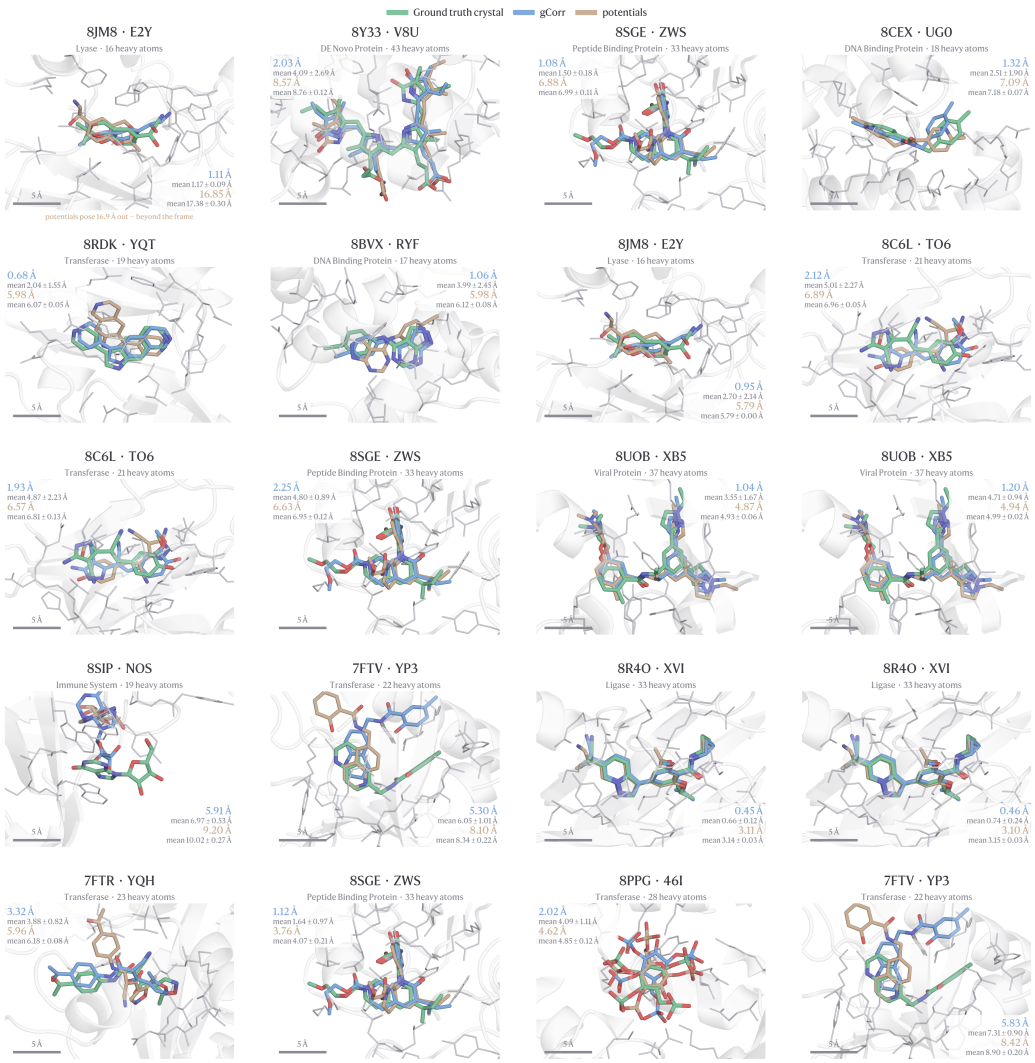


Figure 21: Top 20 complexes that outperform the baseline based on RMSD from the ground truth ligand crystal (green), the baseline DeCAF (orange), and the MARTINI-best [gCorrelation] (blue). The mean and standard deviation of the RMSD for that complex is show across the 5 seeds used to generate the particles. All steering runs here used $k = 15$ particles.

K RELATION TO MULTI-OBJECTIVE OPTIMISATION

Multi-objective optimization treats $R_1, \dots, R_k$ as objectives to be balanced. Here they provide imperfect information about a single oracle reward R^*. We formalize why surrogate Pareto efficiency need not identify candidates preferred by the oracle.

For rewards to be maximized, y dominates x if $R_j(y) \geq R_j(x)$ for every j, with at least one strict inequality. The nondominated feasible candidates form the Pareto set $\mathcal{P}$; their reward vectors form the Pareto front $\mathbf{R}(\mathcal{P})$.

Lemma 4 (Increasing compositions are Pareto efficient)**.** *If g is coordinatewise strictly increasing, every maximizer of $g \circ \mathbf{R}$ over the feasible domain belongs to the surrogate Pareto set $\mathcal{P}$.*

This ordering need not agree with the oracle:

Lemma 5 (The oracle optimum can be dominated)**.** *There exist a base law and rewards with $\bar{v} > 0$ whose unique oracle maximizer is strictly dominated in every surrogate.*

Example. Let $x = (a, b) \in [-1, 1]^2$ and $p_1 = \mathrm{Unif}([-1, 1]^2)$, with

$$R_1(x) = -(a+1)^2 - b^2, \qquad R_2(x) = -(a-1)^2 - b^2, \qquad R^*(x) = -3(a^2+b^2) + 2b - 3.$$

The surrogate Pareto set is the segment $\mathcal{P} = [-1, 1] \times \{0\}$, whereas the oracle has the unique maximizer $x^\star = (0, 1/3)$. The point $(0, 0)$ strictly dominates $x^\star$ on both surrogates, but has lower oracle reward:

$$R_1(0,0) = R_2(0,0) = -1 > -10/9 = R_1(x^\star) = R_2(x^\star),$$
$$R^*(0,0) = -3 < -8/3 = R^*(x^\star).$$

Figure 2 illustrates the distinction.

Removing the monotonicity restriction allows compositions to favor points outside the surrogate Pareto set, but does not recover missing information. Here $\mathbf{R}(a, b) = \mathbf{R}(a, -b)$, so every composition assigns the same score to $x^\star$ and $(0, -1/3)$ even though their oracle rewards differ.

K.1 PROOFS

Lemma 4. If a maximizer x were dominated by y, coordinatewise strict increase would give $g(\mathbf{R}(y)) > g(\mathbf{R}(x))$, a contradiction. □

Lemma 5. In the example, the surrogate values determine

$$a = (R_2 - R_1)/4, \qquad b^2 = -(R_1 + R_2)/2 - 1 - a^2.$$

Under the uniform base law, the sign of b is conditionally symmetric. Hence

$$\mu(x) = -3(a^2+b^2) - 3, \qquad R^*(x) - \mu(x) = 2b, \qquad \bar{v} = 4\mathbb{E}_{p_1}[b^2] = 4/3 > 0.$$

Any point with $b \neq 0$ is strictly dominated by $(a, 0)$. For a point on the segment, its distances to $(-1, 0)$ and $(1, 0)$ sum to 2. A dominating point would have a strictly smaller sum, contradicting the triangle inequality. Thus $\mathcal{P}$ is exactly this segment. Finally,

$$R^*(a, b) = -3a^2 - 3(b - 1/3)^2 - 8/3$$

has the unique maximizer stated above, which $(0, 0)$ strictly dominates in both surrogates. □